T.C.

BURSA ULUDAĞ ÜNİVERSİTESİ

SOSYAL BİLİMLER ENSTİTÜSÜ

EKONOMETRİ ANABİLİM DALI

İSTATİSTİK BİLİM DALI

MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE KREDİ TEMERRÜT RİSKİNİ TAHMİN ETME

(YÜKSEK LİSANS TEZİ)

Toprak Enes TÜTÜNCÜ

BURSA - 2022

T.C.

BURSA ULUDAĞ ÜNİVERSİTESİ

SOSYAL BİLİMLER ENSTİTÜSÜ

EKONOMETRİ ANABİLİM DALI

İSTATİSTİK BİLİM DALI

MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE KREDİ TEMERRÜT RİSKİNİ TAHMİN ETME

(YÜKSEK LİSANS TEZİ)

Toprak Enes TÜTÜNCÜ

Danışman:

Prof. Dr. Sevda GÜRSAKAL

BURSA – 2022

ÖZET

Yazar Adı ve Soyadı	: Toprak Enes TÜTÜNCÜ
Üniversite	: Bursa Uludağ Üniversitesi
Enstitüsü	: Sosyal Bilimler Enstitüsü
Anabilim	: Ekonometri
Bilim/Sanat Dalı	: İstatistik
Tezin Niteliği	: Yüksek Lisans Tezi
Sayfa Sayısı	: x + 88
Mezuniyet Tarihi	: 25/07/2022
Tez Danışmanı	: Prof. Dr. Sevda GÜRSAKAL

MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE KREDİ TEMERRÜT RİSKİNİ TAHMİN ETME

Bankalar ve çeşitli finans kuruluşları tarafından karşılanan kredilerin, müşteri tarafından geri ödenememesi hem kredi veren kuruluşun sermaye kaybını hem de genel ekonomide oluşabilecek çeşitli risk faktörlerini beraberinde getirmektedir. Bu süreçte, oldukça kritik öneme sahip olan kredi riskinin doğru yönetilebilmesi ve uluslararası finans istikrarının sağlanması için Basel Komitesi ve BDDK (Bankacılık Düzenleme ve Denetleme Kurumu) gibi finans denetimi kuruluşları, kredi veren kurumların kredi verme karar aşamasında çeşitli regülasyon politikaları belirlemektedir. Ayrıca, kredi veren kurumlar analitik risk birimleri aracılığıyla kredi değerlendirme modelleri geliştirerek, müşterilere ait kredi risk skorunu hesaplamaktadır.

Bu araştırmada, makine öğrenmesi yöntemiyle kredi skorlama sistemlerinde kullanılabilecek en başarılı tahmini gerçekleştiren algoritmanın belirlenmesi amaçlanmıştır. Bu kapsamda, Gradyan Artırma, Yapay Sinir Ağları, Lojistik Regresyon, Rassal Orman, Karar Ağacı, Destek Vektör Makineleri, K-En Yakın Komşu ve WOE dönüşümleriyle Lojistik Regresyon algoritmaları için modeller kurulmuş ve temerrüde düşen ve temerrüde düşmeyen müşteriler için en iyi sınıflandırma performansı gösteren Gradyan Artırma algoritması olmuştur.

Analitik veri kalitesi ve model geliştirme süreçlerinde SAS Enterprise Guide ve SAS Enterprise Miner yazılım programları kullanılmıştır.

Anahtar Sözcükler: Kredi Riski, Makine Öğrenmesi, Gradyan Artırma, Yapay Sinir Ağları, Lojistik Regresyon, Rassal Orman, Karar Ağacı, Destek Vektör Makineleri, K-En Yakın Komşu

ABSTRACT

Name and Surname	: Toprak Enes TÜTÜNCÜ
University	: Bursa Uludağ University
Institution	: Social Science Institution
Field	: Econometrics
Branch	: Statistics
Degree Awarded	: Master
Page Number	: x + 88
Degree Date	: 25/07/2022
Supervisor	: Prof. Dr. Sevda GÜRSAKAL

PREDICTING DEFAULT PROBABILITY IN CREDIT RISK WITH MACHINE LEARNING ALGORITHMS

Failure to repay the loans provided by banks and various financial foundations by the customer, entails both the capital loss of the lending institution and various risk factors that may occur in the general economy. In this context, financial control institutions such as the Basel Committee and BRSA (Turkish Banking Regulatory and Supervision Agency) have determined various regulatory policies during the phase of lending decision of the lending institutions in order to ensure the appropriate management of loan risk, which have critical importance, and to ensure international financial stability. In addition, lending institutions develop credit evaluation models via analytical risk units and calculate the credit risk score of customers.

In this study, it is aimed to determine the algorithm that makes the most successful estimation that can be used in credit scoring systems with the machine learning method. Within this scope, models for algorithms with Gradient Boosting, Artificial Neural Networks, Logistic Regression, Random Forest, Decision Tree, Support Vector Machines, K-Nearest Neighbor and WOE transformations Logistic Regression were established and Gradient Boosting algorithm has shown the best classification performance for defaulters and non-defaulters.

In analytical data quality and model development processes, SAS Enterprise Guide and SAS Enterprise Miner software programs were used.

Key Words: Credit Risk, Machine Learning, Gradient Boosting, Neural Network, Logistic Regression, Random Forest, Decision Tree, Support Vector Machine, K-Nearest Neighbor

ÖNSÖZ

Yüksek Lisans eğitimime başladığım ilk günümden bu çalışmanın tüm aşamasına kadar desteklerini ve rehberliğini esirgemeyen, kıymetli fikirleriyle katkıda bulunan değerli hocam ve tez danışmanım Prof. Dr. Sevda Gürsakal'a sonsuz teşekkürlerimi sunarım.

Analitik Danışman & Veri Bilimci olarak başladığım iş hayatımda daima yanımda olan ve desteklerini esirgemeyen kıymetli dostlarım Can Lütfü Yılmazer ve Mesut Aytekin'e teşekkürlerimi bir borç bilirim.

Her koşulda yanımda olan sevgili aileme…

Sonsuz teşekkürler...

Toprak Enes TÜTÜNCÜ Bursa, 2022

İÇİNDEKİLER

Sayfa

BİRİNCİ BÖLÜM
MAKİNE ÖĞRENMESİ VE KREDİ RİSKİ

İKİNCİ BÖLÜM
METODOLOJİ

ÜÇÜNCÜ BÖLÜM
VERİ KALİTESİ VE MODEL GELİŞTİRME

TABLOLAR

ŞEKİLLER

GRAFİKLER

KISALTMALAR

ABD	**AMERİKA BİRLEŞİK DEVLETLERİ**
AIRB	**ADVANCED INTERNAL RATINGS BASED**
AUC	**THE AREA UNDER THE CURVE**
BDDK	**BANKACILIK DÜZENLEME VE DENETLEME KURUMU**
BSS	**BİYOLOJİK SİNİR SİSTEMİ**
CART	**CLASSIFICATION AND REGRESSION TREE**
DVM	**DESTEK VEKTÖR MAKİNELERİ**
DT	**DECISION TREE**
EAD	**EXPOSURE AT DEFAULT**
EL	**EXPECTED LOSS**
GB	**GRADIENT BOOSTING**
IGN	**INTERACTIVE GROUPING NODE**
IRB	**INTERNAL RATINGS BASED**
IV	**INFORMATION VALUE**
KNN	**K-NEAREST NEIGHBORS**
LASSO	**LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR**
LGD	**LOSS GIVEN DEFAULT**
LR	**LOGISTIC REGRESSION**
M	**MATURITY**
MBR	**MEMORY-BASED REASONING**
NN	**NEURAL NETWORK**
OOB	**OUT OF BAG**
PD	**PROBABILTY OF DEFAULT**
RELU	**RECTIFIED LINEAR UNIT**
RF	**RANDOM FOREST**
ROC	**RECEIVER OPERATING CHARACTERISTIC**
SAS	**STATISTICAL ANALYSIS SOFTWARE**
SVM	**SUPPORT VECTOR MACHINES**
WOE	**WEIGHT OF EVIDENCE**
YSA	**YAPAY SİNİR AĞLARI**

GİRİŞ

Geleceğin bilinmezliğine olan merak ve kontrol altına alma isteği, insanoğlunun doğal arzularından bir tanesidir. Geleceğe ışık tutmak ise sadece belirli bir düzeyde mevcuttur. Gelecek için risk hala gizemli bir düzeyde kendini saklıyor olacaktır.

Riske karşı ortaya çıkan doğal savunma, geçmişin sonuçlarından öğrenerek yargısal bir tahminleme ile yapılmaktadır. Günümüzde risk boyutunun doğru ölçümlenmesi bilgisayarlar ve istatistik bilimi ile daha objektif temellere dayalı ampirik bir yöntem üzerinden makine öğrenmesi teknikleri kullanılarak gerçekleştirilmektedir.

Sürdürülebilir yaşam ve çalışma hayatının sürekliliğinin sağlanması için sahip olunan kaynakların etkin ve verimli düzeyde kullanılması, yapılacak olan risk tahminleri ile önlem alınmasına bağlıdır. Dolayısıyla riskin yönetimi her sektör için kritik önem arz etmektedir. Bu durumda ortaya çıkan en önemli risk tiplerinden biri de finansal risktir. Bu kapsamda, finans sektörü için sürdürülebilirlik hayati önem arz etmektedir.

Alternatif durumlara bağlı olan finansal riskte, getirinin geleceği bir önlem mekanizmasına ihtiyaç duyar. Finans sektörünün temelini oluşturan bankalar ana gelir kaynaklarını müşterileri için verdiği kredilerin faizi ile oluştururken, müşterilerin kredileri zamanında ödeyebilme durumuna bağlı bir karşılıklı ilişki içerisinde olduklarını bilirler. Bu yüzden bireysel veya kurumsal müşterilerine kredi vermeden önce haklarında nitel ve nicel verileri kullanarak, müşterilerin risk profilini uyguladıkları modeller ile kredi riski bağlamında ortaya çıkarırlar. Bu önlem mekanizmasıyla, riski minimum düzeyde tutarak kredi faaliyetlerini yönetmeye çalışırlar. Risk yönetiminde başarısız oldukları takdirde sadece kendi yapısını etkilemekle kalmaz, finanse ettiği mevduat sahipleri ve fon kaynaklarını da riske maruz bırakmış olurlar. Bu bağlamda, kredi politikalarının doğru yönetilmesi, risk ekosisteminde oldukça önemli yer tutmaktadır.

Artan nüfus sayısı ile tüketimin de artması, beraberinde kredi endüstrisinin genişlemesine sebep olmaktadır. Kredi talebinin artmasına yönelik bankaların kredilendirme faaliyetlerinde hızlı ve etkin kararlar alınmasına olanak sağlayan makine öğrenmesi teknikleri ile risk ekosisteminde optimum modeller inşa edilmektedir. Bu doğrultuda Lojistik Regresyon, Yapay Sinir Ağları, Karar Ağaçları gibi birçok algoritma kullanılmasına rağmen, hangi tekniğin en iyi performansı sağladığına dair bir konsensüse

varılmamıştır. Bu sebeple, algoritmalar ile kurulan modellerin sınıflandırma başarıları, çeşitli istatistiksel ve makine öğrenmesi teknikleriyle ölçülerek, en iyi sınıflandırma performansı sağlayan model, optimum model belirlenmektedir.

Bu çalışmada, müşterilere ait bireysel kredilerin riskini hesaplayarak, temerrüt oranını değerlendirmek amacıyla yedi farklı istatistiksel ve makine öğrenmesi algoritmaları kullanılmıştır. Araştırma kapsamında, modelde öznitelik değişkeni olarak girdi görevi görecek değişkenler farklı değişken indirgeme teknikleri kullanılarak belirlenmiştir. Nihai değişken seçimi için LASSO Regresyonu kullanılmış olup, ilgili tekniğin ceza parametresinin en güçlü sınıflandırıcı öznitelik değişkenlerin keşfedilmesini sağlamıştır. Nihai öznitelik değişkenleri ile temerrüt risk oranının tahminlemesi için Lojistik Regresyon, Yapay Sinir Ağları, Karar Ağacı, Destek Vektör Makineleri, K-En Yakın Komşular, Rassal Orman, Gradyan Artırma ve WOE dönüşümleri gerçekleştirilmiş haliyle Lojistik Regresyon algoritmaları kullanılmıştır. Elde edilen sonuçlar, Doğruluk, Hassasiyet, Özgüllük, Kesinlik, F1 skoru ve ROC eğrisi olmak üzere altı farklı ölçüye göre algoritmalar arasında performans karşılaştırılması yapılmıştır.

Çalışmanın devam eden bölümlerinde sırasıyla; Bölüm 1'de; makine öğrenmesi tekniklerinin temel yapıları ve bankacılıkta kredi risk analitiği ve önemi hakkında literatür taramaları yapılarak ele alınmıştır. Bölüm 2'de; Temerrüt Olasılığı (Probability of Default (PD)) modeli için uygulanacak olan makine öğrenmesi algoritmaları, performans ölçüleri, örneklemin belirlenmesi ve değişken indirgeme teknikleri ile izlenilecek metodolojiye yer verilmiştir. Bölüm 3'de; veri setinin ön işleme süreciyle modele dahil olacak değişkenlerin belirlenmesi için değişken eleme teknikleri ve kredi ödemesinde gecikme olan kitle kadar gecikme olmayan kitle ile dengelenmiş veri kümesi oluşturulmuştur. Dengelenmiş veri kümesine nihai değişken eleme teknikleri uygulanarak, uygun görülen değişkenler ile alternatif modeller kurulmuş ve performans kıyaslamaları gerçekleştirilmiştir. Bu doğrultuda uygulanan makine öğrenmesi algoritmaların performansları hakkında değerlendirmeler, öneriler ile birlikte sonuç bölümünde detaylandırılmıştır.

BİRİNCİ BÖLÜM

MAKİNE ÖĞRENMESİ VE KREDİ RİSKİ

1.1 MAKİNE ÖĞRENMESİ

Başlangıçta bilgisayarların öğrenmesini sağlayan tekniklerin geliştirilmesi için kullanılan makine öğrenmesi, zamanla yapay zekânın bir yöntemi haline gelmiştir.

Dartmouth'da matematik profesörü olan John McCarthy, 1956 yılında verdiği konferansta, yapay zekâyı "akıllı makineler yapma bilim ve mühendisliği" olarak tanımlamıştı. Bu doğrultuda yapay zekâ, makineleri akıllı yapma bilimi ise makine öğrenmesinin de bilgisayarların örneklerden öğrenerek belirli görevleri akıllıca yürütmesine izin veren bir teknoloji olduğu söylenebilir.

Geleneksel programlama yaklaşımları, bir sorunun çözümünü belirleyen adım adım kodlanmış kurallara dayanırken, makine öğrenmesi sistemleri bir görev olarak belirlenir. Dolayısıyla bu sistemler önceden programlanmış kurallara uymak yerine, verilerden öğrenerek karmaşık süreçleri yürütebilme imkânı sağlar. Bu görevi nasıl gerçekleştirebileceğine veya örüntülerin tespit edileceğine örnek olarak büyük bir veri kümesi işleme alınır. Daha sonra sistem istenen çıktıya nasıl ulaşacağını öğrenir.

Diğer bir ifadeyle makine öğrenmesi, bilgisayar algoritmalarının veri ve bilgilerden bağımsız olarak öğrenmek için kullanıldığı yapay zekanın bir alt kümesi olarak düşünülebilir. Makine öğrenmesinde bilgisayarların açıkça programlanması gerekmez, algoritmalarını kendi başlarına değiştirebilir ve geliştirebilirler.

Makine Öğrenme algoritmaları, "eğitim verileri" olarak da bilinen örnek veri setini kullanarak otomatik olarak bir matematiksel model oluşturur ve bu kararları almak için özel olarak programlanma ihtiyacı duymaz. Öğrenmenin en temel örneği verilere düz bir çizginin yerleştirilmesi olabilir, ancak makine öğrenmesi genellikle düz çizgilere göre çok daha esnek modellerle ilgilenir. Bunu yapmasının amacı, modelin öğrenmede kullanılmayan veriler hakkında kendi içinde yeni sonuçlar çıkarmak içindir. Bir modeli 1000 köpek yavrusu resmi verisinden öğrenirsek, model doğru bir şekilde seçilirse, başka bir görüntünün (öğrenme için kullanılan 1000 köpek yavrusu resmi dışında) bir köpek

yavrusu tasvir edip etmediğini söyleyebilir. Bu genelleme olarak bilinir (Lindholm, 2019:7).

Son yıllarda alandaki teknik gelişmeler, verilerin kullanılabilirliğinin artması ve artan bilgi işlem gücünün bir sonucu olarak makine öğrenmesinin yeteneklerinde önemli ilerlemeler görülmüştür. Bu ilerlemelerin bir sonucu olarak, sadece birkaç yıl önce doğru sonuçlar elde etmek için mücadele eden sistemlerin artık belirli görevlerde insanlardan daha iyi performans gösterebileceği kanıtlanmıştır. Günümüzde bazı görevlerde insanlardan daha iyi performans gösterebilen ses ve nesne tanıma sistemleri bulunmaktadır. Örneğin, 2015 yılında araştırmacılar, tek tek el yazısı rakamları tanımaya odaklanan dar bir vizyonla ilgili görevde insan yeteneklerini aşan bir makine öğrenme sistemi oluşturmuşlardır (Markoff, 2015:1).

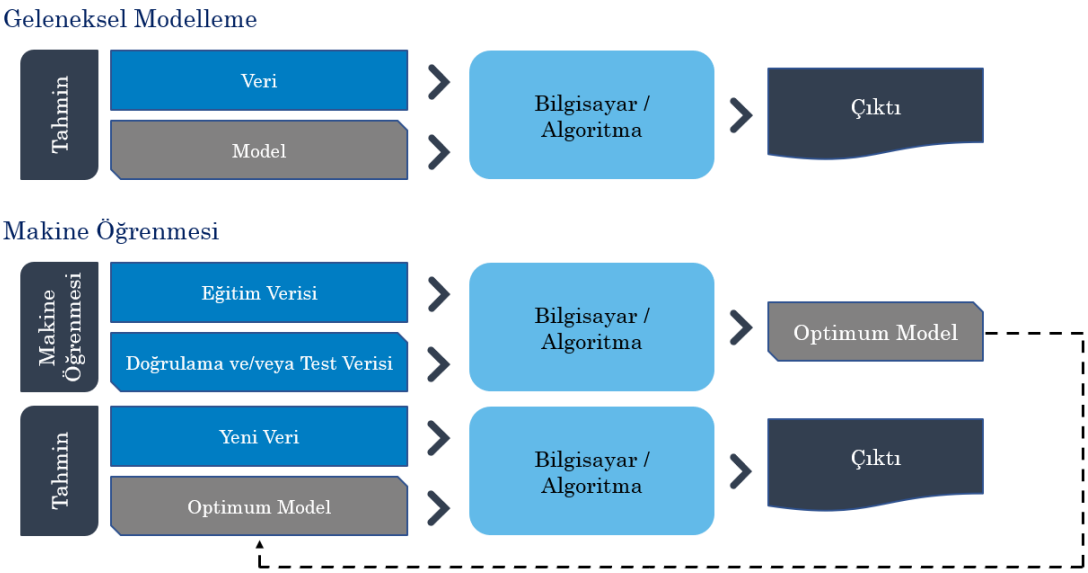
Makine öğrenmesi sağlık hizmeti, finans, insan kaynakları, satış ve pazarlama, lojistik ve üretim gibi birçok alanda kullanılarak sağlığımız, üretkenliğimiz ve refahımız için küresel zorlukları ele almayı ve verimliliği artırarak küresel ekonomiye trilyonlarca dolar eklemeyi vaat ediyor (The Royal Society, 2017:16).

Makine öğrenmesi teknikleri denetimli öğrenme, denetimsiz öğrenme, yarı – denetimli öğrenme ve pekiştirmeli öğrenme olarak dört ana başlıkta incelenebilir.

1.1.1 Denetimli Öğrenme

Denetimli öğrenmedeki esas amaç, öngörülemeyen veya gelecekteki veriler hakkında tahminlerde bulunmamızı sağlayan etiketli eğitim verilerinden bir model öğrenmektir (Raschka, 2015:3).

Denetimli öğrenme, öngörülemeyen verileri tahmin etmek için etiketli eğitim veri setinde bulunan geçmiş bilgilerden yararlanarak öğrenme işlemi gerçekleştirir. Örneğin geçmiş dönemdeki satışlardan oluşan bir veri kümesi ile gelecekteki fiyatları tahmin etmek için kullanılabilir. Denetimli öğrenme de etiketlenmiş eğitim verileri ve istenen çıktı değişkeninden oluşan bir girdi değişkeni mevcuttur. Girdi ile çıktıyı eşleştiren işlevi öğrenmek üzere eğitim verilerini analiz etmek için bir algoritma kullanılır. Bu çıkarımsal işlev, görünmeyen durumlarda sonuçları tahmin etmek için eğitim verilerinden genelleme yaparak yeni ve bilinmeyen örnekleri eşleştirir (Lui, 2017:1).



Şekil 1.1: Denetimli Makine Öğrenmesi Döngüsü ve Geleneksel Modelleme

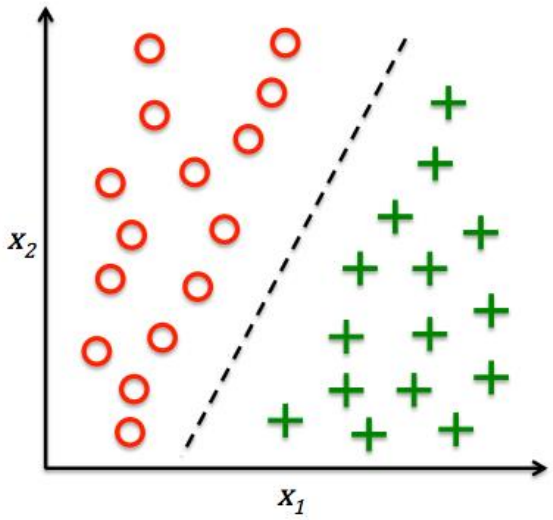
Günümüzde denetimli öğrenme, makine öğrenmesi algoritmalarında en yaygın biçimde kullanılan Doğrusal Regresyon Modeli, Lojistik Regresyon, Karar Ağaçları, Destek Vektör Makineleri, Topluluk Öğrenme Yöntemleri ve Yapay Sinir Ağları gibi öğrenim algoritmalarını içinde barındırır (Niculescu-Mizil, 2005:626). Bu algoritmalar kendi içinde uygulama amaçlarına ve yapılarına göre sınıflandırma, regresyon ve tahmin yaklaşımları şeklinde görevlere ayrılır.

1.1.1.1 Sınıflandırma

Sınıflandırma, hedefin, geçmiş gözlemlere dayalı yeni örneklerin kategorik sınıf etiketlerini öngörmek olduğu, denetimli öğrenmenin bir alt kategorisidir. Bu sınıf etiketleri, örneklerin grup üyelikleri olarak anlaşılabilen ayrık, sıralanmamış değerlerden oluşur (Raschka, 2015:3).

Sınıflandırma amacıyla kullanılan algoritmalar, yapısal veya yapısal olmayan veriler üzerine uygulanarak, gözlemlenen değerlerden bir sonuç çıkarıp yeni gözlemin hangi kategoriye ait olduğunu ikili sınıflandırma, çoklu sınıflandırma veya çoklu etiket sınıflandırma gibi yöntemler kullanarak belirler.

Sınıflandırma yöntemleri örnekler üzerinden ifade edilecek olursa: iki boyutlu sınıflandırma için batan veya batmayan olarak kredi durumu düşünülebilir[1]. Çoklu sınıflandırma için denizde yaşayan canlı türleri (balıklar, kabuklu canlılar, yumuşak gövdeli canlılar vd.) veya ikiden fazla segmente ayrılmış müşterilerin ayrıştırılması olabilir. Sınıflandırma işlemini görsel veri işleme veya yapısal bilgileri kullanarak gerçekleştirmek mümkündür. Bir diğer yöntem olan çoklu etiket sınıflandırması için bir kitabın hem tıp hem spor hem de istatistik ile ilgili olabileceği düşünülerek örneklendirilebilir.



Şekil 1.2: İkili Sınıflandırma

Günümüzde sınıflandırma yöntemlerinde Lojistik Regresyon, Boosting, Karar Ağacı, Rassal Orman, Naive Bayes, En Yakın Komşu ve Destek Vektör Makineleri en çok tercih edilen algoritmalardır.

1.1.1.2 Regresyon

Denetimli öğrenmenin diğer bir alt kategorisi olan regresyon, veri kümesinde bulunan değişkenlerin boyut değerleri arasında bir ilişki arar. Örneğin, ebeveynlerin boy uzunluğu ile çocukların boy uzunluğu arasındaki ilişki veya ikinci el araba fiyatının enflasyon ile arasındaki matematiksel bağıntı bulunabilir.

---

[1] Şekil 1.2'de iki boyutlu bir veri seti için ikili sınıflandırma görevi kavramsal olarak gösterilmektedir.

En yaygın olarak regresyon analizi, öznitelik değişkenleri verildiğinde, hedef değişkeninin koşullu beklentisini, yani öznitelik değişkenleri sabitlendiğinde hedef değişkeninin ortalama değerini tahmin eder (Ouyang, 2018:14). Bu yaklaşımla denetimli öğrenme teknikleriyle her gözlem, eğitim veri setinden öğrendiklerinden yola çıkarak reel bir değer tahmininde bulunur.

Günümüzde regresyon yöntemlerinde Lineer Regresyon, Çoklu Lineer Regresyon, Polinomal Regresyon, Destek Vektör Regresyonu en sık tercih edilen algoritmalardır.

1.1.1.3 Tahmin

Denetimli öğrenmenin bir diğer alt kategorisi olan tahmin, geçmiş ve şimdiki verilere dayanarak, gelecek hakkında tahminler yapma sürecidir. En yaygın olarak eğilimleri analiz etmek için kullanılır. Yaygın bir örnek, şimdiki ve geçmiş yıllardaki satış verilerine dayanarak gelecek yıl için satış tahmini yapılması olabilir (Lui, 2017:1).

1.1.2 Denetimsiz Öğrenme

Denetimsiz öğrenme, etiketlenmemiş verileri kullanan makine öğrenmesinin diğer bir yaklaşım türüdür. Genellikle veri noktalarını birbirine daha fazla veya daha az benzeyen özellikler belirlemeye çalışarak verileri kümeler veya ortak özellikler gibi özet bir formda temsil etmeye çalışır (The Royal Society, 2017:123).

Denetimsiz öğrenme, denetimli öğrenmenin aksine herhangi bir sınıflandırma veya etiketlenmeye maruz kalmamış bir eğitim verisi ile eğitilmez. Denetimsiz öğrenme yöntemi, eğitilmemiş veriler üzerinden bir korelasyon ve ilişki arar. Bulunan bağıntılar sonucu, birbiri ile ilişkisi olan verileri kendi içinde kategorize eder. Girdi verisinin hangi sınıfa ait olduğu, algoritmalar tarafından sınıflandırma işlemleri ile öğrenilir. Algoritma, daha fazla yeni veriyi değerlendirdikçe, sınıflandırma gücü ve performansı artarak, daha rafine sonuçlar üretir.

Makine öğrenmesinin bu dalı, veri görselleştirme, veri sıkıştırma veya veri dengeleme amacıyla veya eldeki verilerdeki korelasyonları daha iyi anlamak için, herhangi bir hedef değişkenin yardımı olmadan, girdi verilerinin ilgili dönüşümlerini bulmaktan oluşur. Denetimsiz öğrenme, veri analitiğinin ekmeği ve tereyağıdır. Bu

yüzden denetimli öğrenme sorununu çözmeye çalışmadan önce veri kümesini daha iyi anlamak için gerekli bir adımdır (Chollet, 2018:94).

Eğitim verisi kompleks bir yapıda ise veriler için denetimsiz öğrenme teknikleri kullanılmalıdır. Böylelikle veri setinin içindeki karmaşıklığı farklı segmentler üzerinden müdahale etme fırsatı oluşturacaktır. Örneğin farklı gruplardaki müşterileri araçları veya yapıları kendi içinde segmentlere ayırarak, spesifik tetkiklerde bulunmak için kullanılabilir.

K-Means Algoritması, Temel Bileşenler Analizi, Birliktelik Kurallarının Algoritmaları (örn. Apriori Algoritması) ve Hiyerarşik Kümeleme gibi öğrenme algoritmaları denetimsiz makine öğrenmesi algoritmaları arasında en yaygın biçimde kullanılan algoritmalardı. Denetimsiz öğrenim algoritmaları kendi içinde uygulama yapılarıyla temel olarak kümeleme ve boyut azaltma gibi görev yapılarından oluşur.

1.1.2.1 Kümeleme

Kümeleme, keşifsel veri analizi için en yaygın kullanılan yöntemlerden biridir. Sosyal bilimlerden biyolojiye ve bilgisayar bilimlerine kadar tüm disiplinlerde, insanlar veri noktaları arasında anlamlı gruplar belirleyerek verileri hakkında ilk sezgiyi elde etmeye çalışırlar. Örneğin, biyologlar, genleri farklı deneylerde ifadelerindeki benzerliklere dayanarak kümelendirir; perakendeciler, müşterileri hedeflenen pazarlama amacıyla, müşteri profili temelinde kümelendirir ve gökbilimciler, yıldızları uzaysal yakınlıklarına göre kümeler (Shwartz, 2014:307).

Analiz sonrasında ortaya çıkabilecek her küme, belirli bir benzerlik derecesini paylaşan ancak diğer kümelerdeki nesnelere daha benzemeyen bir grup nesneyi tanımlar, bu nedenle kümeleme bazen "denetimsiz sınıflandırma" olarak da tanımlanmaktadır (Raschka, 2015:3).

1.1.2.2 Boyut Azaltma

Denetimsiz öğrenmenin bir diğer alt görev alanı, boyutsallığın azaltılmasıdır. Çoğunlukla, yüksek boyutlu veriler için sınırlı depolama alanı ve algoritmaların hesaplama performansında zorluk oluşturabilecek durumlarda kullanılır (Raschka, 2015:7).

Veri yapısının çok sayıda özniteliğe sahip olduğu durumlarda, karmaşıklığı ortadan kaldırmak için daha düşük boyutlu bir yapı genellikle arzu edilir. Boyutsal azaltma (veya manifold öğrenme) tekniklerine ilişkin temel prosedürler şunlardır:

- Hesaplamalı: Veriler üzerindeki işlemleri hızlandırmak için ilk verileri bir önişleme yöntemi ile sıkıştırmak.
- Görselleştirme: Girdi verilerini iki veya üç boyutlu boşluklara senkronize ederek keşif analizi için verileri görselleştirmek.
- Özellik çıkarma: Daha minimal ve daha güçlü veya daha ergonomik bir özellik/öznitelik kümesi oluşmasını sağlamak (Mohri, 2012:347).

Boyut azaltması, verideki gürültüyü temizlemek için özniteliklerin ön işlemesinde kullanılırken, ilgili bilgilerin çoğunu koruyup, verileri daha küçük boyutlu alt bir alana sıkıştırabilen ve belirli algoritmaların tahmin performansının düşmesini engelleyen yaygın bir yaklaşımdır.

Boyut azaltma amacıyla temel olarak sık kullanılan yöntemler arasında Temel Bileşenler Analizi, Faktör Analizi, Çok Boyutlu Ölçekleme ve Isomap yer almaktadır.

1.1.3 Yarı-Denetimli Öğrenme

Bu yaklaşım özellikle çok sayıda etiketlenmemiş veri olması ve verileri etiketleme maliyetinin oldukça yüksek olduğu uygulamalarda tercih edilir. Adından da anlaşılacağı gibi yarı-denetimli öğrenme, denetimli ve denetimsiz öğrenmenin ortasında yer alır. Aslında, yarı-denetimli öğrenme stratejilerinin çoğu, denetimli veya denetimsiz öğrenmeyi diğer öğrenme paradigmasına özgü ek bilgileri içerecek şekilde genişletmeye dayanır (Zhu, 2009:9).

Yarı denetimli öğrenme, uygun bir işlev veya sınıflandırıcı oluşturmak için etiketlenmiş ve etiketlenmemiş verileri birleştirerek, denetlenen algoritmaların performansını artırmak için önerilen algoritmaların bir çerçevesidir (Design, 2004:251).

Literatürde yarı-denetimli öğrenmeye Maeireizo (2004), "*birlikte eğitim*", Yarowsky (1995), "*kendi kendine eğitim*" ve Nigam (2000), "*üretken modeller*" ile farklı yaklaşımlar önermiştir.

1.1.4 Pekiştirmeli Öğrenme

Pekiştirmeli öğrenmede amaç, çevre ile etkileşimlere dayalı olarak kendi performansını artıran bir sistemi geliştirmektir (Raschka, 2015:6). Bir pekiştirmeli öğrenme sistemi, açıkça öğretilmekten ziyade eylemlerinin sonuçlarından öğrenir. Pekiştirmeli öğrenme diğer öğrenme tekniklerinin aksine, her iterasyon bir önceki iterasyondan geri bildirim alarak modelini sürekli olarak geliştirme döngüsündedir. Diğer bir ifadeyle, eylemlerini geçmiş deneyimlerini dikkate alarak gerçekleştirir. Bu durum tıpkı insan dünyasındaki deneme yoluyla öğrenme paradigmasına benzer. Hatalardan ders çıkartarak öğrenme kolaylaşır çünkü ceza (maliyet, zaman kaybı, pişmanlık, acı, vb.) durumuna düşmekten kaçınılır (Mueller, 2016:169). Nitekim pekiştirmeli öğrenme "tecrübeli öğrenme" olarak ifade edilebilir.

Pekiştirmeli öğrenmenin bir önceki iterasyondan geri bildirim alarak eylemlerini gerçekleştirmesi, muhtemelen en iyi şekilde satranç veya bir video oyunun yapısı ile açıklanabilir. Sanal alanda bir oyuncu, farklı koşullar altında çeşitli eylemlerin sonuçlarını tecrübeler ve oyun alanına daha aşina olur. Öğrenilen bu değerler ile sonraki davranışlarını etkileyerek performansını iyileştirir. Satranç durumunda ise yenilgiden kaçınmak da benzer şekilde olumlu bir ödüle dönüşür (Theobald, 2017:15).

