T.C.
MİMAR SİNAN GÜZEL SANATLAR ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KREDİ SKORLAMADA YAPAY ZEKÂ TEKNİKLERİ İLE ÇOK AŞAMALI LOJİSTİK MODELLEMEYİ TEMEL ALAN HİBRİT YAKLAŞIMLAR

DOKTORA TEZİ

Damla İLTER

Tez Danışmanı: Doç.Dr. Eylem DENİZ

OCAK 2021

T.C.
MİMAR SİNAN GÜZEL SANATLAR ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KREDİ SKORLAMADA YAPAY ZEKÂ TEKNİKLERİ İLE ÇOK AŞAMALI LOJİSTİK MODELLEMEYİ TEMEL ALAN HİBRİT YAKLAŞIMLAR

DOKTORA TEZİ

Damla İLTER
(20130897001)

Tez Danışmanı: Doç.Dr. Eylem DENİZ

OCAK 2021

Bu tez çalışması TÜBİTAK BİDEB 2214-A Yurtdışı Doktora Sırası Araştırma Burs Programı tarafından 1059B141800285 numaralı proje ile desteklenmiştir.

ÖNSÖZ

İlk olarak; çalışmalarıma ve bana duyduğu güven ile bana destek olan, değerli fikirlerini, bilgi birikimini paylaşan ve yardımlarını esirgemeyerek katkıda bulunan danışmanım Doç. Dr. Eylem DENİZ' e,
Tez izleme toplantılarında destekleyici konuşmalarından ve tezin şekillenmesinde önemli katkılarından dolayı değerli hocalarım Prof. Dr. M. Aydın ERAR'a, Doç. Dr. Ayça Çakmak Pehlivanlı'ya ve Doç. Dr. Esra AKDENİZ'e,
Emekleri için Doç. Dr. Ozan KOCADAĞLI'ya saygılarımı sunarken, akademisyen kimliğimle MSGSÜ ailesine katılmama vesile olan ve tüm doktora sürecim boyunca manevî desteklerini, bana inançlarını hiç esirgemeyen başta İstatistik bölüm başkanımız Prof. Dr. Gülay BAŞARIR'a, Dr. Öğr. Üyesi M. Levend DURANSOY'a ve diğer bölüm hocalarım ile birlikte çalıştığım bölüm arkadaşlarıma,
Lisans hayatımdan bugünlere kadar gelebilme serüvenimde üzerimde çok büyük emeği olan ve her zaman desteğini, maneviyatını üzerimde hissettiren Prof. Dr. Erol EĞRİOĞLU'na,
Hayatım boyunca sonsuz sevgi ve desteklerini hissettiğim, bana karşı tüm emekleri boyunca hiçbir fedakârlıktan kaçınmadan, tüm sıkıntılarımda yanımda olup, zorlukların üstesinden gelmemde bana yol gösteren, her konuda bana güvenip destek olup ve aldığım her kararda beni cesaretlendiren sevgili Annem ve Babam'a,
Birbirimizden ayrı kaldığımız tüm uzun süreçlerde birlikte vakit geçirebilmek için beni sabırla bekleyen, sevgileri ve varlıkları ile beni mutlu eden, hayatımın gülen yüzleri olan başta sevgili Ablam olmak üzere canım yeğenlerim Duru ve Masal'a,
Manevi destekleri sayesinde huzur bulmamı sağlayan, sevgi ve desteklerini tez dönemi boyunca esirgemeyen, dertlerime ortak olup, yeri geldiğinde üzüntülerimi unutturup beni mutlu eden, iyi ki varlar dediğim mesafelerin ayıramadığı sevgili dostlarıma ve yanıbaşımda olan sevgili dostlarıma,
Son olarak; tanıdığım günden bu yana işime, çalışmalarıma ve tezime olan sabrı, bana karşı bitmeyen anlayışı, sevgi ve saygısı için nişanlım Manar'a ,

sonsuz teşekkürler.

Küçük yeğenim Masal'a

OCAK 2021

Damla İLTER
(Öğretim Görevlisi)

İÇİNDEKİLER

Sayfa

KISALTMALAR

ABD : Amerika Birleşik Devletleri
AIC : Akaike Bilgi Kriteri (Akaike Information Criterion)
AUC : Eğri Altında Kalan Alan (Area Under the Curve)
BIC : Bayesçi Bilgi Kriteri (Bayesian Information Criterion)
CIT : Koşullu Çıkarım Ağaçları (Conditional Inference Trees)
DM : Veri Madenciliği (Data Mining)
DVM : Destek Vektör Makineleri
EKK : En Küçük Kareler
FICO : Fair Isaac Corporation
IDB : İrlanda Dummy Bankası Irish Dummy Bank
KS : Kolmogorov Smirnov
LR : Lojistik Regresyon (Logistic Regression)
MLM : Çok Aşamalı Lojistik Model (Multilevel Logistic Model)
RF : Rastgele Orman (Random Forest)
ROC : Alıcı İşletim Karakteristik (Receiver Operating Characteristic)
RP : Yinelemeli Bölümleme (Recursive Partition)
YSA : Yapay Sinir Ağları (Artificial Neural Networks)
YZT : Yapay Zeka Teknikleri (Artificial intelligence techniques)

ÇİZELGE LİSTESİ

ŞEKİL LİSTESİ

KREDİ SKORLAMADA
YAPAY ZEKÂ TEKNİKLERİ İLE ÇOK AŞAMALI LOJİSTİK MODELLEMEYİ TEMEL ALAN HİBRİT YAKLAŞIMLAR

ÖZET

Gelişen teknolojiyle birlikte finans dünyasındaki verinin akış hızının ve hacminin artması son yıllarda bu alanda kullanılan istatistik ve yapay zekâ tekniklerine olan ilgiyi arttırmıştır. Özellikle süper bilgisayar sistemlerindeki depolama hacmi ve işlemci hızındaki artış yukarıda bahsi geçen tekniklerle beraber hibrit yapay zekâ tekniklerinin kullanımının önünü açmıştır. Bununla birlikte, veri boyutundaki artış verinin derlenmesi, model kestirimi, modeller üzerinden yapılan çıkarımların testi ve güvenilirliği bakımından sorun teşkil etmektedir. Bu sorunlara çözüm ararken izlenmek istenen yol, yapay zekâ teknikleri analizlerinde yer alan alternatif modellerin içerisinde en iyi olan modelin değişkenleri ile kurulan çok aşamalı lojistik modelleme sonuç modelinin standart uyum kriterlerinden Akaike bilgi kriteri ve Bayes bilgi kriteri tarafından değerlendirilmesi ve yorumlanmasıdır. Finansal modelleme yaparken, kredi skorlamada kullanılabilecek etkin bir karar destek sisteminin geliştirilmesi ve verinin yorumlama gücünü artırmak bu çalışmanın amacıdır. Çözüm olabilmesi adına kredi skorlamada yapay zekâ teknikleri ile çok aşamalı lojistik modellemeyi temel alan hibrit yaklaşımlar elde edilmeye çalışılmış ve finansal açıdan veri detaylandırılmıştır. Geliştirilen yaklaşım ile hem yapay zekâ tekniklerinin hem de çok aşamalı lojistik modellemenin hibritleşen model kestirimi elde edilmiştir. Bu amaç doğrultusunda, kredi skorlamada öne çıkmış yapay zekâ teknikleri ile doğruluk oranı yüksek olan, finansal veriyi açıklamada kullanılacak özellik seçimi detaylı olarak incelenmiştir. Elde edilen modelin, finansal yorumlama açısından kredi skorlamasını değerlendirebilme bağlantısının kurulmasında çok aşamalı lojistik modellemenin avantajlarına da yer verilmiştir.

Anahtar kelimeler: Hibrit Yapay Zekâ Yöntemleri, Çok Aşamalı Lojistik Modelleme, Kredi Skorlama, Finansal Modeller, Hiyerarşik Yapılar, Havuzlama

HYBRID APPROACHES BASED ON ARTIFICIAL INTELLIGENCE AND MULTILEVEL LOGISTIC MODEL IN CREDIT SCORING

SUMMARY

The increase in data flow rate and size in the financial world with the developing technology has increased the interest in statistics and artificial intelligence techniques used in this field in recent years. Especially the increase in storage capacity and processor speed in supercomputer systems has caused the way for the use of hybrid artificial intelligence techniques together with the techniques mentioned above. In addition, the excessive increase in data size causes problems in terms of data processing, model estimation, testing, and reliability of inferences made from models. When determining results to these problems, the desired way to follow is the evaluation and interpretation of the Multilevel Logistic Model result model, which is established with the variables of the model that is the best among the alternative models in the artificial intelligence techniques analysis, by the Akaike Information Criterion and the Bayesian Information Criterion. This study aims to develop an effective decision support system that can be used in credit scoring while performing financial modeling and to realize its software. In order to be a solution, hybrid approaches based on artificial intelligence techniques and multilevel logistics modeling in credit scoring were tried to be obtained and the data were detailed financially. With the developed approaches, the hybridizing model estimation of both artificial intelligence techniques and multilevel logistics modeling has been obtained. The selection of features with high accuracy rate, which will be used in explaining financial data, has been analyzed in detail with artificial intelligence techniques that stand out in credit scoring. The advantages of multilevel logistics modeling are given to establish the link to evaluate the obtained model, credit scoring in terms of financial interpretation.

Keywords: Hybrid Artificial Intelligence Techniques, Multilevel Logistic Models, Credit Scoring, Financial Models, Hierarchical Structures, Pooling

1 GİRİŞ

Kredi skorlama, geliştirilen en eski finansal risk yönetimi araçlarından biridir. Aynı zamanda tüketici davranışları açısından incelenen en eski veri madenciliği konularının öncüsü olarak bilinmektedir (Thomas, 1998). Kredi skorlama alanında ilk çalışma Durand tarafından "iyi" ve "kötü" kredi kavramlarının tanımlamaları yapılarak başlatılmıştır (Durand, 1941).

II. Dünya savaşının başlamasıyla tüm finans büroları kredi kararları ile ilgili zorluklarla tanışmaya başladılar. Ordu içerisindeki kredi uzmanları, deneyimlerinden yola çıkarak bilgi vermekteydiler. Bu deneyimler sonucu kredi uzmanları, kredi kullanmak isteyen bireylere yol gösterici olabilmek adına kural listesi tanımladılar (Johnson, 1992). Savaşın ardından kredi verilip verilmemesi durumu üzerine alınacak kararlarda istatistiksel sınıflandırma tekniklerinin ve modellemelerin etkisinin hissedilmesi çok uzun sürmedi (Wonderlic, 1952). Myers ve Forgy, hesaplama gücünün büyümesiyle bireysel yargılardan ve hata oranlarından uzak, daha iyi bir kredi skorlama yöntemi bulmuşlardır (Myers ve Forgy, 1963). 1980'ler ve 1990'larda kredi skorlama başarılı bir şekilde bankalarda uygulanmaya başlandı. 1980'lerde lojistik regresyon ve linear programlama tekniklerinin kredi skorlaması üzerindeki başarısı tanıtıldı. 1990'larda ise bankaların ortak bir dilde karar alabileceği skor kartları oluşturuldu (Lewis, 1992). Gelişen istatistiksel yöntemlerle birlikte bir çok çalışma kredi skorlama teknikleri adı altında tanımlanmaya başladı (Rosenberg ve Gleit, 1994; Henley ve Hand, 1996; Mays, 1998; Thomas ve diğ., 2002).

Günümüzde tüm dünya ekonomilerinin birbirine olan bağımlılığı göz önüne alındığında finansal veriler arasındaki korelasyonun ülkeler, finansal kurumlar, kobiler ve bireysel yatırımcılar üzerinde farklı ve karmaşık yansımaları görülmektedir. Bu yansımaların ekonomik çökmelere ve çeşitli krizlere sebep olduğu bir gerçektir. Kara pazartesi, asya mali krizi, dünya gıda krizi ve küresel ekonomik kriz bunların belli başlıcaları olsa da ülkesel anlamda 2001 Türkiye ekonomik krizi ve 2011

Yunanistan' daki ekonomik kriz örnek verilebilir. Bu krizlerin sistematik ve sistematik olmayan birçok faktöre bağlı olduğu bilinmektedir. Sistematik olmayan riskin iyi bir yönetim ve planlama ile minimize edilebileceği varsayılabilir olsa da sistematik riskin minimizasyonu için aynı varsayımın yapılması pek mümkün görülmemektedir. Nitekim, gerçek yaşam sistemleri olarak adlandırılabilecek doğa olaylarından (deprem, uzay araştırmaları vb.), tıbbî ve finansal alanlardan elde edilen veriler yardımıyla modellemenin temel zorluğu sistematik risktir. İlgili alanlara özgü sistematik riski oluşturabilecek anabileşenlerin saptanması ve bunlara ait gerçek verilerin elde edilmesi çoğu zaman mümkün olamamaktadır. Elde edilebilen veriler veya bunların simülasyonları yardımıyla gerçek sistemin ancak yaklaşık bir modellemesi yapılabilmektedir.

Genel olarak, uzmanlar dünya çapında kredi skorlama modellerinin çok fazla ortak yanı olduğunu kabul etmektedir. Literatürde çeşitli skorlama sistemleri mevcut olsa da Fair Isaac Corporation (FICO) ve VantageScores bunların arasında popüler olan modellerdir. FICO, finans sektörlerinde müşterilerin kredi erişimlerini genişletmek adına bir yol olarak 1960'dan beri kredi riski puanı hesaplamaktadır. Hala Amerika Birleşik Devleri (ABD)'ndeki finansal kurumlar tarafından kullanılan en popüler skorlama yöntemi FICO' dur. FICO kendi içerisinde Equifax, Experian ve TransUnion adı altında 3 tip skor üretir. Her bir ölçeğin sınıflandırılmış skorları mevcuttur. Kredi değerlendirmeleri bu skor aralıklarına göre yapılmaktadır. VantageScore ise FICO'ya göre daha yeni bir puanlama sistemidir ve FICO'ya rekabetçi bir alternatif skorlama yöntemi olarak sunulmuştur. Çevrimiçi olarak sunulan ücretsiz kredi skorlarının çoğu VantageScores'dur. Kredi skorlama için ise Türkiye, Ficth Ratings kullanmaktadır. Fitch Ratings, aynı zamanda Moody's ve Standart Poor's ile beraber dünyanın en büyük ABD Güvenlik ve Döviz Komisyonu tarafından tanınan istatistiksel skorlama kuruluşundan birisidir. Bunların yanısıra, bankalar kendilerine özgü kredi skorlama modellerini de kullanmaktadırlar.

Tüm ülkelerde finans sektöründe bankacılık, ekonominin geneli için pozitif bir faktör olarak daima öne çıkmaktadır. Türkiye açısından da finans sektörünün %80'ini oluşturan bankacılık, güçlü ve sağlıklı yapısı ile reel sektör etkileşimi ile dikkat çekmektedir. Bilindiği üzere küresel kriz sonrası dönemde batı dünyasında pek çok banka doğrudan veya dolaylı olarak devlet desteği almak zorunda kalmıştır.

Gelişen bilgisayar sistemleriyle birlikte veri depolama ve işleme kapasitesi günden güne artmış, mühendislik, istatistik ve yöneylem araştırması alanlarında daha önceleri hız ve maliyet açısından etkin bulunmayan birçok yöntem ve algoritma günümüzde işlevsel hale gelmiştir. Büyük veri bankalarının oluşturulması ve süper bilgisayarların devreye girmesi ile birlikte uzman görüşü, fonksiyon ve parametre vb. kontrolleri gerektiren klasik yöntemler ile yapay zekâ teknikleri (YZT) birlikte hibrit sistemlerle kontrol edilebilmektedirler. Gerçek yaşam sistemlerininin özünde var olan dinamik, raslantısal ve doğrusal olmayan yapıya, verinin barındırdığı kesin olmayan belirsizlikler eklenince bu tür sistemlerin modellenmesi için hibrit yöntemlerin geliştirilmesi kaçınılmaz olmaktadır. Finansal araştırmalar bünyesinde, veri madenciliği (VM) ve çeşitli yöntemler yardımıyla hibrit yapay zekâ tekniklerinin kullanılmasının daha hızlı ve daha doğru çıkarımların yapılabilmesi için gerekliliği öngörülmüştür.

Literatürde, VM ve büyük veri setleriyle ilgili çalışmalarda stokastik doğrusal olmayan programlama, YZT, yapay sinir ağları (YSA), destek vektör makineleri (DVM), regresyon ağaçları, meta-sezgisel yöntemler, uzman sistemleri veya bunların hibrit yaklaşımlarını temel alan birçok metodoloji mevcuttur. Bu yaklaşımlar çoğunlukla veriyi kümeleme, modelleme için kredi skorlamasında başarı ile kullanılan önemli tekniklerdir. Özellikle son yıllarda, yüksek depolama ve hızlı işletim sistemlerine uygun hale getirilen YZT'lerin ön plana çıktığı görülmektedir. Lakin, finans alanındaki verinin büyüklüğü, içerdiği bilginin kesinliği ve belirsizliğine ek olarak oluşturulacak modellerin karmaşıklığı ve bunlar üzerinden gerçekleştirilecek çıkarımlarım güvenilirliği ayrı bir araştırma konusudur.

Finansal modelleme problemlerinin başında kredi skorlaması gelmektedir. Son yıllarda dünya ekonomisinde finansal krize paralel olarak birçok ekonomik faktörün yanısıra takipteki kredilerin sayısı artmış ve böylece kredi skorlama modellerinin önemi artmıştır. Sistematik ve sistematik olmayan riskleri, ödemelerin takibi ve yasal süreçler gibi birçok durumu aynı anda ölçme ve değerlendirme süreci kapsamında kredi skorlama önemli bir araştırma sahasıdır. Erken uyarı mekanizmalarıyla, ekonomik, finansal ve mali risklerin analiz edilerek uluslararası kırılganlıkların önceden belirlenip, risk yaratacak unsurların azaltılması veya tamamen ortadan kaldırılması amaçlanmaktadır.

Finansal kurumların riski en aza indirgeyecek finansal modellemeleri yapabilmeleri ve erken uyarıcı sistemler geliştirebilmeleri için kullanabilecekleri bir çok bilimsel yöntem mevcuttur. Geleneksel istatistiksel ve ekonometrik yöntemler, ön koşullar içermeleri aşırı dinamik doğrusal olmayan durumlarda bile doğrusal fonksiyonel yaklaşımlardan dolayı eleştirildikleri noktada YZT, finansal modelleme için esnek bir yaklaşım sağlamaktadır. Ancak finans sektöründe karar verme veya strateji geliştirme aşamasında verinin yapısı gereği hiyerarşik modellemelerin göz ardı edilmemesi gerekmektedir. Özellikle kredi kullanan kişilerin kredi alırken verdikleri kişisel bilgiler doğrultusunda oluşacak bilgi zincirinde ortak özelliğe sahip bireyler mevcut olabilmektedir. Bu doğrultuda, portföy optimizasyonu, risk ölçümü, kredi skorlama, erken uyarıcı sistemi gibi dinamik ve doğrusal olmayan problemlerin, YZT temel alan ve geleneksel modellerle hibrit hale getirilmesi, finansal kurumlardaki yöneticilerin ve uzmanların karar mekanizmasını hız ve doğruluk bakımından iyileştirmek anlamına gelmektedir. Ekonomik faktörlerin çeşitliliği ve değişkenliği, kredi başvuru sayısı ve verinin büyüklüğü gözönüne alındığında, bu tür otomatik karar mekanizmaları finansal kurumların iş yüküyle birlikte hata riskini de azaltacaktır.

Doğrusal olmayan sistemlerin modellenmesindeki en önemli faktör veride ilgilenilen probleme ilişkin olarak hangi tekniklerin kullanılması gerektiğidir. Bunun yanısıra, model seçim kriterleri ile belirlenen modellerden elde edilen çıkarımların etkinliği, güvenilirlik analizleriyle sınanmalıdır. Literatürde açıklayıcı değişken olarak ikili (0/1, Doğru/Yanlış, Evet/Hayır) olasılık tahmini, kredi değerlendirmesinde yaygın olarak kullanılmıştır. Bu ifadeler herhangi bir kredi başvurusunun onaylanmaya uygun olup olmadığını açıklamaktadır.

Bu çalışmanın amacı, finans sektörünün temel bileşeni haline gelen kredi kullanımından, takibinden, sonuçlandırılmasına kadar geçen süreçteki risk analizlerinin yapılmasında ve yorumlanmasında güvenilirliği yüksek olan yeni bir karar destek sistemi geliştirilmesidir. Bu bağlamda, geliştirilecek algoritmaların bu tür problemlerin çözümünde yapılan hataları en aza indirdiği gibi kullanıcılara da güçlü bir karar destek sistemi sağlaması beklenmektedir. Bunun için dinamik doğrusal olmayan sistemlerin modellenmesinde ve çözülmesinde esnek bir yapıya sahip olduğu kadar etkin çözümler üretebilen hibrit yaklaşımlar elde edilecektir. Önerilen yaklaşımda, yüksek boyutlu doğrusal olmayan problemlerde sıklıkla kullanılan ve

YZT kapsamında değerlendirilerek, istatistiksel ve makine öğrenimi modellemeleriyle yapılan kestirimlerin karmaşıklığını ölçmek ve bu modeller üzerinden gerçekleştirilen minimum hatayı öngören çıkarımların yapılmasına yoğunlaşılmıştır. Son olarak finans sektörünün çok önemli bir parçası olan kredi skorlama için öngörülen çıkarım modelindeki skorlamada etkili faktörler, havuzlanmış (pooling) MLM ile zamana bağlı olarak tekrar modellenip açıklanmıştır. Finans alanında büyük önemi olan kredi kullanımı, kredi skorlaması, kredi değerlendirmesi gibi müşteriye, kuruma veya kullanılan kredi türüne göre etkileyen faktörlere dair literatürde çalışmalar mevcut olmasına karşın, sistemdeki hiyerarşik yapıyı araştırmaya ve bu hiyerarşik yapıda mevcut olan rastgele, sabit ve karışık etkilere göre yorumlanan bir araştırmaya daha önce rastlanılmamıştır.

Birinci Bölüm'de literatürde mevcut olan tekniklerden hangilerinin daha etkin olduğu araştırılmış ve değerlendirilmiştir. Bu aşama, geliştirilecek yeni yaklaşımlar için bir referans olduğu gibi mevcut çalışmaların güçlü ve eksik yanlarının saptanması bakımından fayda sağlayacağı ve geliştirilmesi düşünülen yeni yöntemlerin ne gibi yenilikler içermesi gerektiği konusunda da yol gösterici olabileceği düşünülmektedir.

İkinci Bölüm'de ilgili problem için YZT kavramlarından bahsedilerek, kredi skorlamada öne çıkmış modellemeler detaylı olarak anlatılmıştır.

Üçüncü Bölüm'de MLM anlatılarak, YZT ile hibritlenebilmesi ve entregrasyonunda kredi skorlamasını değerlendirebilme bağlantısının kurulmasına, modellemenin avantajlarına yer verilmiştir.

Dördüncü Bölüm'de uygulamaya yönelik olarak, geliştirilen yöntemin uygulanabilirliği test edilip performans bakımından mevcut yaklaşımlarla karşılaştırılmıştır. Bunun yanısıra, geliştirilen yöntemlerle kestirilen modellerin kullanımı ile elde edilen sonuçların etkinlik ve güvenilirlik analizleri bu bölümde verilmiştir. Bu aşama için bir kredi bankası olan Irish Dummy Bankası'na ait riske bağlı olarak kullandırdıkları veri kümesi setiyle amaca yönelik YZT ile veri üzerinde değişken seçimleri yapılmasına yer verilerek, yanıt değişkeni ikili olan modellerin değerlendirilmesinde Kolmogorov Smirnov (KS) ve gini endeksi kullanılmıştır. Ardından etkili değişkenlerin kredi verilip verilmemesindeki etkisi MLM analizleri ile yorumlanmıştır. Gerçekleştirilen

analizlerin değerlendirilmesi için R İstatistiksel programlama dilinden (R-Studio, 2019), (versiyon 4.0.3) faydalanılmıştır.

Beşinci Bölüm olan sonuç bölümünde ise yapılan uygulamanın sonucunda elde edilen bilgilerin ve bulguların yorumlanıp, tartışılmasına yer verilmiştir.

2 GENEL BİLGİLER

2.1 KREDİ SKORLAMA için YAPAY ZEKÂ TEKNİKLERİ TARİHÇESİ

Kredi skorlama, finansal modellemeler için oldukça önem arz etmektedir. Son yıllarda dünya ekonomisindeki finansal krizin yanı sıra birçok ekonomik faktör, kredileri artırmış ve dolayısıyla kredi skorlama modellerinin önemini ortaya çıkartmıştır. Kredi kullanacak bireylerin önceden finansal kurum tarafından belirlenmiş puan eşiğinin üzerinde olması beklenir. Kredi riski kaçınılmaz bir gerçektir, çünkü finansal kuruluşlar herhangi bir kredi başvurusunu onaylamak ve değerlendirmek için birçok faktörü aynı anda değerlendirirler. Böyle bir durumda kredi kullanımına izin verilirken, puan eşiğinin altında kalan kullanıcılara kredi verilmemektedir. Tam da bu yüzden finansal banka veya kurum tarafından kredi kullanacak bireyler doğru bir şekilde değerlendirilmelidir. Son yıllarda bu anlamda doğrusal olmayan yaklaşımlar ve veri madenciliği çalışmalarıyla hatayı minimize etmeyi amaçlamışlardır. İstatistiksel yöntemler, parametrik olmayan yöntemler, VM ve YZT yaklaşımları önerilmiştir (Ong, 2005; Keramati ve Yousefi, 2011; Bunker ve diğ., 2017; Kuznetsov, 2020).