1.2 BANKACILIKTA KREDİ RİSKİ VE ÖNEMİ

BDDK'ya göre kredi riski, kredi müşterisinin yapılan sözleşme gereklerine uymayarak yükümlülüğünü kısmen veya tamamen zamanında yerine getirememesinden dolayı bankanın maruz kalabileceği zarar olasılığı olarak tanımlanmıştır (2012:1). Jorion, "Financial Risk Manager Handbook, Wiley Finance Series" adlı eserinde kredi riski için karşı tarafın sözleşmeden doğan yükümlülüklerini yerine getirememesinden kaynaklanan ekonomik kayıp risk olarak tanımlamıştır (2009:431). Mandacı, "Türk Bankacılık Sektörünün Taşıdığı Riskler ve Finansal Krizi Asmada Kullanılan Risk Ölçüm Teknikleri" çalışmasında kredi riskini, ödenmeme veya geç ödemeden dolayı net kar ve özvarlığın piyasa değerindeki olası değişim olarak ifade etmiştir. (2003: 71). Coyle, "Introduction to Currency Risk" adlı eserinde kredi riski için kredi verenin, borcun ödenmemesi veya geç ödenmesi sonucu zararla karşılaşma olasılığı tanımını yapmıştır

(2000:6). Sinkey'e göre ise kredi riski, "Commercial Bank Financial Management" adlı eserinde borç ödemesindeki belirsizlik olarak ifade edilmiştir (1998: 190).

Yanlış müşteri seçimi, sözleşmedeki eksiklikler, müşterinin mali gücünün sorumluluklarını yerine getiremeyecek kadar yetersiz olması, gelir/borç ödeme dengesine uymayacak kadar yüksek kredi limiti tahsisi, alınan teminatların yetersiz olması ve ekonomik faktörler nedeniyle krediler tahsil edilememe riski taşımaktadır (Bhargava, 2000:8).

Kredi riski, banka ile müşteri arasında yapılan kontrat gereği müşterinin üstlendiği yükümlülüklerini sözleşmede belirlenen süre zarfında eksiksiz yerine getirememe durumudur. Diğer bir ifadeyle, müşterinin bankadan almış olduğu faizli kredinin sözleşme yükümlülüklerine göre zamanında bankaya geri ödeyememe (temerrüde düşme) durumudur.

Bankaların finansal getirileri göz önünde bulundurulduğunda, kredi riski için ayrılan likidite ile diğer risk tipleri için ayrılan likidite arasında yüksek fark vardır. Dolayısıyla banka için kredi riskinin diğer risklere kıyasla daha önem arz ettiği söylenebilir. Kredi riski, bankaların karşılaştığı en büyük risktir (Apostolik, 2009:18).

2008 küresel ekonomik krizi, kredi riskinin kritik öneme sahip olduğuna emsal bir durumdur. Nitekim 2008'de gerçekleşen küresel ekonomik krizinin başlangıcına sebep olduğu düşünülen mortgage piyasasının, yanlış müşteri seçimi ile aniden değer kaybetmesi ve ipotekli satışın yapılmasıyla da kişisel iflasların artmasını tetiklemiştir.

Yanlış kredi risk politikalarının zincirleme etkileri ile ekonomide durgunluktan işsizliğin artmasına kadar kötü sonuçlar doğurabilmektedir. 2008 yılında ABD'nin yanlış kredi risk politikaları sebebiyle küresel ekonomi çökmeye kadar ilerlemiştir. Yapılan yanlış kredi risk politikalarıyla, düşük kredi notuna sahip müşteriler için mortgage kredisi verilerek, temerrüde düşme oranı yüksek olan bu müşterilerin nihayet kredilerini temerrüde düşürmeye başlamıştır. Bankalar, temerrüde düşen müşterilerin mülklerine haciz koyarak, emlak piyasasında fiyatların düşmesine sebep olmuştur. Küresel olarak bankalar, sigorta şirketleri, yatırımcılar ve birçok finansal ve finansal olmayan kurumlar bu fonlara olan yatırımlarıyla büyük mali kayıplar vererek ekonomide durgunluğun yaşanmasına neden olmuştur.

Yanlış kredi risk politikalarının, finansal ve finansal olmayan sektörler üzerinde ciddi izler bırakması sonucu risk ölçümünün önemi ön plana çıkarak, bankacılık risklerine yönelik düzenlemelerle ilgili uluslararası çalışmaların hız kazanmasına sebep olmuştur.

Bankacılık denetim ve düzenleme konularında ortak çalışma imkanları ve uluslararası finansal istikrarın sağlanması amacıyla, G10 ülkelerinin merkez bankası yöneticileri tarafından üye 17 ülke ile birlikte 1974 yılında İsviçre'nin Basel kentinde Basel Komitesi kurulmuştur. 1988 yılında başlatılan Basel I anlaşması ile komite, kredi riskine odaklanmış ve yayınladığı Sermaye Ölçümü ve Sermaye Standartlarının Uluslararası Düzeyde Uyumlaştırılması ile üye ve üye olmayan ülkelerin bankalarının 1992 yıl sonuna kadar birlikte uyum sağlamaları gereken, sermaye yeterlilik rasyosunun minimum %8 oranını koruması gerektiği belirlenmiştir.

Basel I'in sınırlamaları ortadan kaldırmak için komite, Haziran 2004'te Basel II anlaşmasını gerçekleştirmiştir. Basel I'de sadece kredi riskine odaklı düzenlemeler yapılırken, Basel II'de operasyonel ve piyasa riskini de kapsayan yeni bir düzenleme sağlanmıştır. Kredi riskinin tahmini için Basel II düzenlemeleriyle Standart Yaklaşım, İçsel Derecelendirmeye Dayalı Yaklaşım ve Gelişmiş İçsel Derecelendirmeye Dayalı Yaklaşım ile üç ayrı metodoloji tanımlanmıştır.

Basel komitesinin temel amacı ve stratejisi, denetleyici bir yaklaşımdan en iyi uygulamaya yönelik gelişmiş içsel derecelendirmeye dayalı yaklaşıma geçen bankalara sermaye teşvikleri sunmaktır (Brown, 2014:4). Bu bağlamda İçsel Derecelendirmeye Dayalı Yaklaşım (Internal Ratings Based - IRB) ve Gelişmiş İçsel Derecelendirmeye Dayalı Yaklaşım (Advanced Internal Ratings Based – AIRB) versiyonları ile bankalar, kendi iç risk derecelendirmelerini farklı derecelerde geliştirmelerine ve kullanmalarına izin vermektedir.

İçsel Derecelendirmeye Dayalı yaklaşım aşağıdaki dört temel parametreye dayanmaktadır:

i. Temerrüt Olasılığı (Probabilty of Default - PD): Bir kredinin 12 aylık süre zarfında geri ödenmeme olasılığıdır.

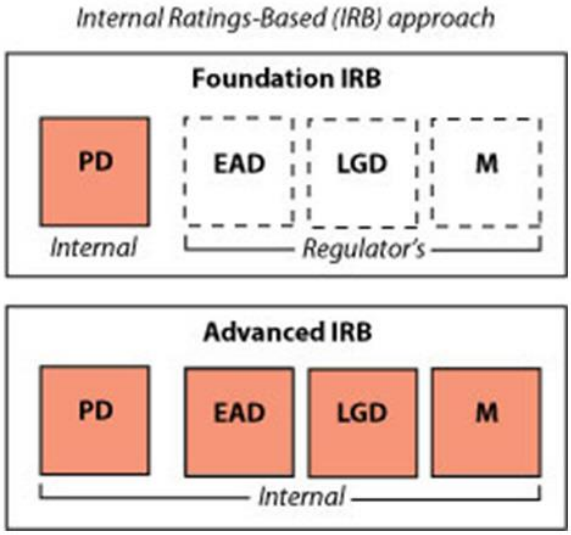
ii. Temerrüt Halinde Kayıp (Loss Given Default - LGD): Borçlunun temerrüde düşmesi halinde maruz kalma yüzdesi olarak ifade edilen tahmini ekonomik kayıptır.

iii. Temerrüt Halinde Risk Tutarı (Exposure at default - EAD): Borçlunun temerrüde düşmesi halinde bankaya ödemek zorunda olduğu beklenen brüt (ekonomik) miktardır.

iv. Vade (Maturity - M): Bir kredinin veya başka bir finansal aracın nihai ödeme tarihine kadar geçen süredir (Brown, 2014:4).

Yukarıdaki parametrelerinin yardımları ile banka, beklenen kredi zararını hesaplayabilmektedir.

$$Beklenen\ Kredi\ Zararı\ (Expected\ loss - EL) = PD \times LGD \times EAD \qquad (1.1)$$



Şekil 1.3: İçsel Değerlendirmeye Dayalı Yaklaşım Türleri (Brown, 2014:4)

Finansal kurumlar için temel ve gelişmiş olarak iki İçsel Değerlendirmeye Dayalı Yaklaşım seçeneği mevcuttur (Basel Bankacılık Denetimi Komitesi, 2001a:34). İki yaklaşım arasındaki fark, parametrelerin banka tarafından ölçülme derecesidir. Temel yaklaşımda (IRB), denetleyici incelemeye tabi olarak banka tarafından sadece PD dahili olarak tahmin edilirken, Gelişmiş IRB yaklaşımında, dört parametrenin tamamı banka tarafından hesaplanacak ve denetim incelemesine tabi tutulacaktır (Schuermann, 2004:3).

Çalışmaya konu olan Temerrüt Olasılığı parametresi, 12 ay içinde borçlunun temerrüde düşme olasılığı incelemektedir. Bu bağlamda temel yaklaşım olan İçsel Değerlendirmeye Dayalı Yaklaşım esas alınarak PD tahmini hesaplanacaktır.

Çalışma kapsamında, kredi başvurusunda bulunan her bir müşteri için gelecek 12 aylık süre zarfında, temerrüt olasılığının tahminlemesi için kullanılan algoritmalar arasında performans ölçülerinin yardımıyla en uygun model ve tekniğin belirlenmesi amaçlanmıştır.

1.3 LİTERATÜRDE KREDİ RİSK ANALİTİĞİ

Geçen yüzyıla kadar uzanan kredi risk analitiği üzerine yapılan araştırma ve geliştirmeler, günümüzde de finans alanında kritik öneme sahip araştırma konusu olmaya devam etmektedir. Küresel mali krizler sonucunda önem kazanan düzenleyici odaklar sebebiyle, kredi risk analitiği süreci akademik ve iş dünyasında da rağbet görmeye devam etmektedir.

Kredi risk analitiğindeki genel yaklaşım, geçmiş ve şimdiki müşteriye ait özellikler ve potansiyel başarısızlıkları arasındaki ilişkiyi analiz ederek sınıflandırmaktır. Bu bağlamda yeni başvuru sahiplerinin veya mevcut müşterilerin iyi veya kötü olarak sınıflandırılmasında uygulanabilecek sınıflandırıcıların belirlenmesi için kullanılabilir (Wang, 2005:820).

Geleneksel olarak kredi risk analitiğinde temerrüt olasılığının tahmininde Lojistik Regresyon ve Diskriminant Analizi gibi teknikler kullanılmaktadır. Destek Vektör Makineleri, kredi kartı müşterilerinin sınıflandırılmasında kimin temerrüde düşeceğinin hesaplamasında başarılıdır. Ayrıca test edildiğinde ve geleneksel tekniklerle kıyaslandığında temerrüt riskini belirlemede en önemli özellikleri keşfetmede rekabetçi oldukları bulunmuştur (Bellotti, 2009:3302).

Destek Vektör Makinelerinin kredi puanlamasında önemli ölçüde daha iyi sonuçlar verdiği gösterilmiştir (Gestel, 2003:11). Destek Vektör Makineleri, regresyon modelinden önemli ölçüde daha iyi performans göstermiştir (Yao, 2017: 687).

Kredi puanlama tekniklerinde sınıflandırıcı algoritmaların Lojistik Regresyondan daha önemli düzeyde iyi performanslar gösterdiği görülmüştür. Ayrıca, Yapay Sinir

Ağları kredi puanlama veri setlerinde, aşırı öğrenen makinelerden daha iyi performans gösterdiği bulunmuştur (Lesssmann, 2015:124).

Kredi skorlamada Diskriminant Analiz, Lojistik Regresyon, Yapay Sinir Ağları, sınıflandırma ağaçları ve bayes sınıflandırıcı gibi birçok algoritmadan yararlanılmaktadır. Yapay Sinir Ağlarının diğer beş yönteme göre daha doğru sonuçlar gerçekleştirdiği görülmüştür (Yeh, 2009:2479).

1994 yılında Altman ve meslektaşları, geleneksel istatistiki stres ve iflas tahmini ile alternatif bir sinir ağı algoritması arasında ilk karşılaştırmalı analiz yöntemini gerçekleştirerek, iki yöntem için birleşik yaklaşımın doğruluğunun önemli ölçüde artırdığını saptadılar (1994:527).

Zhou ve Wang daha iyi tahmin için karar ağaçlarına ağırlık tahsis etmeyi önermektedir (2012:1523). Hamori ve arkadaşları, PD analizinde sinir ağı yöntemleriyle Torbalama (Bagging), Rassal Orman ve Artırma (Boosting) ile tahmin doğruluğu ve sınıflandırma yeteneğini incelemiş ve karşılaştırmıştır. Çalışmada makine öğrenme algoritmaları arasında Artırmanın daha iyi performans sağladığını buldular (2018:12).

Temerrüt olasılığının düşük olduğu portföyler, düşük risk olarak kabul edilirken, temerrüde düşen sınıflar arasında bir dengesizlik problemi ile karşılaşılabilir. Sınıf dengesizliği oluşturan portföyler için Gradyan Artırma ve Rassal Karar Ormanları sınıflandırıcı tekniklerinin iyi performans gösterdiği bulunmuştur (Brown, 2012:3453).

Doğruluk oranı söz konusu olduğunda K-En Yakın Komşu, Rassal Orman ve Yapay Sinir Ağları algoritmaları iyi performans gösterir (Zhang, 2017:372).

Torbalama, Artırma ve Rassal Orman benzer prosedürleri içermesine rağmen, Rassal Orman genellikle daha iyi doğruluk ve hata oranları üretmiştir (Barboza, 2017:415).

Kavcıoğlu, kurumsal kredileri skorlamada klasik yöntemler ile yapay sinir ağlarını karşılaştırarak, eğitim veri setinde yapay sinir ağlarının lojistik regresyona kıyasla daha başarılı sonuçlar ürettiğini saptamıştır. Verinin boyutu ve kalitesini dikkate alarak Yapay Sinir Ağları gibi makine öğrenmesi algoritmalarının daha iyi performans gösterdiği bulgusuna ulaşılmıştır (2019:241).

Yeşilyurt ve Şeker, kredi skorlama algoritmalarının karşılaştırmaları için yapmış oldukları literatür araştırmaları sonucu elde edilen algoritma karşılaştırma tablosuna aşağıda yer verilmiştir.

Tablo 1.1: Algoritma Başarılarının Karşılaştırılması (Yeşilyurt, 2018:11)

YSA (Yapay Sinir Ağları)			>	Karar Ağacı
C4.5 (Karar Ağacı)			>	YSA
Lojistik Regresyon			>	Çoklu Diskriminant Analizi
DVM (Destek Vektör Makineleri)			>	Lojistik Regresyon
YSA			>	Lojistik Regresyon
YSA	≅	Lojistik Regresyon	>	Doğrusal Diskriminant Analizi
YSA	>	Genetik Programlama	>	DVM

Demirbulut ve meslektaşları, istatistiksel ve makine öğrenmesi algoritmalarıyla kredi skorlama yöntemlerini ele alarak karşılaştırma analizleri yapmışlardır. Sınıflandırma başarısı AUC (Area Under Curve) değeriyle ölçülerek, YSA modelinin en başarılı algoritma olduğu bulgusuna ulaşılmıştır (2017:283).

Literatürde kredi temerrüt riskinin skorlanması için birçok farklı istatistiksel ve makine öğrenmesi algoritmalarının karşılaştırmaları mevcut olmakla birlikte, bu çalışmalar bulgularında en başarılı algoritmalar, sektör bilgisiyle incelenerek, çalışmaya konu olan karşılaştırma algoritmaları belirlenmiştir.

İKİNCİ BÖLÜM

METODOLOJİ

2.1 ÇALIŞMADA KULLANILAN SINIFLANDIRMA ALGORİTMALARI

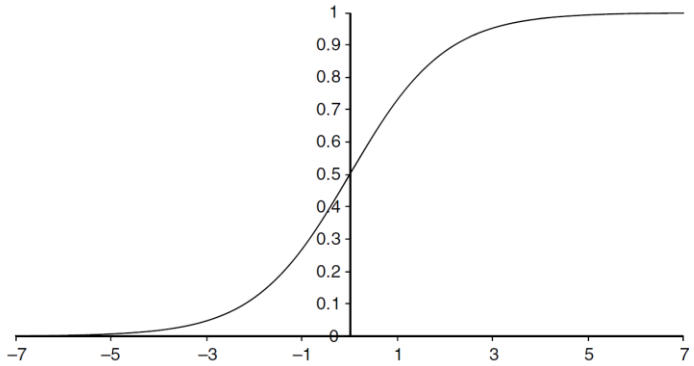
Çalışmanın bu bölümünde, temerrüt riskinin sınıflandırılması kapsamında, denetimli makine öğrenmesi algoritmalarından sınıflandırma amacıyla kullanılanlar ele alınacak olup, ilgili algoritmaların sınıflandırma yeteneklerinin arkasında bulunan teknik farklılıklara değinilecektir.

2.1.1 Lojistik Regresyon (Logistic Regression)

Lojistik Regresyon, hedef değişkeninin olası iki değer alması (dikotom) durumunda, öznitelik değişkenlerinin bir lojistik fonksiyon olarak tanımlanması ve hedef değişkeni ile arasındaki ilişkinin regresyon analizi yardımıyla incelenerek, hedef değişkeni için sonucun olasılık değerinin logaritmasının tahmin edilmesidir. Öznitelik değişkenlerinin lojistik fonksiyon olarak tanımlanması için aşağıdaki sınırlayıcı fonksiyondan yararlanılmaktadır.

$$\pi = \frac{1}{1 + e^{-\beta.x}} \tag{2.1}$$

Her olası π değeri için sonuç her zaman Şekil 2.1'deki gibi 0 ile 1 aralığında sınırlı bir olasılık olacaktır.



Şekil 2.1: Lojistik Regresyon Sınıflandırma Grafiği

Bu bağlamda, regresyon denklemi sınırlayıcı fonksiyon yardımı ile aşağıdaki gibidir.

$$logit\ (\pi) \equiv \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n = \beta . x^T \qquad (2.2)$$

Lojistik Regresyon denkleminde, π bir olayın gerçekleşme olasılığını, β_0 ve β_n model sabit katsayısı ve öznitelik değişkenlerine ait model katsayılarını, x_n regresyon modelinde yer alan girdi özniteliklerine ait değerleri ifade eder.

Bir olayın gerçekleşme olasılığının gerçekleşmeme olasılığına olan oranı $(\frac{\pi}{1-\pi})$ diğer bir ifadeyle odds oranı değeri $(0, +\infty)$ arasındaki değerleri aldığı için $log(\frac{\pi}{1-\pi})$ dönüşümü uygulanarak $(-\infty, +\infty)$ arasındaki değerler alması sağlanır. Eşitlikteki her iki tarafın üstel fonksiyonu alınarak (2.7)'daki olasılık fonksiyonuna ulaşılır.

$$log\left(\frac{\pi}{1-\pi}\right) = \beta . x \qquad (2.3)$$

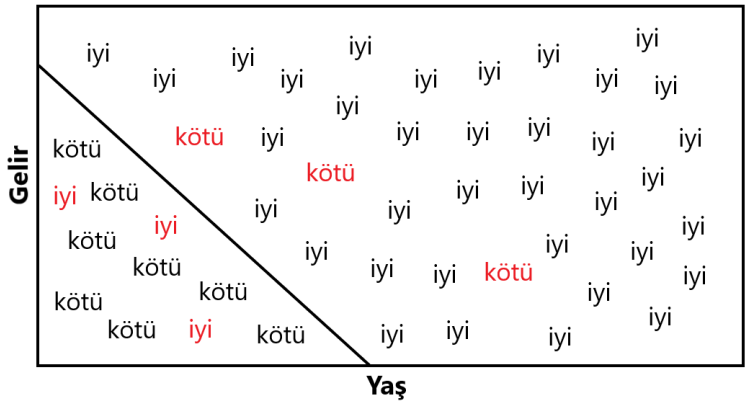
$$\left(\frac{\pi}{1-\pi}\right) = e^{\beta . x} \qquad (2.4)$$

$$\pi = (1-\pi)e^{\beta . x} \qquad (2.5)$$

$$\pi(1 + e^{\beta . x}) = e^{\beta . x} \qquad (2.6)$$

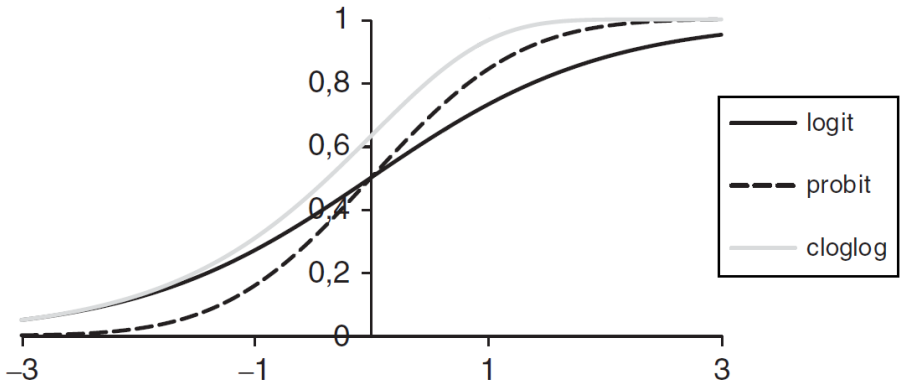
$$\pi = \frac{e^{\beta . x}}{1 + e^{\beta . x}} = \frac{1}{1 + e^{-\beta . x}} \qquad (2.7)$$

Log dönüşümü oranları (logit) doğrusal yapıda olduğu için iki sınıfa ait ayrışımı Şekil 2.2'deki gibi doğrusal bir karar sınırı ile tahmin eder. Burada "iyi" temerrüde düşmeyen müşterileri temsil ederken, "kötü" temerrüde düşen müşterileri ifade etmektedir.



Şekil 2.2: Lojistik Regresyon Karar Sınırı

Yukarıda bahsedilen logit dönüşümüne alternatif olarak literatürde probit ve cloglog dönüşümleri de önerilmektedir.



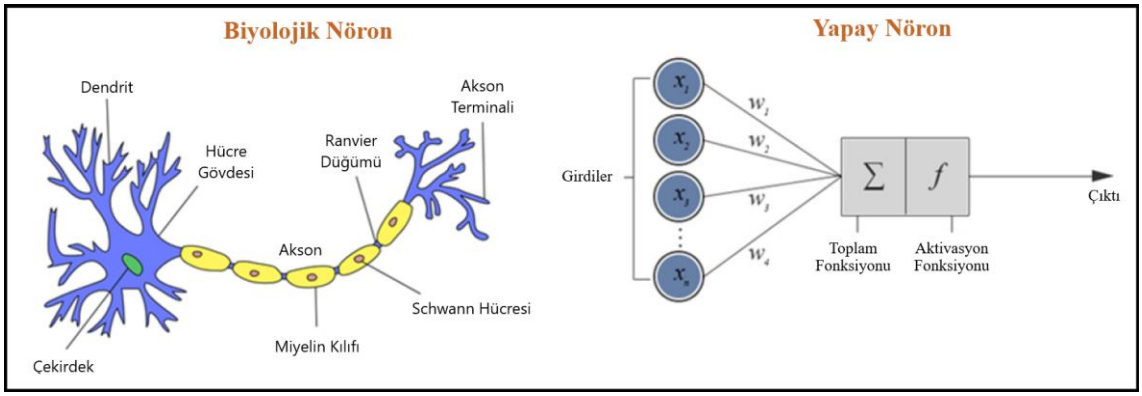
Şekil 2.3: Popüler Dönüşümler

Lojistik Regresyonun yapısı lineer regresyona benzerlik gösterse de teknik olarak ayrık sınıfları öngören bir sınıflandırma aracıdır. Bu bağlamda, Lojistik Regresyon, özellikle tüketici kredisi endüstrisinde yaygın olarak kullanılan bir algoritmadır (Hand, 2009:1541).

Temerrüt riskinin tahmininde, bir müşterinin kredi ödemelerindeki davranış yapısının iyi veya kötü ödeme durumuna bağlı olarak ikili durumu incelenir. Bu iki yanıt modeli için, hedef değişken y iki olası değerden birini alabilir: müşteri kötü bir ödeyici ise y=1; iyi bir ödeyici ise y=0 (Brown, 2014:30).

2.1.2 Yapay Sinir Ağları (Neural Network)

Yapay sinir ağları, hedef ve öznitelik değişkenleri arasındaki ilişkiyi analiz katmanları aracılığıyla işlemek için daha esnek bir tasarım sunan bir makine öğrenme algoritmasıdır. Temel olarak girdi katmanı, gizli (ara) katman ve çıktı katmanından oluşan bir yapıdadır. YSA, biyolojik sinir ağlarını taklit eden sentetik yapılardır (Eğrioğlu, 2009:10590). Dolayısıyla belirlenmiş bir modelin parametrelerini tahmin etmek yerine, beyin gibi biyolojik sinir sistemlerinin bilgi işlem biçiminden ilham alınarak tasarlanmıştır. Bu tasarım esasen insan beyinleri değil, hayvan beyinlerinin paralel mimarisi dikkate alınarak modellenmiştir (Bell, 2014:91).



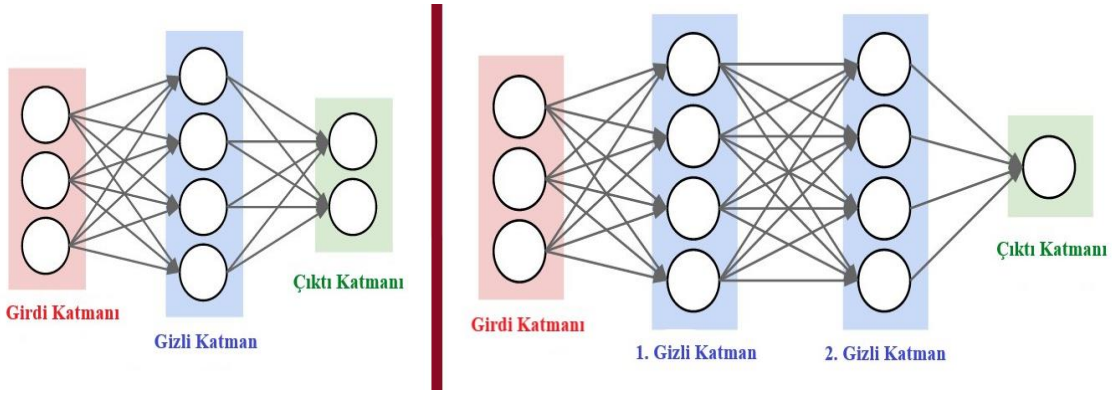
Şekil 2.4: Biyolojik Sinir Ağı ve Yapay Sinir Ağı Görseli (Dangeti, 2017:241)

Biyolojik Sinir Ağları ile Yapay Sinir Ağları arasındaki terminolojiler aşağıdaki gibidir;

Tablo 2.1: BSS'nin YSA'daki Terminolojik Karşılıkları (Öztürk, 2018:28)

Biyolojik Sinir Sistemi (BSS)	Yapay Sinir Ağı (YSA)
Nöron	İşlem Elemanı
Dentrit	Toplama Fonksiyonu
Hücre Gövdesi	Aktivasyon Fonksiyonu
Akson	Eleman Çıkışı
Sinaps	Ağırlıklar

Yapay Sinir Ağlarının yapısı, girdiler, ağırlıklar, toplam fonksiyonu, aktivasyon fonksiyonu ve çıkış fonksiyonu olarak beş bölümden oluşmaktadır. YSA üzerindeki her bir nöron, bir işleme alınma durumunu ifade etmektedir. Ağın öğrenme işlemini gerçekleştirmek için dışarıdan veya gereksinime göre diğer hücrelerden nöronlara gelen girdiler (X_j), işlenerek bir sonraki nöron için çıkış değerini iletir. Girdilerin çıktı üzerindeki etkisinin hesaplanabilmesi için her bir girdinin eğitim sırasında geldikleri bağlantıların ağırlığıyla çarpılarak bir parametre (ağırlık/W_j) üretilir. Çıkan sonuç, girdi değerlerinden bağımsız ve modelin fit edilmesine yardımcı olan bias (b) değişkeni ile toplanarak, Tablo 2.2'de bulunan toplama fonksiyonları (toplam, çarpım, maksimum, minimum, vd.) ile beslenir. Toplama fonksiyonları aracılığıyla, elde edilen net girdi, Tablo 2.3'de bulunan aktivasyon fonksiyonları (sigmoid, tanjant hiperbolik, doğrusal, relu, vd.) ile beslenerek bir veri çıktısı elde edilir.



Şekil 2.5: Solda Tek Gizli Katmanlı ve Sağda Çok Katmanlı Sinir Ağı Yapısı

Bir gizli katmandaki nörona ait çıktının fonksiyonel olarak gösterimi aşağıdaki gibidir;

$$h_i = f(b_i + \sum_{j=1}^{n} w_{ij} x_j) \tag{2.8}$$

(2.8) denkleminde h_i gizli katman nöronuna ait çıktıyı, b_i bias değerini, w_{ij} girdi değişkenlerine ait ağırlıkların gizli nöron i ile bağlantılı ağırlık matrisini, x_j her bir girdi değerini ve f aktivasyon fonksiyonunu ifade eder.

Tablo 2.2: Toplama Fonksiyonları Örnekleri (Öztemel, 2006:50)

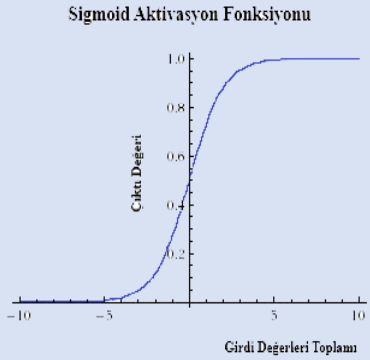
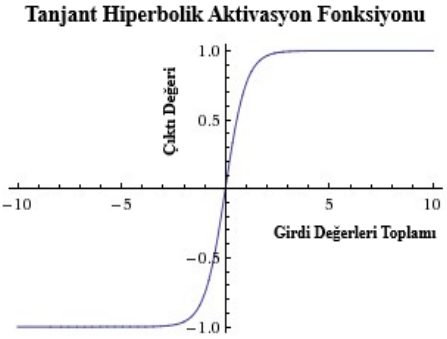
Toplama Fonksiyonu		Açıklama
Toplam	$net = \sum_{j=1}^{n} w_j x_j$	Ağırlık değerleri ile girdi değerlerinin çarpımlarından sonra hesaplanan her bir değerin birbirleriyle toplanması ile net girdinin elde edilmesidir.
Çarpım	$net = \prod_{j=1}^{n} w_j x_j$	Ağırlık değerleri ile girdi değerlerinin çarpımlarından sonra hesaplanan her bir değerin birbirleriyle çarpılması ile net girdinin elde edilmesidir.

Maksimum	$net = \max(w_j x_j)$	Ağırlık değerleri ile girdi değerlerinin çarpımlarının sonucu en büyük çarpım değeri net girdi olarak kabul edilir.
Minimum	$net = \min(w_j x_j)$	Ağırlık değerleri ile girdi değerlerinin çarpımlarının sonucu en küçük çarpım değeri net girdi olarak kabul edilir.
Çoğunluk	$net = \sum_{j=1}^{n} Sgn(w_j x_j)$	Ağırlık değerleri ile girdi değerlerinin çarpımlarından sonra bulunan pozitif ile negatif değerlerin sayısı arasından büyük olan sayının net girdi olarak kabul edilmesidir
Kümülatif Toplam	$net = (eski) + \sum_{j=1}^{n} w_j x_j$	Daha önce toplama işlemiyle hesaplanan girdi değerlerine yeni hesaplanan değerlerin de eklenerek kümülatif olarak toplanmasıyla elde edilen net girdi değeridir.

Çalışmaya konu olan temerrüt riski tahmininde net girdinin hesaplanması için Tablo 2.2'de bulunan toplama fonksiyonları arasından ağırlıklı olarak kullanılan toplam fonksiyonu kullanılmıştır.

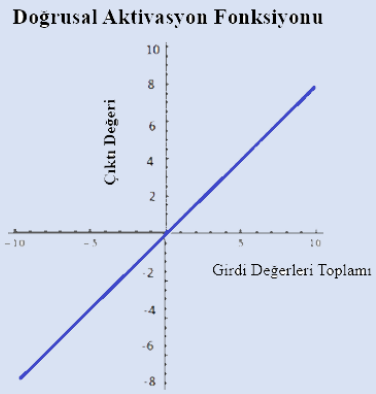
Çıkış katmanındaki aktivasyon fonksiyonu ile bir yanıt olasılığı elde etmek için Tablo 2.3'de bulunan aktivasyon fonksiyonları arasından ikili bir tahminleme olan sigmoid aktivasyon fonksiyonu kullanılmıştır.