Kredi skorlama sürecini iyileştirmenin ve doğruluk oranlarının artırılması için kullanılan karar ağaçlarının kontrolünün minimum hataya yer verecek şekilde yapılmasında fayda vardır. Finansal kuruluşlar istatistiksel karar alırken oluşturdukları karar ağaçları, örüntüleme, kümeleme veya sınıflandırma gibi aşamalarda çeşitli sorunlarla karşılaşabilirler. Bu tarz sorunları çözebilmek için yapılmış veya yapılmaya devam çalışmalar halen literatürde mevcuttur. Kredi skorlama ve kredi değerlendirme için finansal kuruluşların denetimli veya denetimsiz olmak üzere bünyesinde birçok teknik barındıran YZT kullandıkları görülmüştür. Kullanımdaki amaç ise tüm olası problemlerden arındırılan değişkenler arasındaki ilişkiyi anlamlı kılan, karar yapısını öğrenebilmek ve gerçek hayat verilerine uygulayabilmektir.

Khasman (2010) çalışmasında, YSA kullanarak kredi başvurularının otomatik olarak işlenmesinde verimli ve hızlı bir şekilde kullanılabilmesi için hangi sinir ağının, hangi öğrenme şeması altında önerilen kredi riski değerlendirme sisteminde optimum performans sağlayabileceğine cevap aramıştır. Ghodselahi ve Amirmadhi (2011), yine YSA kullanarak kredi riskinin finansal firmalar için en büyük risk olduğunu çalışmasında göstermişlerdir.

Ionescu (2018), tamamen ekonomik kuruluşlar tarafından kullanılabilen hızlı, şeffaf ve karmaşık iş kararlarını, akıllı dijital makine kavramı adı altında YZT tekniklerinin kullanılarak detaylandırabilmesi ile bireysel kararlarının müdahalesi olmadan işlem yapımının mümkün olabileceğini göstermiştir. Olaniyan ve Maheswan (2017) yaptıkları çalışmada, bulut sistemlerin sahip olduğu çok sayıda karmaşık, farklı ve heterojen yapıdaki bilgi kümelerinin birlikte bulunmasının yarattığı hesaplama, kaynak yönetimi arasındaki farklılıkları veya benzerlikleri gibi durumlara çeşitli programlama yaklaşımları ile çözüm aramışlardır. Zeng ve diğ. (2017), önerdikleri çalışmada açıklayıcı değişkenlerin, lojistik regresyondaki monoton fonksiyonların dönüşümlerinin genellikle veriye uyumunun araştırılmasının yanısıra maksimum olasılık tahminleriyle bilgi değerlerinin karmaşıklığını azalttığını ispatlamıştır. Bhatia ve diğ. (2017), yaptıkları çalışmada kredi skorlama hesaplamalarında müşterilerin kredi taleplerini incelemek için hibrit yapay zekâ tekniklerini kullanmışlardır. Goh ve Lee (2019), çalışmalarında kredi puanlamasında başarılı bir performans gösteren YZT'ini kullanarak, daha sonra hibrit modelleme için yaklaşım önermişlerdir.

Kredi skorlama için kullanılan Bayes yöntemlerinin YSA'larda sınıflandırma ve regresyon problemlerinde başarılı sonuç verdiği görülmüştür (MacKey, 1992; Neal, 1996). Abdou ve diğ. (2018), yaptıkları çalışmada kredi skorlama da YSA'nın klasik tekniklere göre daha iyi performans elde etmesi ile alternatif bir çözüm olarak düşünülmesini sağlamışlardır. Belotti ve Crook (2009), çalışmalarında geleneksel yöntemlere DVM yöntemlerini karşılaştırarak, DVM'lerin bir özellik seçim yönteminin temeli olarak kullanılabileceğini savunmuşlardır. Yine Chen ve diğ. (2009), kredi skorlama için DVM yöntemlerinin performans gücünü, sınıflandırma, regresyon ağacı, çok değişkenli uyarlamalı regresyon eğrileri ve model parametrelerinde göstermişlerdir. Oreski ve diğ. (2012), YSA'nın sınıflandırma oranlarını artıran bir optimal modeli bulabilmek için bir özellik seçim yöntemi

önermişlerdir. Alaraj ve Abbod (2016) ve Xiao ve diğ. (2016) hibritleştirilmiş YZT ve son zamanlarda kullanılan temel sınıflandırıcılar dahil olmak üzere farklı topluluk yöntemleriyle kredi skorlaması yapmışlardır. Optimal model tahmini ve minimum hata terimi elde etmek için de Bayes yaklaşımlar kullanılmaktadır. Bu anlamda Ilter ve Kocadağlı (2019), önerdikleri model için iki ayrı kredi skorlama verisi üzerinde, çapraz entropi ve bulanık ilişkiler ile YSA'yı ve geleneksel istatistik modellerini kıyaslamışlardır. Roy ve Urolagin (2019) müşterilerin güvenilirliğini, doğru performansını, rasgele orman ve DVM kullanarak göstermişlerdir.

Kredi skorlamasında, yanıt değişkeni ikili olan modellerin değerlendirilmesinde eğri altında kalan alan (AUC), KS ve gini endeksi değerleri baz alınarak tercih edilecek model seçiminde kullanılmıştır. Abellan ve Castellano (2017), çalışmalarında kredi skorlama değerlendirmesinde, iflas tahmini üzerine iyileştirmeler yapmışlardır. Gini endeksi, geleneksel karar ağaçlarında bilgi ölçüsünü maksimize eden değişkeni seçmek için kullanılmaktadır. Ayrıca ekonomik veriler için Gini endeksi, verinin dağılımını değerlendirebilmek amaçlı da kullanıma uygundur. KS, modelin güvenilirliğini ve skorlama fonksiyonunu belirlemek için sıklıkla kullanılan bir yöntemdir. KS sonuçları yardımı ile kredi müşterileri için kullanılan istatistiksel fonksiyonları iyi veya kötü müşteri diye ayırmak mümkündür. Kredi veri setlerinde kullanılan en iyi sınıflandırıcı seçimlerini güçlendirerek, AUC doğruluk oranlarının daha iyi bir dengeye ulaştığını göstermişlerdir. Yufei ve Liu (2017), önerdikleri modelde müşterilerin kredi taleplerini değerlendirmek için doğruluk hata oranını, AUC, AUC-H ve Brier ölçülerini kullanmışlardır. Soui (2019), çalışmasında kredi riskinin üzerinde durarak, riski, üretilen çözümün karmaşıklığını en aza indirgemek adına model doğruluğunu en üst düzeye çıkararak, arama tabanlı bir optimizasyon problemi için çözüm önermiştir.

2.2 KREDİ SKORLAMA için İSTATİSTİK ve YAPAY ZEKÂ TEKNİKLERİ

2.2.1 Lojistik Regresyon

Lojistik regresyon (LR), karmaşık yapıda ve ikili bağımlı değişkeni olan veriler için bir lojistik fonksiyonu kullanarak oluşturulan istatistiksel bir modeldir. LR, logistic modelin parametrelerini tahmin eden popüler istatistiksel tahminleme yöntemlerinden

biridir. LR, verileri tanımlamak , bir bağımlı değişken ile bir veya daha fazla nominal (göstermelik değişkenler), sıra, aralık veya oran düzeyinde bağımsız değişkenin arasındaki ilişkiyi açıklamak için kullanılmaktadır. Ekonomi ve finans verilerini açıklama da sıklıkla kullanılan bir yöntemdir.

- **Avantajları**
  - ✓ LR uygulanması, eğitilmesi ve yorumlanması oldukça kolaydır.
  - ✓ Özellik uzayında sınıfların dağılımı için herhangi bir varsayım yoktur.
  - ✓ Çok terimli regresyon ve sınıf tahminlerinin doğal olasılıklı bir görünümüne kolayca genişletilebilir.
  - ✓ Birçok veri kümesi için doğruluk oranı belirlemede ve veri doğrusal olduğunda oldukça iyi performans gösterir.
  - ✓ Model katsayılarını, değişkenin öneminin göstergeleri olarak yorumlayabilir.
  - ✓ Bir modeldeki katsayının ne kadar uygun olduğunun bir ölçüsünü sağlamakla kalmaz, aynı zamanda ilişkilendirme yönünü (pozitif veya negatif) gösterir.
- **Dezavantajları**
  - ✓ Doğrusal sınırlar oluşturur.
  - ✓ Gözlem sayısı değişken sayısından az ise kullanılmamalıdır.
  - ✓ Lineer olmayan problemler LR ile çözülemezler.

2.2.2 Sınıflandırma için Basit ve Bayes Yinelemeli (İteratif) Bölümleme

Yinelemeli bölümleme (RP), çok değişkenli analizler için kullanılan istatistiksel bir yöntemdir. RP, bir sınıfın popülasyonu bağımsız değişkene ait alt kümeleme yaparak bir karar ağacı oluşturmaya yardımcı olur. Yinelemeli olarak adlandırılmasının sebebi her bir alt popülasyon, belirli bir durdurma kriteri tanıtılmazsa bölme süreci sona erene kadar sonsuz sayıda bölünmeye devam eder. Bölümlemenin amacı ise her bir düğümdeki yanıt değişkeninin homojenliğini aşamalı olarak artırmaktır. RP yöntemleri, verilerin uyumu için farklı algoritmalar kullanılarak ve hibritleştirilerek geliştirilmiştir. RP aynı zamanda yanlış sınıflandırmalar yapılmasını önlemek için değişen algoritma modellemelerine göre de oldukça hassastır. Bayes yaklaşımında ise,

özellikle parametre ve model belirsizliğinin ele alınmasıyla ilgili olarak, çıkarımdaki belirsizliği detaylandırabilme özelliğine sahiptir. Yanıt değişkeni kategorik ise, yanıt sınıfını tahmin etmek için bir sınıflandırma ağacı kullanılır ve homojenliğin değerlendirilmesi, bir düğüm içindeki gözlemlerin doğru yerleşimini esas alır. Eğer cevap sürekli ise, bir düğüm içindeki ortalama yanıtı tahmin eder ve homojenliğin değerlendirilmesi, karşılık gelen varyans, sapma veya benzer ölçülerle açıklanan bir regresyon ağacıdır.

- **Avantajları**
  - ✓ Hesaplama yapılmasını gerektirmeyen daha çok sezgisel modeller oluşturur.
  - ✓ Daha hassas veya daha özgül bir karar ağacı oluşturmak için yapılmış yanlış sınıflandırmaların hatalarını görmeye yardımcıdır.
  - ✓ Çoğunlukla kurulan model açıklayıcıdır.
- **Dezavantajları**
  - ✓ Sürekli verilerde çok iyi çalışmaz.
  - ✓ Veride aşırılığa neden olabilir.

2.2.3 Genel ve Koşullu Rastgele Orman

Rastgele orman (RF), sınıflandırma ve regresyon problemleri için eğitim aşamasında çok sayıda karar ağacı oluşturarak probleme göre sınıf veya sayı tahmini yapan bir toplu öğrenme yöntemidir. Denetimli bir sınıflandırma algoritmasıdır. Aynı eğitim örnekleriyle farklı sınıflandırmalar da yapmak mümkündür. Oluşan sınıflar arası değerlendirmede çeşitliliği görmek ve özellik seçimine fayda sağlayacak şekilde yapıyı belirlemek mümkün olacaktır (Breiman, 2001).

- **Avantajları**
  - ✓ Büyük veri setleri için idealdir.
  - ✓ Öğrenme hızlı ve genellikle yüksek doğruluk oranına sahiptir.
  - ✓ Aynı anda birden fazla değişken ile ilgilenebilir.
  - ✓ Bu algoritma için aşırı öğrenme söz konusu değildir.

- **Dezavantajları**

  ✓ Algoritma bir dizi ağaç oluşturduğundan ve en iyi çıktıyı üretmek için çıktılarını birleştirdiğinden, daha fazla hesaplama süresi ve kaynak gerektirir. Bu anlamda yapıdaki karmaşıklık büyük bir sorundur.

  ✓ Algoritma oluşum süreci zaman alıcıdır.