Tablo 2.3: Bazı Aktivasyon Fonksiyonları

Aktivasyon Fonksiyonu	Açıklama
Sigmoid Fonksiyonu $$f(net) = \frac{1}{1 + e^{-net}}$$ 	Doğrusal olmayışı sebebiyle YSA uygulamalarında sıkça kullanılan bir aktivasyon fonksiyonudur. Sürekli ve türevi alınabilir bir fonksiyon olmasıyla beraber kayıp aktivasyon değeri oluşturmaz. Fakat gradyan kaybı problemi mevcuttur. Aralık değeri (0,1) arasındadır.
Tanjant Hiperbolik Fonksiyonu $$f(net) = \frac{e^{net} - e^{-net}}{e^{net} + e^{-net}}$$ 	Sigmoid fonksiyonuna benzer bir yapıya sahiptir. Ancak fonksiyon aralığı (-1,1) arasındadır. Dolayısıyla daha çok değer alarak türevinin daha dik bir yapıda olmasına, böylelikle daha hızlı öğrenme ve sınıflandırma işlemi gerçekleştirmesine sebebiyet vermektedir. Sigmoid fonksiyonu gibi gradyan kaybı vermektedir. Aralık değeri (-1,1) arasındadır.

Doğrusal Fonksiyon

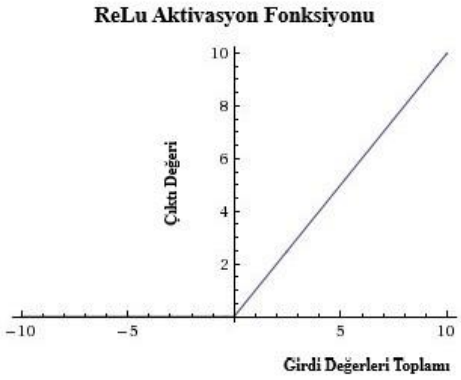
$$f(net) = net$$



Doğrusal problemler çözmek amacıyla kullanılan bu fonksiyon, sigmoid fonksiyonu gibi ikili değerler üretmeyerek birden fazla çıkışa izin verir. Ancak türevi sabit olduğu için modelin eğitiminde gerçekleştirilen geriye yayılıma (backpropagation) izin vermemektedir. Dolayısıyla giriş nöronlarındaki hangi ağırlıkların daha iyi bir tahmin sağlayabileceğini anlamak mümkün değildir. Aralık değeri (-∞, ∞) arasındadır.

ReLu

$$f(net) = \begin{cases} net < 0 \; ise \; 0 \\ net \geq 0 \; ise \; net \end{cases}$$



(Rectified Linear Unit – ReLu) Doğrultulmuş Lineer Birim pozitif eksende doğrusal fonksiyon ile aynı özelliklere sahip gibi görünse de aslında doğrusal olmayan bir yapıdadır. Fonksiyon aralıkları itibariyle negatif değerleri sıfıra dönüştürerek verilerin eğitiminin azalmasına ve böylelikle öğrenmenin zayıf kalmasına neden olabilmektedir. Sigmoid ve Tanjant Hiperbolik fonksiyonlardaki gibi gradyan kaybı vermemektir. Aralık değeri [0, ∞) arasındadır.

Toplam fonksiyonu ve aktivasyon fonksiyonu seçimleri sonrası nihai YSA denklem yapısı aşağıdaki gibidir.

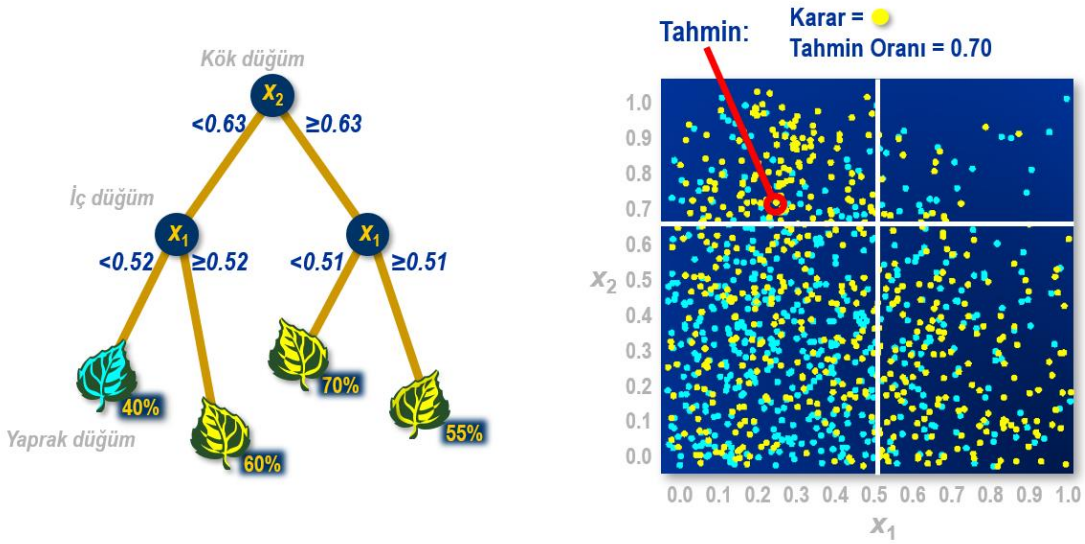
$$\pi = f^{sigmoid}(b + \sum_{j=1}^{n_h} v_j h_j) \qquad (2.9)$$

Denklem (2.9)'de; π çıktı değerini, $f^{sigmoid}$ sigmoid (lojistik) fonksiyonunu, b bias sabitini, n_h gizli katmandaki nöron sayısını, v_j ağırlık vektörünü, h_j gizli nöronun çıktı değeri üzerindeki ağırlığını temsil eder.

2.1.3 Karar Ağacı (Decision Tree)

Denetimli bir öğrenme tekniği olan karar ağaçları hem sınıflandırma hem de regresyon için kullanılabilen, ağaç benzeri bir yapıda düzenlenmiş bir dizi kural tabanlı hiyerarşik bir sınıflandırma algoritmasıdır. Karar ağaçları, heterojen yapıdaki nicel veya nitel veri setlerini, hedef değişkene dayalı homojen alt gruplara ayırarak hedef değişkene ait değerlerin tahminini gerçekleştiren, uygulanabilir bir model kurmayı hedefler. Örneğin, bir kredi riski vaka çalışmasında, kredi başvurusunda bulunan müşteriye ait borç, yaş, gelir ve medeni durum bilgileri bulunabilir. Karar ağacı, her bir başvuranı iyi veya kötü bir kredi riski olarak tahmin edebilen (sınıflandırabilen) bir dizi metin kuralı veya grafiksel bir ağaç olarak bir model oluşturur (Zhang, 2002:11).

Karar ağaçları için grafik gösterimi son derece sezgisel olduğundan, kullanıcılar diğer algoritmalara göre karar ağacındaki verileri kolaylıkla inceleyebilir ve yorumlayabilir.



Şekil 2.6: Karar Ağacı Örneği

Şekil 2.6'deki gibi karar ağaçları temel olarak kök karar düğümü, iç karar düğümleri (üst ve alt düğümler) ve yaprak düğümünden oluşmaktadır. Bir başlangıç noktası görevi gören kök karar düğümü, belirlenen kritere (karar kuralı) göre dallanarak alt gruplara ayrılır ve iç karar düğümlerini oluşturur. İç karar düğümleri içerisindeki veriler homojen olana kadar dallanma işlemine devam edilerek alt gruplara ayrılır. Nihai uç yaprağa ulaşıldığında test işlemi sona ererek bir çıktı elde edilir.

Dolayısıyla karar ağaçlarının oluşumundaki eğitim süreci kural indüksiyon algoritmalarına benzeyen bir tür tümevarım yöntemidir. Diğer bir ifadeyle, veri setinin hedef değişken üzerinden sahip olduğu tüm kombinasyonları için kural indüksiyonları yardımıyla gerçekleştirdiği bir tümevarım yöntemidir.

Her bir karar düğümünde kullanılan bölme kriteri, verimli bir karar ağacının dizayn edilmesi için oldukça önem arz etmektedir. Bu bağlamda karar düğümlerindeki dallanmalar, verilerin safsızlığına göre belirlenir. Safsızlık, her bir düğümdeki verilerin, her sınıfa ne kadar iyi böldüğünü gösteren bir ölçüdür. Safsızlığı ölçmede kategorik hedefler için genellikle Entropi ve Gini ölçüleri kullanılırken, sürekli hedefler için varyans azaltma veya F testine ait p-değeri gibi ölçüler kullanılır. İlgili çalışma dahilinde kullanılan hedef değişkenin kategorik olması sebebiyle Entropi ölçüsü incelenecektir.

Bilgi teorisinden gelen ve farklı sınıflar arasındaki verilerin varyans ölçüsünü açıklayan entropi denklem (2.10)'da gösterilmektedir.

$$Entropi(S) = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{2.10}$$

Shannon'un (2.10)'daki entropi denkleminde S entropiyi, n sınıf sayısını ve p_i gerçekleşebilecek her bir sınıfa ait olasılığı temsil eder. Bir karar düğümünde bulunan veriler %100 homojen olarak sınıflandırıldıysa, entropi değeri sıfır olacaktır. Ancak sınıflandırma eşit yapıda bölünmüş ise entropi seviyesi 1 olacaktır. Nitekim entropi seviyesinin düşük değerde olması, sınıflandırmanın iyi olduğu göstergesidir.

Entropinin alternatifi olan Gini safsızlığı ise bir yanlış sınıflandırma ölçüsüdür. Diğer bir ifadeyle sınıflandırmanın heterojenliğini ölçen Gini safsızlığı aşağıdaki formül ile hesaplanmaktadır.

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2 \tag{2.11}$$

Burada i sınıf sayısını ve p_i sınıflandırma olasılığını temsil eder.

Gini safsızlığında logaritmik hesaplama uygulanmadığı için entropiye göre daha hızlı hesaplama olanağı sunar ve 0 ile 1 arasında değerler alır. Yanlış sınıflandırma oranı

yükseldikçe bu değer 1'e yaklaşırken, tüm öğelerin belirli bir sınıfa ait olduğu veya yalnızca bir sınıfa sahip olduğu zaman bu oran 0 olacaktır. Eğer sınıflar eşit olarak dağıtılmış ise Gini safsızlığı 0.5'i gösterecektir. Nitekim Gini değerinin düşük olması, doğru sınıflandırma oranının yüksek olduğu anlamına gelmektedir.

Entropi ve Gini gibi safsızlık ölçülerinin sınıflandırmadaki önemi dikkate alındığında, ağacın başlangıç düğümü olan kök karar düğümünde konumlanacak özelliğin belirlenmesi, ağacın verimliliği için oldukça önemlidir. Bu bağlamda kök düğümünde konumlanacak özelliğin belirlenmesi için veri setinin bir özellik üzerinde en iyi sınıflandırmayı sağlayan, bilgi kazancı değerine ihtiyaç vardır. Bilgi kazancı ise tüm popülasyona ait safsızlık (Entropi veya Gini) değeri ile alt düğümlerde bulunan veri kümesinin özelliklerine ait safsızlık değeri arasındaki farktır. Entropiye dayalı bilgi kazancına ilişkin denklem aşağıdaki gibidir.

$$Bilgi\ Kazancı\ (D_p, f) = S(D_p) - \sum_{i=1}^{n} \frac{N_j}{N} S(D_j) \tag{2.12}$$

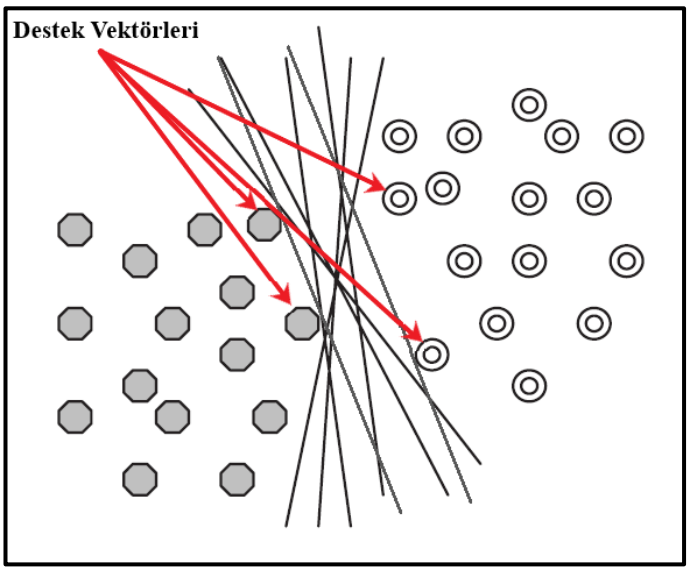
(2.12)'deki denklemde f bölünmeyi gerçekleştiren özelliği, D_p üst düğüme ait veri kümesini, D_j j'ninci alt düğüme ait veri kümesini, S entropiyi, N toplam gözlem sayısını, N_j j'ninci alt düğüme ait gözlem sayısını ifade etmektedir.

Her özellik için hesaplanan bilgi kazanç değeri arasından en yüksek olan özellik, kök düğüm olarak kabul edilir. Verimli bir ağacın dizayn edilmesi için safsızlık ve bilgi kazancı ölçüleri ile Karar Ağacı eğitilirken, kök düğümden son bölünmeye kadar safsızlık seviyesinin düşürülmesi hedeflenir.

2.1.4 Destek Vektör Makineleri (Support Vector Machine)

Yüksek boyutlara sahip veri türleri için sınıflandırma veya regresyon analizini gerçekleştirmede ekstra avantajlara sahip denetimli öğrenme yöntemi olan Destek Vektör Makineleri, iki sınıfın optimal bölünmesinde karar fonksiyonunun tahmin edilmesi, diğer bir ifadeyle n boyutlu bir uzayda optimal bir sınıflandırma için hiperdüzlemin belirlenmesine dayanmaktadır. Destek Vektör Makineleri, doğrusal ayrılabilen ve doğrusal ayrılamayan veri türleri için farklı prensipler sunmakla birlikte, ilgili çalışma kapsamında kullanılan doğrusal ayrılabilen DVM incelenecektir.

Destek Vektör Makinelerinde veri kümelerini ayırmak için Şekil 2.7'teki gibi sonsuz hiperdüzlemler çizmek mümkündür. Ancak hangi hiperdüzlemin en ideal sınıflandırma koşulunu sağladığını belirlemek için hiperdüzlemler ile destek vektörleri arasındaki mesafe incelenir. Bu mesafe marj olarak adlandırılıp, marjın maksimum olduğu hiperdüzlem, optimum hiperdüzlem olarak kabul edilir.



Şekil 2.7: İki Sınıflı Bir Problem için Hiperdüzlemler (Kavzoğlu, 2010:76)

DVM ile sınıflandırmalar genellikle -1 ve +1 etiketleri ile iki ayrı sınıfı temsil etmektedir. Bu bağlamda, iki sınıflı doğrusal olarak ayrılabilen bir sınıflandırma işleminde, eğitim verisini (x_i, y_i) en iyi şekilde ayıracak hiperdüzlemin bulunması kapsamında sınırlara ait eşitsizlikler aşağıdaki gibidir:

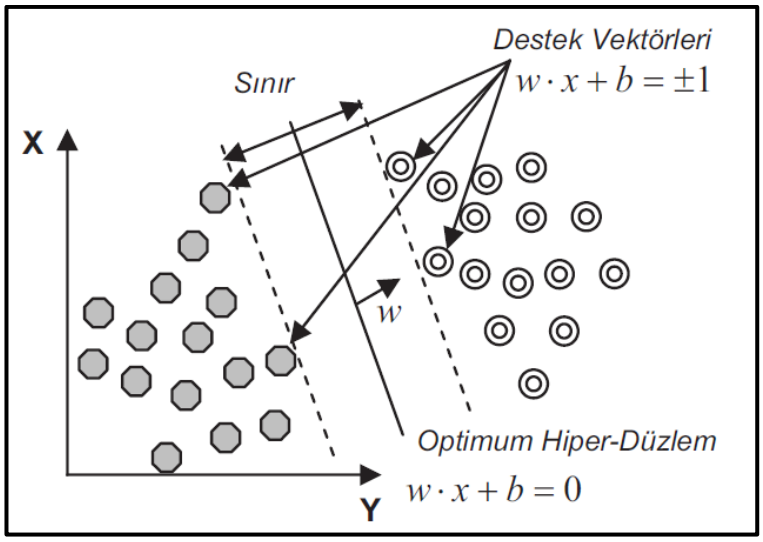
$$w \cdot x_i + b \geq +1, \quad her \quad y_i = +1 \ için \tag{2.13}$$

$$w \cdot x_i + b \leq -1, \quad her \quad y_i = -1 \ için \tag{2.14}$$

Burada $x \in R^n$ olup n-boyutlu girdi vektörünü, $y_i \in \{-1, +1\}$ sınıf etiketlerini, w ağırlık vektörünü ve b eğilim değerini göstermektedir. Bu eşitsizlikler, Eşitlik (2.15)'te verildiği gibi tek bir eşitsizlikte birleştirilebilir (Ayhan, 2014:180).

$$\forall i \ için \ y_i(\langle w, x_i \rangle + b) - 1 \geq 0 \tag{2.15}$$

Destek vektörleri ve eşitsizliklerin yardımı ile birbirine paralel olarak sınırlayıcı hiperdüzlemlerin belirlenmesi sağlanarak, Şekil 2.8'daki gibi doğrusal olarak ayrılabilen optimum hiperdüzlem elde edilir.



Şekil 2.8: Doğrusal Olarak Ayrılabilen Veri Setleri için Hiper-Düzlemin Belirlenmesi (Kavzoğlu, 2010:76)

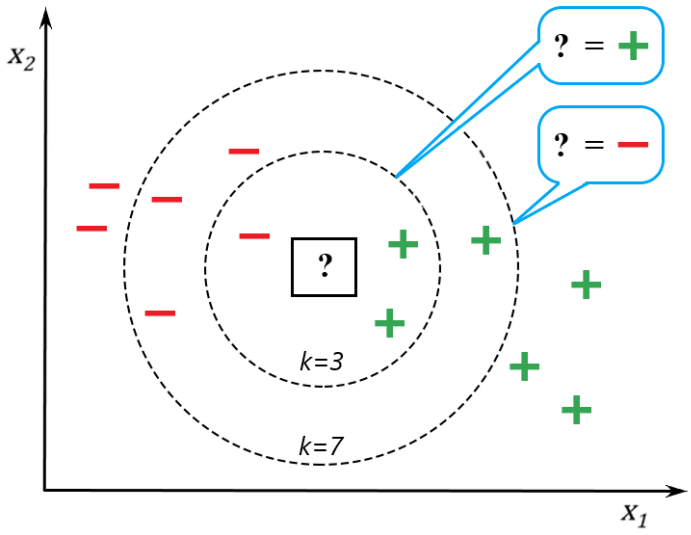
Sonuç olarak, doğrusal olarak ayrılabilen iki sınıflı bir problem için karar fonksiyonu aşağıdaki şekilde yazılabilir (Osuna, 1997:8).

$$f(x) = sign\left(\sum_{i=1}^{n} \lambda_i y_i (x \cdot x_i) + b\right) \tag{2.16}$$

2.1.5 K-En Yakın Komşu (K-Nearest Neighbors)

K-En Yakın Komşu algoritması, sınıflandırma ve regresyon için kullanılan parametrik olmayan bir denetimli öğrenme tekniğidir. Eğitim sürecini verilerin kaydını tutarak gerçekleştirdiğinden diğer makine öğrenme yöntemlerine göre daha tembel bir öğrenme tekniğine sahiptir. Basit yapılarına rağmen, uydu görüntüleri, el yazıları, elektrokardiyogram görüntüleri gibi birçok sınıflandırma çalışmalarında başarılı sonuçlar elde etmektedir. Algoritma temel olarak, eğitim verilerinin özellik uzayındaki belirli bir k hiperparametresine olan uzaklığına göre oluşturulan sınıflandırmanın, örnek tabanlı öğrenmesine dayanmaktadır.

Diğer bir ifadeyle, K-En Yakın Komşu algoritması, “Bana arkadaşını söyle, sana kim olduğunu söyleyeyim.” veya “Üzüm üzüme baka baka kararır.” atasözlerindeki nesne veya insanların, birbirine olan yakınlığından kaynaklanan yapısal benzeşme durumunun mantığını benimsemektedir.



Şekil 2.9: K-En Yakın Komşu Örneği

K-En Yakın Komşu algoritmasının optimum sonuç üretmesi için en uygun uzaklık matrisinin hesaplanmasına ve optimal sınıflandırmayı sağlayacak komşu sayısına (k hiperparametresine) ihtiyaç vardır. Eğitim veri setindeki gözlemler arasındaki mesafenin tanımlanması ve optimal uzaklık matrisinin hesaplanması için Öklid, Manhattan, Minkowski, Mahalanobis, Kosinüs, Jaccard, Hamming gibi yöntemler kullanılmaktadır. Yaygın kullanımı olan Öklid uzaklığı, iki gözlem arasındaki doğrusal uzaklığı ifade edip aşağıdaki gibi tanımlanmaktadır.

$$d(y,x) = d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \qquad (2.17)$$

$$d_{öklid} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (2.18)$$

Burada x_i ve y_i iki ayrı gözlemi ifade etmektedir. Öklid uzaklığının genellemesi olan Minkowski uzaklığının ölçümüne aşağıda yer verilmiştir.

$$d_{minkowski} = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p} \qquad (2.19)$$

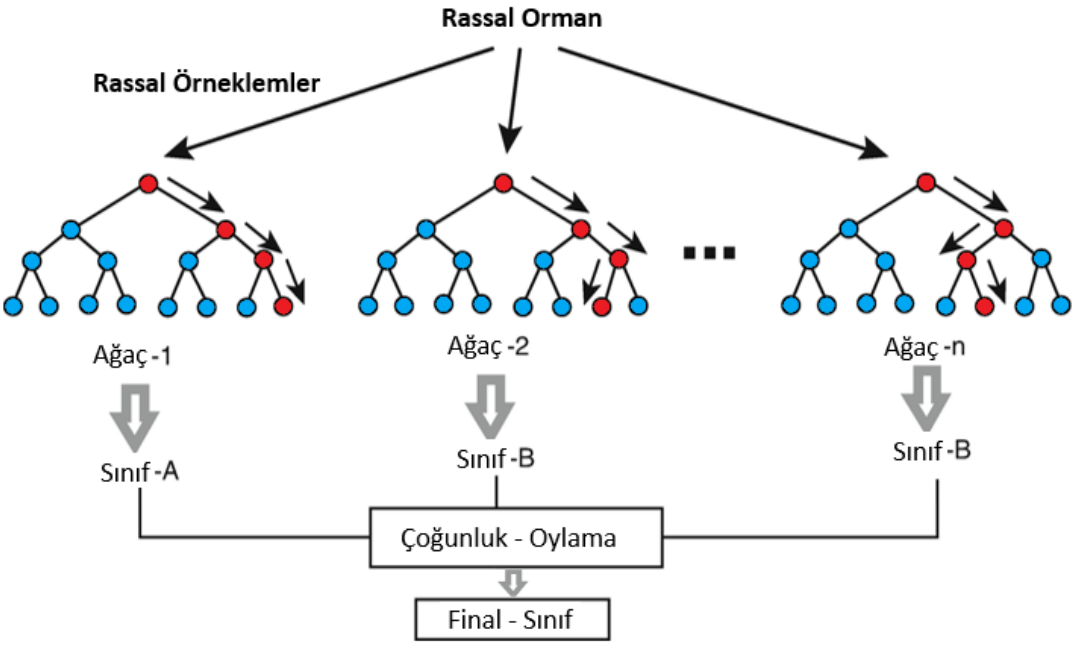
Yukarıdaki denklemde bulunan p değeri 2 olduğunda Öklid uzaklık ölçüsü elde edilirken, 1 olduğunda ikili sınıflandırma problemlerinde yaygın kullanımı olan Manhattan uzaklık ölçüsüne eşdeğer olmaktadır.

$$d_{manhattan} = \sum_{i=1}^{n} |x_i - y_i| \tag{2.20}$$

Optimum sınıflandırma için bir diğer unsur k hiperparametresinin belirlenmesi olup, yüksek değer alması yanlılığı artırırken, düşük değer alması varyansın artmasına neden olmaktadır. İlgili çalışma kapsamında uzaklık matrisinin hesaplanmasında, Öklid uzaklık ölçüsü ve optimal sınıflandırma için k-hiperparametresi deneysel olarak farklı değerler ile belirlenmiştir.

2.1.6 Rassal Orman (Random Forest)

Topluluk Öğrenmesine dayanan Rassal Orman (RO) algoritması, Torbalama (Bootstrap Aggregation) tekniğinden yararlanarak, herhangi bir bilgi ölçüsü metodu kullanmadan seçilen rastgele örneklemler ve değişkenlerle inşa edilen karar ağaçları topluluklarıdır. Rassal Orman, torbalama tekniği ile seçilen rastgele örneklemlere ek olarak, ağacın her düğümünde rastgele seçilen değişkenler ile karar ormanını oluşturur. Dolayısıyla Rassal Orman, torbalama yönteminin gelişmiş bir şekli olarak kabul edilebilir (Breiman, 2001:28).



Şekil 2.10: Karar Ormanı Diyagramı (Sniatala, 2020:160)

Karar ormanını oluşturan ağaçlar, ağacı maksimum boyutta geliştirmek için CART (Classification and Regression Tree) algoritmasından faydalanarak, kök karar düğümünde hangi özelliğin konumlanacağını bilgi kazancı (2.12) ile belirler. Kök karar düğümünde konumlanacak özelliğin belirlenmesinden sonra sınıflandırmaların safsızlığı için gini indeksinden (2.11) yararlanır. Bu prosedür kullanıcı tarafından belirlenen N tane ağaç sayısı oluşturulana kadar yinelenir. Her bir karar düğümünde kullanılacak öznitelik değişkenleri, kullanıcının belirlemesi gereken diğer bir hiperparametre olup toplam öznitelik sayısının karekökü kadar olması tavsiye edilmektedir.

Rastgele örneklemler ile inşa edilen her bir ağacın sınıflandırma sonuçları, oylama tekniğine benzer yaklaşımla ortalaması alınarak nihai sınıf kararına ulaşılır. Tahminlerin ortalamasının dikkate alınması, son sınıflandırıcıda varyansın ve sapmanın azalmasına olanak sağlamaktadır. Bu durumla güçlü ve yüksek performansta sınıflandırmalar elde edilmektedir.

Ek olarak, RO algoritması eğitim veri setinin belirli bir kısmını ağaçlardan elde edilen sonuçların hata oranlarını değerlendirmek için kullanır. (Out-of-Bag, OOB) olarak adlandırılan bu veri seti, ormana ait genel sınıflandırma başarısını da hata skoruyla tespit ederek, model doğruluğunu ölçümleyebilmektedir. Sahip olduğu bu teknikler ile aşırı öğrenme (overfitting) durumuna dayanıklı olsa da gürültülü verilerin yoğunluğuna göre aşırı öğrenme eğiliminde olabilirler.

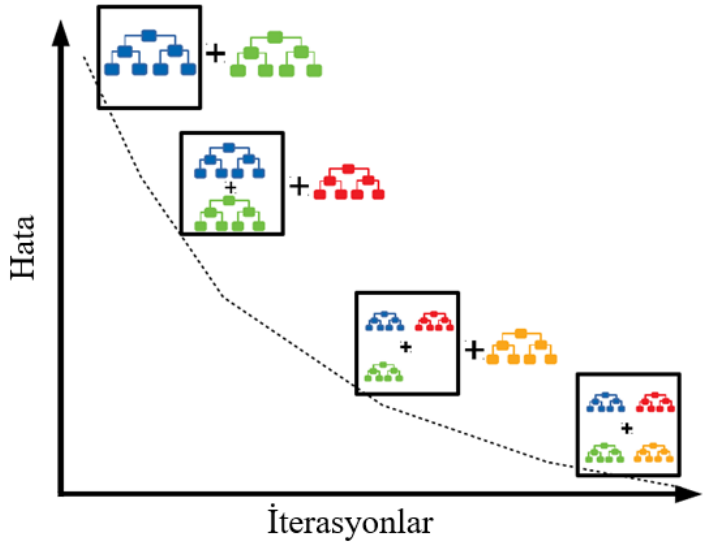
RO, tıpkı Karar Ağaçları ve Gradyan Artırma algoritmaları gibi öznitelik değişkenlerinin hedef değişken üzerindeki sınıflandırıcı gücünü kullanarak öznitelik önemliliğini hesaplamaktadır. Gini indeksinin büyüklüğüne göre belirlenen öznitelik önemliliği topluluk öğrenmelerinde aşağıdaki gibi hesaplanmaktadır.

$$VI(X_j) = \frac{1}{N}\sum_{t}(hata_t(D) - hata_t(\widetilde{D}_j)) \tag{2.21}$$

Burada N, RO'da bulunan toplam ağaç sayısını, t her bir ağacı, D eğitim verilerini ve $\widetilde{D}_j$, X_j özniteliğinin sahip olduğu verileri temsil etmektedir. Sınıflandırma problemlerinde hata 1 – Doğruluk (Yanlış Sınıflandırma Oranı) ile ölçülürken, regresyon problemlerinde ortalama kare hata (Mean Squared Error) olarak değerlendirilmektedir.

2.1.7 Gradyan Artırma (Gradient Boosting)

Gradyan Artırma, regresyon veya sınıflandırma modellerinde tahmin edilen değerler ile gözlemlenen değerler arasındaki uyumun iyileştirilmesi için kayıp fonksiyonun türevleri ile model optimizasyonuna izin veren bir makine öğrenme algoritmasıdır. Model optimizasyonunda kullanılan kayıp fonksiyonu için regresyon modellerinde ortalama hata kare kullanılırken, sınıflandırma modellerinde logaritmik kayıp kullanılmaktadır.



Şekil 2.11: Gradyan Artırma Algoritmasının Yaygın Bir Örneği (Vasiloudis, 2019,1)

Algoritmanın ardındaki yeteneği, artıklardaki örüntüleri model tahmininde parametreleştirilmiş bir yapıda kullanarak, zayıf tahmincileri iyileştirmesi ve algoritmik olarak kayıp fonksiyonunu optimize etmesidir. Bu bağlamda, artık değerlerin bir kayıp fonksiyonunun türevleri olarak toplu etkileşimli iterasyonlarla hem varyansı hem de sapmayı azaltacak şekilde bir öngörü fonksiyonunun doğruluğunu artıran kolektif bir algoritmadır.

Gözlemlenen değerler ile tahmin değerleri arasındaki farkın (artıklar) her bir iterasyonda bir parametre görevi görerek tahmindeki sapmayı azaltması aşağıda bulunan adımlar ile gerçekleşmektedir.

Tablo 2.4: Gradyan Artırma Sözde Teknik Kodu

1. Temel Modelin Oluşturulması

$$f_0(x) = argmin_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$$

2. Artıkların Hesaplanarak Model Dahil Edilmesi

$\{(x_i y_i)\}_{i=1}^{n}$ Eğitim veri seti, $L(y, F(x))$ türevlenebilir kayıp fonksiyonu ve M iterasyon sayısını ifade etmek üzere,

- Artıkların hesaplanması:

$$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}} \qquad i = 1, 2, \dots, N$$

- Sözde artıkları parametreleştirip $h_m(x)$ yeniden eğitilmesi:

$$\gamma_m = arg\ \min_\gamma \sum_{i=1}^{n} L(y_i, f_{m-1}(x_i) + \gamma h_m(x))$$

- Modelin güncellenmesi:

$$f_m(x) = f_{m-1}(x) + \gamma_m h_m(x)$$

3. Nihai Çıktı

$$\hat{f}(x) = f_M(x)$$

Yukarıda optimum modelin elde edilmesi için gerçekleştirilen iterasyonlar doğrultusunda nihai gradyan artırma modeli temel olarak aşağıdaki gibi gösterilmektedir.

$$f_M(x) = G_0 + \beta_1 T_1(x) + \beta_2 T_2(x) + \cdots\cdots + \beta_u T_u(x) \qquad (2.22)$$

Burada G_0 serideki ilk değeri ifade ederken, T_1,,, T_u değerleri kalıntıları içeren ağaçları temsil etmektedir. β_1,,, β_u değerleri ağaçlardaki düğümler için katsayıları ifade etmektedir.

2.2 SINIFLANDIRMALAR İÇİN PERFORMANS ÖLÇÜLERİ

Çalışmaya konu olan kredi temerrüt risk tahmininde, temerrüde düşen ve temerrüde düşmeyen olarak ikili sınıflandırmaları gerçekleştiren algoritmaların performansları için karmaşıklık matrisinden elde edilen sınıflandırma ölçüleri ve ROC eğrisinden yararlanılmaktadır.

2.2.1 Karmaşıklık Matrisi

Karmaşıklık matrisi, algoritmaların doğrulama veya test veri setleri üzerinden gerçekleştirdiği sınıflandırmaların, dört farklı perspektif (A, B, C, D) üzerinden, gerçek gözlemlenen değerler ile karşılaştırıldığı özet matris tablosudur.

Tablo 2.5: Karmaşıklık Matrisi

	(Tahminlenen) Negatif	**(Tahminlenen) Pozitif**	**Toplam Gözlemlenen Olasılık**
(Gözlemlenen) Negatif	A (Doğru Negatif)	B (Yanlış Pozitif)	$\frac{A+B}{A+B+C+D}$
(Gözlemlenen) Pozitif	C (Yanlış Negatif)	D (Doğru Pozitif)	$\frac{C+D}{A+B+C+D}$
Toplam Tahminlenen Değer	A + C	B + D	$A+B+C+D$

Sınıflandırma çıktıları birbirleriyle ilişkilendirilerek aşağıdaki performans ölçüleri türetilir.

Tablo 2.6: Sınıflandırma Ölçüleri

Sınıflandırma Ölçüsü	**Formül**
Doğruluk (Doğru Sınıflandırma Oranı)	$\frac{A+D}{A+B+C+D}$
Hassasiyet (Gerçek Pozitif Oranı)	$\frac{D}{C+D}$
Özgüllük (Gerçek Negatif Oranı)	$\frac{A}{A+B}$
1-Özgüllük (Yanlış Pozitif Oranı)	$\frac{B}{A+B}$
Kesinlik	$\frac{D}{D+B}$
F1 Skoru	$2 * \frac{\frac{D}{(C+D)}}{\frac{D}{(D+B)}}$

Çalışma kapsamında temerrüde düşenler pozitif, temerrüde düşmeyenler ise negatif olarak değerlendirilmiştir. Bu bağlamda, ilgili performans ölçülerinin çalışma kapsamındaki tanımlamalara aşağıda yer verilmiştir.

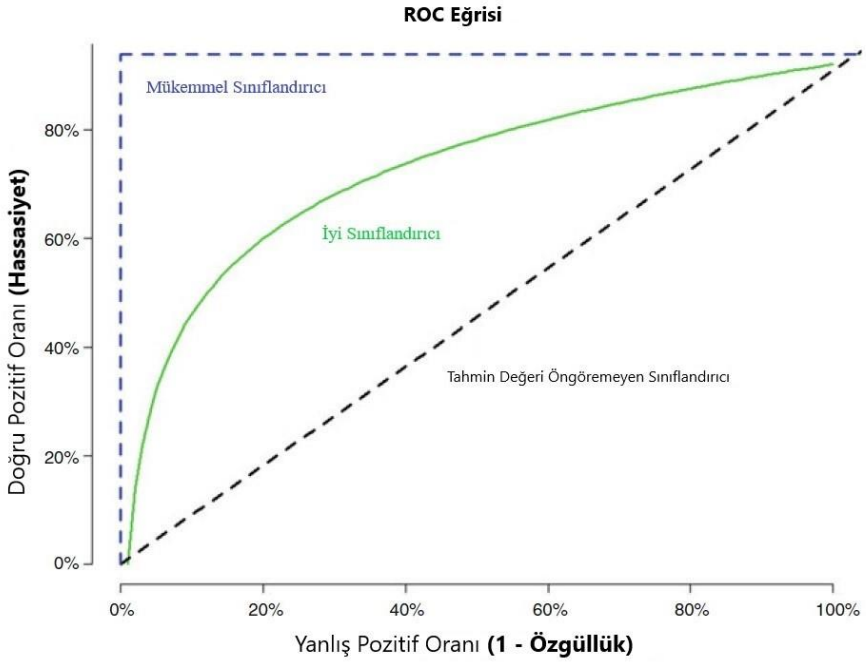
Doğruluk oranı, temerrüde düşen ve temerrüde düşmeyen olarak gözlemlenen değerler için genel olarak ne oranda doğru tahmin edildiğinin bilgisini veren ölçüttür. Yanlış sınıflandırma oranı ise doğru sınıflandırma oranının tersi olarak, tahmin değerlerinin gözlemlenen gerçek değerlerden ne oranda yanlış sınıflandırmalar yaptığını açıklar. Diğer bir ifadeyle, gerçekte temerrüde düşenlerin, tahmin değerlerinde temerrüde düşmediğini ve gerçekte temerrüde düşmeyenlerin, tahmin değerlerinde temerrüde düştüğünün genel oranıdır. Pozitiflerin genel tahmin başarısını Tip-II hata perspektifiyle inceleyen hassasiyet oranı, gerçekte temerrüde düşenlerin, tahmin değerlerinde ne oranda temerrüde düşenler olarak sınıflandırıldığının bir ölçüsüdür. Temerrüde düşenler için Tip-I hata perspektifiyle hesaplanan diğer bir ölçüt ise, kesinlik oranı olup, tahmin değerinde temerrüde düşen olarak sınıflandırılan müşterilerin gerçekte kaçının temerrüde düştüğü bilgisini doğrulamaktadır. Negatif vakalar için hesaplanan özgüllük oranı, gerçekte temerrüde düşmeyenlerin, tahmin değerlerinde ne oranda temerrüde düşmeyenler olarak sınıflandırıldığının bir ölçüsüdür. ROC eğrisinin yatay ekseninde kullanılan ve tahmin değerlerinin kayıp oranı olan yanlış pozitif oranı (1–Özgüllük), gerçekte temerrüde düşmeyenlerin tahmin değerlerinde ne oranda temerrüde düşenler olarak sınıflandırıldığı ölçüdür.

Sınıflandırıcı ölçüler tek boyutlu olmadığı için performans değerlendirmesi de tek bir ölçüt üzerinden değerlendirilmemektedir. Özellikle kredi riski perspektifinde temerrüde düşen ve temerrüde düşmeyenlerin yanlış sınıflandırılması farklı risk yapılarını ortaya çıkardığı için yukarıda tanımlanan ölçüler tek başına değerlendirilmesi kayıp risk perspektifine neden olacaktır.

Bu bağlamda, yukarıda hesaplanan hassasiyet ve kesinlik ölçütlerinin harmonik ortalamasıyla elde edilen bir diğer ölçü ise F1 skoru olup, öncelikle Tip-I ve Tip-II hata perspektifleriyle hesaplanan hassasiyet ve kesinlik ölçütlerinin sınıflandırma performansını karşılaştırmak için kullanılmaktadır. Ayrıca, sınıflandırıcı algoritmaların performanslarının karşılaştırılmasında yaygın olarak kullanılan F1 skoru, modelin genel başarısını da karakterize etmektedir.

2.2.2 ROC (Receiver Operating Characteristic) Eğrisi

Dikotom bir ölçümü tahmin etmek için kullanılan ROC eğrisi, karmaşıklık matrisinden elde edilen gerçek pozitif oranı ve yanlış pozitif oranı ölçülerinden yararlanarak, optimum kesim skorunu belirlemeye yardımcı olur. İki ölçü arasında içsel bir etkileşim bulunduğundan, dikey eksende gerçek pozitif fraksiyonu ve yatay eksende yanlış pozitif fraksiyonu olmak üzere, tipik olarak Şekil 2.12'de görüldüğü gibi bir eğri elde edilmektedir.



Şekil 2.12: ROC Eğrisi Örneği (Dinov, 2018,488)

ROC eğrisinde sınıflandırma başarısı için gerçekte pozitif olarak gözlemlenen değerlerin, tahminlenen pozitif oranlar (gerçek pozitif oran ve yanlış pozitif oran) arasında gerçek pozitif oranda birikmesi istenilmektedir. Dolayısıyla, gerçekte negatif olan değerlerin, pozitif olarak tahminlenmesinin (yanlış pozitif oran) düşük oranda olması beklenir. Bu bağlamda, Şekil 2.12'de bulunan kesikli mavi çizgi, yanlış pozitif fraksiyonun bulunmadığını, bu yüzden %100 olarak gerçek pozitifler ile mükemmel bir sınıflandırıcıyı ifade etmektedir. Yeşil çizgi, gerçek verilerle eğitilmiş bir modelin genelde benzer yapıda olduğu sınıflandırıcıyı ve kesikli siyah diyagonal çizgi tahmin değeri öngöremeyen bir sınıflandırıcıyı temsil etmektedir.

Çalışma kapsamında, farklı sınıflandırıcıların ve girdi değişkenlerin temerrüde düşen ve temerrüde düşmeyen müşteriler için yaptıkları sınıflandırmaların ayırt edici gücünü kavramak için kullanılan Gini Katsayısı, PD modellerinde yaygın olarak kullanılan bir ölçüt olup, 0 ile 1 arasında değerler almaktadır.

$$Gini = 2 * (\text{AUC} - 1) \tag{2.23}$$

ROC eğrisinin altında kalan alan (AUC) ile hesaplanan Gini Katsayısı için eğrinin altında kalan alan büyüdükçe, test verisi için sınıflandırmanın doğruluğu da artarak 1'e yaklaşır.

2.3 ÖRNEKLEMİN BELİRLENMESİ

Finansal kurumlar, maruz kalacağı riskleri tam olarak belirlemeleri için değerli bilgileri kapsayan, doğru ve güçlü modellerin kurulmasına ihtiyaç duyar. Bu kontekste, verilerden mümkün olabildiğince bilgilendirici iç görüyü keşfetmek için örnekleme ve veri kalitesi ve temizliği teknikleri uygulanır.

Kredi riski perspektifinden örnekleme ihtiyacı, büyük hacimli verilerin işlenmesinin zaman alıcı olmasından, dengesiz sınıf dağılımına sahip olmasından (temerrüde düşen/düşmeyen) ve hedef kitleyi en iyi temsil edecek zamansal aralığın belirlenmesinden kaynaklanmaktadır. Özellikle sınıf dağılımının dengesiz yapıda olduğu bir veri seti ile model oluşturmanın, kitledeki yoğunluğun davranışına eğilimli yönde sonuçlar ürettiği görülmektedir. Buna benzer yanlı tahminlerin oluşumuna engel olmak için sınıf dağılımı dengede tutularak yanlılık ortadan kaldırılmaktadır. Diğer yandan örneklemin belirlendiği zaman çizelgeleri de mevsimsellik açısından aynı derece önemlidir. Küresel ekonomik koşullar veya bayramlar gibi harcama eğilimi farklı eğilimler gösterebilmektedir. Bu doğrultuda normal iş dönemi belirlenerek örneklem yapısı dengelenmelidir.

Yukarıda bulunan yaklaşımlar doğrultusunda, hedef kitleyi temsil edecek bir örneklem kümesi elde etmek için tabakalı örnekleme yöntemi kullanılarak, birbirine benzer alt gruplar oluşturulmaktadır. Bu adımla veri kümesindeki denge sağlanarak, tahminlerde hatayı azaltmak amaçlanmaktadır.

2.4 DEĞİŞKEN İNDİRGEME METOTLARI

Yapılan kredi risk tanımlarından yola çıkarak, müşteriye ait demografik bilgiler, tarihsel olarak banka ile ilişkili ürün sayısı ve boyut bilgileri, ödeme performansları, temerrüt geçmişi ve dış kaynaklardan elde edilen kredi puanları gibi birçok parametre, müşterinin kredi risk profilinin oluşturulmasında kullanılan risk faktörleridir. Bu doğrultuda veri kümesinde bulunan tüm değişkenler içinde, hedef profili tanımlayıcı güce sahip olan değişkenlerin belirlenmesinde, istatistiksel teknikler kullanılarak değişken indirgeme yöntemleri gerçekleştirilmektedir.

Çağımızda sıklıkla karşılaştığımız büyük boyutlu veri setlerindeki karmaşıklığın azaltılması için veri kalitesi ve temizliği işlemleri, sütun olarak değişkenler için de uygulanma ihtiyacı duymaktadır. Bu işlemler model performansını artırırken işlem süresinde de iyileştirmeler sağlayabilmektedir.

2.4.1 Kayıp ve Aykırı (Uç) Değerler Tespiti

Büyük veri kümelerinde doğal olarak ortaya çıkan problem, verilerdeki eksik değerlerin varlığıdır. Bunun nedeni, insan atıf hatası, bilgilerin işlenebilir olmaması ve kişisel bilgilerin gizliliği kapsamında açıklanmaması gibi bir dizi nedenden kaynaklanabilmektedir.

Karar ağaçları gibi algoritmalar bu tür kayıp veriler ile doğrudan başa çıkarak anlamlı yaklaşımlar sergileyebilirler. Ancak diğer algoritmalar için aynı durum söz konusu olmayabilir. Bu noktada kayıp değerler için bir ön işleme ihtiyacı duyulur. Geçmişten günümüze kayıp veriler için birçok ön işleme metodu geliştirilmiştir. Kayıp verilere müdahale etmeden analizde yer verme, silerek analiz dışı bırakma, yaklaşık değer atama veya istatistiksel metotlar ile değer atama, kayıp verileri ön işlemede kullanılan metotlardan bazılarıdır.

Kayıp değerler özellikle finansal risk konularında yüksek derecede öneme sahip olabilirler. Örneğin müşteriye ait eksik bir bilgi kendi içinde dolandırıcılık davranışı gösterebilir. Kayıp değerleri içinde barındıran bir bilginin silinmesi için kayıp değerlerin yoğunluğu ile karar verilebilir. Diğer yandan kayıp değerlere yaklaşık veya istatistiksel metotlar ile değer atama tekniklerinde tutarlı olmak gerekmektedir. Bu kontekste, %70 oranında kayıp veri yoğunluğuna sahip değişkenler için bir eleme eşik değeri

belirlenmiştir. Eşik değerin altında kalan ve kayıp veri barındıran değişkenler için Karar Ağacı algoritması ile kayıp veriler için bir tahminsel atama işlemi esas alınmıştır.

Büyük veri kümelerinde ortaya çıkan bir diğer problem ise aykırı değerlerin görülmesidir. Bunun nedeni tıpkı kayıp verilerdeki insan atıf hatasından kaynaklanabileceği gibi gözlemler içinde geçerli ancak boyut dışı bir değere sahip olmasından kaynaklanır. Örneğin bir şirkette çalışanların maaşları analiz edilirken, yöneticinin maaşı ile ofis personellerin maaşları arasında olağandışı bir gözlem olarak kabul edilebilir. Bu tür olağandışı gözlemler ile model eğitmenin sakıncaları olabileceği için aykırı değerler yaklaşık değer atama yöntemleriyle dönüştürülebilir veya analiz dışı bırakılabilirler.

2.4.2 Varyans Eşiği

Değişkenlerin dağılımına göre belirlenen varyans eşikleme, değişkenlerin indirgenmesine yönelik en temel yaklaşımlardan biridir. Büyük boyutlu veri setlerinde varyans eşikleme tekniğinin, değişken indirgeme sürecinin ilk adımlarında uygulanması önerilerek, özellikle model süre performansına pozitif etki ettiği söylenebilmektedir.

Tekniğin motive noktası, düşük varyansa sahip olan değişkenlerin, yüksek varyansa sahip olan değişkenlere göre model için daha az kullanışlı olmasıdır. Nitekim bir değişkene ait değerlerin neredeyse tamamının aynı değere sahip olması, yaklaşık sıfır varyans özelliğini taşıdığı anlamına gelebileceği için model için anlamlı bir etkisinin olmayacağı öngörülmektedir. Bu bağlamda, değişkenlerin her birine ait varyans değeri hesaplanarak, belirlenen eşik değerine göre değişken indirgeme işlemi gerçekleştirilir.

2.4.3 Kanıt Ağırlığı Dönüşümü

Başlangıçta krediler için skor kartı geliştirmelerinde kullanılan Kanıt Ağırlığının (Weight of Evidence), son yıllarda kredi riski modellerinde segmentasyon ve değişken seçiminde kullanımı yaygınlaşmıştır. Özellikle PD modellemede temerrüt riskini tahmin etmeye yönelik kullanılan Kanıt Ağırlığı, öznitelik değişkeni içindeki niteliklerin (grupların) hedef değişken üzerinden tahmin yeteneğini ifade eder. WOE değeri, iyi ve kötü riskin ayrıştırılması için gruplandırmanın gücünü kullanarak, benzer yapıda bulunan nitelikleri aynı grupta birleştirir. Nitelik bazında hesaplanan WOE dönüşümü aşağıdaki formül ile elde edilmektedir.

$$WOE_i = \ln\left(\frac{p_i^{Temerrüde\ Düşmeyenler}}{p_i^{Temerrüde\ Düşenler}}\right) = \ln\left(\frac{\frac{N_i^{Temerrüde\ Düşmeyenler}}{N_{Toplam}^{Temerrüde\ Düşmeyenler}}}{\frac{N_i^{Temerrüde\ Düşenler}}{N_{Toplam}^{Temerrüde\ Düşenler}}}\right) \quad (2.24)$$

$N_i^{Temerrüde\ Düşmeyenler}$ = Gruptaki temerrüde düşmeyenlerin sayısı

$N_{Toplam}^{Temerrüde\ Düşmeyenler}$ = Toplam temerrüde düşmeyenlerin sayısı

$N_i^{Temerrüde\ Düşenler}$ = Gruptaki temerrüde düşenlerin sayısı

$N_{Toplam}^{Temerrüde\ Düşenler}$ = Toplam temerrüde düşenlerin sayısı

WOE perspektifi log-odds yaklaşımına dayandığı için Lojistik Regresyon algoritması için oldukça uygun bir yapıdadır. Logit dönüşümü tüm olayların (iyi ve kötü) olasılıklarını ifade ettiği için WOE dönüşümleri her bir niteliği standartlaştırır. Bu durum Lojistik Regresyonda kullanılan parametrelerle karşılaştırma imkânı sunar. Müşteri yaşına göre WOE hesaplanmasının örneğine Tablo 2.6'te yer verilmiştir.

Tablo 2.7: WOE Hesaplaması Örneği

Aralık	Grup	Temerrüde Düşmeyenler	Temerrüde Düşenler	Temerrüde Düşmeyenler Oranı	Temerrüde Düşenlerin Oranı	WOE	IV
18-25	1	185	202	%9	%45	-1.568	0.556
26-35	2	305	143	%15	%32	-0.722	0.118
36-45	3	550	75	%28	%17	0.513	0.057
46≥	4	941	31	%48	%7	1.933	0.785
Toplam		1981	451				**1.516**

2.4.4 Bilgi Değeri

Bilgi Değeri (Information Value), öznitelik değişkenlerinin hedef değişken üzerindeki tahmin performanslarına göre modeldeki önemi sıralanarak, değişkenin ayırt edici tanısal bilgi miktarını ifade etmektedir. Bilgi değeri ne kadar büyükse, ayırt etme özelliği de o kadar güçlüdür. Tüm bilgi değerleri için alt sınır 0 olmakla beraber üst sınırı yoktur.

Kredi riski modellemesinde Kullback sapma ölçüsü olarak adlandırılan bilgi değeri, kredilerini ödeme durumuna göre belirlenen iyi ve kötü müşteriler arasındaki dağılım farkını ölçümler. Değerlendirilen her niteliğe ait indekse (i) ve toplam niteliklerin sayısına (k) göre aşağıdaki gibi hesaplanmaktadır.

$$IV = \sum_{i=1}^{k} \left(\frac{N_i^{Temerrüde\ Düşmeyenler}}{N_{Toplam}^{Temerrüde\ Düşmeyenler}} - \frac{N_i^{Temerrüde\ Düşenler}}{N_{Toplam}^{Temerrüde\ Düşenler}} \right) * WOE_i \qquad (2.25)$$

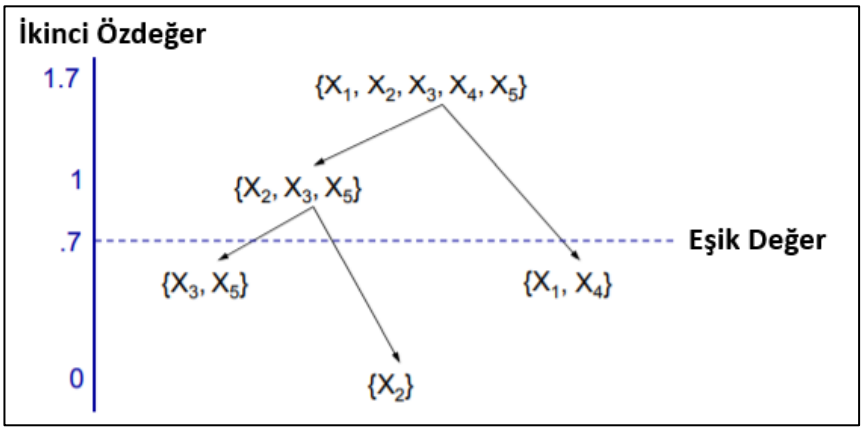
Temel olarak 0.02'den küçük olan bir bilgi değeri modelleme için kullanışlı olmayabilir. 0.02 ila 0.1 arasındaki bilgi değeri zayıf bir ayırt etme özelliğine sahip olduğunu gösterirken, 0.1 ila 0.3 arasındaki bilgi değeri modelleme için kabul edilebilir. 0.3 veya daha büyük bir bilgi değeri ayırt etme özelliğinin güçlü olduğunu ifade eder. Değer 1'e yaklaştıkça gerçek olmayacak kadar güçlü bir durum söz konusu olduğunda şüpheli yaklaşılmalıdır.

2.4.5 Değişken Kümeleme

Değişkenlerin indirgemesinde kullanılan bir diğer yaklaşım, değişken kümeleme tekniğidir. Özellikle değişkenler arasındaki yüksek korelasyonu veya kovaryansın tanımlanmasında oldukça yetenekli olan bu teknik, birbirine benzer yönde hareket eden değişkenleri bir kümede toplayarak, değişkenlerden ayrık veya hiyerarşik kümeler elde eder. Her küme içerisinden tahmin gücü yüksek düzeyde olan değişken belirlenerek, küme içerisinde bulunan değişkenleri analitik yapısı gereği temsilen seçilir. Dolayısıyla değişken kümeleme tekniği kullanılarak hem değişkenlerin indirgenmesi hem de çoklu bağlantının önlenmesi sağlanmış olur.

Algoritmanın uygulanmasında SAS'ın VARCLUS prosedürü ile elde edilen kümelerin toplam varyansını maksimize etmesi beklenmektedir. Prosedür, temel bileşenler analizinden faydalanarak, birbirine benzer yönde hareket eden değişkenleri bir kümede ve ilişkili olmayan değişkenleri ayrı bir kümede tutar. Bu işlemle beraber gerekli bulunmayan değişkenlerin elenerek boyutun indirgenmesi sağlanır. VARCLUS prosedürü ile yapılan kümeleme işleminden sonra alt boyutların incelenmesi için belirlenen eşik değerini kullanarak, ikinci özdeğerin bu değerin altında olması beklenir. Eğer ikinci özdeğer bu eşik değerden yüksek ise küme birden fazla boyuta sahiptir ve bölünerek tekrar eşik değer kontrolü yapılır. Bu durum Şekil 2.13'de 5 değişkenli ilk

küme, 4 ayrı bölünmeyle gösterilmiştir. İlk bölünmede X1 ve X4 açıklanan belirli bir varyasyon yüzdesine sahip olarak eşik değerin altında bir küme oluşturmuştur. Diğer değişken kümesi (X2, X3, X5) eşik değerin üstünde kaldığı için tekrar bölünme uygulanmış ve eşik değerin altında iki ayrı küme elde edilmiştir. Bu durum aynı zamanda maksimum küme sayısına ulaşıldığını ifade etmektedir.



Şekil 2.13: VARCLUS Kümeleme Prosedürü Örneği

VARCLUS prosedüründen elde edilen çıktı tablosu, maksimum küme sayısını, her kümeye ait değişken bilgisini ve değişken seçimi yapılabilmesi için kümenin R2'sini, en yakın kümenin R2'sini ve kümenin 1–R2'sini vermektedir.

$$1 - R^2 = \frac{1 - R^2_{kendi\ kümesi}}{1 - R^2_{en\ yakın\ küme}} \tag{2.26}$$

Değişken seçim sürecinde, küme içinde en yüksek korelasyona sahip olan değişkenin aynı zamanda diğer kümeler arası korelasyonunun düşük olması tercih edilmelidir. Bu kontekste, en düşük 1–R2 oranı kümeyi en iyi temsil eden değişken olarak kabul edilebilir. Ancak kredi riski perspektifinde, hedef değişkeni ile daha net ilişkisel yapıda olan değişken sektör bilgisiyle sürece dahil edilebilir.

2.4.6 LASSO (En Küçük Mutlak Daralma ve Seçim Operatörü)

Literatürde ilk defa 1996 yılında Robert Tibshirani tarafından kullanılan LASSO yöntemi, regresyon modeline dahil edilen λ ceza parametresiyle, tahmin gücüne bağlı olarak aday tahmincilerin katsayılarının mutlak boyutunu daraltmaktadır. Katsayılara uygulanan cezalandırma işlemi, tahmin gücü yüksek düzeyde olan değişkenlerin katsayılarını küçültürken, düşük düzeydeki tahmincilerin katsayılarını sıfıra kadar

indirebilmektedir. Bu yöntem ile katsayıları sıfıra indirilen değişkenlerin model performansına katkı sağlamadığı kabul edilerek, değişken indirgeme işlemi de gerçekleştirilmiş olur.

$$LASSO = \min\left[\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right] \tag{2.27}$$

LASSO regresyonu içerisinde bulunan sabit ceza parametresi (λ), k-katlamalı çapraz doğrulama yöntemi ile elde edilmektedir. Sabit ceza parametresinin belirlenmesi için uygulanan çapraz doğrulama işlemi, ek olarak modelin aşırı öğrenmesinin de önüne geçmektedir. Bu bağlamda LASSO yöntemi, nihai değişkenlerin belirlenmesi, modelin geliştirilmesi ve nihai modelinin seçilmesinde oldukça faydalı bir teknik olarak değerlendirilmektedir.

ÜÇÜNCÜ BÖLÜM

VERİ KALİTESİ VE MODEL GELİŞTİRME

3.1 VERİ SETLERİNİN TANIMLANMASI

Analizlerde modele girdi olacak veri kümesi, Kaggle'ın açık erişiminde bulunan Home Credit tarafından sağlanmaktadır. Merkezi Hollanda'da bulunan şirket, 1997'de Çek Cumhuriyeti'nde kurulmuş ve Slovakya, Rusya, Çin, Vietnam, Kazakistan, Hindistan, Endonezya ve Filipinler olmak üzere 9 ülkede faaliyet göstermektedir. Şirketin kredi verme kitlesi öncelikli olarak kredi geçmişi çok az olan veya hiç olmayan kişilerden oluşmaktadır. Bu politikayla daha fazla müşteriyi hedefleyen şirket ne kadar çok kredi sunarsa o kadar yüksek riskli konut kredisine maruz kalacaktır. Bir müşterinin temerrüde düşme olasılığını belirlemek, bu noktada şirket için kritik bir karar olmaktadır. Hedef kitlenin temerrüt olasılığını belirlemek için çok çeşitli geçmiş bilgilerinden yararlanılmaktadır.

Analitik girdileri kapsayan temel veri kümesi, yedi farklı veri setinden meydana gelmektedir. Başvuru veri seti, tüm başvuru sahiplerinin daha önce temerrüde düşüp düşmediği gibi statik verilerin ve demografik bilgilerin bulunduğu diğer bazı bilgileri içerir. Diğer veri setleri kredi bürosundaki önceki başvuru bilgileri, kredi kartı bakiye ve ödeme bilgileri, daha önce alınan kredi bilgileri, müşterilerin daha önce almış olduğu konut kredilerine ait davranışsal bilgileri, önceki konut kredilerine ait ödeme geçmişi bilgileri ve kredi bürosundan gelen aylık davranışsal bilgilerden oluşmaktadır. Bu veri setlerinden elde edilen 1460 açıklayıcı (öznitelik) değişken, 307510 gözleme sahiptir. Çalışma kapsamında, hedef değişken değerlerindeki temerrüde düşenler 1, temerrüde düşmeyenler 0 olarak tanımlanmıştır.

3.2 KULLANILAN YAZILIM VE PROGRAMLAMA DİLLERİ

Birçok farklı sektörün analitik yazılım ihtiyacını karşılayan SAS'ın, özellikle finans sektöründeki ürünlerinin kullanımı oldukça yaygın olup, kredi skorlama modelleri için sahip olduğu modüller sayesinde analitik ve bütünlük avantajı sağlamaktadır. Uygulama kapsamında, SAS Enterprise Guide ürünü ile analitik girdi veri seti

oluşturulmuş ve diğer analitik geliştirmeler ve makine öğrenmesi algoritmaları SAS Enterprise Miner modülleri kullanılarak gerçekleştirilmiştir.

3.3 ÖZNİTELİK SEÇİMİ (DEĞİŞKEN İNDİRGEME)

Analitik modellerde girdi olarak kullanılacak değişkenlerin belirlenmesi, model tahmin gücünü doğrudan etkileyen faktördür. Aynı zamanda girdi olarak kullanılacak gözlemler içinde bu durum geçerlidir. Bu iki perspektifle, temel veri seti üzerinden hem satır (gözlem) hem de sütun (değişken) bazlı indirgeme işlemi bir boyut azaltma tekniğidir. Diğer bir ifadeyle, temel veri seti üzerinden örneklem belirleme ve değişken indirgeme işlemleri, yatay ve dikey perspektifle boyut azaltma tekniklerdir.

Bu kontekste, aşağıda uygulanan istatistiksel veri analizleri ile modelin tahmin performansını güçlendirecek değişkenler ve örneklem elde edilmiştir.

3.3.1 Kayıp Değer Oranı ile Öznitelik Seçimi

Örneklem öncesi yapılan değişken indirgeme sürecine kayıp değer elemesiyle başlanmaktadır. Buradaki öncelik analitik değerlendirilebilirliğine olanak sağlayacak bir model veri kümesi elde etmektir. Bu yaklaşımla, kayıp değerler analizi uygulanmış ve değişken bazında kayıp değer oranları elde edilmiştir. Mevcutta sahip olunan 1460 değişken içerisinden 461 değişkenin, eşik değeri olarak belirlenen %70'in üstünde kayıp gözleme sahip olduğu tespit edilerek analiz dışı bırakılmıştır.

Tablo 3.1: Kayıp Değer Oranı ile Öznitelik Seçimi

Sahip Olunan Öznitelik Sayısı	Belirlenen Eşik Değeri	Eleme Sonrası Öznitelik Sayısı
1460	%70'in Altında Kayıp Değer Oranı	999

3.3.2 Varyans Eşiği ile Öznitelik Seçimi

Bir diğer değişken indirgeme metodu değişkenlerin dağılımı ile ilgilidir. Eğer bir değişkenin dağılımı yaklaşık sıfır varyans özelliğini taşıyorsa, bu değişkenin hedef değişken üzerinde anlamlı bir etkisi beklenmemektedir. Diğer bir ifadeyle, değişkene ait değerlerin neredeyse tamamı aynı değere sahip ise değişkenin modele katkısının yüksek olmayacağı öngörülmektedir. Bu bağlamda, değişkenlerin dağılımları için 0.01'lik varyans eşik değeri belirlenmiş ve her bir değişkenin kendi içinde %99 oranda aynı

değerlere sahip olması durumunda analiz dışı bırakılmasına karar verilmiştir. İncelenen 999 değişken içinden 86 değişkenin %99 oranında aynı değerlere sahip olduğu analiz edilmiştir. Eleme sonrası kalan değişken sayısına Tablo 3.2'de yer verilmiştir.

Tablo 3.2: Varyans Oranı ile Öznitelik Seçimi

Sahip Olunan Öznitelik Sayısı	Belirlenen Eşik Değeri	Eleme Sonrası Öznitelik Sayısı
999	0.01'in Üstünde Varyans Oranı	913

3.3.3 Bilgi Değeri (IV) ile Öznitelik Seçimi

Bilgi değeri, öznitelik değişkenlerinin hedef değişkeni ayırt edici tahmin gücünü değerlendirmek için kullanılan kullanılırken, bu bölümde değişkenlerin indirgenmesi için kullanılmıştır. SAS Enterprise Miner ortamında Interactive Grouping düğümü yardımıyla gerçekleştirilen bu işlem ile her bir açıklayıcı değişkenin ayırt edici Bilgi Değeri elde edilmiştir. Düğüm, ayrıca Bilgi Değerine alternatif olarak her açıklayıcı değişken için Gini katsayısını hesaplar. Değişkenlerin anlamlılık gücünü ölçmek için her iki istatistiksel analiz de kullanılabilmektedir. Analiz sonucunda düğüm, temerrüt riskini ayrıştırmak için her bir açıklayıcı değişkeni, hedef değişken üzerinden farklı risk gruplarına göre ağırlıklandırır. Temerrüde düşenler ile düşmeyenler oranını karşılaştırmaya dayanan gruplandırma işleminden bir WOE (Kanıt Ağırlığı) değeri elde edilir. Çalışmanın bu sürecinde, Gini katsayısı ve WOE'li halleriyle gruplandırılmış değişkenler bir ölçü olarak incelenmiştir.

Değişkenlerin açıklayıcı ölçütü olarak Bilgi Değeri kullanılmış ve çalışma konusu olan temerrüt olasılığı kapsamında, öznitelik değişkenlerinin iyi ve kötü kredileri ayırt etme yeteneğine göre önem sırası oluşturulmuştur. Değişkenlerin ayrıştırıcı gücü için belirlenen 0.1'lik eşik değerinin altında kalan 866 değişken elenerek analiz dışı bırakılmıştır.

Tablo 3.3: Bilgi Değeri (IV) ile Öznitelik Seçimi

Sahip Olunan Öznitelik Sayısı	Belirlenen Eşik Değeri	Eleme Sonrası Öznitelik Sayısı
913	0.1'in Üstünde Bilgi Değeri Oranı	47

3.3.4 Aykırı (Uç) Değerlerin Elemesi

Tahmin çıktılarının kalitesi, büyük ölçüde girdi değerlerinin kalitesi tarafından belirlenmektedir. Bu yaklaşımla, aykırı değerlerin geliştirdiğimiz modelde istatistiksel varsayımları ihlal etmemesi için veri kümesinden kaldırılması veya dönüştürülmesi son derece önemlidir. Bununla beraber, kayıp değerlerin tahmini veya doldurulması gibi analitik süreçlerde de sapmalara neden olmaması için aykırı değerlerin temizliği kritik öneme sahiptir.

Çalışma kapsamında aykırı değerlerin tespiti için her bir açıklayıcı değişkene ait ortalama, minimum, maksimum ve 90'ıncı yüzdelik dilimi incelendiğinde, gözlemler arasındaki değişkenliğin yüksek olduğu belirlenmiştir. Aykırı değerlerin, örnekleme olan etkisini azaltması için 90'ıncı yüzdelik dilimin üzerinde bulunan 11692 gözlem, aykırı değer olarak kabul edilerek analiz dışı bırakılmıştır. Eleme sonrası kalan gözlem sayısına Tablo 3.4'de yer verilmiştir.