2.2.4 Koşullu Çıkarım Ağaçları

Koşullu çıkarım ağaçları (CIT), yinelemeli bölümlemeye alternatif olarak gösterilmektedir (Hothorn ve diğ., 2006). Parametrik olmayan bir karar ağaçları sınıfıdır ve tarafsız özyinelemeli bölümleme olarak da bilinir. Koşullu bir çıkarım çerçevesinde sürekli ve çok değişkenli yanıt değişkenleri için yinelemeli bir bölümleme yaklaşımıdır. CIT, korelasyonların değerine dayalı olarak bağımlı değişkenlerin özyinelemeli bölümlemesini kullanan farklı bir karar ağacı türüdür. İstatistiksel bir eşiğin geçtiği en güçlü ilişki, bu duruma karşılık gelen girdi değişkeninde ikili bölünme gerçekleştirir; aksi takdirde mevcut düğüm bir geçiş düğümdür. CIT, tahmin performansını sağlarken, girdi değişkenlerine yönelik ön yargılı ağaç oluşturma durumundan kaçınır.

- **Avantajları**

  ✓ Makine öğreniminde diğer algoritmalardan farklı olarak yanlı olmayı engeller. Böylece verilerdeki sorunlar için daha esnek olunmasını sağlar.

  ✓ Bir değişkeni bölmek ve tekrarlamak için ortak değişken seçen permütasyon testi olan bir önem testi kullanır.

  ✓ Anlamlılık testi, algoritmanın her başlangıcında gerçekleştirilir.

- **Dezavantajları**

  ✓ Öğrenme için eksik değerleri olan veriler için uygun bir algoritma değildir.

2.2.5 Destek Vektör Makineleri

Destek vektörleri (DV), sınıflandırma problemleri için kullanılır ve hiper düzlem oluştururlar. DVM'ler optimal çözüm sağlayan doğrudan etkiye sahiptirler. Karar, eğitim örneklerinin (genellikle çok küçük) bir alt kümesi olan destek vektörleri

tarafından belirlenir. DVM yaygın kullanılan bir algoritmadır. Genellikle iyi bir algoritma performansına sahiptirler ve aynı algoritma için çok fazla parametre yorumlamasına gerek kalmadan ele alınan problemi çözebilirler. Aslında DVM, bir dizi çekirdek olarak tanımlanan matematiksel fonksiyonlardan oluşmaktadır. Çekirdeğin işlevi, girdi olarak alınan verileri doğrusal çözüme olanak verecek gerekli yapıya dönüştürmektir. Farklı DVM farklı türlerde çekirdek işlevleri ile kullanılır. Kullanılan çekirdek matematiksel fonksiyonlar ise doğrusal, doğrusal olmayan, polinom, radyal temel işlevli ve sigmoid yapıdaki farklı fonksiyonlardır.

- **Avantajları**
  - ✓ Veri üzerinde herhangi bir bilgi sahibi olmadan DVM algoritmaları çalıştırılabilir.
  - ✓ DVM, sınıflar arası net bir ayrım olduğunda çok iyi çalışan algoritmalardır.
  - ✓ Yüksek boyutlu verilerde daha iyi çalışırlar.
  - ✓ DVM modelleri genelleme yapabilme gücü vardır. YSA modelleriyle karşılaştırıldığında, DVM'ler daha iyi sonuçlar verir.
- **Dezavantajları**
  - ✓ Uygun çekirdeği seçebilmek çok kolay değildir.
  - ✓ Büyük veriler için çok uzun eğitilme süresine sahiptir.
  - ✓ Karmaşık yapılarda sonuç modelini anlamak ve yorumlamak zor olabilir.
  - ✓ DVM algoritma yapısının etkisini hiper parametre özelliğinden dolayı görselleştirmek oldukça zordur.

2.2.6 LASSO

Lasso, en az mutlak büzülme ve seçim operatörü anlamına gelir. İstatistikte ve makine öğrenmesinde değişken seçimini gerçekleştiren bir regresyon yöntemidir. Modelin tahmin doğruluğunu ve yorumlanabilirliği artırmak için kullanılır. Maliyet fonksiyonuna ceza terimi ekleyen yapısı vardır. Bu terim, katsayıların mutlak toplamıdır. Katsayıların değeri 0'dan yükseldikçe, bu terim cezalandırır, modelin katsayıların değerini düşürmesine neden olur, böylece bilgi kaybını azaltır.

- **Avantajları**

  ✓ Yanıt değişkeni ile ilişkili olmayan değişkenleri ortadan kaldırarak modelin yorumlanabilirliğini artırmaya yardımcı olur.

  ✓ Aşırı uyumu önler.

  ✓ Güçlü bir tahminleme etkisi vardır.

  ✓ Ceza faktörünü seçmek için çapraz doğrulama yaparak, modelin gelecekteki veri örneklerine iyi bir şekilde genelleştirilmesini sağlamaya yardımcı olur.

- **Dezavantajları**

  ✓ Veriyi normalleştirdikten ve normalleştirmeden önce karşılaştırdığımızda Lasso, tutarlı sonuç vermeyebilir.

  ✓ İkili korelasyonların çok yüksek olduğu bir grup değişken varsa, LASSO keyfi olarak gruptan yalnızca bir değişken seçme eğilimindedir.

3 ÇOK AŞAMALI LOJİSTİK MODELLEME

İstatistiksel yöntemler uygulanacak veri türüne göre şekillenmektedirler. Kredi verileri, genellikle bireylerin, zamanın etkisinde yorumlanabilirliği oldukça yaygın olan araştırma konularıdır. Çok aşamalı problemler, son zamanlarda model içerisindeki teorinin ve etkilerinin detaylı açıklanabilirliğinden dolayı sıkça kullanılan yöntemlerden biridir. Örnek olarak, eğer bir bireyin üzerinde açıklanabilir etkiler varsa, bu etkileri özelliklere bağlı olarak belli prosedürlerle belirlenmelidir. Çok aşamalı modellerde analizlerin detaylı ilerlemesi ve açık olması, değişkenlerin kendi içerisindeki uygunluğunun belirlenebilmesi açısındanda önemlidir. Farklı aşamalardaki değişkenlerin sayısı çok olduğu zaman, muhtemel aşamalar arası etkileşim oldukça fazla olacaktır. Sonuç olarak çok aşamalı bir düzende hangi değişkenlerin hangi seviyeye ait olması gerektiğini ve hangi doğrudan etkilerin ve seviyelerarası etkileşimin beklenebileceğini belirtmelidir.

Hiyerarşik yapılarda gözlemler belli bir sırayla birbirlerini izleyen seviyeler içerisine yerleşiktirler. Aynı seviye içerisindeki birimlerin korelasyonu ile birbirinden farklı seviyelerde bulunan birimlerin korelasyonu için hesaplama teknikleri farklı olacaktır. Özellikle farklı seviyelerdeki birimleri değerlendirmede bu veriler için bağımsızlık varsayımını gerektiren durumlarda en küçük kareler (EKK) gibi yöntemleri kullanmak uygun olmayacaktır (Osborne, 2000). Bu sebeplerden dolayı çok aşamalı modellerde parametre tahmininde en yaygın olarak en çok olabilirlik tahmin metodu kullanılmaktadır. Bu methodun en önemli özelliği varsayımlardan sapmalardan etkilenmemesi ve asimtotik olarak etkin ve tutarlı tahminleme yapmasıdır. Aynı zamanda amaç, iki farklı regresyon denklemi sabit paremetrelerini, varyans ve kovaryans bileşenlerinin kombinasyonu olan olabilirliği maksimum yapmaktır.

Çok aşamalı analizlerdeki amaç, belli bir hiyerarşik yapıda ve yüksek doğruluk, güvenilirlik ile model oluşturmaktır. Bu sebepten dolayı son dönemlerde sıklıkla

kullanılan modelleme yöntemlerinden olmuştur (Ciarleglio ve Makuch, 2007). Bu modellemelerin en önemli kısmı verideki değişken etkilerinin doğru bir şekilde açıklanabilmesidir. Bu sebeple MLM, farklı aşamalardan geçerek oluşturulmuş yapılardaki değişkenlerin, sahip oldukları etkileri belirlemek amaçlı bir istatistik modelin analiz sonuçlarını yorumlamaya çalışmaktadır. Lindley ve Smith tarafından, doğrusal modeller için Bayes tahmin edicileri belirlemek için 1972 yılında MLM ilk kez ortaya konmuştur (Lindley ve Smith, 1972). Ardından hiyerarşik veri yapıları için MLM üzerine uygulamalar yapılmıştır (Dempster ve diğ., 1977; Longford, 1995). Aitkin ve diğ. (1981) MLM'nin sosyal bilimler alanındaki ilk çalışmalarını yapmışlardır (Aitkin ve diğ., 1981). Mason ve diğ. (1983) MLM'nin yatay veriler için uygulanabilirliğini göstermişlerdir (Mason ve diğ., 1983).

- **Avantajları**

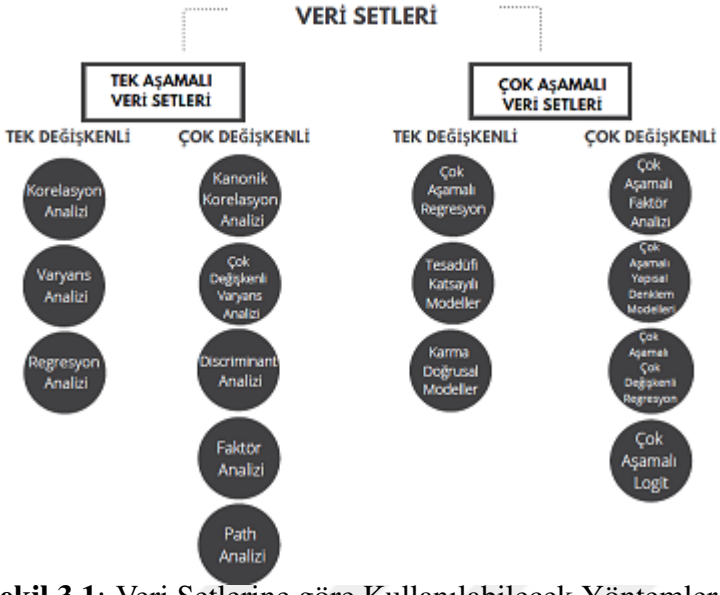
  - ✓ Hiyerarşik yapıların içiçe olma özelliğinden dolayı doğrusal regresyon modelleri uygulandığında hata terimleri arasındaki korelasyon EKK ile çözümlenemez, hiyerarşik yapıdaki veriler için MLM oldukça uygundur (Moerbeek ve diğ., 2003).
  - ✓ Aşama aşama oluşacak yapı için, her yeni aşama da eski aşamanın hatasını minimize ederek kontrollü bir şekilde ilerlemektedir.
  - ✓ Daha az parametreye ihtiyaç duyar.
  - ✓ Parametre sayısını indirgeme özelliği yapıda karmaşıklığın giderilmesi için oldukça önemlidir.
  - ✓ Grup içi veya grup dışı modelleme özelliğinden dolayı yapılar içerisinde ki ölçümler kontrol edilebilip tek düzeyli modellere göre daha kapsamlı yorumlar yapılabilmektedir.

- **Dezavantajları**

  - ✓ Karmaşık bir yapıda olması.
  - ✓ Büyük örneklemlerle çalışılması.
  - ✓ Lineer olmayan problemler MLM ile çözülemezler.

Kullanılacak veri tiplerinin tek aşama veya birden fazla aşamadan meydana gelmesine göre kullanılacak alternatif yöntemleri aşağıdaki şekildeki gibi özetlenebilir (Tezergıl ve Bülbül, 2018):

MLM'de hiyerarşik bir yapıya sahip olup aynı zamanda bağımlı değişkeni ikili olan veri tiplerine uygulanmaktadır. MLM, parametrelerin bazılarına veya tamamına sabit etki yerine rastgele etki gibi davranılarak en basit haldeki lojistik regresyon modelinin genişletilmesidir. MLM mevcut aşama sayısına göre iki aşamalı, üç aşamalı, ..., k aşamalı lojistik model adını almaktadır. MLM ile bağımsız değişkenlerin model üzerindeki etkilerinin yanı sıra bağımlı değişken için de tahminleme yapılmaktadır. Geleneksel doğrusal modellerde bir bağımsız değişkenin bağımlı değişken üzerindeki etkisini yorumlamaya yardımcı beta katsayıları bulunurken, MLM için bu katsayılardan güvenilir oranda yardım alabilmek adına bağımlı değişkeni bir



Şekil 3.1: Veri Setlerine göre Kullanılabilecek Yöntemler

olasılığa dönüştürmek için üssel dönüşüm uygulanmaktadır. Üssel dönüşüm bu tarz hesaplamalarda sıklıkla kullanılan matematiksel bir yapıdır. Ardından bu dönüşüm odds oranları kullanılarak yorumlanır. MLM'de özellikle ikili bağımsız değişkeni olan veriler için çok önemli olan odds oranları, başarı olasılığınınn başarısızlık olasılığına oranı olarak bilinmektedir.