Tablo 3.4: Aykırı (Uç) Değerlerin Elemesi

Sahip Olunan Gözlem Sayısı	Belirlenen Eşik Değeri	Eleme Sonrası Gözlem Sayısı
307510	90. Yüzdelik Dilim	295818

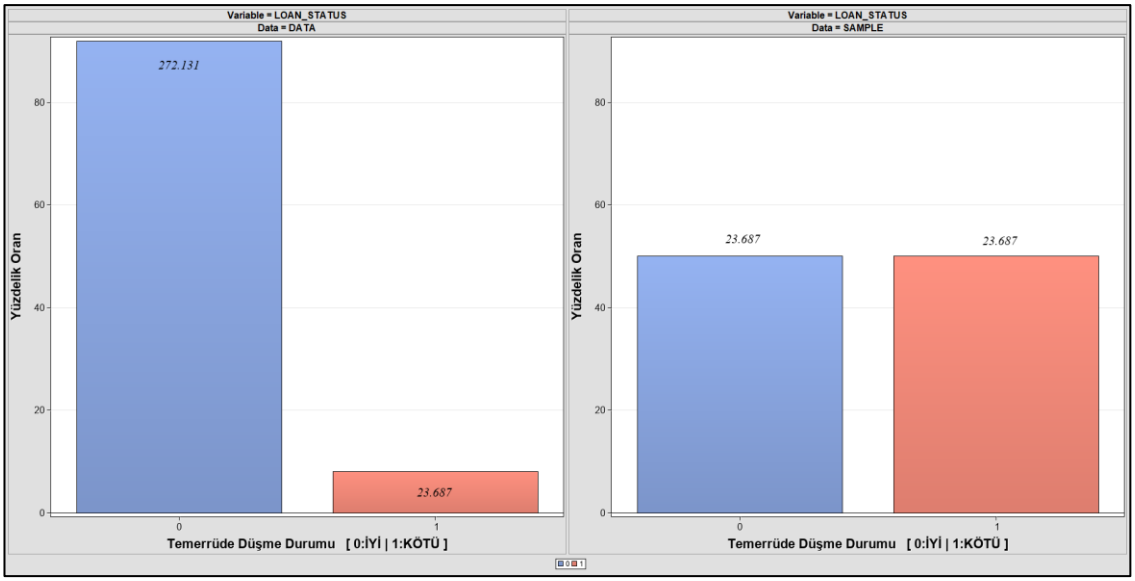
3.3.5 Örneklemin Belirlenmesi

Kredi riski modellemesinde, temerrüt sayısı genellikle temerrüt olmayan popülasyona göre önemli ölçüde düşüktür. İyi ve kötü kredilerin sınıf dağılımındaki bu dengesiz durum, model geliştirme sürecinde verimsizliğe neden olabileceğinden, genellikle temerrüde düşenler ile düşmeyenler arasında tabakalı örneklem yöntemiyle dengeli bir yol izlenmektedir.

Bu bağlamda ham veri setinde bulunan 295818 müşteriye ait sınıf dağılımı incelenmiş ve temerrüde düşmeyenlerin sayısının temerrüde düşenlere oranla yaklaşık 12 kat daha fazla olduğu tespit edilmiştir. Tüm popülasyonun %92'sini oluşturan iyi kredilerin, model geliştirme sürecinde sapmaya neden olmaması için tabakalı örneklem metodolojisiyle her iki sınıftan eşit sayıda örneklem belirlenmiştir. Elde edilen veri kümesine düşen gözlem sayısına Tablo 3.5'te, örneklem öncesi popülasyonun ve nihai olarak elde edilen popülasyonun grafik dağılımına Grafik 3.1'de yer verilmiştir.

Tablo 3.5: Örneklemin Belirlenmesi

Sahip Olunan Gözlem Sayısı	Belirlenen Örnekleme Yöntemi	Örneklem Sonrası Gözlem Sayısı
295818	1:1'lik (%50 İYİ + %50 KÖTÜ) Tabakalı Örnekleme	47374



Grafik 3.1: Örneklem Öncesi ve Örneklem Sonrasına Ait İYİ-KÖTÜ Dağılımı

Veri kalitesi ve model geliştirme bölümünde, nihai örneklem seti üzerinden kayıp değerlerin tahmini ve değişken indirgeme işlemlerine devam edilmiş olup, makine öğrenmesinin temel alt yapısını oluşturan verilerin bölünme işlemi gerçekleştirilmiştir. Bu kontekste, modellerin eğitilmesi ve en iyi ağırlıkların elde edilmesi için veri setinin %60'ı eğitim, eğitilen modellerin aşırı öğrenmesini önlemek ve sınıflandırma performansını değerlendirmek için %20'si doğrulama ve modelin tahmin yeteneğini ölçmek için %20'si test veri seti olarak ayrılmıştır.

3.3.6 Kayıp Değerlerin Atamasında Ağaç Tabanlı Yaklaşım

Analitik model verisini zenginleştirmek için kayıp değerlerin doldurulması veya tahmin edilme işlemi, değişkenlerin yapısına doğrudan etki ettiği için bu süreçte tutarlı olmak oldukça önemlidir. Kredi riski perspektifinden, kayıp değerlerin doğru atamasının yapılmaması, müşterilerin risk sınıfını doğrudan etkileyeceği gibi elde edilecek kâr ve müşteri kaybı gibi risk sonuçlarını doğurabilir. Bu bağlamda, kayıp değerler için yapılacak tahmin atamalarındaki sapmanın minimum düzeyde olması için müşterilerin statik ve davranışsal bilgilerinden benzerliklerine göre tahminsel atama gerçekleştiren

ağaç tabanlı atama tekniği uygulanmıştır. Her değişken için kayıp değerler yerine atanacak değeri, veri setindeki diğer değişkenlere bağlı olarak tahmin eden bir Karar Ağacının, ortalama, medyan veya olasılık dağılımı gibi yöntemlere göre daha tutarlı değerler ürettiği görülmüştür. Örneklem içinde bulunan 47 değişkene ait kayıp değerlerin oranları tekrar incelenmiş ve belirlenen %30'luk eşik değerinin üstünde kayıp değere sahip olan değişkenler, ağaç tabanlı kayıp değerlerin atamasında daha tutarlı sonuçlar üretmesi için girdi olarak kullanılmamıştır.

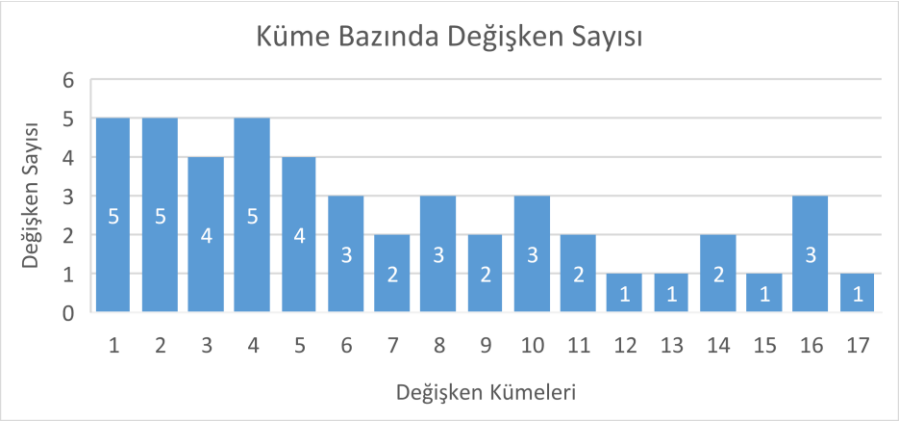
Analitik model verisinin geliştirilme sürecine, örneklem veri setindeki kayıp verilerin atama işlemiyle birlikte, değişken indirgeme adımları ile devam edilmiştir.

3.3.7 Değişken Kümeleme Tekniği ile Öznitelik Seçimi

Değişken kümeleme, aralarında yüksek düzeyde ilişki barındıran değişkenlerin elenmesi için oldukça iyi sonuçlar üreten güçlü bir tekniktir. Analitik model veri kümesinin temel yapısını, sınıflandırma gücü yüksek düzeyde olan değişkenler ile ortaya çıkaran bu teknik, oluşabilecek çoklu bağlantı sorununun da önüne geçmektedir.

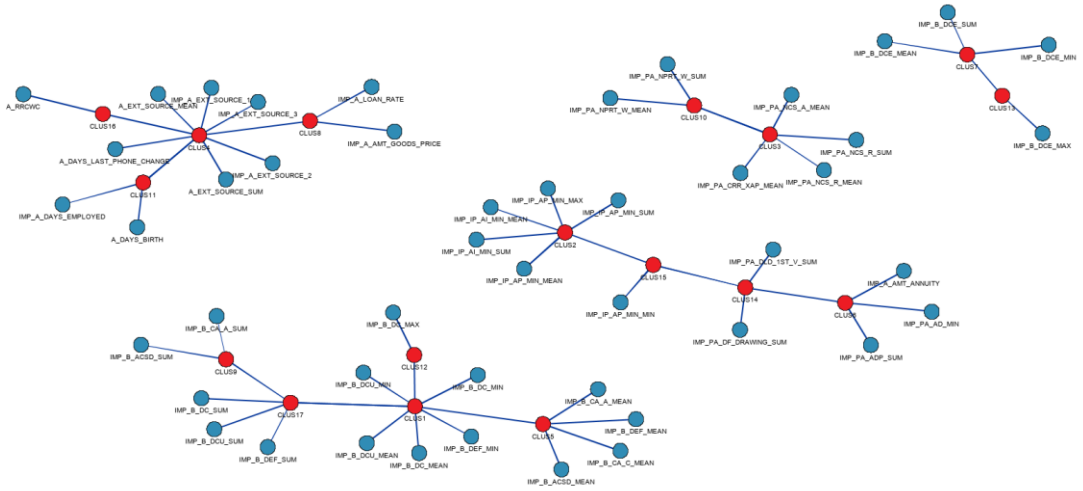
İlgili çalışma kapsamında, SAS Enterprise Miner'ın Variable Clustering düğümü kullanılarak, birbirine benzer yönde hareket eden değişkenler, aralarındaki korelasyonu dikkate alarak bir kümede konumlandırılmıştır. Aynı kümede bulunan değişkenler birbirleri arasında yüksek düzeyde korelasyona sahip iken, diğer kümelerde bulunan değişkenler ile düşük düzeyde bir ilişkiye sahiptirler.

Gerçekleştirilen değişken kümeleme analizi sonucunda elde edilen küme sayısı ve kümelere düşen değişken sayısına Grafik 3.2'de yer verilmiştir.



Grafik 3.2: Değişken Kümeleme Tekniği ile Küme Bazında Değişken Sayısı

Değişken kümeleme tekniği ile elde edilen kümelerin, değişken bazında gösterimine Grafik 3.3'de yer verilmiştir.



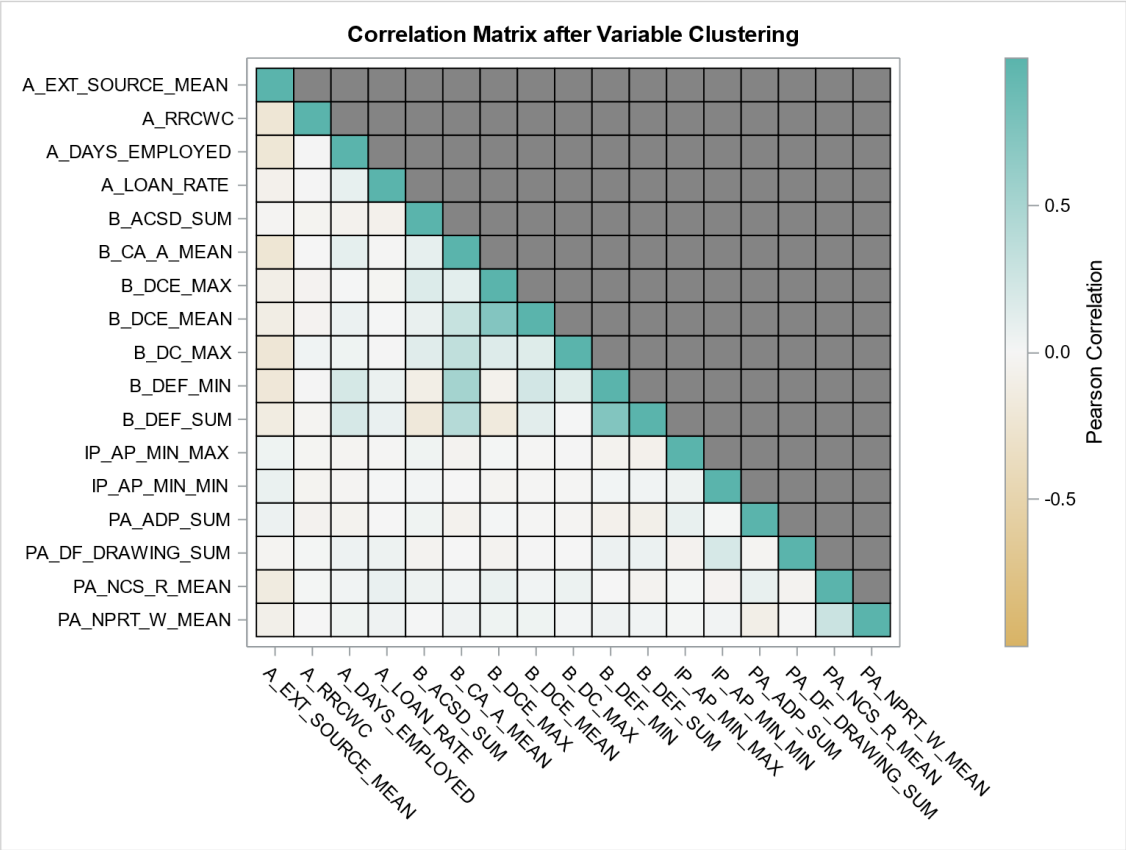
Grafik 3.3: Değişken Kümeleme Grafiği

Analiz sonucunda düğüm, söz konusu kümelere ve değişkenlere ilişkin, R^2 oranını, diğer en yakın kümenin R^2 oranını ve 1-R^2 oranıyla 3 farklı istatistik bilgisi vermektedir. Sahip olunan bu bilgiler doğrultusunda, istatistiki olarak benzer bilgi değeri taşıyan kümelenmiş değişkenler içinden temsilci değişkenler 1-R^2 oranı ile belirlenmiştir. Her kümede minimum 1-R^2 oranına sahip olan değişken, temsilci değişken olarak seçilmiştir.

Tablo 3.6: Değişken Kümeleme ile Öznitelik Seçimi

Sahip Olunan Öznitelik Sayısı	Belirlenen Eleme Kriteri	Eleme Sonrası Öznitelik Sayısı
47	Korelasyon ile Değişken Kümeleme	17

Değişken kümeleme tekniği ile elde edilen 17 değişkene ilişkin korelasyon matrisi incelenmiş olup, aralarındaki ilişkinin %75'in altında olduğu görülmüş ve korelasyon sebebiyle herhangi bir elemeye gidilmemiştir.

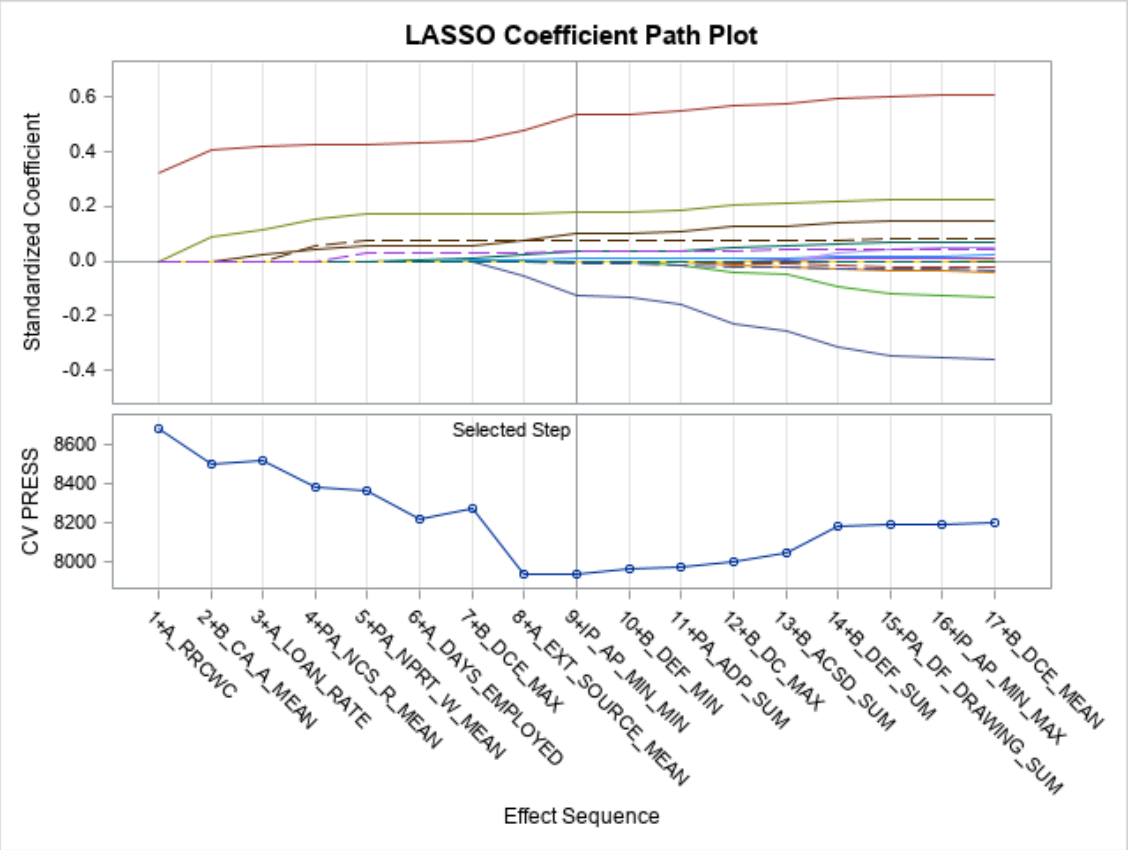


Grafik 3.4: Değişken Kümeleme Sonrası Korelasyon Matrisi

3.3.8 LASSO ile Nihai Öznitelikerin Belirlenmesi

Uygulanan birçok farklı değişken indirgeme teknikleriyle elde edilen 17 değişken arasından, modelleme öncesi nihai değişkenlerin belirlenme işlemi LASSO tekniği ile gerçekleştirilmiştir. K-katlamalı çapraz doğrulama yöntemiyle model performansına ve makine öğrenmesi tekniğine pozitif katkı sağlayan LASSO, geleneksel yaklaşımlara göre makine öğrenmesi tekniklerinde çoklu fayda sağlayan bir teknik haline gelmiştir. Bu bağlamda, eğitim veri seti 10 parçaya bölünerek k-katlamalı çapraz doğrulama ile modelin tahmine dayalı performansı değerlendirilmiştir.

Sahip olduğu λ ceza parametresiyle katsayıların mutlak boyutunu daraltan LASSO tekniği, uygulama kapsamında düşük düzeydeki tahmincilerin elenerek analiz dışı bırakılmasını sağlamıştır. Toplamda 17 değişken ile gerçekleştirilen LASSO analizine ait katsayıların daraltılma adımlarına Grafik 3.5'te yer verilmiştir.



Grafik 3.5: Katsayıların Daraltılma Grafiği

Değişkenlerin modelde yer almasında göre katsayı değerlerinin adım adım değişikliğini gösteren grafiğin (Standardized Coefficient), dikey çizgisi, eğitim ve doğrulama verisi üzerinden optimal modele karşılık gelen adımı ifade etmektedir. Ek olarak, eğitim veri setinin 10 parçaya bölümlenmesiyle gerçekleştirilen k-katlamalı çapraz doğrulamadan elde edilen çıktılara iç grafikte (CV PRESS) yer verilmiştir. Minimum kalıntı kareler toplamına sahip modeli, optimal model olarak seçen k-katlamalı çapraz doğrulama tekniği, yine grafiğin dikey çizgisine karşılık gelen adımında elde etmiştir.

Modellemenin optimal noktasında sahip olunan 9 öngörücü ve nihai model girdi değişkenlerine ilişkin açıklamalarına Tablo 3.7'de yer verilmiştir. Diğer tüm girdi değişkenleri, katsayı değerlerinin sıfır olması sebebiyle elenerek analiz dışı bırakılmıştır. Ek olarak, nihai girdi değişkenleri için nihai korelasyon kontrolü gerçekleştirilmiş olup, aralarındaki ilişkinin %30'un altında olduğu görülmüş ve herhangi bir elemeye gidilmemiştir.

Tablo 3.7: Nihai Öznitelikler ve Açıklamaları

Öznitelik	Açıklama
A_DAYS_EMPLOYED	Başvuru sahibinin ilgili başvurudan kaç gün önce bir iş yerine istihdam edildiği gün sayısıdır
A_EXT_SOURCE_MEAN	İlgili başvuru için harici veri kaynağından elde edilen başvuru sahibine ait skor puanlarının ortalaması
A_LOAN_RATE	İlgili başvuru için istenilen kredi tutarının verilen kredi tutarına oranı
A_RRCWC	İlgili başvuru sahibinin yaşadığı şehri dikkate alarak Home Credit'in bölgedeki reyting oranı
B_CA_A_MEAN	İlgili başvuru sahibinin Kredi Kayıt Bürosundaki aktif olan ortalama kredi sayısının, ödenmemiş kredilerin sayısına oranı
B_DCE_MAX	İlgili başvuru için Kredi Kayıt Bürosundaki aktif olan kredi ürünlerinin ödemelerine kalan maksimum gün sayısı
IP_AP_MIN_MIN	İlgili başvuru sahibinin bir önceki kredide gerçekleştirdiği minimum ödeme miktarının, aktif kredilerine yaptığı minimum ödeme miktarına oranı
PA_NCS_R_MEAN	Önceki başvuruların sözleşmelerine ilişkin ortalama reddedilme oranı
PA_NPRT_W_MEAN	İlgili başvuru sahibinin başvuruda bulunduğu önceki bankada düzenli hesabı bulunmama durumunun ortalaması

3.4 MODEL GELİŞTİRME

Çalışmanın bu bölümünde değişken indirgeme teknikleri ile elde edilen nihai öznitelik değişkenleri kullanılarak, kredilerin temerrüt riski (0-1) yukarıda ele alınan sınıflandırma algoritmaları ile belirlenmeye çalışılacak ve kullanılan algoritmaların sınıflandırma performansları karşılaştırılacaktır.

3.4.1 Lojistik Regresyon ile Model Geliştirme

Nihai öznitelik değişkenleri ile kurulan Lojistik Regresyon modelinde backward (geriye doğru eleme) metodu kullanılarak, %95 güven düzeyinde değişkenlerin

anlamlılığı test edilmiştir. Tekniğin ilk adımında nihai değişkenler arasından *B_DCE_MAX* değişkenine ait p değerinin 0.5422 olması sebebiyle elenerek modelden çıkartılmıştır. İkinci adımda kurulan modelde tüm değişkenlerin önemlilik değerleri 0.05'ten küçük olması sebebiyle backward yöntemi optimal Lojistik Regresyon modelini aşağıdaki değişkenler ile elde etmiştir.

Tablo 3.8: Lojistik Regresyon Katsayıları

Parametre	Tahmini Katsayı	Standart Hata	Wald X^2	Pr > X^2
Intercept	1.8792	0.0909	426.92	<.0001
A_EXT_SOURCE_MEAN	-4.5121	0.093	2354.49	<.0001
A_RRCWC	0.1466	0.0265	30.68	<.0001
A_DAYS_EMPLOYED	0.000099	6.91E-06	207.3	<.0001
A_LOAN_RATE	-2.009	0.5996	11.23	0.0008
B_CA_A_MEAN	0.3665	0.046	63.54	<.0001
IP_AP_MIN_MIN	-0.00002	1.95E-06	96.7	<.0001
PA_NCS_R_MEAN	0.7549	0.0693	118.72	<.0001
PA_NPRT_W_MEAN	0.5806	0.0704	67.92	<.0001

Yukarıdaki kurulan Lojistik Regresyon modeli kullanılarak eğitim ve doğrulama veri setleriyle gerçekleştirilen sınıflandırma oranlarına aşağıda yer verilmiştir.

Tablo 3.9: Eğitim ve Doğrulama Veri Setleri ile Lojistik Regresyona Ait Sınıflandırma Oranları

Gerçek Değer	Tahmin Değeri	Eğitim	Doğrulama
Temerrüde Düşmeyen	Temerrüde Düşmeyen	%33.97	%33.71
Temerrüde Düşen	Temerrüde Düşmeyen	%17.09	%16.91
Temerrüde Düşmeyen	Temerrüde Düşen	%16.03	%16.29
Temerrüde Düşen	Temerrüde Düşen	%32.91	%33.09

B_DCE_MAX değişkeninin modelden çıkarılarak elde edilen optimal Lojistik Regresyon modeline göre eğitim veri setleriyle gerçekleştirilen sınıflandırma oranlarındaki doğruluk başarısı, gerçekte temerrüde düşmeyenlerin tahmin değerinde temerrüde düşmeyenler olarak sınıflandırılmasında ve gerçekte temerrüde düşenlerin tahmin değerlerinde temerrüde düşen olarak sınıflandırılmasında yanlış sınıflandırma oranlarına nispeten daha iyi sonuç vermiştir. Benzer oranlar doğrulama veri setiyle gerçekleştirilen tahminlerde de görülmekte olup, kurulan model ile eğitim ve doğrulama veri setleri arasında başarılı bir uyum olduğu söylenebilir. Ek olarak, Lojistik Regresyon modeli eğitim veri setiyle gerçekleştirdiği tahminlerin %66.88'inde doğru sınıflandırma oranı elde ederken, doğrulama veri setiyle %66.80 oranında doğru sınıflandırma başarısı göstermiştir.

3.4.2 Yapay Sinir Ağları ile Model Geliştirme

Yapay Sinir Ağları modeli, nihai 9 değişken ile 2 katman üzerinden 20 nöron oluşturacak şekilde, tanjant hiperbolik aktivasyon fonksiyonu kullanılarak, 300 iterasyon ile eğitim ve doğrulama veri setleri üzerinden elde edilmiştir. Doğrulama veri setiyle belirlenen optimum model, 139. iterasyonda minimum doğrulama hatasını elde ederek durdurulmuştur. Elde edilen optimum YSA modeline ait girdi değişkenleri ve gizli nöronların girdi ve çıktı ağırlıklarına Tablo 3.10'da yer verilmiştir.

Tablo 3.10: YSA Modeline Ait Girdi ve Çıktı Ağırlıkları

Parametre	X	Y	Rol	Katman
LOAN_STATUS=1	400	175	Hedef Değişken	Output
A_EXT_SOURCE_MEAN	0	35	Öznitelik	0
A_RRCWC	0	70	Öznitelik	0
A_DAYS_EMPLOYED	0	105	Öznitelik	0
A_LOAN_RATE	0	140	Öznitelik	0
B_CA_A_MEAN	0	175	Öznitelik	0
B_DCE_MAX	0	210	Öznitelik	0
IP_AP_MIN_MIN	0	245	Öznitelik	0
PA_NCS_R_MEAN	0	280	Öznitelik	0
PA_NPRT_W_MEAN	0	315	Öznitelik	0
H1	133.33	31.82	Gizli Nöron	1
H2	133.33	63.64	Gizli Nöron	1

H3	133.33	95.45	Gizli Nöron	1
H4	133.33	127.27	Gizli Nöron	1
H5	133.33	159.09	Gizli Nöron	1
H6	133.33	190.91	Gizli Nöron	1
H7	133.33	222.73	Gizli Nöron	1
H8	133.33	254.55	Gizli Nöron	1
H9	133.33	286.36	Gizli Nöron	1
H10	133.33	318.18	Gizli Nöron	1
H11	266.67	31.82	Gizli Nöron	2
H12	266.67	63.64	Gizli Nöron	2
H13	266.67	95.45	Gizli Nöron	2
H14	266.67	127.27	Gizli Nöron	2
H15	266.67	159.09	Gizli Nöron	2
H16	266.67	190.91	Gizli Nöron	2
H17	266.67	222.73	Gizli Nöron	2
H18	266.67	254.55	Gizli Nöron	2
H19	266.67	286.36	Gizli Nöron	2
H20	266.67	318.18	Gizli Nöron	2

Yukarıda optimum YSA modeliyle eğitim ve doğrulama veri setiyle elde edilen sınıflandırma oranlarına aşağıda yer verilmiştir.

Tablo 3.11: Eğitim ve Doğrulama Veri Setleri ile YSA Sınıflandırma Oranları

Gerçek Değer	Tahmin Değeri	Eğitim	Doğrulama
Temerrüde Düşmeyen	Temerrüde Düşmeyen	%33.68	%33.47
Temerrüde Düşen	Temerrüde Düşmeyen	%16.07	%15.82
Temerrüde Düşmeyen	Temerrüde Düşen	%16.32	%16.53
Temerrüde Düşen	Temerrüde Düşen	%33.93	%34.18

İlgili eğitim ve doğrulama veri setleriyle gerçekleştirilen sınıflandırma oranlarına incelendiğinde, Lojistik Regresyona göre nispeten yakın değerlere sahip olsa da YSA'nın her iki veri setiyle temerrüde düşenler için gerçekleştirdiği sınıflandırma başarısının daha iyi değerlere sahip olduğu görülmektedir. Ek olarak, eğitim veri setiyle gerçekleştirilen tahminlerin toplam doğruluk oranı %67.61 iken, doğrulama veri setleriyle bu oran %67.65 kadardır.

3.4.3 Karar Ağacı ile Model Geliştirme

Karar Ağacı algoritmasında bulunan düğümlerin bölünme kuralı, hedef değişkenin dikotom yapıda olması sebebiyle entropi ölçüsüyle gerçekleştirilmiştir. Bu kontekste, algoritmanın uygulanmasında maksimum dallanma için 2, maksimum düğüm derinliği için 10, her yaprak düğümünde eğitim için gözlem sayısının minimum 5 olacak şekilde hiperparametreler belirlenmiş olup, eğitim ve doğrulama veri setleriyle elde edilen önemlilik oranlarına aşağıda yer verilmiştir.

Tablo 3.12: Eğitim ve Doğrulama Veri Setleri ile Karar Ağacına Göre Değişkenlerin Önemlilik Oranları

Öznitelik	Eğitim Hata Kareler Toplamı	Eğitim Önemlilik Oranı	Doğrulama Hata Kareler Toplamı	Doğrulama Önemlilik Oranı
A_EXT_SOURCE_MEAN	43.65	1	25.05	1
A_DAYS_EMPLOYED	11.50	0.26	5.71	0.23
A_LOAN_RATE	11.21	0.26	5.09	0.20
B_CA_A_MEAN	8.82	0.20	3.71	0.15
PA_NCS_R_MEAN	8.63	0.20	4.82	0.19
IP_AP_MIN_MIN	8.49	0.19	3.50	0.14
B_DCE_MAX	6.43	0.15	1.48	0.06
PA_NPRT_W_MEAN	4.76	0.11	2.77	0.11
A_RRCWC	3.07	0.07	1.35	0.05

Karar Ağacının entropi ile değişkenlerin sınıflandırma başarısına göre önemlilik oranlarının hesapladığı yukarıdaki değerler, hedef değişken üzerinden sahip olduğu tüm kombinasyonlarla birlikte, belirli eşik değerler üzerinden kural indüksiyonlarıyla her bir düğümde hesaplanma örneğine aşağıda yer verilmiştir.

```
*-------------------------------------------------------------*
NODE = 105
*-------------------------------------------------------------*
(Imputed: DAYS_EMPLOYED < -1640)
AND (Imputed: PREVIOUS_NAME_CONTRACT_STATUS_Refused_mean >= 0.11)
AND (EXT_SOURCE_MEAN < 0.50134067)
AND MISSING(Imputed: DAYS_EMPLOYED) OR (Imputed: DAYS_EMPLOYED >= -2648)
AND (Imputed: IN_CLIENT_IN_LOAN_AMT_PAYMENT_min_min >= 3429)
AND (EXT_SOURCE_MEAN < 0.56464817)
AND MISSING(EXT_SOURCE_MEAN) OR (EXT_SOURCE_MEAN >= 0.45386004)
    PREDICTED VALUE IS 1
    PREDICTED 1 = 0.7119( 42/59)
    PREDICTED 0 = 0.2881( 17/59)
*-------------------------------------------------------------*
NODE = 154
*-------------------------------------------------------------*
(Imputed: IN_CLIENT_IN_LOAN_AMT_PAYMENT_min_min >= 8001)
AND MISSING(Imputed: PREVIOUS_NAME_CONTRACT_STATUS_Refused_mean) OR (Imputed: PREVIOUS_NAME_CONTRACT_STATUS_Refused_mean >= 0.13)
AND MISSING(Imputed: DAYS_EMPLOYED) OR (Imputed: DAYS_EMPLOYED >= -1640)
AND (Imputed: PREVIOUS_NAME_CONTRACT_STATUS_Refused_mean >= 0.11)
AND (EXT_SOURCE_MEAN < 0.50134067)
AND MISSING(Imputed: DAYS_EMPLOYED) OR (Imputed: DAYS_EMPLOYED >= -2648)
AND (Imputed: IN_CLIENT_IN_LOAN_AMT_PAYMENT_min_min >= 3429)
AND (EXT_SOURCE_MEAN < 0.56464817)
AND MISSING(EXT_SOURCE_MEAN) OR (EXT_SOURCE_MEAN >= 0.45386004)
    PREDICTED VALUE IS 0
    PREDICTED 1 = 0.3827( 31/81)
    PREDICTED 0 = 0.6173( 50/81)
*-------------------------------------------------------------*
NODE = 110
```

Şekil 3.1: Düğüm Kuralları Örneği

Yukarıdaki kural indüksiyonlarıyla gerçekleştirilen Karar Ağacı algoritmasının eğitim ve doğrulama veri setleriyle gerçekleştirdiği sınıflandırma oranlarına aşağıda yer verilmiştir.

Tablo 3.13: Eğitim ve Doğrulama Veri Setleri ile Karar Ağacına Ait Sınıflandırma Oranları

Gerçek Değer	Tahmin Değeri	Eğitim	Doğrulama
Temerrüde Düşmeyen	Temerrüde Düşmeyen	%34.31	%33.11
Temerrüde Düşen	Temerrüde Düşmeyen	%15.28	%15.82
Temerrüde Düşmeyen	Temerrüde Düşen	%15.69	%16.89
Temerrüde Düşen	Temerrüde Düşen	%34.72	%34.18

Sınıflandırma oranları incelendiğinde, Karar Ağacı algoritmasının eğitim veri setleriyle gerçekleştirdiği tahminlerin Lojistik Regresyon ve YSA'ya göre daha başarılı olduğu gözlemlenmektedir. Eğitim veri setiyle gerçekleştirilen tahminlerin doğruluk oranı %69.03 iken, doğrulama için %67.29'dur. Bu bağlamda, doğrulama veri setindeki oran YSA'nın doğrulama veri setiyle elde edilen doğruluk oranının altında kaldığı gözlemlenmiştir.