MLM'nin örneklemesi, bağlantı fonksiyonu ve yapısal modeli aşağıdaki gibi tanımlanır (Tezergıl ve Bülbül, 2018):

Örnekleme :

$$Y_{ij}/\varphi_{ij} \sim B(m_{ij}, \varphi_{ij}) \tag{3.1}$$

$$E(Y_{ij}/\varphi_{ij}) = m_{ij}\varphi_{ij} \tag{3.2}$$

$$E(Y_{ij}/\varphi_{ij}) = m_{ij}\varphi_{ij}(1-\varphi_{ij}) \tag{3.3}$$

Buradaki;

i: bireyleri,

j: grupları,

m_{ij} : j. gruptaki i. birey için deneme sayısı,

φ_{ij} : j. gruptaki i. birey için başarı olasılığını,

Y_{ij} : Kredi başvurusunda bulunan j. gruptaki i.birey için kredinin "iyi kredi = 1" veya

"kötü kredi = 0" olarak sonuçlanmasını,

$E(Y_{ij}/\varphi_{ij})$: m_{ij} ve φ_{ij} verildiğinde Y_{ij}' nin beklenen değerini ifade etmektedir.

Bernoulli durumunda $m_{ij} = 1$ olduğunda Y_{ij} sıfır veya bir değerini alır ki bu da Bernoulli dağılımı olarak bilinen Binom dağılımının özel bir halidir.

Bağlantı Fonksiyonu :

$$\eta_{ij} = log\left(\frac{\varphi_{ij}}{1-\varphi_{ij}}\right) \tag{3.4}$$

Buradaki;

η_{ij} : başarı oddsunun logaritmasını ifade etmektedir.

Yapısal Model :

$$\eta_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + ... + \beta_{pj}X_{pij} \tag{3.5}$$

Buradaki;

j. gruptaki i. birey için $Y_{ij} = 1$ olasılığı φ_{ij} olarak tanımlanır. Ayrıca $\varphi_{ij}/[1-\varphi_{ij}]$ olarak belirlenirken Z_{pj} ise modelin 2. aşamasındaki bağımsız değişkenin j. grubunun skorudur.

X_{pij} : Modelin 1. aşamasındaki p. bağımsız değişkenin j. gruptaki i. kişi için değeri iken $Y_{ij} = 1$ ' in odds olasılığı η_{ij} kullanarak aşağıdaki şekilde hesaplanır. Ayrıca burada η_{ij}' nin her değeri için φ_{ij}, 0 ile 1 arasında değer alır.

$$\varphi_{ij} = \left(\frac{exp(\eta_{ij})}{1+exp(-\eta_{ij})}\right) \tag{3.6}$$

2. aşama modeli hiyerarşik modellemelerdeki 2. aşama modelleriyle benzerdir.

$$\beta_{qj} = \gamma_{qo} + \sum_{s=1}^{sq} \gamma_{qs}W_{sj} + u_{qj} \tag{3.7}$$

Buradaki;

$beta_{qj}$: j. bireyin q. aşama tesadüfî katsayısını,

$gamma_{qo}$: q. aşamadaki her sabit etkiyi,

W_{sj} : $beta_{qj}$ üzerindeki açıklayıcı değişkeni,

u_{qj} : "0" ortalamalı T varyans-kovaryans matrisli çok değişkenli normal dağılıma sahip, tesadüfî etkiler anlamına gelmektedir.

3. aşama modelinin yapısı da yine 2. aşama modelleriyle benzerdir. Çoklu aşamalar bu şekilde devam etmektedir.

MLM' de oluşacak son modelin, aşamaları, süreci ve adımları tamamen araştırmacıya bağlıdır.

MLM seçimi için takip edilmesi gereken yol aşağıdaki gibidir:

- **Koşulsuz Model**

  Amaç:

  İlk olarak, birinci aşama veya 2. aşama açıklayıcı değişkeni olmayan koşulsuz model tahmin edilerek bağımlı değişken için 2. aşama birimleri arasındaki değişkenliğin araştırılması.

  Model:

  $\eta_{ij} = \beta_{0j}$= başarı için log - odds oranı

  $u_{0j} \sim N(0, \tau_{00})$

  $$\beta_{0j} = \gamma_{00} + u_{0j} \tag{3.8}$$

  Buradaki;

  γ_{00} : başarı için log - odds oranının ortalamasıdır.

  u_{0j} : tesadüfî etkidir.

  Yorum:

  Bağımlı değişkenin ikili olması durumunda η_{ij} ile başarı olasılığı arasında doğrusal olmayan ilişki, 2. aşama birimi başarı olasılığı ve anakütledeki başarı olasılığı arasında farklılık olması ile sonuçlanır.

- **Koşullu Model**

  Amaç:

  1. aşamada model kurabilmek için "başarının" bağımsız değişkenleri değerine bağlı olan yapısal bir model kurulur. 1. aşama modelindeki sabit terim ve eğim

katsayıları 2. aşama modellerinin sonuç değişkeni oluşturulması.

Model:

✓ 1. aşama

$$\eta_{ij} = \beta_{0j} + \beta_{1j}X_{1j} + ... + \beta_{pj}X_{pij} \tag{3.9}$$

✓ 2. aşama

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j} \\ &... \\ \beta_{pj} &= \gamma_{p0} + u_{pj} \end{aligned} \tag{3.10}$$

Buradaki;

β_{pj} : Eğim katsayılarıdır.

γ_{p0} : Gruplar için sabit terimdir.

u_{pj} : Sabit terim için hatadır.

Yorum:

1. aşama koşullu modelinin çoklu iterasyonları, u_{pj}' yi varyansı çok büyük olmayan özgün etki herhangi bir 2. aşama denkleminden çıkarılarak 1. aşama koşullu modelleri test edilebilir. 2. aşama denklemindeki özgün veya tesadüfî etkinin varlığı 1. aşama değişkeninin veya sabit teriminin etkisinin ele alınan 2. aşama değişkenlerine göre değiştiğini ifade etmesiyle sonuçlanır.

MLM için tahminleme adımları;

1. hiyerarşideki aşamaların sayısı,
2. her bir aşamadaki açıklayıcı değişkenler,
3. her bir aşama için kullanılacak olasılık dağılımları,
4. sonuç değişkenlerin beklenen değeriyle açıklayıcı değişkenleri ilişkilendiren en uygun bağlantı fonksiyonu,

belirlemektir.

MLM'de 1. aşama da normal dağılımlı olmayan veri ve doğrusal olmayan bağlantı fonksiyonu ele alınmakta ve daha ileriki aşamalarda tesadüfî etkilerin çok değişkenli normal dağıldığı varsayılmaktadır (Raudenbush ve Bryk, 1992) . Ayrıca 1. ve 2. aşama hatalarının korelasyonsuz olduğuda varsayılmaktadır (Sullivan ve diğ., 2004).

Uygulanacak algoritmaların en optimal, yüksek doğruluk performansına en yakın ve en güvenilir sonuç vermesi beklenmektedir. Bir değişken bazen sabit, bazen rastgele etki, bazen her iki etki altında kalabilir. Genel anlamda sabit etki için bağımlı değişken üzerinde bir etkiye sahip olması beklenir. Bunlar doğrusal regresyonda açıklayıcı değişkenlere de karşılık gelmektedir. Rastgele etki ise kontrol etmeye calıstığımız gruplarla ilgilidir. Sürekli değişkenler üzerinden rastgele bir etki aranmaz bu yüzden rastgele etki genellikle kategorik değişken üzerinden değerlendirilmektedir. Durum, temsili olarak örneklemeye dayalı sonuçların tüm popülasyona genellenmesinden ibarettir. Her iki etkinin görüldüğü modeller verinin detaylanmasına elverişlidir. Tüm verinin kullanımına ve veriler arasındaki korelasyonların hesaba katılmasına izin verir. Ayrıca, daha az parametrenin tahmin edilmesini sağlayıp, regresyonları kullanırken karşılaşacağımız çoklu karşılaştırmalardaki sorunları önlemeye elverişlidir.

3.1 HAVUZLAMA

Veri analistleri, havuzlama verisi adı verdikleri veri türüyle de ilgilenmişlerdir. Havuzlanmış veri, bir zaman serisi verisi içerisindeki bir örneklemde herhangi bir değişkeni tanımlayan veri tabanı anlamına gelmektedir. Havuzlama verisi olarak kullanılan ilk çalışmalar aşağıdaki gibidir:

- Bir pazarlama verisi, belirli marka ürünlerin zaman içindeki satışını tanımlaması (Backwith, 1972),
- Bir tıbbî araştırma verisi, bir dizi hasta için periyodik zaman aralıklarıyla ilaç dozajı ve daha sonra kan şekeri düzeyi ölçümleri için bir veri tabanı geliştirilmesi (Sheiner ve diğ., 1972),
- Arabanın boyutuna göre zaman içinde otomobillere olan talebin değerlendirilmesi (Carlson, 1978),

- Bir çam ağacı dikimi için verim oluşturmak adına, zaman içindeki ayrı parsellerdeki hacmin ölçülmesi (Ferguson ve Leech, 1978),
- Bir talep fonksiyonu geliştirmek amaçlı, farklı şehirlerde zaman içinde benzine olan talebin ölçülmesi (Mehta ve diğ., 1978),
- Hanehalkı tarafından saatlik elektrik talebinin belirlenmesi (Granger ve diğ., 1979),
- Bir finansal piyasa veri tabanının, bir menkul kıymet değerleri için aylık getiri oranlarını içermesi (Dielman, 1979)

Havuzlanmış veri tabanında yer alan bilginin potansiyel değerini anlamak için, analistin aşağıdaki işlevleri kapsayan istatistiksel bir metodolojiye ihtiyacı vardır:

1. Veri tabanında tanımlanan tek bir bireyin performansının incelenmesi ve bu performansın uygun açıklayıcı değişkenlerle ilişkilendirilmesi.
2. Örneklemlerdeki bireysel ilişkilerden elde edilen özet ve bu özet istatistiklerden de genelleştirilmiş çıkarımların elde edilmesi.

En basit haliyle iki seviyeli bir hiyerarşik modelde basit bir havuzlama yapısından bahsedilirse, tüm 2. aşama birimlerinden gelen verileri havuzlar ve normal bir dağılımla tanımlanabileceğini varsayar. Sonuç olarak, parametrelerdeki tahmini varyans, bu normal dağılımın parametrelerini tahmin etmeye indirgenir.

Daha önceden de bahsettiğimiz gibi tek aşamalı model tahmini tüm değişkenlerin varlığıyla karmaşık yapıda olabilir. Havuzlanmış bir model tahmininde bu karmaşıklık aynı zamanda oluşturulacak hiyerarşik yapıyla en aza indirgenmeye çalışılmaktadır. Havuzlanmış tahminleme stratejisi kullananılan çalışmalarda, her aşama için bir dizi kontrol mekanizması mevcuttur. Havuzlanmış modellemenin kontrol mekanizması araştırmacının seyrinde olduğu unutulmamalıdır. Oluşturulan havuzlanmış modellerden verideki önemli özelliklerin tahminleriyle yüksek güvenilirlikle eşleşen tahminler elde etmek mümkündür.

4 UYGULAMA

MLM'ler hiyerarşik yapının kullanılabileceği tüm veriler için uygun bir kullanım yapısına sahiptir. Önerilen yaklaşımın uygulaması için finans sektöründe çok önemli yeri olan kredi skorlama verilerine daha önceki yıllarda uygulanandan farklı bir yaklaşım uygulanmıştır. Uygulamada, 8 yıllık kredi skorlama verilerine YZT ile en etkin değişken seçimi yol haritası izlenmiş, etkin değişkenlerin havuzlanmış MLM ile veri üzerinde yorumlama yapılmaya çalışılmıştır. Kredi skorlama için öngörülen etkin modeldeki, skorlamada etkili faktörler, havuzlanmış MLM ile zamana bağlı olarak tekrar modellenip açıklanmıştır. Kredi kullanan bireylerin, kredi kullanırken etkileyen faktörlere dair zamanın etkisiyle oluşan hiyerarşik yapı, araştırmaya dahil edilmiştir. Önerilen yaklaşım için Kaggle'dan alınan "İrlanda Dummy Banka"sı (Kaggle, 2018) verisi kullanılmıştır. Veri, 2008-2015 yılları arasında bankadan kredi kullanmış 886776 bireysel müşterinin bilgilerini içermektedir. Belirtilen yıllar içerisinde farklı sayıda bireysel müşterinin ay bazında kullandırılan kredilerin *kabul edilen (iyi)* veya *kabul edilmeyen (kötü)* krediler olarak sonuçları veride mevcuttur. Veriler oluşturulurken müşteri tabanlı alınan bilgiler doğrultusunda, kredi verilen *ay*, *yıl*, kredi alınan *şehir*, *ev sahipliği durumu*, *gelir seviyesi*, bankanın müşteriye verdiği *kredi notu*, *kredi kullanma sebebi*, kredinin *faiz tipi*, *yıl olarak meslek sahipliği*, *ay olarak kredi geri ödeme süresi*, kredi *faiz oranı*, *toplam geri ödeme*, *kredi miktarı*, *gelir*, *aylık kredi ödeme miktarı* verinin değişkenleri olarak kullanılmıştır.

Uygulamada aşağıdaki 4 ana başlık takip edilmiştir:

1. Veriyi hazırlama ve veri hakkında özet bilgi süreci
2. Sınıflandırma ve değişken seçim süreci
3. Model oluşturma süreci
4. MLM ile veriyi açıklama süreci

Bu süreçler içerisinde analizlerin değerlendirilmesi için sonuçlar R İstatistiksel programlama dili (R-Studio, 2019), (versiyon 4.0.3) ile elde edildi.

4.1 VERİYİ HAZIRLAMA ve VERİ HAKKINDA ÖZET BİLGİ SÜRECİ

Verimizde bağımlı değişken, kredi müşterilerinin banka tarafından kullandıkları kredi sonucunda kredilerinin *iyi* olarak değerlendirilmesi *1*, *kötü* olarak değerlendirilmesi *0* olarak kodlanmıştır. Bağımsız kategorik değişkenler; kredi verilen *ay*, *yıl*, kredi alınan *şehir*, *ev sahipliği durumu*, *gelir seviyesi*, bankanın müşteriye verdiği *kredi notu*, *kredi kullanma sebebi*, kredinin *faiz tipi*, *yıl olarak meslek sahipliği*, *ay olarak kredi geri ödeme süresi* olarak belirlenirken bağımsız sürekli değişkenler kredi *faiz oranı*, *toplam geri ödeme*, *kredi miktarı*, *gelir*, *aylık kredi ödeme miktarı* olarak belirlenmiştir.

İlk aşama olarak veri hazırlama sürecinde verimizi tanımamız adına bireysel kredi müşterilerine ait bilgiler aşağıda sunulmuştur. Bu bilgilerin görsel grafik dağılımları yine Ekler bölümünde verilmiştir.

886776 kredi başvuru müşterilerinin kredi durum dağılımlarında 819505 başvurunun *iyi = 1* kredi, 67271 başvurunun *kötü = 0* kredi olarak değerlendirildiği görülmüştür. Yüzde olarak %92.41'i *iyi=1* sonuçlanan krediye sahip bireyleri %7.59'u *kötü = 0* sonuçlanan krediye sahip bireyleri göstermektedir (7.1).

886776 kredi başvuru müşterilerinin *2008 - 2015 yılları* arasında kredi kullanan müşteri dağılımları (7.2)'de gösterilmiştir. Aynı zamanda (7.3)'te *2008 Ocak* ayından itibaren *2015 Aralık* ayına kadar tüm müşterilerin dağılımını görmek mümkündür. Yine (7.4)'te "yıllara göre ay bazında" müşterileri dağılımlarından görüleceği gibi yıllar ve aylar geçtikçe kredi kullanımının arttığını söyleyebiliriz.

886776 kredi başvuru müşterilerinin krediyi kullandıkları *şehir*lere göre dağılımları; 214441 (%24.18) kredi başvuru müşterisi "Leinster", 208677 (%23.53) kredi başvuru müşterisi "Ulster", 204156 (%23.02) kredi başvuru müşterisi "Nothern Iri", 154951 (%17.47) kredi başvuru müşterisi "Cannught", 104551 (%11.79) kredi başvuru müşterisi "Munster" eyaletinden başvurmuştur (7.5).

886776 kredi başvuru müşterilerinin *ev sahipliği durumu*na göre dağılımları; 443394 (%50) kredi başvuru müşterisi "ipotek", 355779 (%40.12) kredi başvuru müşterisi "kira", 87418 (%9.86) kredi başvuru müşterisi "ev sahibi", 185 (%0.02) kredi başvuru müşterisi "diğer" şeklindedir (7.6).

886776 kredi başvuru müşterilerinin *gelir seviye*leri durumuna göre dağılımları; 729093 (%82.22) kredi başvurusu müşterisinin "düşük", 140917 (%15.89) kredi başvuru müşterisinin "orta", 16766 kredi başvurusu müşterisinin "diğer" şeklindedir (7.7).

886776 kredi başvuru müşterilerinin banka tarafından tanımlanan *kredi not*ları durumuna göre dağılımları (7.8)'de gösterilmiştir.

886776 kredi başvuru müşterilerinin *krediyi kullanma sebepleri*ne göre dağılımları; 206095 (%23.24) kredi müşterisi krediye "kredi kartı" ödeme, 8850 (%1) kredi müşterisi krediye "araba" alma, 10321 (%1.16)kredi müşterisi krediye "küçük işyeri" açma, 42781 (%4.82) kredi müşterisi krediye "düğün" masrafları, 2338 (%0.26) kredi müşterisi krediye "kişisel borç" ödeyebilme, 524008 (%59.09) kredi müşterisi rediye "ev tadilatı" yaptırabilme, 517793 (%5.84) kredi müşterisi krediye "satınalım" gücü, 17267 (%1.95) kredi müşterisi krediye "sağlık" giderleri, 8531 (%0.96) kredi müşterisi krediye "taşınma" masrafları, 5399 (%0.61) kredi müşterisi krediye "tatil" harcamaları, 4732 (%0.53) kredi müşterisi krediye "ev" alımı, 3699 (%0.42) kredi müşterisi krediye "yenilenebilir enerji", 575 (%0.06) kredi müşterisi krediye "eğitim" giderleri, 387 (%0.04) kredi müşterisi krediye "diğer" masraflar amacıyla başvurmuştur (7.9)

886776 kredi başvuru müşterilerinin, 464900 (%52.43) kişisi krediyi "düşük", 421876 (%47.57) kişisi krediyi "yüksek" faizle kullanmışlardır (7.10).

886776 kredi başvuru müşterilerinin *kaç yıllık çalışan* olduklarının durumuna göre dağılımları (7.11)'de gösterilmiştir.

886776 kredi başvuru müşterilerinin; 620522 (%69.98) kişisi krediyi "30" ay, 266254 (%30.02) kişisi krediyi "60" ay *geri ödeme süresi* olarak seçmişlerdir (7.12).

Kredi değerlendirmesinde, bir başvuru sahibinin niteliklerini temsil eden sayısal değerler önemli ölçüde farklı yapılara sahiptir ve kullanılacak model gereksinimlerinde normalleştirme süreci için tüm veri setine basit bir normalleştirme işlemi uygulanması, bazı yararlı bilgilerin kaybına neden olabilir. Bağımsız değişken yapısındaki farklılıklar sebebiyle (7.13) veri hazırlama sürecinde her değişken ayrı ayrı değerlendirilip farklı normalleştirme süreci uygulanmıştır. Bu süreçte kullanılan veri dönüşümü, orjinal verilerin bir veya daha fazla doğrusal regresyon varyasımını ihlal etmesi durumunda verileri doğrusal regresyonla modellemeye uygun hale getirmek için iyileştirici bir yöntem olarak kullanılmıştır. Bu yöntem için standartlaştırma, logaritma, signed logaritma vb. algoritmalar gerektiğinde iç içe gerektiğinde tekrarlı

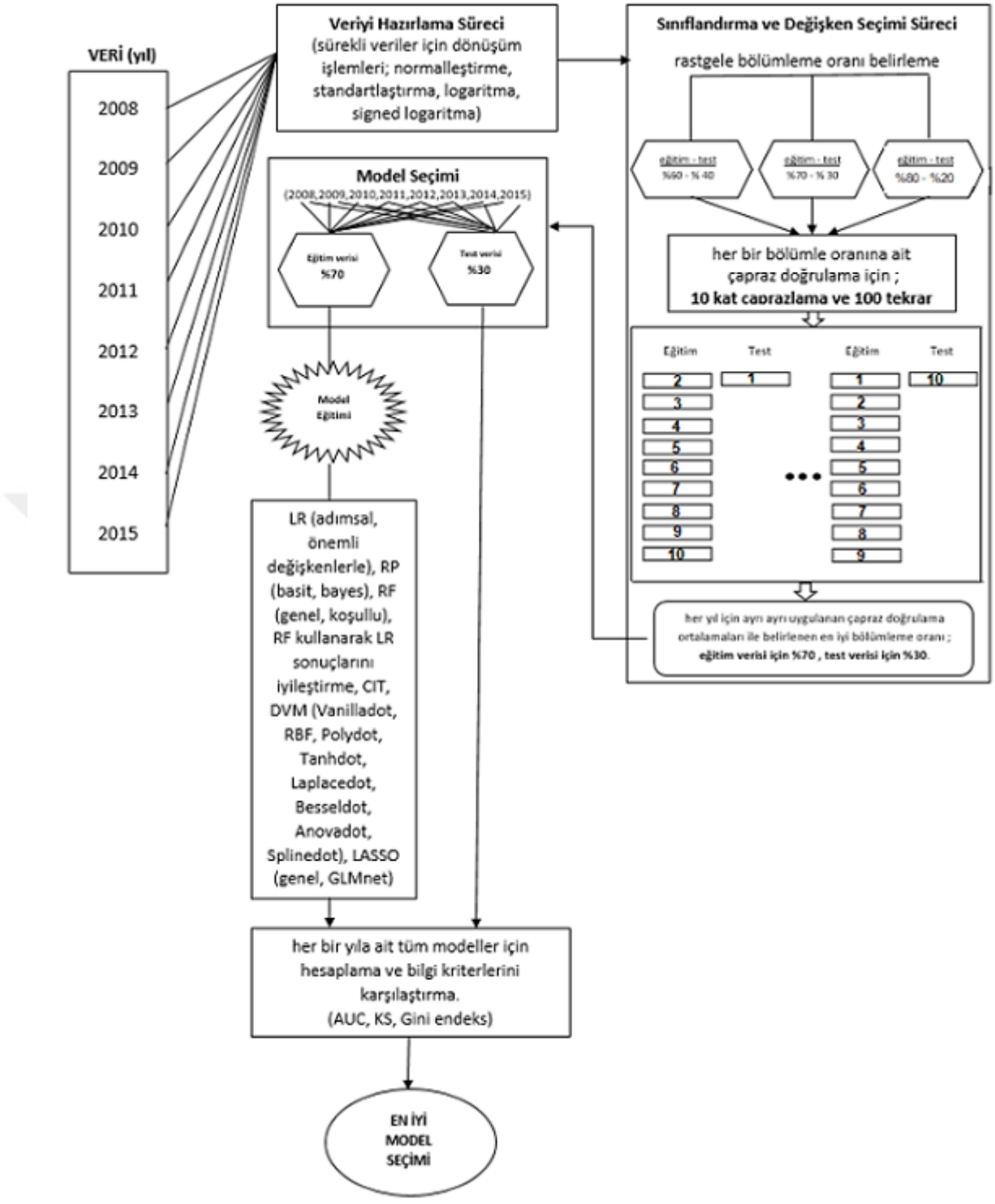
olarak değişkenlere uygulanarak normal bir dağılıma sahip olmaları sağlanmıştır (7.14).

4.2 SINIFLANDIRMA ve DEĞİŞKEN SEÇİM SÜRECİ

Etkin bir model elde etmek için verilerin önceden işlenmesi oldukça önemlidir. Veriyi hazırlama, işleme ve temizleme gibi işlemler yapıldıktan sonra veriyi düzenleme kısmında anlamsız değişkenlerin kullanılmaması, bazı değişkenlerin finansal anlamda dönüştürülmesi ve verinin görselleştirilmesi kurulacak model için bir önhazırlık aşaması olup, yüksek güvenilirlik açısından da bir kontrol mekanizması olarak değerlendirilebilir.

Doğrusal olmayan sistemlerin optimizasyonu ile elde edilen tahminlerin zamanla daha etkili sonuçlar sağlayacak model performansları sağladıkları bilinmektedir. Doğrusal olmayan bir yapı ile modellemek için, literatürde "ikili sınıflandırma modelleri" ve "çok sınıflı sınıflandırma modelleri" olmak üzere iki tür sınıflandırma modeli vardır.