3.4.4 Destek Vektör Makineleri ile Model Geliştirme

Destek Vektör Makineleri ile hedef değişkene ait iki sınıf arasındaki marjı maksimize etmek için doğrusal çekirdek fonksiyonu kullanılarak, 16. iterasyonda optimum model elde edilmiştir. İlgili nihai değişkenlerin etkileşimleriyle, iç çarpım değeri, sapma değeri, oluşturulan destek vektörlerin sayısı, marjın maksimum olduğu hiperdüzlemdeki vektörlerin sayısı gibi optimum modele ait çıktılara aşağıda yer verilmiştir.

Tablo 3.14: SVM Optimum Model Oranları

Etkileşime Giren Değişken Sayısı	9
Ağırlıkların İç Çarpımı	22.40476
Sapma	-0.63578
Toplam Bolluk (Kısıt İhlalleri)	21115.17
En Uzun Vektör Normu	2.275462
Destek Vektörlerin Sayısı	21158
Marj Üzerindeki Destek Vektörlerinin Sayısı	21123

Eğitim veri setiyle, yukarıdaki oranlar ile elde edilen optimal modelin doğrulama veri seti oranlarıyla birlikte sınıflandırma başarısına aşağıda yer verilmiştir.

Tablo 3.15: Eğitim ve Doğrulama Veri Setleri ile SVM Sınıflandırma Oranları

Gerçek Değer	Tahmin Değeri	Eğitim	Doğrulama
Temerrüde Düşmeyen	Temerrüde Düşmeyen	%34.29	%34.00
Temerrüde Düşen	Temerrüde Düşmeyen	%17.37	%17.02
Temerrüde Düşmeyen	Temerrüde Düşen	%15.71	%16.00
Temerrüde Düşen	Temerrüde Düşen	%32.63	%32.99

SVM algoritmasının sınıflandırma oranları incelendiğinde, eğitim veri setinde doğru sınıflandırma oranı %66.92 iken, doğrulama veri setinde bu oran %66.99'dur. SVM ve Lojistik Regresyon algoritmasının doğru sınıflandırma oranları incelendiğinde birbirlerine oldukça yakın değerlere sahip olduğu görülmektedir.

3.4.5 K-En Yakın Komşu ile Model Geliştirme

Algoritmanın eğitim veri seti üzerinden öğrenme becerisi için SAS Enterprise Miner'ın MBR (Memory-Based Reasoning) düğümünün yeteneklerinden faydalanarak, bellek tabanlı akıl yürütme ile boyut indirgeme ağacı (RD-Tree) metodu kullanılmıştır. Veri setinden kümülatif olarak azalacak şekilde elde edilen alt kümelerin, çok boyutlu uzayda ikili ağaçlara bölünmesiyle gerçekleştirilen RD-Tree metodu, en yakın komşuları bulmak için pratik ve hızlı bir yöntemdir. Bu bağlamda, maksimum bölünme için 100 küme grubu oluşturulurken, gözlemlerin tahmini için komşu sayısı (k) 200 olarak belirlenmiştir. Ek olarak, hedef değişken ile nihai değişkenler arasındaki korelasyonunun mutlak değeri ağırlıklandırılarak optimum model elde edilmiştir.

İlgili eğitim veri setinin öğrenme becerisi ve doğrulama veri setiyle gerçekleştirilen sınıflandırmalara oranlarına aşağıda yer verilmiştir.

Tablo 3.16: Eğitim ve Doğrulama Veri Setleri ile KNN Algoritmasına Ait Sınıflandırma Oranları

Gerçek Değer	Tahmin Değeri	Eğitim	Doğrulama
Temerrüde Düşmeyen	Temerrüde Düşmeyen	%27.36	%26.83
Temerrüde Düşen	Temerrüde Düşmeyen	%17.81	%17.99
Temerrüde Düşmeyen	Temerrüde Düşen	%22.64	%23.17
Temerrüde Düşen	Temerrüde Düşen	%32.19	%32.01

Algoritma, eğitim veri setiyle toplamda %59.55 doğrulma veri setiyle toplamda %58.84 oranında doğru sınıflandırma oranına sahiptir. K-En Yakın Komşu algoritmasının her iki veri setiyle gerçekleştirdiği doğru sınıflandırmaların diğer algoritmalara görece daha başarısız performansa sahip olduğu görülmektedir.

3.4.6 Rassal Orman ile Model Geliştirme

Rassal orman algoritması, maksimum 50 ağaç ve 10 düğüm derinliği ile her ağaçta veri setindeki rastgele gözlemlerin %60'ı kullanılmış olup, Kayıp Azaltma (Loss Reduciton) tekniği ile değişken önemliliği belirlenmiştir. Bu doğrultuda geliştirilen Rassal Orman modeline ilişkin çıktılara aşağıda yer verilmiştir.

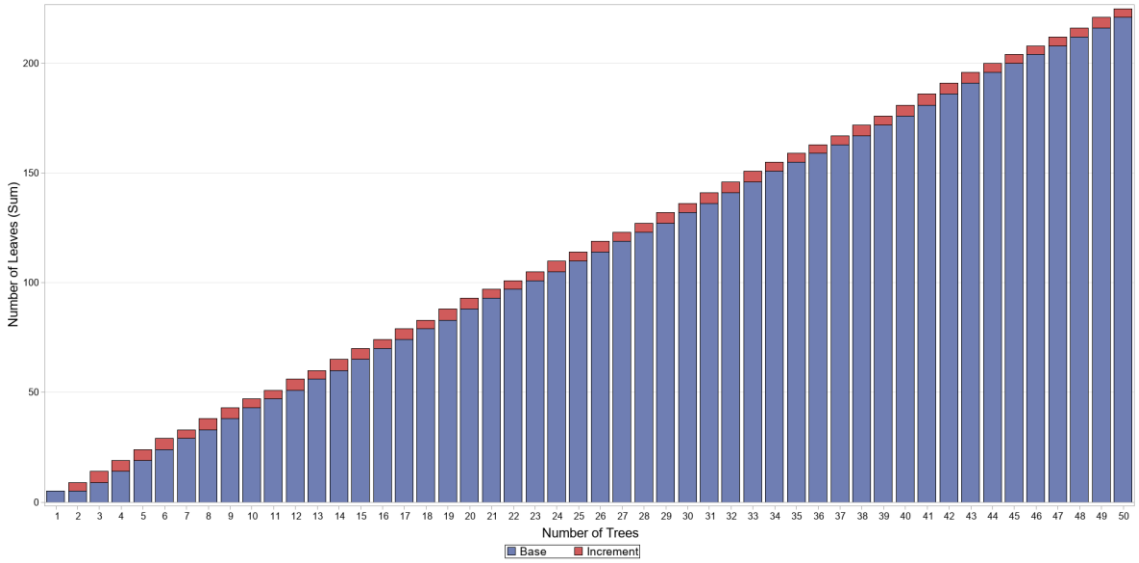
Tablo 3.17: Rassal Orman Algoritmasına Göre Değişkenlerin Önemlilik Oranları

Öznitelik	Bölme Kural Sayısı	Eğitim: Gini İndirgeme	Eğitim: Marj İndirgeme	OOB: Gini İndirgeme	OOB: Marj İndirgeme	Doğrulama: Gini İndirgeme	Doğrulama: Marj İndirgeme
A_EXT_SOURCE_MEAN	62	0.034060	0.068119	0.033156	0.067198	0.033509	0.067658
A_DAYS_EMPLOYED	28	0.003701	0.007402	0.003415	0.007131	0.003882	0.007548
B_CA_A_MEAN	17	0.002363	0.004726	0.002078	0.004441	0.002497	0.004857
A_LOAN_RATE	22	0.001094	0.002187	0.000952	0.002037	0.001033	0.002106
PA_NCS_R_MEAN	13	0.001094	0.002188	0.001067	0.002144	0.001033	0.002098
IP_AP_MIN_MIN	21	0.000846	0.001693	0.000537	0.001397	0.000848	0.001716
PA_NPRT_W_MEAN	9	0.000752	0.001504	0.000770	0.001525	0.000632	0.001380
B_DCE_MAX	3	0.000073	0.000147	0.000006	0.000083	0.000062	0.000141
A_RRCWC	0	0	0	0	0	0	0

Tablo 3.17'de eğitim %60, OOB (Out of Bag) %40 ve doğrulama veri setleriyle gerçekleştirilen girdi değişkenlerinin Rassal Ormanlar ile sınıflandırma sonuçlarına göre önemlilik oranları bulunmaktadır. "*A_RRCWC*" değişkeninin Rassal Orman algoritmasına göre hedef değişken üzerinde herhangi bir önemliliği olmadığı, bölünme kuralına sahip olmamasıyla anlaşılmaktadır. Bu bağlamda, girdi değişkeni olarak kullanılmayarak modelden çıkarılmıştır.

Tabloda bulunan "Gini Reduction" alanları, modelin sınıflandırma başarısını Gini oranlarıyla verirken, "Margin Reduction" alanları, gerçek sınıfın olasılığı ile diğer sınıfların maksimum olasılığının çıkarılmasıyla elde edilen marj oranını ifade etmektedir. Her iki oranda yüksek değerler tercih edilmekle beraber, çalışma kapsamında veri setlerinin oranlarındaki uyum dikkate alınarak, model geliştirme sonlandırılmıştır. Ek olarak, Rassal Ormanların, değişkenlerin açıklayıcı oranlarına göre seçilebilir yapısıyla, veri madenciliği algoritmaları mimarisine oldukça uygun olduğu söylenebilir.

Algoritmanın, eğitim verisiyle öğrenme başarısının ölçümü için bir diğer teknik, her yeni ağaçta oluşan yaprak düğüm sayısının yaklaşık olarak eşit oranlarda artış göstermesidir. Bu kapsamda, Rassal Ormanlar ile elde edilen ağaçların sahip olduğu yaprak düğümü sayısına, aşağıdaki grafikte artış oranlarıyla birlikte yer verilmiştir.



Grafik 3.6: Yaprak Grafiği

Grafik 3.6 incelendiğinde, yaprak düğüm sayısındaki artış oranları yaklaşık olarak eşit oranlara sahip olduğu görülmektedir.

Tablo 3.18: Eğitim ve Doğrulama Veri Setleri ile Rassal Orman Algoritmasına Ait Sınıflandırma Oranları

Gerçek Değer	Tahmin Değeri	Eğitim	Doğrulama
Temerrüde Düşmeyen	Temerrüde Düşmeyen	%35.15	%34.63
Temerrüde Düşen	Temerrüde Düşmeyen	%18.71	%18.22
Temerrüde Düşmeyen	Temerrüde Düşen	%14.84	%15.37
Temerrüde Düşen	Temerrüde Düşen	%31.29	%31.78

Rassal ormanların, eğitim ve doğrulama veri setleriyle temerrüde düşmeyenler için gerçekleştirdiği doğru tahmin oranları, her ne kadar başarılı olsa da toplam doğru sınıflandırma oranı K-En Yakın Komşu haricinde diğer algoritmalara görece daha düşük kalmıştır. Eğitim veri seti için bu oran toplamda %66.44 olurken, doğrulama veri setinde %66.41'dir.

3.4.7 Lojistik Regresyon (WOE) ile Model Geliştirme

Nihai 9 değişkenin, hedef değişkene ait iyi ve kötü kredi dağılımları üzerinden açıklayıcı gücünün hesaplandığı WOE değerleriyle Scorecard düğümü yardımıyla

Lojistik Regresyon modeli geliştirilerek, ham haliyle kurulan Lojistik Regresyon modeli kıyaslanmıştır. Nihai değişkenlerin WOE'li halleriyle gerçekleştirilen Lojistik Regresyon algoritmasında backward tekniği kullanılmıştır. Bu kapsamda, ilgili değişkenlerin modeldeki anlamlılığı ile eğitim ve doğrulama veri setleriyle gerçekleştirilen sınıflandırma oranlarına aşağıda yer verilmiştir.

Tablo 3.19: Lojistik Regresyon (WOE) Katsayıları

Parametre	Tahmini Katsayı	Standart Hata	Wald X^2	Pr > X^2
Intercept	-0.00069	0.013	0	0.9574
WOE_PA_NCS_R_MEAN	-0.4966	0.0544	83.37	<.0001
WOE_PA_NPRT_W_MEAN	-0.3833	0.0675	32.27	<.0001
WOE_A_EXT_SOURCE_MEAN	-0.8322	0.0181	2109.12	<.0001
WOE_A_RRCWC	-0.4222	0.0666	40.24	<.0001
WOE_B_CA_A_MEAN	-0.3432	0.0487	49.69	<.0001
WOE_A_DAYS_EMPLOYED	-0.528	0.0371	202.08	<.0001
WOE_A_LOAN_RATE	-0.6329	0.0442	205.23	<.0001
WOE_B_DCE_MAX	-0.3843	0.063	37.26	<.0001

Yukarıda değişkenlerin WOE'li halleriyle kurulan Lojistik Regresyon modelinde nihai 9 değişkenin modelde anlamlı olduğu görülmekte olup herhangi bir elemeye gidilmemiştir. Eğitim verileri üzerinden geliştirilen modelin tahmin oranları ile doğrulama verisine ait tahmin oranlarına Tablo 3.20'de yer verilmiştir.

Tablo 3.20: Eğitim ve Doğrulama Veri Setleri ile Lojistik Regresyon (WOE) Algoritmasına Ait Sınıflandırma Oranları

Gerçek Değer	Tahmin Değeri	Eğitim	Doğrulama
Temerrüde Düşmeyen	Temerrüde Düşmeyen	%34.24	%33.73
Temerrüde Düşen	Temerrüde Düşmeyen	%16.63	%16.36
Temerrüde Düşmeyen	Temerrüde Düşen	%15.76	%16.27
Temerrüde Düşen	Temerrüde Düşen	%33.37	%33.64

Değişkenlerin WOE'li halleriyle kurulan Lojistik Regresyon modelinin, eğitim ve doğrulama veri setleriyle gerçekleştirdiği sınıflandırma oranları gözlemlendiğinde, ham halleriyle kurulan Lojistik Regresyon modelinin sınıflandırma oranlarına göre daha başarılı olduğu görülmektedir. Eğitim veri setinin toplam doğru sınıflandırma oranı %67.61 olurken, doğrulama verinde bu oran %67.37'dir. Ek olarak, WOE'li Lojistik Regresyon modelinin YSA modeline ait sınıflandırma performansında benzer oranlara sahip olduğu söylenebilir.

3.4.8 Gradyan Artırma ile Model Geliştirme

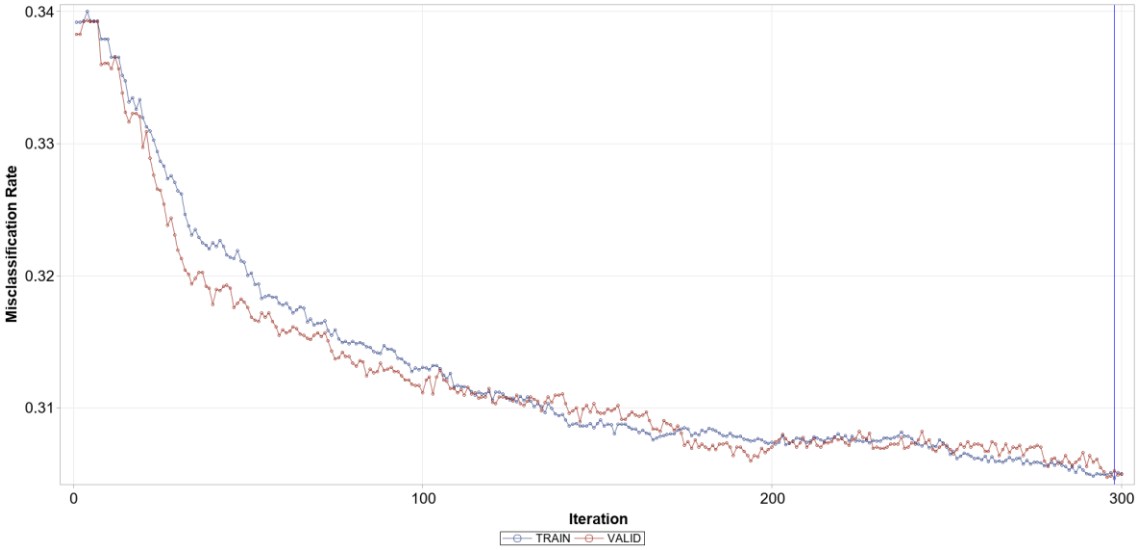
Gradyan Artırma algoritması ile model geliştirme sürecinde, eğitim veri setinin %70'i, 0.1 oranında öğrenme hızına sahip olacak şekilde daralma değeri tanımlanmıştır. Ek olarak, maksimum 30 dallanma kuralı ve 10 düğüm derinliği ile sınırlandırılarak 300 iterasyon ile algoritma optimum modeli elde etmiştir.

Belirlenen hiperparametreler üzerinden eğitim ve doğrulama veri seti ile elde edilen değişkenlerin önemlilik oranlarına Tablo 3.21'de yer verilmiştir.

Tablo 3.21: Gradyan Artırma Algoritmasına Göre Değişkenlerin Önemlilik Oranları

Öznitelik	Bölme Kural Sayısı	Eğitim: Önemlilik Oranı	Doğrulama: Önemlilik Oranı
A_EXT_SOURCE_MEAN	114	1	1
A_LOAN_RATE	152	0.54456251	0.457866369
A_DAYS_EMPLOYED	88	0.317967324	0.259472756
IP_AP_MIN_MIN	73	0.249623648	0.208416438
B_CA_A_MEAN	24	0.19997376	0.180501344
PA_NCS_R_MEAN	14	0.178122265	0.168224794
B_DCE_MAX	65	0.25380576	0.148651308
A_RRCWC	8	0.091933353	0.139179555
PA_NPRT_W_MEAN	11	0.118126583	0.088401028

Yukarıda önemlilik oranlarıyla bulunan nihai 9 değişken ile eğitim ve doğrulama veri setiyle gerçekleştirilen hatalı sınıflandırmaların grafiksel gösterimine aşağıda yer verilmiştir.



Grafik 3.7: Eğitim ve Doğrulama Veri Setleri ile Gradyan Artıma Algoritmasına Ait Yanlış Sınıflandırma Grafiği

Grafiğe göre eğitim ve doğrulama veri seti ile gerçekleştirilen tahminlerin yanlış sınıflandırma oranları arasında iyi oranda uyumlu olduğu gözlemlenmektedir. Ayrıca, 298. iterasyon ile yanlış sınıflandırma oranı 0.04 oranında azalarak, optimum modeli 0.31'in altında bir değer ile elde etmiştir.

Bu kontekste, geliştirilen model ile eğitim ve doğrulama veri setlerinin tahmin başarısı, aşağıdaki sınıflandırma oranlarıyla detaylandırılmıştır.

Tablo 3.22: Eğitim ve Doğrulama Veri Setleri ile Gradyan Artırma Algoritmasına Ait Sınıflandırma Oranları

Gerçek Değer	Tahmin Değeri	Eğitim	Doğrulama
Temerrüde Düşmeyen	Temerrüde Düşmeyen	%35.08	%34.93
Temerrüde Düşen	Temerrüde Düşmeyen	%15.84	%15.53
Temerrüde Düşmeyen	Temerrüde Düşen	%14.91	%15.07
Temerrüde Düşen	Temerrüde Düşen	%34.17	%34.47

Temerrüde düşenler ve temerrüde düşmeyenler için gerçekleştirilen tahmin oranlarının, Gradyan Artırma algoritmasıyla eğitim ve doğrulama veri setiyle en iyi oranlara sahip olduğu gözlemlenmektedir. Toplamda eğitim veri setinin doğru sınıflandırma oranı %69.51 olurken, doğrulama veri setiyle bu oran %69.53'tür.

3.5 PERFORMANS DEĞERLENDİRME

Belirlenen nihai değişkenler ile eğitim, doğrulama ve test veri setleri üzerinden algoritmalar iyi ve kötü müşterileri sınıflandırarak, karmaşıklık matrisleri elde edilmiştir. Test veri seti üzerinden her bir algoritmaya ait karmaşıklık matrisinden elde edilen sınıflandırma ölçülerinin sonuçlarına Tablo 3.23'de yer verilmiştir.

Tablo 3.23: Algoritmalara Ait Sınıflandırma Sonuçları

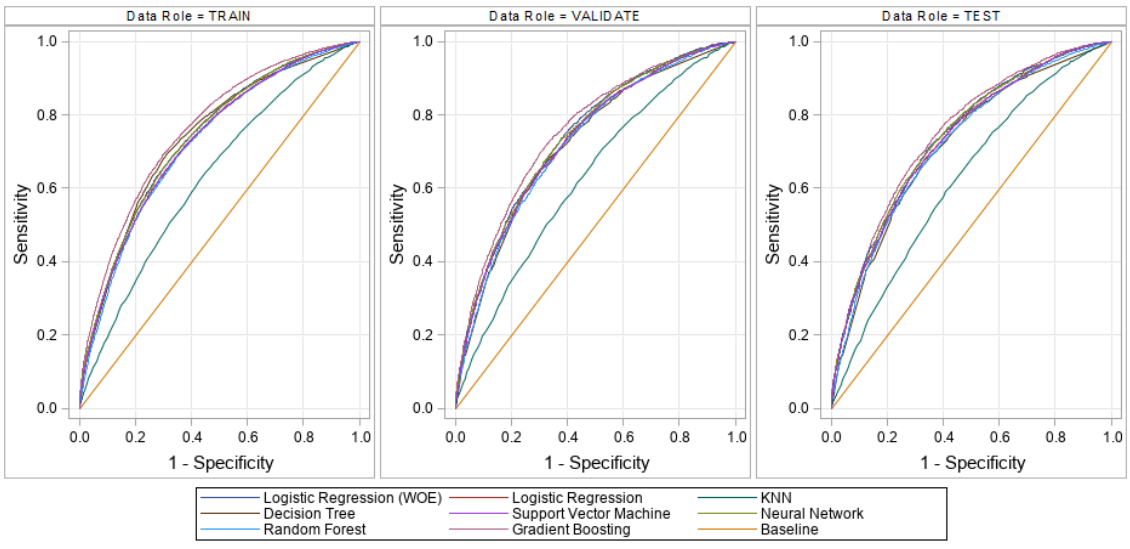
Model Algoritmaları	Doğruluk	Hassasiyet	Özgüllük	Kesinlik	F1 Skoru
Gradyan Artırma	%68.59	%68.09	%69.09	%68.77	%68.43
Lojistik Regresyon (WOE)	%67.74	%67.10	%68.39	%67.97	%67.53
Yapay Sinir Ağları	%67.71	%67.73	%67.69	%67.70	%67.71
Lojistik Regresyon	%67.11	%65.79	%68.43	%67.57	%66.67
Destek Vektör Makineleri	%67.03	%65.20	%68.85	%67.57	%66.41
Rassal Orman	%66.31	%62.39	%70.23	%67.69	%64.93
Karar Ağacı	%66.56	%66.82	%66.30	%66.47	%66.65
K-En Yakın Komşu	%58.85	%63.13	%54.57	%58.15	%60.53

Yukarıdaki oranlar incelendiğinde, K-En Yakın Komşu algoritması temerrüde düşenler ile temerrüde düşmeyenlerin sınıflandırılmasında diğer algoritmalara görece daha başarısız sonuçlar vermektedir. Sınıflandırma başarısının genel performans bilgisini veren doğruluk oranı için K-En Yakın Komşu algoritması haricinde diğer algoritmalar birbirlerine yakın değerlere sahip olsa da Gradyan Artırma algoritmasının en başarılı sınıflandırma oranına sahip olduğu söylenebilir. Tip-II hata perspektifiyle, gerçekte temerrüde düşen müşteriler için algoritmaların sınıflandırma başarısını ölçen hassasiyet oranı için Gradyan Artırma algoritması en başarılı sonucu verse de Yapay Sinir Ağları

algoritması ve WOE'li Lojistik Regresyon algoritmalarıyla yakın oranlara sahiptir. Diğer hata perspektifi (Tip-I) ile algoritmaların temerrüde düşen olarak gerçekleştirdiği sınıflandırma tahmininde ne oranda başarılı olduğu kesinlik ölçüsünde, Gradyan Artırma algoritması diğer algoritmalara görece daha başarılı bir sonuç vermiştir. Temerrüde düşmeyen müşterilerin Tip-I hata perspektifi ile hesaplanan özgüllük ölçüsünde, Rassal Orman algoritmasının sınıflandırma performansı yüksek olsa da genel sınıflandırma başarısı düşük kalmıştır. Ek olarak, bu ölçüde Destek Vektör Makineleri ve Gradyan Artırma algoritmalarının sınıflandırma yeteneklerinin de iyi sonuçlar verdiği görülmüştür. Karmaşıklık matrisinden elde edilen yukarıdaki ölçüler yardımıyla, sınıflandırma başarısının değerlendirilmesinde sıklıkla kullanılan F1 skor oranı hesaplanmış olup, Gradyan Artırma algoritmasının en iyi oranla sınıflandırma yeteneğine sahip olduğu gözlemlenmiştir.

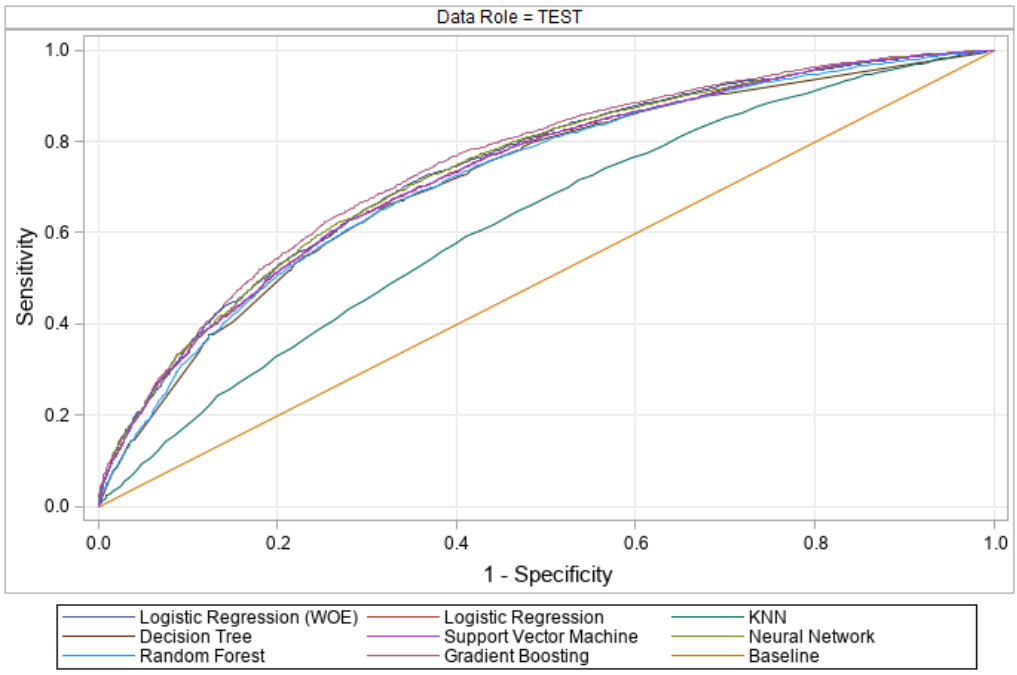
Ek olarak, WOE'li halleriyle kurulan Lojistik Regresyon modelinin, değişkenlerin ham haliyle inşa edilmiş Lojistik Regresyon modeline göre daha başarılı sınıflandırma oranlarına sahip olduğu görülmektedir.

Algoritmaların sınıflandırma performanslarının incelenmesinde kullanılan bir diğer ölçü ROC eğrisi olup, ilgili algoritmalara ait eğitim, doğrulama ve test verisinden elde edilen ROC eğrilerine Grafik 3.8'de yer verilmiştir.



Grafik 3.8: Eğitim, Doğrulama ve Test Verilerinin Algoritmalar için ROC Eğrisi

Algoritmaların, eğitim, doğrulama ve test veri setleriyle temerrüde düşen ve temerrüde düşmeyen müşteriler için gerçekleştirmiş olduğu sınıflandırma yetenekleri, ROC eğri oranlarınca incelenmiş olup, modellerin sınıflandırma uyumunda aşırı öğrenme veya eksik öğrenme davranışı görülmemiştir. Modellerin performans gücü değerlendirmeleri, test verisinden elde edilen ROC eğrisinin altında kalan alanın büyüklüğüne göre belirlenmiştir.



Grafik 3.9: Test Verisinin Algoritmalar için ROC Eğrisi

Bu kapsamda, Hassasiyet ve 1-Özgüllük oranlarının farklı eşik değerlerine göre hesaplanan ROC eğrisi incelendiğinde, K-En Yakın Komşu algoritmasının diğer algoritmalara görece daha az başarılı bir sınıflandırma oranına sahip olduğu görülmektedir. İlgili görseldeki algoritmaların ROC eğri oranlarına Tablo 3.24'de yer verilmiştir.

Tablo 3.24: Algoritmalara Ait ROC Eğrisi Oranları

Model Algoritmaları	ROC Eğrisi
Gradyan Artırma	0.75
Lojistik Regresyon (WOE)	0.74
Yapay Sinir Ağları	0.74
Lojistik Regresyon	0.73

Destek Vektör Makineleri	0.73
Rassal Orman	0.72
Karar Ağacı	0.72
K-En Yakın Komşu	0.62

ROC eğrisine göre, K-En Yakın Komşu haricinde diğer algoritmaların model performans gücü nispeten birbirine yakın değerlere sahip olsa da Gradyan Artırma algoritmasının sınıflandırmadaki başarısının daha yüksek olduğu söylenebilir. Yapay Sinir Ağları ve WOE'li Lojistik Regresyon modeli benzer sınıflandırma performansı gösterirken, WOE'li Lojistik Regresyon analizine ait ROC değerinin, değişkenlerin ham haliyle gerçekleştirilen Lojistik Regresyon analizinin ROC değerine göre de daha başarılı olduğu görülmektedir. Ek olarak, Destek Vektör Makineleri ile geleneksel Lojistik Regresyon modeli birbirlerine yakın oranlarda sonuçlar üretirken, ağaç tabanlı algoritmaların (Rassal Orman, Karar Ağacı) genel sınıflandırma başarısı nispeten daha düşük kalmıştır. Özellikle Rassal Orman algoritmasının temerrüde düşmeyen müşteriler için gerçekleştirmiş olduğu sınıflandırma başarısı dikkat çekse de diğer sınıflandırma ölçülerinde ki yeteneğinin zayıf oranlarda olduğu gözlemlenmiştir.

SONUÇ

Yapılan çalışma kapsamında, Home Credit müşterilerine ait kredi kullanım bilgilerini içeren veriler kullanılarak, veri kalitesi ve model ön işleme çalışmaları gerçekleştirilmiştir.

Yeni başvuru sahiplerinin veya aktif kredilerin temerrüt risk oranını değerlendirmek amacıyla, ilgili istatistiksel ve makine öğrenmesi algoritmaları uygulanmış olup, algoritmaların iyi ve kötü müşterileri ayrıştırıcı gücünün belirlenmesi performans ölçüleriyle sağlanmıştır.

Öznitelik seçimi kapsamında istatistiksel hatalardan arınma ve sınıflandırma gücü yüksek olan değişkenlerin belirlenmesi için çeşitli değişken indirgeme tekniklerinden faydalanılmıştır. İlgili algoritmalarda girdi değişkeni olarak kullanılacak nihai özniteliklerin belirlenmesi için LASSO Regresyonu değişken indirgeme tekniği olarak kullanılmış olup, değişken seçiminde geleneksel Regresyon tekniklerinden (bacward, forward, stepwise) farklı bir yol izlenmiştir. LASSO Regresyonun sahip olduğu ceza parametresi sayesinde, hedef değişken üzerindeki sınıflandırma gücü yüksek olan değişkenler modelde kalmayı başararak nihai girdi olarak belirlenmiştir.

Ek olarak, nihai değişkenlerin IGN düğümü yardımıyla elde edilen WOE'li dönüştürülmüş yapılarıyla, Scorecard düğümü üzerinden Lojistik Regresyon modeli kurulmuştur. Böylelikle değişkenlerin WOE'li yapıları üzerinden gerçekleştirilen Lojistik Regresyon modeli, sınıflandırma performansları değerlendirilecek yedi algoritmanın karşılaştırılmasında sürece dahil edilmiştir.

Ham veri kümesinin %60'ı eğitim, %20'si doğrulama ve %20'si test veri seti olacak şekilde ayrılarak, makine öğrenmesi teknikleri bu veri setleri üzerinden gerçekleştirilmiştir. Ayrıca, veri kalitesi ve değişken indirgeme çalışmalarının yaklaşık %60'ı ham veri kümesi üzerinden gerçekleştirilirken, %40'ı eğitim ve doğrulama veri kümeleri ile gerçekleştirilmiştir. Algoritmalar için optimum performansı sağlayan hiperparametreler, eğitim ve doğrulama veri setleri üzerinden belirlenirken, nihai model geliştirme ve performans karşılaştırmaları test veri seti üzerinden incelenmiştir.

Yukarıdaki bulgular doğrultusunda, karmaşıklık matrisinden elde edilen doğruluk, hassasiyet, özgünlük, kesinlik ve F1 skor ölçüleri, ROC eğrisiyle birlikte

incelenerek, sekiz algoritmanın makine öğrenmesi yetenekleriyle gerçekleştirdiği performansları değerlendirilmiştir. Home Credit müşterilerinin temerrüt riskini değerlendirmek için en iyi sınıflandırma başarısını her bir veri setinde sağlayan modelin, Gradyan Artırma algoritması olduğu gözlemlenmiştir. Gradyan Artırma algoritmasına alternatif olarak en yakın performans başarısı gösteren WOE'li Lojistik Regresyon ve Yapay Sinir Ağlarının kullanılması uygun olsa da modelin yorumlanabilirlik özelliği açısından WOE'li Lojistik Regresyon modeli önerilmektedir.

Gradyan Artırma algoritmasının performans başarısı, nihai değişkenlere ait artıklardaki örüntü bilgisini parametreleştirip, algoritmik olarak kayıp fonksiyonu optimize etmesinden kaynaklanmaktadır. Bu kontekste, Gradyan Artırma algoritmasının çalışma kapsamındaki performansını dikkate alarak, ikili sınıflandırma problemlerini çözmek için oldukça başarılı olan XGBoost, AdaBoost ve CatBoost gibi artırma (Boosting) algoritmaları, gelecek çalışmalarda karşılaştırmaya dahil edilebilir.

KAYNAKLAR

ALTMAN, E.I., MARCO, G., & VARETTO, F., (1994), "Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience)", Journal of Banking & Finance 18, 505-529.

APOSTOLIK, R., DONOHUE, C., WENT, P., and Global Association of Risk Professionals, *Foundations of Banking Risk: An Overview of Banking, Banking Risks, and Risk-Based Banking Regulation*, Hoboken, New Jersey: John Wiley & Sons, Inc., (2009).

AYHAN, S., & ERDOĞMUŞ, Ş., (2014), "Destek Vektör Makineleriyle Sınıflandırma Problemlerinin Çözümü İçin Çekirdek Fonksiyonu Seçimi", Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilimler Dergisi, Nisan, 9(1), s.175-201.

BARBOZA, F., KIMURA, H., & ALTMAN, E., (2017), "Machine learning models and bankruptcy prediction", Expert Systems with Applications 83: 405–417.

BASEL COMMITTEE ON BANKING SUPERVISION, (2001a). The New Basel Capital Accord. Jan. Available at: http://www.bis.org/publ/bcbsca03.pdf. s.34.

BDDK, (2012), "Bankaların İç Denetim ve Risk Yönetimi Sistemleri Hakkında Yönetmelik", https://www.resmigazete.gov.tr/eskiler/2012/06/20120628-17.htm (Erişim Tarihi: 24 Haziran 2020).

BDDK, (2016), Bankaların Kredi Yönetimine İlişkin Rehber, BBDK Kurul Kararı Sayı: 6827, 2016, s.14

BELL, J., *Machine Learning Hands-On for Developers and Technical Professionals*, John Wiley & Sons, Inc., Indianapolis, Indiana, (2014).

BELLOTTI, T., & CROOK, J., (2009), "Support Vector Machines for Credit Scoring and Discovery of Significant Features", Expert Systems with Applications, 3302–3308.

BHARGAVA, A., (Şubat 2000), "Credit Risk Management Systems in Banks", ICICI Bank, s.8., www.garp.com / library/Meets/bhargava.pdf, (27.11.2005).

BREIMAN, L., (2001), "Random Forests, Machine learning", Kluwer Academic Publishers, 45(1), 5-32.

BROWN, I., & MUES, C., (2012), "An experimental comparison of classification algorithms for imbalanced credit scoring data sets", Expert Systems with Applications 39: 3446–3453.

BROWN, I., *Developing Credit Risk Models Using SAS Enterprise MinerTM and SAS/STAT: Theory and Applications*, Cary, NC: SAS Institute Inc, (2014).

CHOLLET, F., *Deep Learning with Python*, Manning Publications Co., Shelter Island, NY, USA, (2018).

COYLE, B., *Introduction to Currency Risk*, Financial World Publishing, UK, (2000).

DANGETI, P., *Statistics for Machine Learning, Build supervised, unsupervised, and reinforcement learning models using both Python and R*, Packt Publishing, Birmingham UK, (2017).

DEMİRBULUT, Y., AKTAŞ, M., KALIPSIZ, O., & BAYRACI, S. (2017). "İstatistiksel ve Makine Öğrenimi Yöntemleriyle Kredi Skorlama", CEUR-WS (s. 273-284). Antalya: Turkish National Software Engineering Symposium.

DESIGN I. T., GABRYS B., PETRAKIEVA L., (2004), "Combining labelled and unlabelled data", International Journal on Approximate Reasoning, vol. 35, p. 251-273.

DİNOV, IVO D., *Data Science and Predictive Analytics: Biomedical and Health Applications Using R*, Springer, Cham, Switzerland, (2018).

EĞRİOGLU, E., ALADAĞ, C.H., YOLCU, U., USLU, V.R., & BAŞARAN, M.A., (2009), "A new approach based on artificial neural networks for high order multivariate fuzzy time series", Expert Systems with Applications, 36(7), 10589-10594.

GESTEL, V., TONY, I., BAESENS, B., GARCIA, I.J., & DIJCKE. P.V., (2003), "A support vector machine approach to credit scoring", 73–82. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.93.6492&rep=rep1&type=pf (Erişim Tarihi: 7 Haziran 2018).

HAMORİ, S., KAWAİ, M., KUME, T., MURAKAMİ, Y., & WATANABE, C., (2018), "Ensemble Learning or Deep Learning? Application to Default Risk Analysis", Journal of Risk and Financial Management 11: 12.

HAND, D., & ZHOU, F., (2009), "Evaluating models for classifying customers in retail banking collections", Journal of the Operational Research Society, 61, 1540–1547.

JORION, P., Financial Risk Manager Handbook, Wiley Finance Series, 5. Baskı, s.431 (2009).

KAVCIOĞLU, Ş. (2019). "Kurumsal kredi skorlamasında klasik yöntemlerle yapay sinir ağı karşılaştırması", İstanbul İktisat Dergisi - Istanbul Journal of Economics, 69(2), 207-245.

KAVZOĞLU, T., & ÇÖLKESEN, İ., (2010), "Destek Vektör Makineleri ile Uydu Görüntülerinin Sınıflandırılmasında Kernel Fonksiyonlarının Etkilerinin İncelenmesi", Harita Dergisi Temmuz 2010 Sayı 144, s.73-82.

LESSMANN, S., BAESENS, B., SEOW, H.V., & THOMAS, L.C., (2015), "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research", European Journal of Operational Research, Vol.247, 124–136.

LINDHOLM, A., WAHLSTRÖM, N., LINDSTEN, F., SCHÖN, T. B., (2019), "Supervised Machine Learning", Version (12 May 2019), s.7 http://www.it.uu.se/edu/course/homepage/sml/literature/lecture_notes.pdf (Erişim Tarihi: 13 Mayıs 2019).

LUI, H., (2017), "Which Machine Learning Algorithm Should I Use?", https://blogs.sas.com/content/subconsciousmusings/2017/04/12/machine-learning-algorithm-use/ (Erişim Tarihi: 12 Mayıs 2020).

MAEIREIZO B., LITMAN D., HWA R., (2004), "Co-training for predicting emotions with spoken dialogue data", Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions, Companion Volume to the proceeding of 42nd Annual Meeting of the Association for Computational Linguistics (ACL), July, Barcelona, Spain.

MANDACI P.E., (2003), "Türk Bankacılık Sektörünün Taşıdığı Riskler ve Finansal Krizi Asmada Kullanılan Risk Ölçüm Teknikleri", Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, s.71.

MARKOFF J., (2015), "A Learning Advance in Artificial İntelligence Rivals Human Abilities", New York Times, https://www.nytimes.com/2015/12/11/science/an-advance-in-artificial-intelligence-rivals-human-vision-abilities.html (Erişim Tarihi: 22 Mart 2019).

MOHRI, M., ROSTAMIZADEH, A., TALWALKAR, A., *Foundations of Machine Learning*, Second Edition, The MIT Press, London, (2012).

MUELLER, J.P., MASSARON, L., *Machine Learning For Dummies*, John Wiley & Sons, Inc., Hoboken, New Jersey, (2016).

NICULESCU-MIZIL, A., & CARUANA, R., (2005), "Predicting Good Probabilities With Supervised Learning", Proceedings of the 22nd international conference on Machine learning, 07–11 August 2005, Bonn, 625-632.

NIGAM K., MCCALLUM A. K., THRUN S., MITCHELL T., (2000), "Text classification from labeled and unlabeled documents using EM", International Journal of Machine Learning, vol. 39 no. 2-3, s. 103-134.

OSUNA, E.E., FREUND, R., GİROSİ, F., (1997), "Support Vector Machines: Training and Applications", A.I. Memo No. 1602, C.B.C.L. Paper No. 144, Massachusetts Institute of Technology and Artificial Intelligence Laboratory, Massachusetts.

OUYANG, Y., HU, M., HUET, A., LI, Z., *Mining Over Air: Wireless Communication Networks Analytics*, Springer, Cham, Switzerland, (2015).

ÖZTEMEL, E., *Yapay Sinir Ağları*, Papatya Yayıncılık, İstanbul, (2006).

ÖZTÜRK, K., ŞAHİN, M.E., (2018), "Yapay Sinir Ağları ve Yapay Zekâ'ya Genel Bir Bakış", Takvim-i Vekayi, Cilt: 6 No: 2 Sayfa: 25-36.

RASCHKA, S., *Python Machine Learning*, Packt Publishing Ltd., Birmingham, UK., (2015).

SCHUERMANN, T., (2004), "What do we know about loss given default?", Wharton Financial Institutions Center, Vol.Feb. s.3

SHWARTZ, S. S., BEN-DAVID, S., *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, UK, (2014).

SINKEY J.F.Jr., *Commercial Bank Financial Management*, Prentice Hall, U.S.A., (1998).

SNİATALA, P., HADİ AMİNİ M., BOROOJENİ K.G, *Fundamentals of Brooks-lyengar Distributed Sensing Algorithm*, Springer, Switzerland, (2020).

THE ROYAL SOCIETY, (2017) "Machine Learning: The Power And Promise Of Computers That Learn By Example", s.16-21, www.royalsociety.org/machine-learning (Erişim Tarihi: 5 Ocak 2019).

THEOBALD, O., *Machine Learning for Absolute Beginners*, Second Edition, (2017).

VASİLOUDİS, T., http://tvas.me/articles/2019/08/26/Block-Distributed-Gradient-Boosted-Trees.html (Erişim Tarihi: 2019/10/05)

WANG, Y., WANG, S., ve LAI, K.K., (2005), "A New Fuzzy Support Vector Machine to Evaluate Credit Risk", IEEE Transactions on Fuzzy Systems, Vol.13: 820-831.

YAO, X., CROOK, J., & ANDREEVA, G., (2017), "Enhancing two-stage modelling methodology for loss given default with support vector machines", European Journal of Operational Research 263: 679-689.

YAROWSKY D., (1995), "Unsupervised word sense disambiguation rivaling supervised methods", Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics.

YEŞİLYURT, A., ŞEKER, Ş. E., (2018), "Skorlama Algoritmaları", YBS Ansiklopedi, Cilt 5, Sayı 1, Mayıs 2018 7-13.

YEH, I. C., & LIEN, C., (2009), "The Comparisons of Data Mining Techniques for the Predictive Accuracy of Probability of Default of Credit Card Clients", Expert Systems with Applications, Vol.36, 2473-2480.

ZHANG, C., ZHANG, S., *Association Rule Mining - Models and Algorithms*, Springer, Berlin, (2002).

ZHANG, W., (2017), "Machine Learning Approaches to Predicting Company Bankruptcy", Journal of Financial Risk Management 6: 364-374.

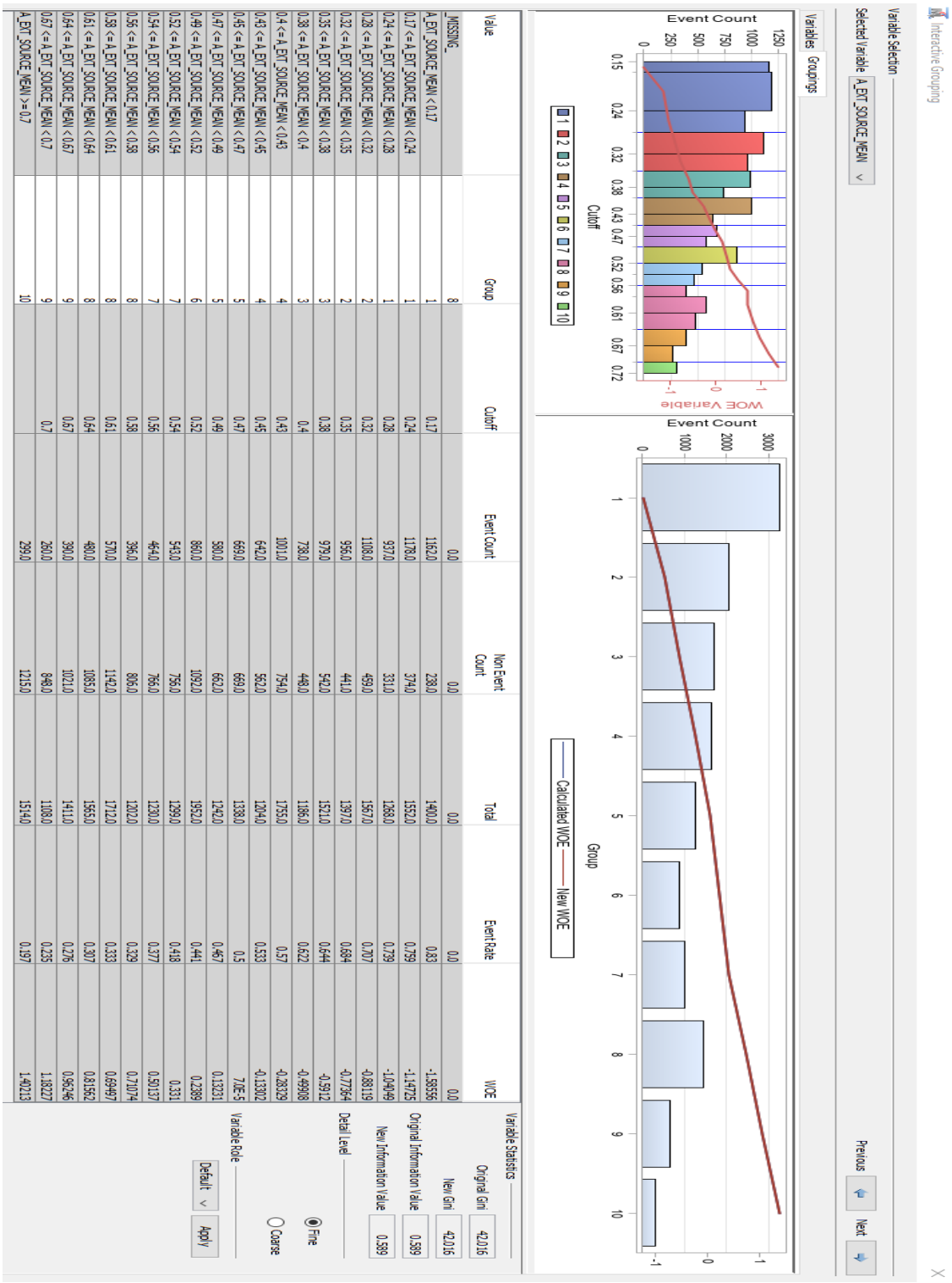
ZHOU, L., & WANG, H., (2012), "Loan Default Prediction on Large Imbalanced Data Using Random Forests", TELKOMNIKA Indonesian Journal of Electrical Engineering, Vol.10, No.6, October 2012, 1519-1525

ZHU, X., GOLDBERG, A. B., *Introduction to Semi-Supervised Learning*, Morgan & Claypool Publishers, (2009).

EKLER

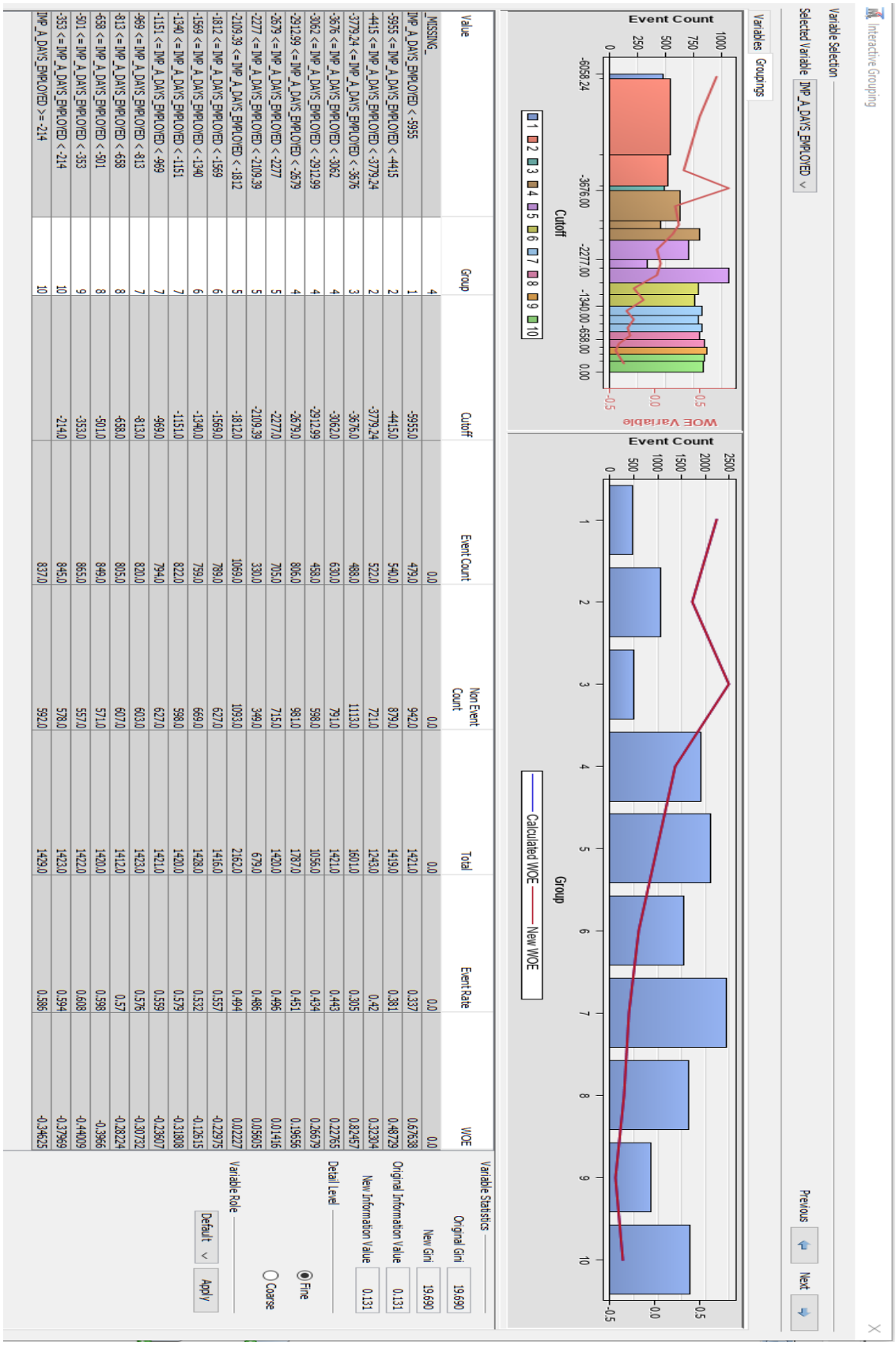
Ek I. IGN Düğümü Yardımıyla Nihai Değişkenlerin WOE Gruplandırılması

A_EXT_SOURCE_MEAN Değişkeni için WOE Dönüşümü



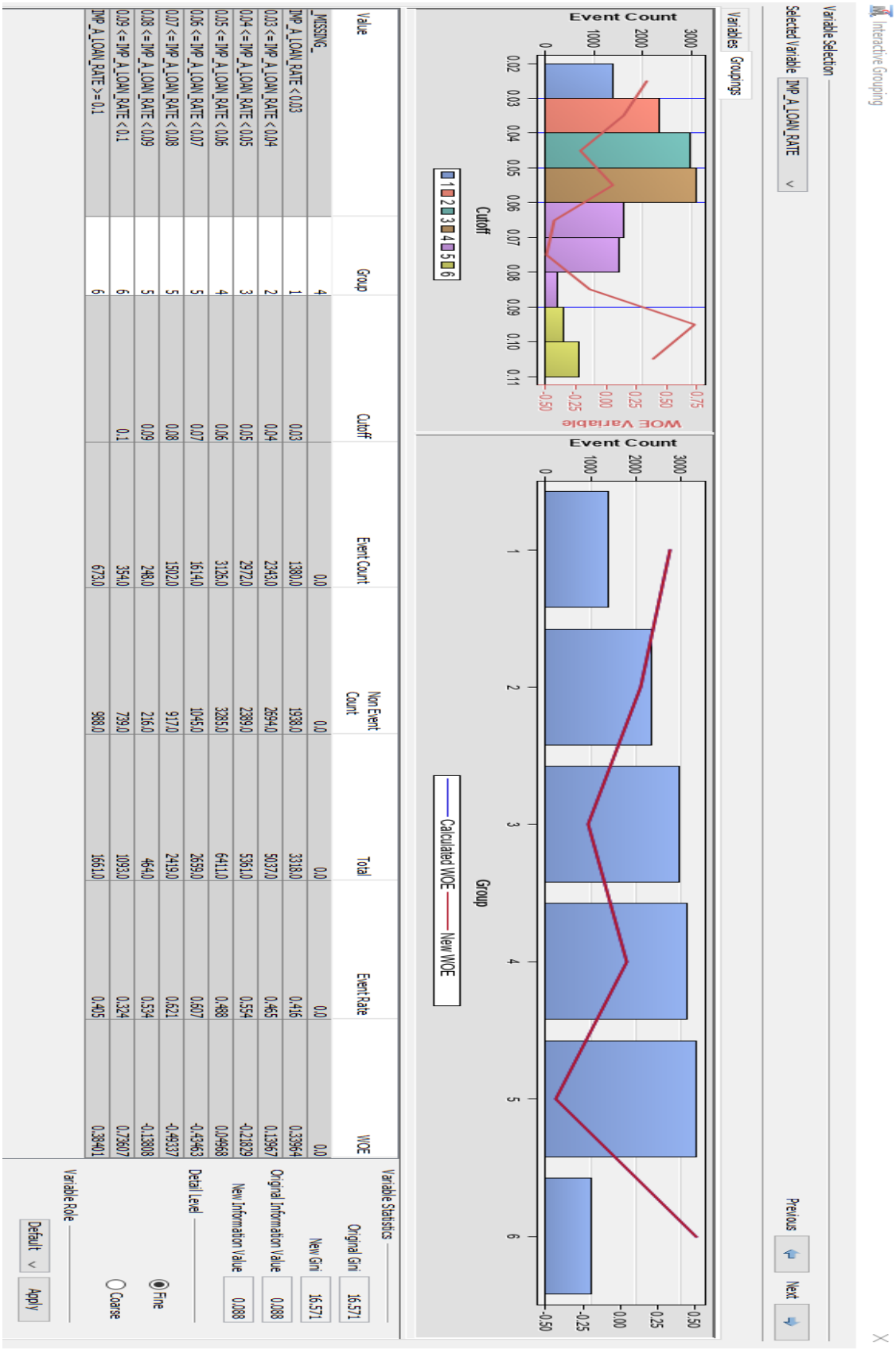
Value	Group	Cutoff	Event Count	Non Event Count	Total	Event Rate	WOE
MISSING	8		0.0	0.0	0.0	0.0	0.0
A_EXT_SOURCE_MEAN < 0.17	1	0.17	1162.0	238.0	1400.0	0.83	-1.88356
0.17 <= A_EXT_SOURCE_MEAN < 0.24	1	0.24	1178.0	374.0	1552.0	0.759	-1.14725
0.24 <= A_EXT_SOURCE_MEAN < 0.28	1	0.28	937.0	331.0	1268.0	0.739	-1.04049
0.28 <= A_EXT_SOURCE_MEAN < 0.32	2	0.32	1108.0	459.0	1567.0	0.707	-0.88119
0.32 <= A_EXT_SOURCE_MEAN < 0.35	2	0.35	956.0	441.0	1397.0	0.684	-0.77364
0.35 <= A_EXT_SOURCE_MEAN < 0.38	3	0.38	979.0	542.0	1521.0	0.644	-0.5912
0.38 <= A_EXT_SOURCE_MEAN < 0.4	3	0.4	738.0	448.0	1186.0	0.622	-0.49908
0.4 <= A_EXT_SOURCE_MEAN < 0.43	4	0.43	1001.0	754.0	1755.0	0.57	-0.28339
0.43 <= A_EXT_SOURCE_MEAN < 0.45	4	0.45	642.0	562.0	1204.0	0.533	-0.13302
0.45 <= A_EXT_SOURCE_MEAN < 0.47	5	0.47	669.0	669.0	1338.0	0.5	7.0E-5
0.47 <= A_EXT_SOURCE_MEAN < 0.49	5	0.49	580.0	662.0	1242.0	0.467	0.13231
0.49 <= A_EXT_SOURCE_MEAN < 0.52	6	0.52	860.0	1092.0	1952.0	0.441	0.2389
0.52 <= A_EXT_SOURCE_MEAN < 0.54	7	0.54	543.0	756.0	1299.0	0.418	0.331
0.54 <= A_EXT_SOURCE_MEAN < 0.56	7	0.56	464.0	766.0	1230.0	0.377	0.50137
0.56 <= A_EXT_SOURCE_MEAN < 0.58	8	0.58	396.0	806.0	1202.0	0.329	0.71074
0.58 <= A_EXT_SOURCE_MEAN < 0.61	8	0.61	570.0	1142.0	1712.0	0.333	0.69497
0.61 <= A_EXT_SOURCE_MEAN < 0.64	8	0.64	480.0	1085.0	1565.0	0.307	0.81562
0.64 <= A_EXT_SOURCE_MEAN < 0.67	9	0.67	390.0	1021.0	1411.0	0.276	0.96246
0.67 <= A_EXT_SOURCE_MEAN < 0.7	9	0.7	260.0	848.0	1108.0	0.235	1.18227
A_EXT_SOURCE_MEAN >= 0.7	10		299.0	1215.0	1514.0	0.197	1.40213

A_EXT_SOURCE_MEAN Değişkeni için WOE Dönüşümü



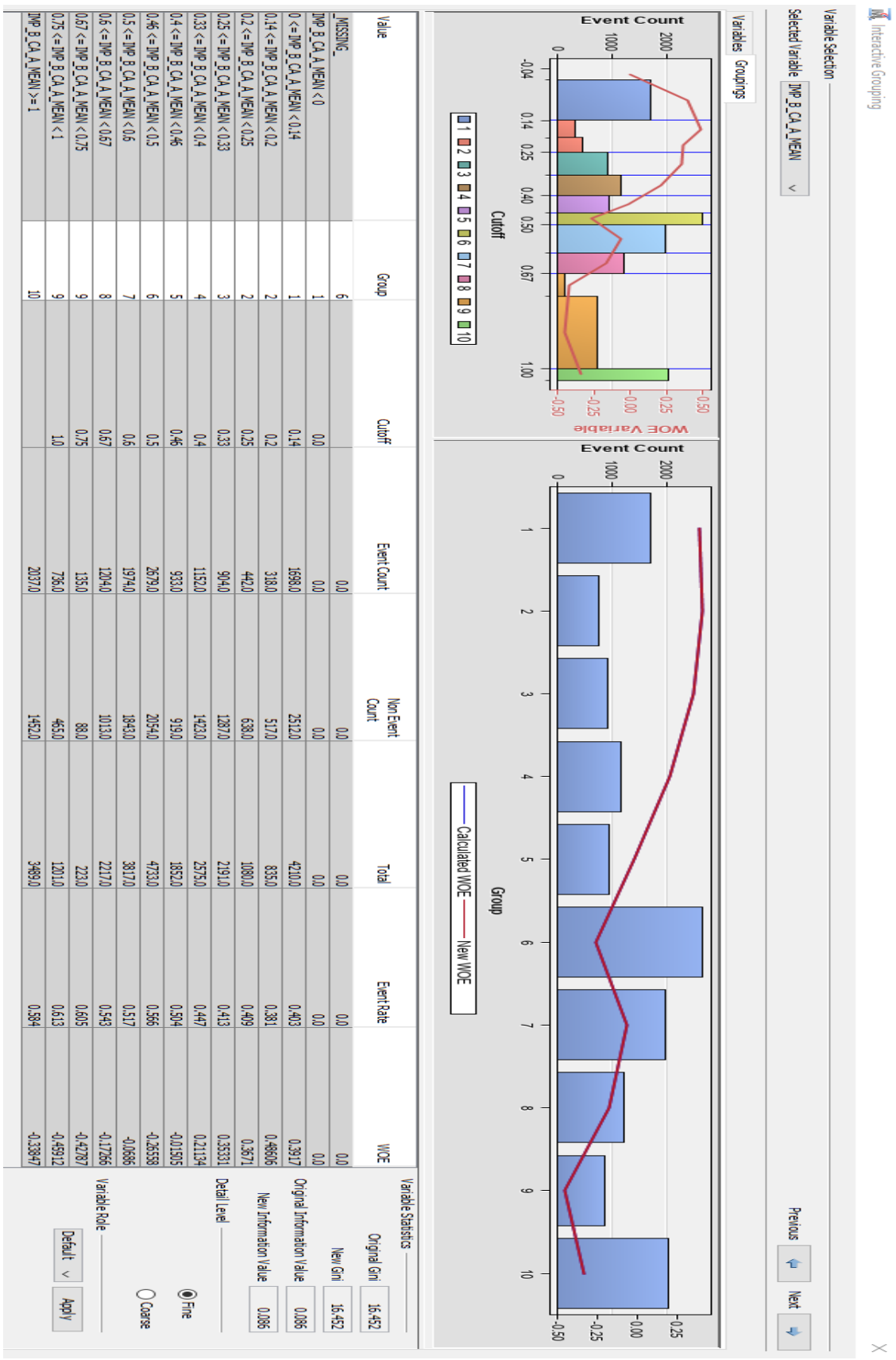
Value	Group	Cutoff	Event Count	Non Event Count	Total	Event Rate	WOE
MISSING	4		0.0	0.0	0.0	0.0	0.0
DMP_A_DAYS_EMPLOYED < -5955	1	-5955.0	479.0	942.0	1421.0	0.337	0.67638
-5955 <= DMP_A_DAYS_EMPLOYED < -4415	2	-4415.0	540.0	879.0	1419.0	0.381	0.48729
-4415 <= DMP_A_DAYS_EMPLOYED < -3779.24	2	-3779.24	522.0	721.0	1243.0	0.42	0.32304
-3779.24 <= DMP_A_DAYS_EMPLOYED < -3676	3	-3676.0	488.0	1113.0	1601.0	0.305	0.82457
-3676 <= DMP_A_DAYS_EMPLOYED < -3062	4	-3062.0	630.0	791.0	1421.0	0.443	0.22765
-3062 <= DMP_A_DAYS_EMPLOYED < -2912.99	4	-2912.99	458.0	598.0	1056.0	0.434	0.26679
-2912.99 <= DMP_A_DAYS_EMPLOYED < -2679	4	-2679.0	806.0	981.0	1787.0	0.451	0.19656
-2679 <= DMP_A_DAYS_EMPLOYED < -2277	5	-2277.0	705.0	715.0	1420.0	0.496	0.01416
-2277 <= DMP_A_DAYS_EMPLOYED < -2109.39	5	-2109.39	330.0	349.0	679.0	0.486	0.05605
-2109.39 <= DMP_A_DAYS_EMPLOYED < -1812	5	-1812.0	1069.0	1093.0	2162.0	0.494	0.02227
-1812 <= DMP_A_DAYS_EMPLOYED < -1569	6	-1569.0	789.0	627.0	1416.0	0.557	-0.22975
-1569 <= DMP_A_DAYS_EMPLOYED < -1340	6	-1340.0	759.0	669.0	1428.0	0.532	-0.12615
-1340 <= DMP_A_DAYS_EMPLOYED < -1151	7	-1151.0	822.0	598.0	1420.0	0.579	-0.31808
-1151 <= DMP_A_DAYS_EMPLOYED < -969	7	-969.0	794.0	627.0	1421.0	0.559	-0.23607
-969 <= DMP_A_DAYS_EMPLOYED < -813	7	-813.0	820.0	603.0	1423.0	0.576	-0.30732
-813 <= DMP_A_DAYS_EMPLOYED < -658	8	-658.0	805.0	607.0	1412.0	0.57	-0.28224
-658 <= DMP_A_DAYS_EMPLOYED < -501	8	-501.0	849.0	571.0	1420.0	0.598	-0.3966
-501 <= DMP_A_DAYS_EMPLOYED < -353	9	-353.0	865.0	557.0	1422.0	0.608	-0.44009
-353 <= DMP_A_DAYS_EMPLOYED < -214	10	-214.0	845.0	578.0	1423.0	0.594	-0.37969
DMP_A_DAYS_EMPLOYED >= -214	10		837.0	592.0	1429.0	0.586	-0.34625