Dengesiz bir sınıflandırma problemi, örneklerin bilinen sınıflar arasında dağılımının önyargılı veya çarpık olduğu bir sınıflandırma problemine bir örnektir. Sınıflandırma için kullanılan algoritmalarının çoğu, her sınıf için eşit sayıda örnek varsayımı etrafında tasarlandığından, dengesiz sınıflandırmalar tahmine dayalı modelleme için bir zorluk teşkil etmektedir. Sınıf dağılımındaki dengesizlik problemlere göre değişkenlik gösterebilir, ancak ciddi bir dengesizliğin modellenmesi daha zordur ve özel teknikler gerektirebilir. Hem pasif hem de aktif öğrenme yöntemleri için, bir performans hedefine ulaşmak için gereken açıklayıcı örneklem büyüklüğünü tahmin etmeye ihtiyaç vardır.



Şekil 4.1: Akış Diyagramı

Şekil 4.1'deki süreçte; optimum parametreler, her algoritma için bazı sınıflandırma doğruluğu ölçüleriyle belirlendi. Sınıflandırma algoritmaları için, kullanılan veri setinin özellik matrisi normalleştirildikten sonra bölünme oranlarını seçildi. Eğitim ve test olarak iki gruba bölünen veri setini oluşturmak, bir algoritmanın performansını problem üzerinde hızlı bir şekilde değerlendirmek için kullanılan bir

yöntemdir. Eğitim veri kümesi, modeli hazırlamak ve onu eğitmek için kullanılır. Eğitimli modelden gelen tahminlerle, test veri setindeki model için bir performans hesaplamamıza olanak tanır. Uygulamada sırasıyla %60-%40, %70-%30 ve %80-%20 bölünmeleri denendi. Değerlendirme çapraz doğrulama ile yapıldı. Çapraz doğrulama, bir yöntemin görünmeyen veriler üzerindeki becerisini tahmin etmenin başka bir yöntemidir. Çapraz doğrulama, veri kümesinin birden çok alt kümesinde birden çok modeli sistematik olarak oluşturur ve değerlendirir. Bu da bir dizi performans ölçümü sağlar. Prosedürün ortalama olarak ne kadar iyi performans gösterdiğine yanıt bulmak için ölçümlerin ortalaması hesaplandı. Bu durum aynı zamanda kullanılacak algoritmayı ve veri hazırlama prosedürleri seçiminde, bir performansın diğeriyle karşılaştırmasını yapmaya olanak tanır. Uygulama için çapraz doğrulama kullanarak, istatistiksel bir modeli doğrulamak için eğitim ve test setlerini belirlendi. Çapraz doğrulamada, veriye 10 kat çaprazlama ve 100 tekrar yapılarak en yüksek doğrulama oranı elde edildi. Her turda doğrulama için test verisini ayırdıktan sonra kalan veriyi eğitim için kullanıldı. Sınıflandırıcı eğitimden sonra, çapraz doğrulamanın doğruluğunu elde etmek için 100 tekrar üzerinden ortalaması alındı. Hem eğitim-test bölümlemeleri hem de k-kat çapraz doğrulama için yeniden örnekleme yöntemleri olarak adlandırılabilir. Yeniden örnekleme yöntemleri, bir veri kümesini örneklemek ve bilinmeyen bir miktarı tahmin etmek için istatistiksel prosedürlerdir. Çapraz doğrulama yaklaşımı ile veriye 10 kat çaprazlama ve 100 tekrar yapılarak değerlendirilen ve bu bölünme oranlarının karşılaştırılmasında en iyi sonuçları veren eğitim - test bölünme oranının %70-%30 olduğu tespit edildi (9.1).

Kredi skorlama bilgilerini sağlamak için istatistiksel yöntemler, parametrik olmayan yöntemler ve YZT gibi yaklaşımlar önerilmiştir. Bu bilgiler doğrultusunda, sınıflandırmadaki bölünme oranı tespit edildikten sonra uygulama içi modelleme analizinde LR (adımsal, önemli değişkenlerle), RP (basit, bayes), RF (genel, koşullu), RF kullanarak LR sonuçlarını iyileştirme, CIT, DVM (Vanilladot, RBF, Polydot, Tanhdot, Laplacedot, Besseldot, Anovadot, Splinedot), LASSO (genel, GLMnet) gibi yöntemler kullanılmıştır.

Öte yandan, karmaşıklık önerilen yaklaşım için önemli bir konudur. Karmaşıklık bilgisi, en iyi istatistiksel modeli seçmek için modellemede bir karar ağacı oluşturmaya yardımcı olur. Bu nedenle önerilen yaklaşımda AUC, KS ve Gini endeksi verilerdeki tüm yıllar boyunca doğru ve güvenilir bir modelleme çerçevesi

olarak değerlendirilmiştir.

Sınıflandırma metotlarının uygulandığı tüm algoritmalar için AUC (8.1), KS ve Gini endeks değerleri gösterilmiştir. Çizelge 4.1.'de 2008 - 2015 yılları için kullanılan tüm sınıflandırma algoritmaları arasından, DVM (RBF) yönteminin 2013 yılı verilerinin 3 ayrı ölçme yöntemi olarak kullandığımız sırasıyla AUC: %98, KS: %90.17 ve Gini: %96 değerlerine göre en iyi performansa sahip olduğu açıktır.

Çizelge 4.1: Yıllara göre Model Performanslarının Karşılaştırılması

Yıl	Model	n	AUC	KS	Gini
2008	Lasso (GLMnet)	2393	97.76	89.49	95.52
2009	DVM (RBF)	5884	97.54	89.08	95.08
2010	DVM (RBF)	12537	97.27	87.22	94.54
2011	DVM (RBF)	21721	97.49	88.44	94.98
2012	DVM (RBF)	53367	97.55	89.04	95.01
2013	**DVM (RBF)**	**134755**	**98**	**90.17**	**96**
2014	DVM (RBF)	235628	97.67	87.55	95.34
2015	DVM (RBF)	420491	97.45	88.61	94.89

Kullanılan sınıflandırma algoritmalarına göre verinin açıklanabileceği en iyi değişken seçimlerini de görmek mümkündür. Çizelge 4.2.'de verinin tanımlayıcı özellikleri yıllara göre ayrı ayrı verilmiştir. Yine bu çizelgeye göre yıllar için belirlenen özelliklerin iki tanesinin ortak olduğu görülmektedir. Birincisi, toplam ödeme tutarını temsil eden *toplam geri ödeme* değişkeni ve ikincisi, aylık taksit tutarını temsil eden *aylık ödeme miktarı* değişkenidir.

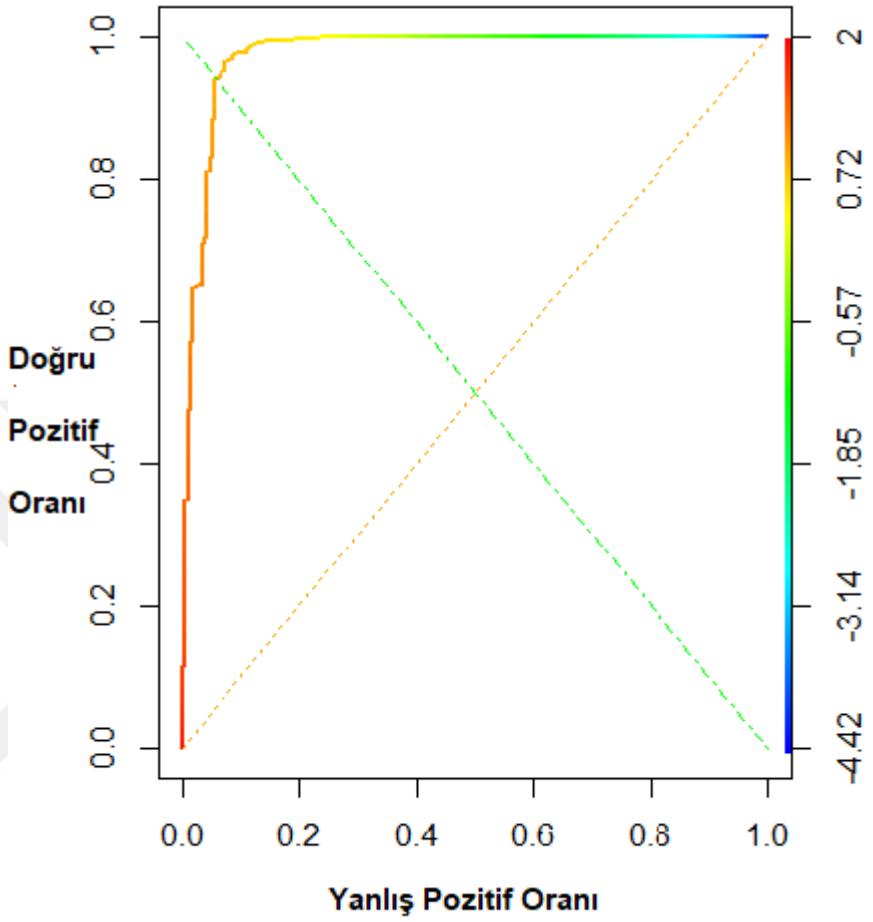
]

Çizelge 4.2: En İyi Model Seçimi için Yıllara göre Belirlenen Önemli Değişkenler

2008	2009	2010	2011
toplam geri ödeme	gelir	yıl	yıl
aylık ödeme miktarı	kredi miktarı	ay	ay
şehir	**toplam geri ödeme**	meslek sahipliği (yıl)	meslek sahipliği (yıl)
	aylık ödeme miktarı	gelir	gelir
		kredi miktarı	kredi miktarı
		geri ödeme süresi(ay)	geri ödeme süresi(ay)
		kredi kullanma sebebi	kredi kullanma sebebi
		faiz oranı	faiz oranı
		toplam geri ödeme	**toplam geri ödeme**
		aylık ödeme miktarı	**aylık ödeme miktarı**

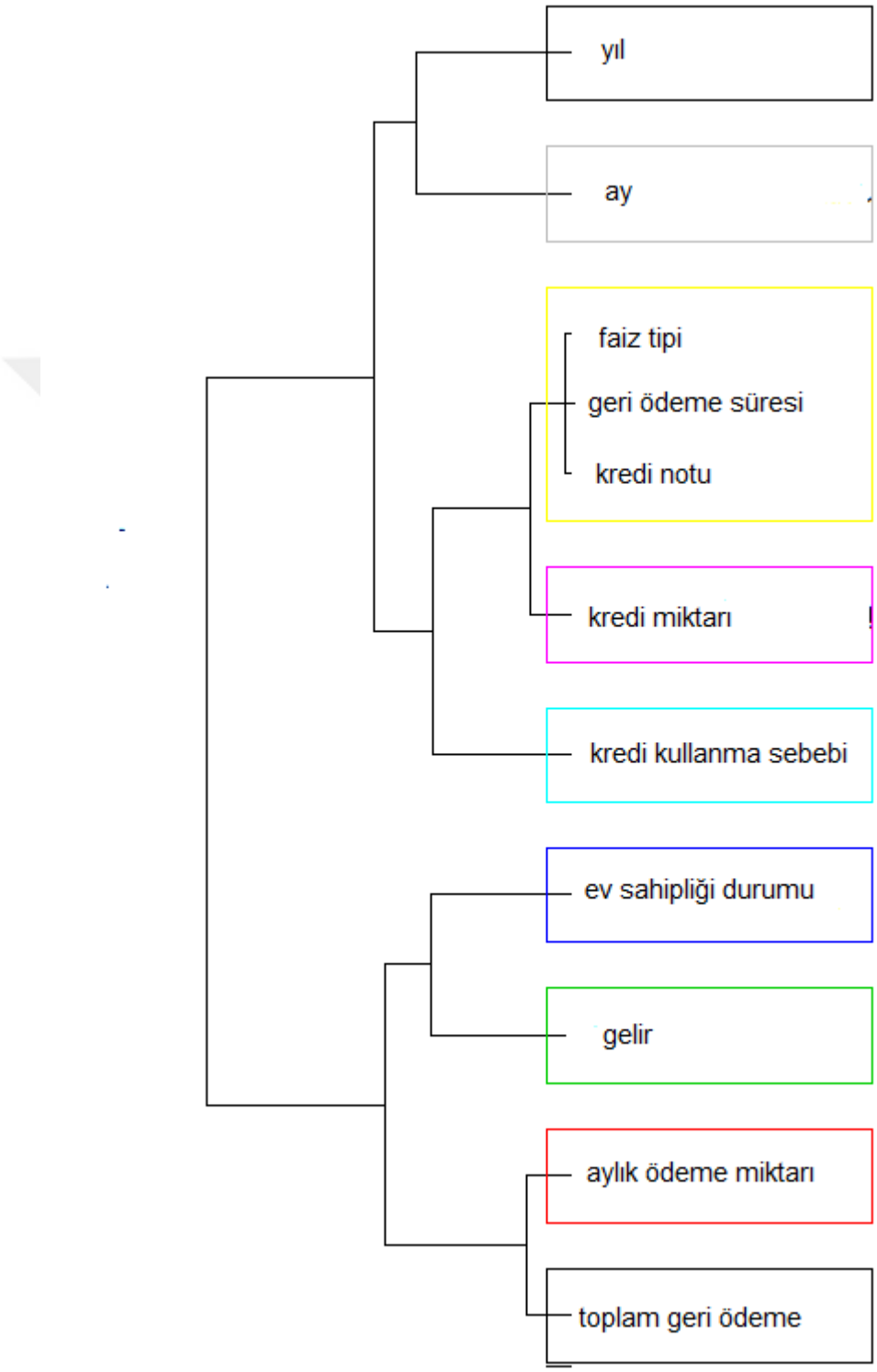
2012	2013	2014	2015
yıl	yıl	yıl	yıl
ay	ay	ay	ay
meslek sahipliği (yıl)	ev sahipliği durumu	meslek sahipliği (yıl)	ev sahipliği durumu
gelir	gelir	ev sahipliği	gelir
kredi miktarı	kredi miktarı	gelir	kredi miktarı
kredi kullanma sebebi	geri ödeme süresi (ay)	kredi miktarı	geri ödeme süresi (ay)
faiz oranı	kredi kullanma sebebi	geri ödeme süresi (ay)	kredi kullanma sebebi
faiz tipi	faiz tipi	faiz oranı	faiz oranı
toplam geri ödeme	kredi notu	şehir	faiz tipi
aylık ödeme miktarı	**toplam geri ödeme**	**toplam geri ödeme**	şehir
	aylık ödeme miktarı	**aylık ödeme miktarı**	**toplam geri ödeme**
			aylık ödeme miktarı