A_LOAN_RATE Değişkeni için WOE Dönüşümü



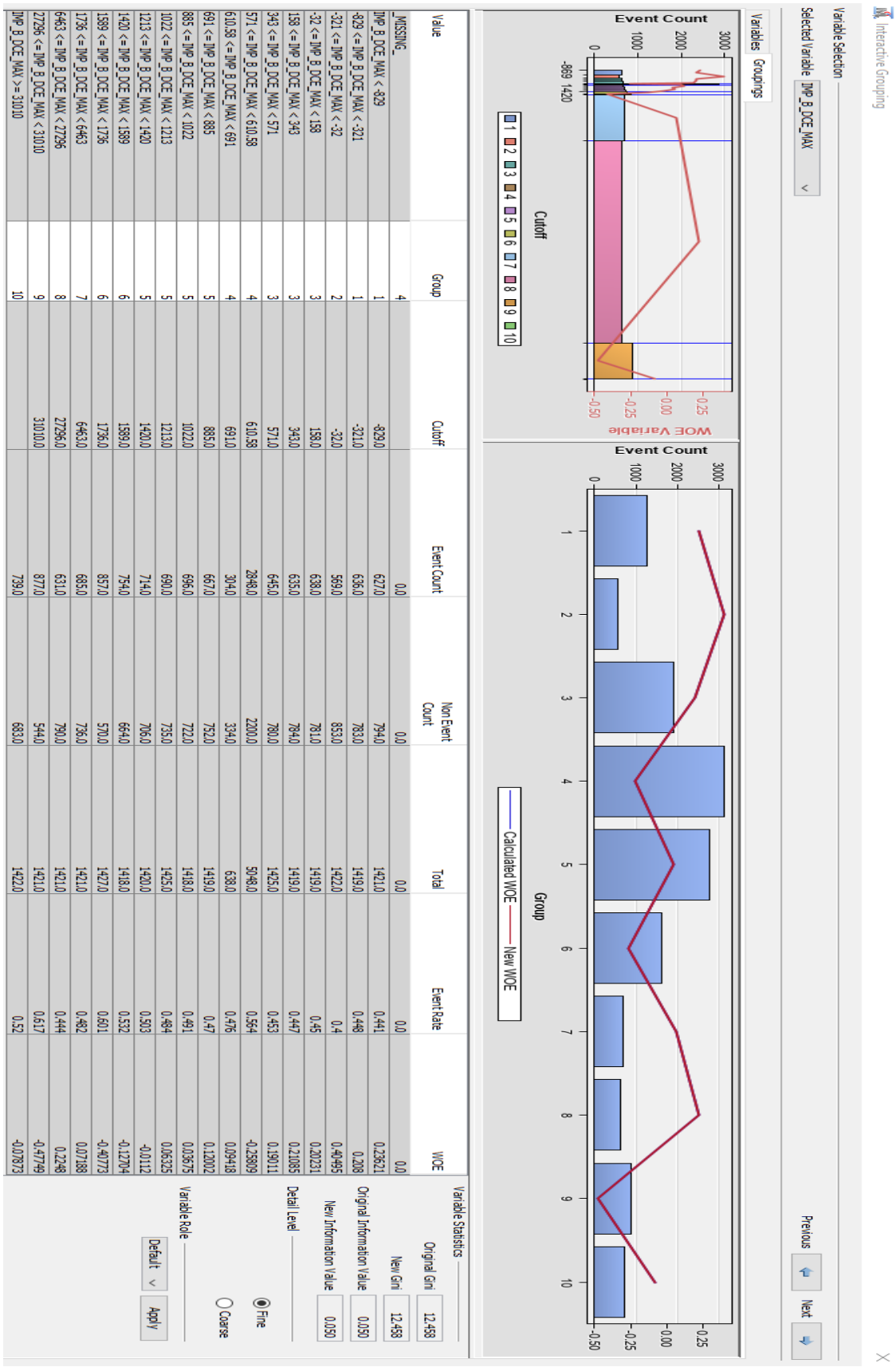
Value	Group	Cutoff	Event Count	Non Event Count	Total	Event Rate	WOE
MISSING	4		0.0	0.0	0.0	0.0	0.0
IMP_A_LOAN_RATE < 0.03	1	0.03	1380.0	1938.0	3318.0	0.416	0.33964
0.03 <= IMP_A_LOAN_RATE < 0.04	2	0.04	2343.0	2694.0	5037.0	0.465	0.13967
0.04 <= IMP_A_LOAN_RATE < 0.05	3	0.05	2972.0	2389.0	5361.0	0.554	-0.21829
0.05 <= IMP_A_LOAN_RATE < 0.06	4	0.06	3126.0	3285.0	6411.0	0.488	0.04968
0.06 <= IMP_A_LOAN_RATE < 0.07	5	0.07	1614.0	1045.0	2659.0	0.607	-0.49463
0.07 <= IMP_A_LOAN_RATE < 0.08	5	0.08	1502.0	917.0	2419.0	0.621	-0.49337
0.08 <= IMP_A_LOAN_RATE < 0.09	5	0.09	248.0	216.0	464.0	0.534	-0.13088
0.09 <= IMP_A_LOAN_RATE < 0.1	6	0.1	354.0	739.0	1093.0	0.324	0.73607
IMP_A_LOAN_RATE >= 0.1	6		673.0	988.0	1661.0	0.405	0.38401

B_CA_A_MEAN Değişkeni için WOE Dönüşümü



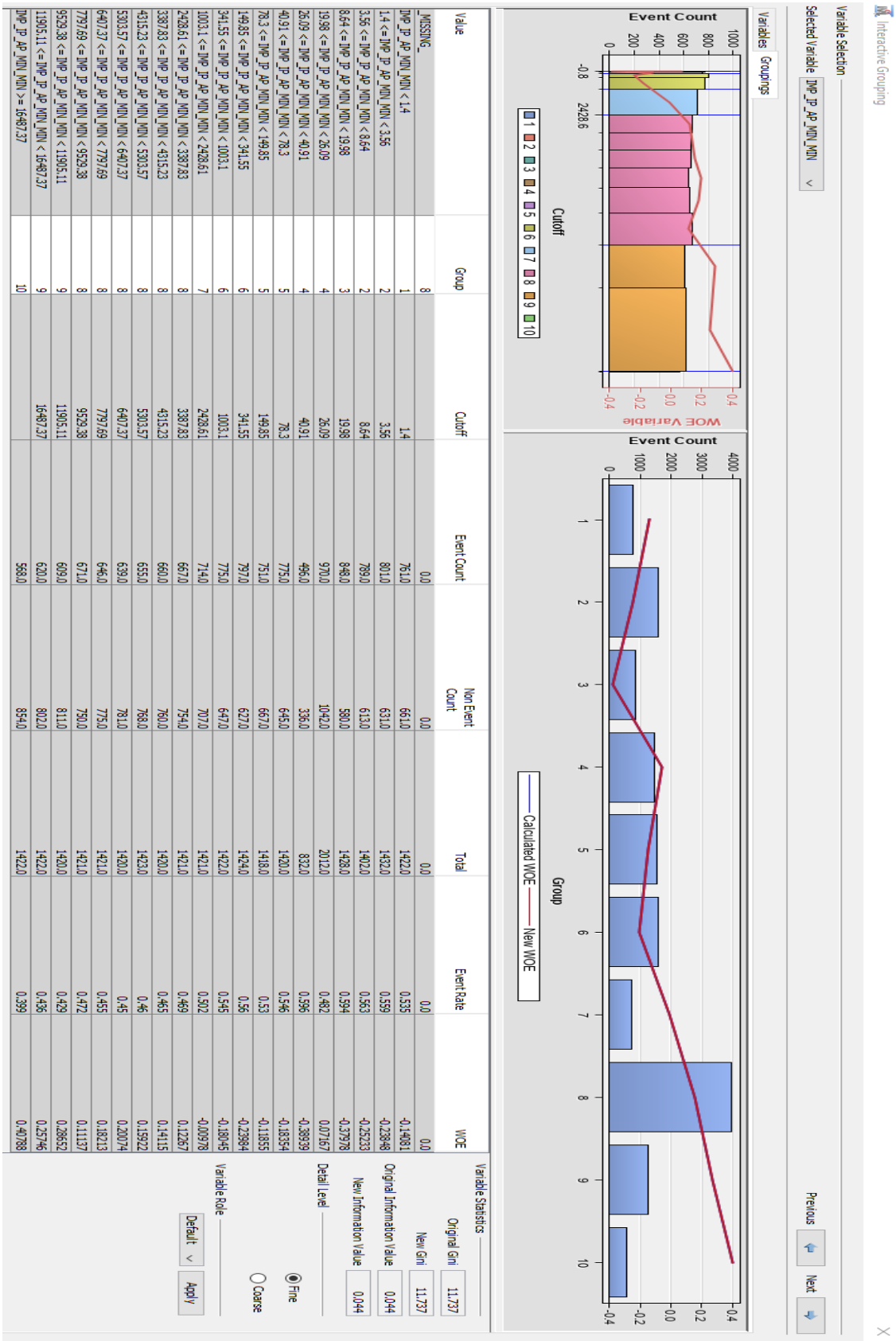
Value	Group	Cutoff	Event Count	Non Event Count	Total	Event Rate	WOE
MISSING	6		0.0	0.0	0.0	0.0	0.0
INP_B_CA_A_MEAN < 0	1	0.0	0.0	0.0	0.0	0.0	0.0
0 <= INP_B_CA_A_MEAN < 0.14	1	0.14	1698.0	2512.0	4210.0	0.403	0.3917
0.14 <= INP_B_CA_A_MEAN < 0.2	2	0.2	318.0	517.0	835.0	0.381	0.48606
0.2 <= INP_B_CA_A_MEAN < 0.25	2	0.25	442.0	638.0	1080.0	0.409	0.3671
0.25 <= INP_B_CA_A_MEAN < 0.33	3	0.33	904.0	1287.0	2191.0	0.413	0.335331
0.33 <= INP_B_CA_A_MEAN < 0.4	4	0.4	1152.0	1423.0	2575.0	0.447	0.211134
0.4 <= INP_B_CA_A_MEAN < 0.46	5	0.46	933.0	919.0	1852.0	0.504	-0.01505
0.46 <= INP_B_CA_A_MEAN < 0.5	6	0.5	2679.0	2054.0	4733.0	0.566	-0.26558
0.5 <= INP_B_CA_A_MEAN < 0.6	7	0.6	1974.0	1843.0	3817.0	0.517	-0.0686
0.6 <= INP_B_CA_A_MEAN < 0.67	8	0.67	1204.0	1013.0	2217.0	0.543	-0.17266
0.67 <= INP_B_CA_A_MEAN < 0.75	9	0.75	135.0	88.0	223.0	0.605	-0.47787
0.75 <= INP_B_CA_A_MEAN < 1	9	1.0	736.0	465.0	1201.0	0.613	-0.45912
INP_B_CA_A_MEAN >= 1	10		2037.0	1452.0	3489.0	0.584	-0.33847

B_DCE_MAX Değişkeni için WOE Dönüşümü



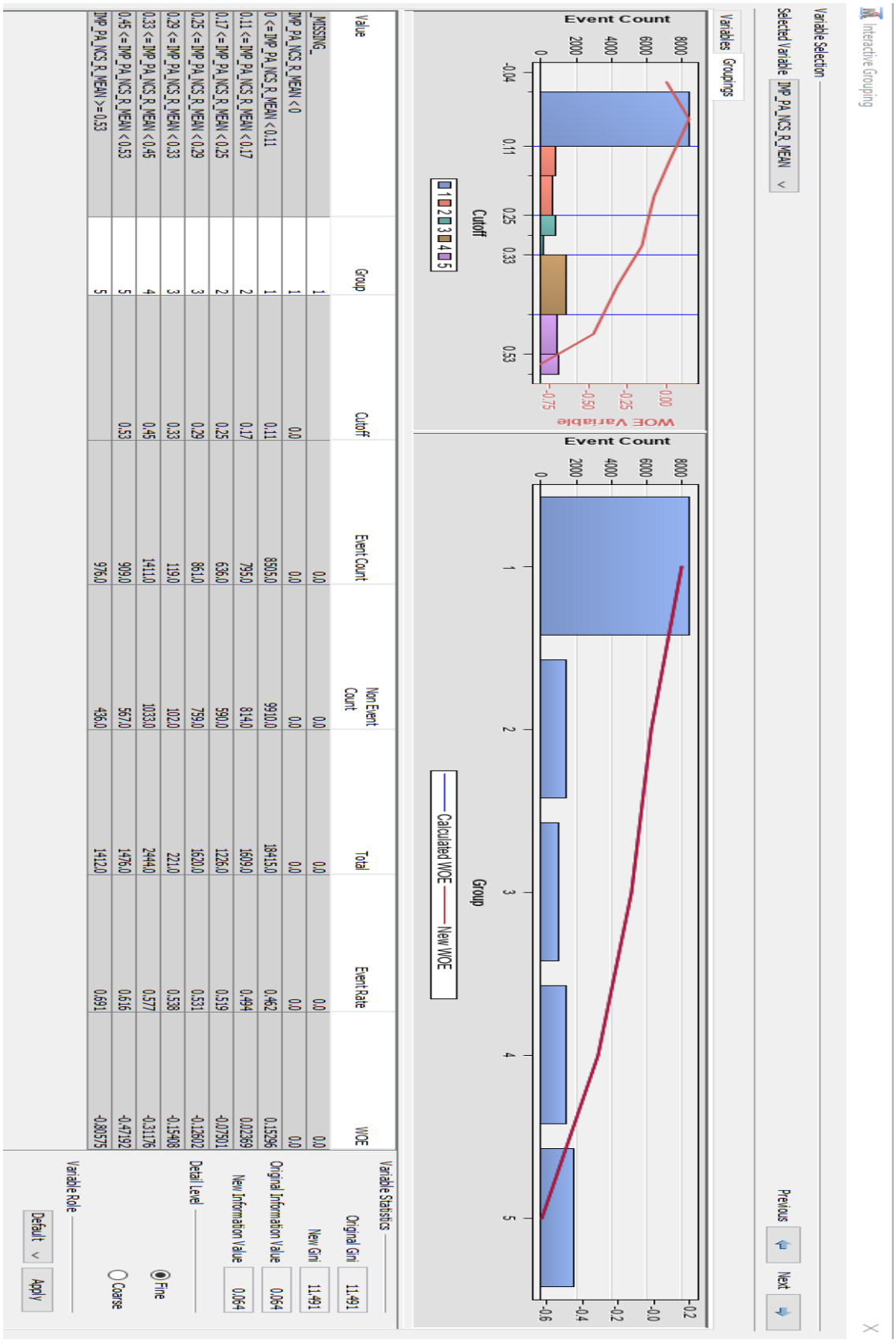
Value	Group	Cutoff	Event Count	Non Event Count	Total	Event Rate	WOE
MISSING	4		0.0	0.0	0.0	0.0	0.0
IMP_B_DCE_MAX < -829	1	-829.0	627.0	794.0	1421.0	0.441	0.23621
-829 <= IMP_B_DCE_MAX < -321	1	-321.0	636.0	783.0	1419.0	0.448	0.208
-321 <= IMP_B_DCE_MAX < -32	2	-32.0	569.0	853.0	1422.0	0.4	0.40495
-32 <= IMP_B_DCE_MAX < 158	3	158.0	638.0	781.0	1419.0	0.45	0.20231
158 <= IMP_B_DCE_MAX < 343	3	343.0	635.0	784.0	1419.0	0.447	0.21085
343 <= IMP_B_DCE_MAX < 571	3	571.0	645.0	780.0	1425.0	0.453	0.19011
571 <= IMP_B_DCE_MAX < 610.58	4	610.58	2848.0	2200.0	5048.0	0.564	-0.25809
610.58 <= IMP_B_DCE_MAX < 691	4	691.0	304.0	334.0	638.0	0.476	0.09418
691 <= IMP_B_DCE_MAX < 885	5	885.0	667.0	752.0	1419.0	0.47	0.12002
885 <= IMP_B_DCE_MAX < 1022	5	1022.0	696.0	722.0	1418.0	0.491	0.03675
1022 <= IMP_B_DCE_MAX < 1213	5	1213.0	690.0	735.0	1425.0	0.484	0.06325
1213 <= IMP_B_DCE_MAX < 1420	5	1420.0	714.0	706.0	1420.0	0.503	-0.0112
1420 <= IMP_B_DCE_MAX < 1589	6	1589.0	754.0	664.0	1418.0	0.532	-0.12704
1589 <= IMP_B_DCE_MAX < 1736	6	1736.0	857.0	570.0	1427.0	0.601	-0.40773
1736 <= IMP_B_DCE_MAX < 6463	7	6463.0	685.0	736.0	1421.0	0.482	0.07188
6463 <= IMP_B_DCE_MAX < 27296	8	27296.0	631.0	790.0	1421.0	0.444	0.2248
27296 <= IMP_B_DCE_MAX < 310.10	9	310.10.0	877.0	544.0	1421.0	0.617	-0.47749
IMP_B_DCE_MAX >= 310.10	10		739.0	683.0	1422.0	0.52	-0.07873

IP_AP_MIN_MIN Değişkeni için WOE Dönüşümü



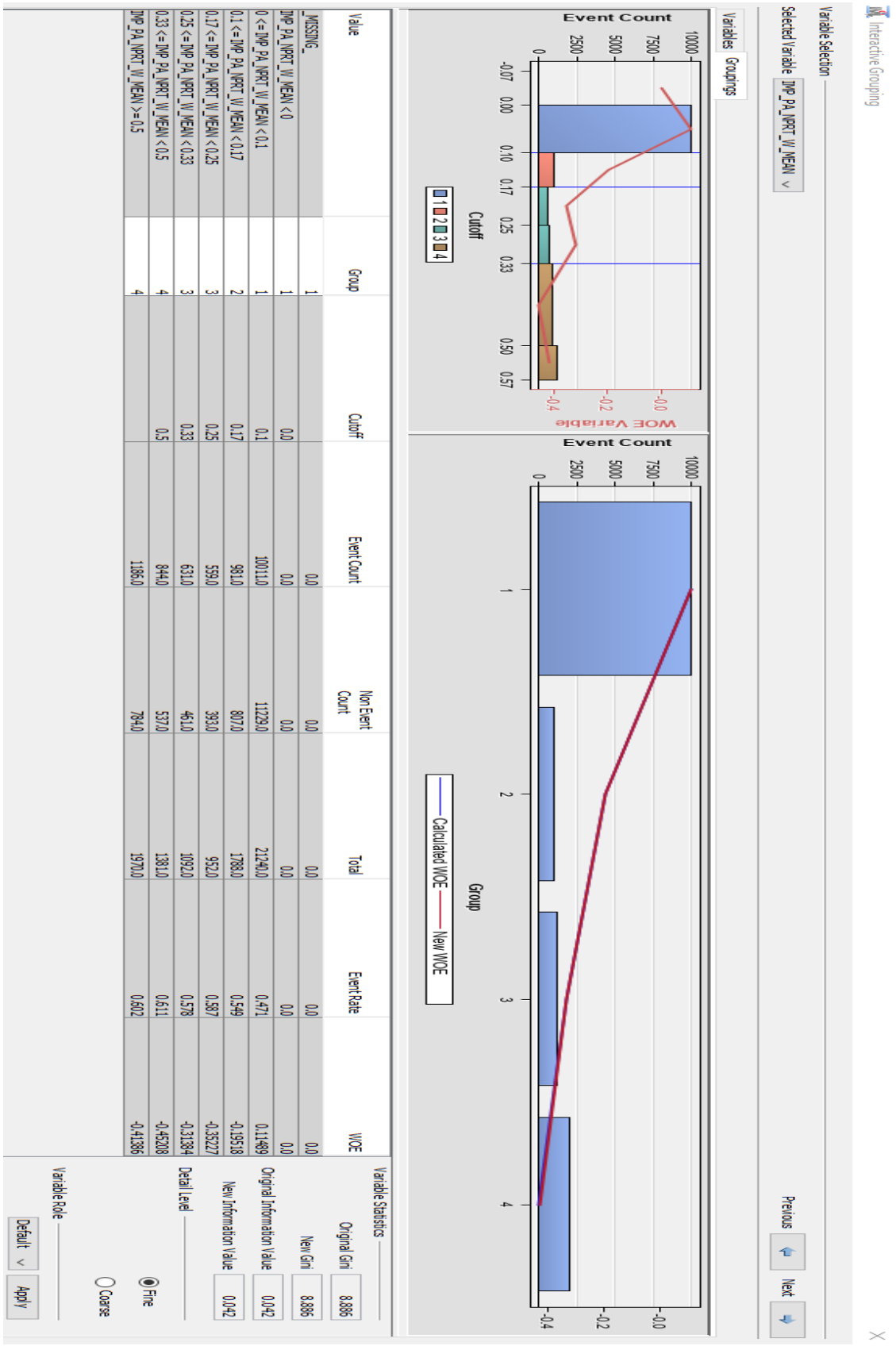
Value	Group	Cutoff	Event Count	Non Event Count	Total	Event Rate	WOE
MISSING	8		0.0	0.0	0.0	0.0	0.0
IMP_IP_AP_MIN_MIN < 1.4	1	1.4	761.0	661.0	1422.0	0.535	-0.14061
1.4 <= IMP_IP_AP_MIN_MIN < 3.56	2	3.56	801.0	631.0	1432.0	0.559	-0.23848
3.56 <= IMP_IP_AP_MIN_MIN < 8.64	2	8.64	789.0	613.0	1402.0	0.563	-0.25233
8.64 <= IMP_IP_AP_MIN_MIN < 19.98	3	19.98	848.0	580.0	1428.0	0.594	-0.37978
19.98 <= IMP_IP_AP_MIN_MIN < 26.09	4	26.09	970.0	1042.0	2012.0	0.482	0.07167
26.09 <= IMP_IP_AP_MIN_MIN < 40.91	4	40.91	496.0	336.0	832.0	0.596	-0.38939
40.91 <= IMP_IP_AP_MIN_MIN < 78.3	5	78.3	775.0	645.0	1420.0	0.546	-0.18354
78.3 <= IMP_IP_AP_MIN_MIN < 149.85	5	149.85	751.0	667.0	1418.0	0.53	-0.11855
149.85 <= IMP_IP_AP_MIN_MIN < 341.55	6	341.55	797.0	627.0	1424.0	0.56	-0.23984
341.55 <= IMP_IP_AP_MIN_MIN < 1003.1	6	1003.1	775.0	647.0	1422.0	0.545	-0.18045
1003.1 <= IMP_IP_AP_MIN_MIN < 2428.61	7	2428.61	714.0	707.0	1421.0	0.502	-0.00978
2428.61 <= IMP_IP_AP_MIN_MIN < 3387.83	8	3387.83	667.0	754.0	1421.0	0.469	0.12267
3387.83 <= IMP_IP_AP_MIN_MIN < 4315.23	8	4315.23	660.0	760.0	1420.0	0.465	0.14115
4315.23 <= IMP_IP_AP_MIN_MIN < 5303.57	8	5303.57	655.0	768.0	1423.0	0.46	0.15922
5303.57 <= IMP_IP_AP_MIN_MIN < 6407.37	8	6407.37	639.0	781.0	1420.0	0.45	0.20074
6407.37 <= IMP_IP_AP_MIN_MIN < 7797.69	8	7797.69	646.0	775.0	1421.0	0.455	0.18213
7797.69 <= IMP_IP_AP_MIN_MIN < 9529.38	8	9529.38	671.0	750.0	1421.0	0.472	0.11137
9529.38 <= IMP_IP_AP_MIN_MIN < 11905.11	9	11905.11	609.0	811.0	1420.0	0.429	0.28652
11905.11 <= IMP_IP_AP_MIN_MIN < 16487.37	9	16487.37	620.0	802.0	1422.0	0.436	0.25746
IMP_IP_AP_MIN_MIN >= 16487.37	10		568.0	854.0	1422.0	0.399	0.40788

PA_NCS_R_MEAN Değişkeni için WOE Dönüşümü



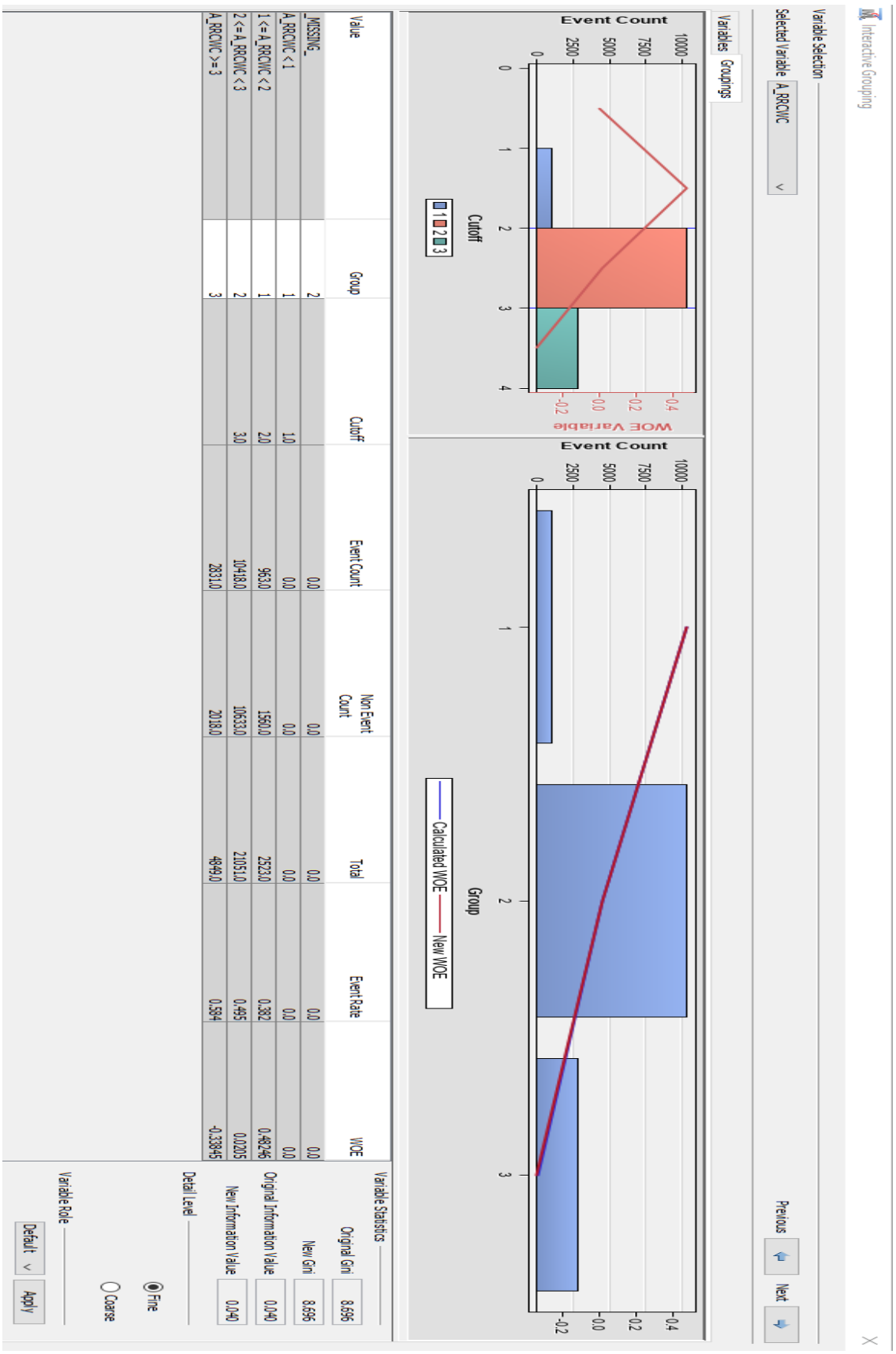
Value	Group	Cutoff	Event Count	Non Event Count	Total	Event Rate	WOE
MISSING	1		0.0	0.0	0.0	0.0	0.0
IMP_PA_NCS_R_MEAN < 0	1	0.0	0.0	0.0	0.0	0.0	0.0
0 <= IMP_PA_NCS_R_MEAN < 0.11	1	0.11	8505.0	9910.0	18415.0	0.462	0.15296
0.11 <= IMP_PA_NCS_R_MEAN < 0.17	2	0.17	795.0	814.0	1609.0	0.494	0.02369
0.17 <= IMP_PA_NCS_R_MEAN < 0.25	2	0.25	636.0	590.0	1226.0	0.519	-0.07501
0.25 <= IMP_PA_NCS_R_MEAN < 0.29	3	0.29	861.0	759.0	1620.0	0.531	-0.12602
0.29 <= IMP_PA_NCS_R_MEAN < 0.33	3	0.33	119.0	102.0	221.0	0.538	-0.15408
0.33 <= IMP_PA_NCS_R_MEAN < 0.45	4	0.45	1411.0	1033.0	2444.0	0.577	-0.31176
0.45 <= IMP_PA_NCS_R_MEAN < 0.53	5	0.53	909.0	567.0	1476.0	0.616	-0.47192
IMP_PA_NCS_R_MEAN >= 0.53	5		976.0	436.0	1412.0	0.691	-0.80575

PA_NPRT_W_MEAN Değişkeni için WOE Dönüşümü



Value	Group	Cutoff	Event Count	Non Event Count	Total	Event Rate	WOE
MISSING	1		0.0	0.0	0.0	0.0	0.0
IMP_PA_NPRT_W_MEAN < 0	1	0.0	0.0	0.0	0.0	0.0	0.0
0 <= IMP_PA_NPRT_W_MEAN < 0.1	1	0.1	10011.0	11229.0	21240.0	0.471	0.11489
0.1 <= IMP_PA_NPRT_W_MEAN < 0.17	2	0.17	981.0	807.0	1788.0	0.549	-0.19518
0.17 <= IMP_PA_NPRT_W_MEAN < 0.25	3	0.25	559.0	393.0	952.0	0.587	-0.35227
0.25 <= IMP_PA_NPRT_W_MEAN < 0.33	3	0.33	631.0	461.0	1092.0	0.578	-0.31384
0.33 <= IMP_PA_NPRT_W_MEAN < 0.5	4	0.5	844.0	537.0	1381.0	0.611	-0.45208
IMP_PA_NPRT_W_MEAN >= 0.5	4		1186.0	784.0	1970.0	0.602	-0.41386

B_RRCWC Değişkeni için WOE Dönüşümü



Ek II. Algoritmalar için Hiperparametreler

Gradyan Artırma

Train	
Variables	
Series Options	
N Iterations	300
Seed	12345
Shrinkage	0.1
Train Proportion	70
Splitting Rule	
Huber M-Regression	No
Maximum Branch	30
Maximum Depth	10
Minimum Categorical Size	5
Reuse Variable	1
Categorical Bins	30
Interval Bins	100
Missing Values	Use in search
Performance	RAM
Node	
Leaf Fraction	0.05
Number of Surrogate Rules	0
Split Size	.
Split Search	
Exhaustive	5000
Node Sample	20000
Subtree	
Assessment Measure	Misclassification
Score	
Subseries	Best Assessment Value
Number of Iterations	1
Create H Statistic	No
Variable Selection	Yes
Report	
Observation Based Importance	No
Number Single Var Importance	5

Lojistik Regresyon (WOE)

Train	
Variables	
Scorecard Points	
Score Ranges	
Analysis Variables	WOE
Freeze Scorecard Points	None
Publish Score Code	
Output Variables	Complete
Scaling Options	
Intercept Based Scorecard	No
Reverse Scorecard	No
Odds	50.0
Scorecard Points	200.0
Points to Double Odds	20.0
Scorecard Type	Summary
Precision	0
Bucketing Method	Min/Max Distribution
Number of Buckets	25
Use Indeterminate Values	No
Revenue Accepted Good	1000
Cost Accepted Bad	50000
Current Approval Rate	70.0
Current Event Rate	2.5
Generate Characteristic Analysis	No
Adverse Characteristic Options	
Method	Neutral Score
Display Value	No
Generate Report	No
Number of Characteristics	3
Adverse Variables	
Model Selection	
Selection Model	Backward
Criterion	Cross Validation Misclassification
Model Ordering	
Use Selection Defaults	Yes
Entry Significance Level	0.05
Stay Significance Level	0.05
Start Variable Number	0
Stop Variable Number	0
Force Candidate Effects	0
Maximum Number of Steps	0

Yapay Sinir Ağları

Train	
Variables	
Use Inverse Priors	No
Create Validation	No
Network Options	
Input Standardization	Range
Architecture	Two Layers
Number of Hidden Neurons	20
Number of Hidden Layers	3
Hidden Layer Options	
Direct Connections	No
Target Standardization	Range
Target Activation Function	Identity
Target Error Function	Normal
Number of Tries	2
Maximum Iterations	300
Use Missing as Level	No

Lojistik Regresyon

Train	
Variables	
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Backward
Selection Criterion	Cross Validation Misclassification
Use Selection Defaults	Yes
Selection Options	
Optimization Options	
Technique	Congra
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour
Convergence Criteria	
Uses Defaults	Yes
Options	
Output Options	
Confidence Limits	No
Save Covariance	No
Covariance	No
Correlation	No
Statistics	No
Suppress Output	No
Details	No
Design Matrix	No

Destek Vektör Makineleri

Train	
Variables	
Maximum Iterations	80
Use Missing as Level	No
Tolerance	1.0E-6
Penalty	1.0
Optimization Method	
Optimization Method	Interior Point
Interior Point Options	
Active Set Options	

Rassal Orman

Train	
Variables	
Tree Options	
Maximum Number of Trees	50
Seed	12345
Type of Sample	Proportion
Proportion of Obs in Each Sample	0.6
Number of Obs in Each Sample	.
Splitting Rule Options	
Maximum Depth	50
Missing Values	Use In Search
Minimum Use In Search	1
Number of Variables to Consider in Split	.
Significance Level	0.05
Max Categories in Split Search	30
Minimum Category Size	5
Exhaustive	5000
Node Options	
Method for Leaf Size	Proportion
Smallest Percentage of Obs in Node	0.1
Smallest Number of Obs in Node	1
Split Size	.
Use as Modeling Node	Yes
Score	
Variable Selection	Yes
Variable Importance Method	Loss Reduction
Number of Variables to Consider	25
Cutoff Fraction	0.01

Karar Ağacı

Train	
Variables	
Splitting Rule	
Interval Target Criterion	Variance
Nominal Target Criterion	Entropy
Interval Bins	100
Minimum Distance	0.01
Significance Level	0.2
Bonferroni	No
Missing Values	Largest
Use Input Once	No
Maximum Branch	2
Maximum Depth	10
Minimum Categorical Size	5
Node	
Leaf Size	5
Surrogate Rules	0
Validation	
Create Validation	No
Validation	0.15
Partition Seed	12345
Split Search	
Exhaustive Search Comparisons	500000
Fast Search Comparisons	1000000
Subtree	
Subtree Method	Cost-Complexity
Selection Method	Automatic
Confidence	0.25
Nominal Target Assessment	Entropy
Minimum Subtree	No
Assessment Threshold Value	1.0
Number of Leaves	1
Cross Validation Folds	10
Cross Validation Seed	12345
Score	
Variable Selection	Yes
Node and Leaf Role	Segment

K-En Yakın Komşu

Train	
Variables	
Method	RD-Tree
Number of Neighbors	200
Epsilon	0.0
Number of Buckets	100
Weighted	Yes
Create Nodes	No
Create Neighbor Variables	Yes