Çizelge 4.1.'de DVM (RBF) yönteminin 2013 yılında diğer kullanılan sınıflandırma algoritmalarına karşı kullanılan ölçme yöntemlerine göre en iyi performansa sahip olması, Şekil.4.2'de de 2013 yılına ait DVM (RBF) ROC eğrisi ile gösterilmiştir.



Şekil 4.2: 2013 Yılı DVM (RBF) ROC Eğrisi

Özellik değişkenleri hakkında yorum yapabilmek amaçlı yine 2013 yılı verilerine yapılan kümeleme ile Şekil 4.3.'de benzer özelliklere sahip değişkenlerin bir araya getirdiği gruplar belirlenmiştir.



Şekil 4.3: 2013 Yılı Özellik Gruplama Diyagramı

Literatür olarak verilen çalışmalarda kullanılan ***farklı veriler*** üzerindeki sınıflandırma performanslarının, önerilen yöntemdeki veriye göre en iyi bulunan sınıflandırma performansı ile karşılaştırılmıştır. Önerilen sınıflandırma yönteminin daha yüksek performans gösterdiği açıkca görülmektedir.

Çizelge 4.3: Mevcut Yaklaşımların Farklı Veriler Üzerindeki Performansları

Yazar(lar)	Veri	Model	Doğruluk (%)
Alaraj ve diğ., 2016	Avustralya kredi verisi	Lojistik Regresyon	92.96
		Destek Vektör Makineleri	89.44
Abdou ve diğ., 2018	Mısır bankası kredi verisi	Lojistik Regresyon	88.3
Liu and Pan., 2018	Avustralya kredi verisi	Lojistik Regresyon	85.7
	Almanya kredi verisi		72.4
Yufei ve Liu, 2017	Avustralya kredi verisi	Lojistik Regresyon	86.77
		Karar Ağaçları	84.51
		Rastgele Orman	87.41
		Destek Vektör Makineleri	85.54
Abellan ve Castellano, 2017	Avustralya kredi verisi	Destek Vektör Makineleri	91.86
İlter, 2021	**İrlanda Dummy Bankası kredi verisi**	**Destek Vektör Makineleri**	**98**

Kaggle tarafından elde edilen İrlanda Dummy Bankası kredi verisinin literatür olarak çalışmalarda kullanılan sınıflandırma performanslarının, önerilen yöntemdeki sınıflandırma modellerine göre en iyi bulunan sınıflandırma performansı ile karşılaştırılmıştır. Önerilen sınıflandırma yönteminin ***aynı veriye*** ait performans sonuçlarının da yine daha yüksek performans gösterdiği açıkca görülmektedir.

Çizelge 4.4: Mevcut Yaklaşımların İrlanda Dummuy Bankası Kredi Veri Seti Üzerindeki Performansları

Yazar(lar)	Model	Doğruluk (%)
Pete Jourgensen, 2019	Rastgele Orman	92
Seung Won Kim, 2020	Yarı Denetimli Öğrenme	87
	Rastgele Orman	82
İlter, 2021	**Destek Vektör Makineleri**	**98**

Siteye Son Giriş Tarihi : 09.02.2021

4.3 MODEL OLUŞTURMA SÜRECİ

MLM'de ilk aşama, koşulsuz model oluşturma aşamasıdır. Uygulama için söz konusu koşulsuz model Bölüm 4.2.'den elde ettiğimiz en yüksek güvenilirlikli değişkenlerle oluşturulacak modeldir. Veride var olan 16 değişkenden, uygulanan YZT sonrası toplam 12 değişkenle verinin açıklandığı Çizelge 4.2.'de gösterilmiştir. Burada önemli olan analize devam edilip edilemeyeceğinin karar verileceği aşama olan ilk aşamadır. Eğer, her aşama için geçerli olan tüm değişkenler arasında anlamlı farklılık mevcut değilse buna çok aşamalı modelleme yapılması diğer bir deyişle hiyerarşik bir yapı oluşturmak mümkün değildir. Bu sebeple analize de devam edilemez. Veriye göre oluşturulmak istenen koşulsuz model için kredi değerlendirilmesinde, kredi müşterilerinden alınan bilgilerin müşteriler arası anlamlı farklılık gösterip göstermediği araştırılmaktadır. Buna göre hipotezi;

H_0= Kredi değerlendirilmesinde, kredi müşterileri arasında anlamlı farklılık yoktur.
H_a=Kredi değerlendirilmesinde, kredi müşterileri arasında anlamlı farklar vardır.

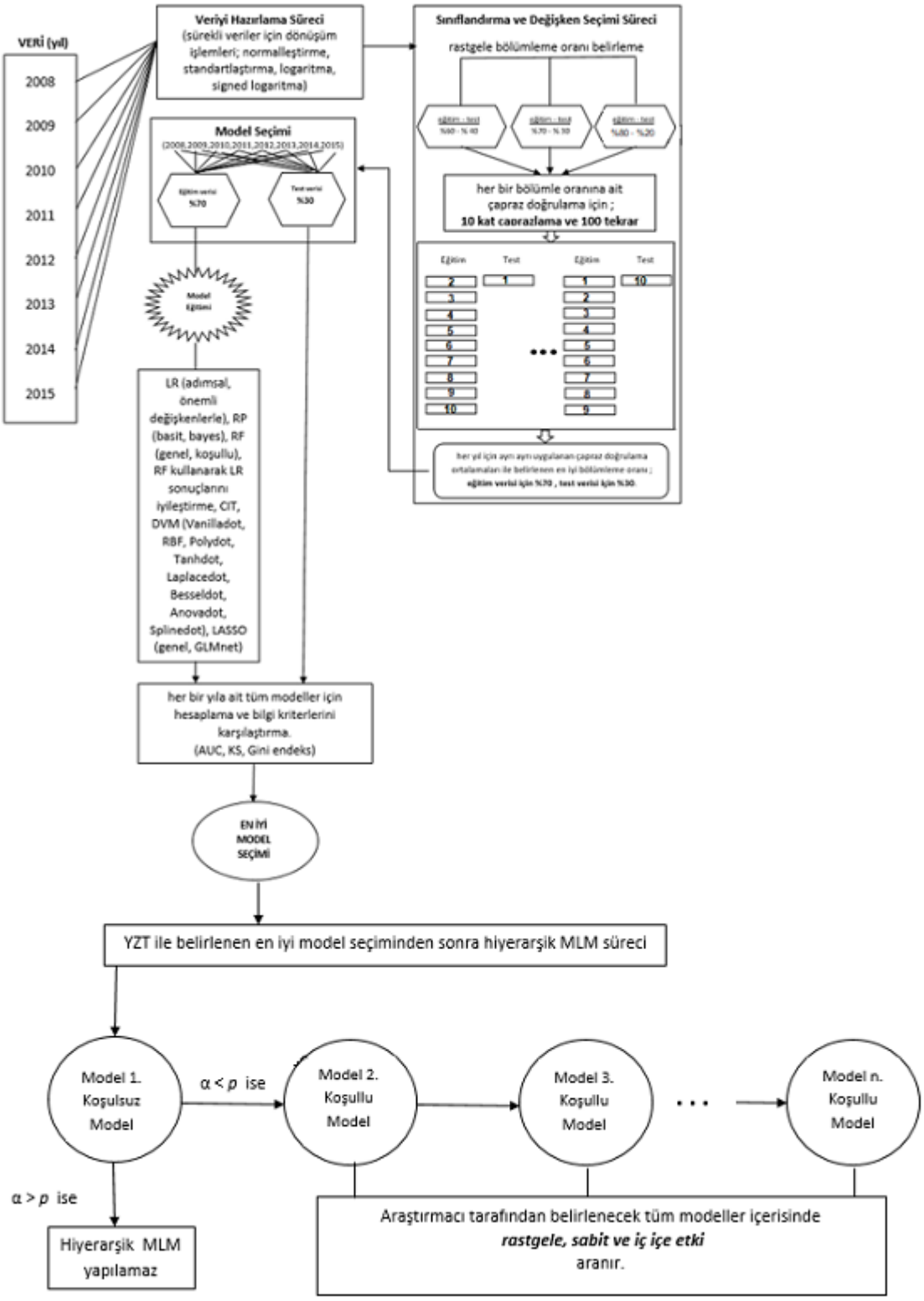
şeklinde kurulur.

Model analizine başlamadan önce MLM için geçerli olan, aykırı değer, çoklu bağlantı, doğrusallık, normallik, homojenlik varsayımları geriye kalan 12 değişken üzerinden kontrol edilmiş ve gerekli düzenlemeler ile sorunlar giderildikten sonra model oluşumuna başlanmıştır.

Belirlenecek hiyerarşik yapı finansal kurumlarda tamamen finans yetkilisi uzmanlığında belirlenmektedir. Değişkenlerin kullanımı yüksek güvenilirlikli fayda faktöründe oldukça önemlidir. MLM yapıları, aynı anda tüm ilişkileri araştıran gruplanmış verilerin hiyerarşik seviyelerin içinde ve arasında analizler yaparak, değişkenler için mevcut analizlerden farklı olarak daha verimli hale getirmektedir.

Bu uygulamada soru yapıları; "Hangi banka hizmetinden, nasıl bir müşteri, hangi zaman aralığında kullandığı kredi sonucu *iyi* olarak değerlendirilebilir?" veya "düşük faizli bir banka kredisi kullanan az gelirli bir kredi müşterisinin kredi sonucunun *iyi* olarak onaylanma durumu nedir?" yönünde cevap aranılarak değerlendirilmiştir. Literatürde bu tanımlama havuzlamaya karşılık gelmektedir. İç içe geçmiş modeller havuzlanmış model biçimleriyle değerlendirilirler. Tüm bu değerlendirilmelerde

rastgele , sabit ve karışık etkiler yol göstericidir. Modeller için etkilerin karşılaştırmalarında kullanılan kriterler AIC ve BIC'dir. MLM sürecinde parametre, varyans ve kovaryans bileşenlerinin kombinasyonları için en çok olabilirlik tahmin yöntemi uygulanmıştır. Verinin bağımlı değişkeni ikili bağımlı değişken olduğundan kredi durumu her seviye için çıktı olarak kullanılmıştır.



Şekil 4.4: Hibrit Yaklaşım için Hiyerarşik MLM Akış Diyagramı

Koşulsuz modelde;

1. Aşama : $\eta_{ij} = \beta_{0j} + r_{ij}$

2. Aşama : $\beta_{0j} = \gamma_{00} + u_{0j}$; $u_{0j} \sim N(0, \gamma_{00})$

Buradaki;

η_{ij} : j. kredi müşterisine ait i. etkinin doğrusal bileşenidir.

r_{ij} : 0 ortalamalı ve sabit σ^2 varyanslı normal dağılıma sahip modelin hata terimidir.

β_{0j} : log-odds değeridir. γ_{00} : j tane kredi müşterisinin *iyi* sonuçlanan kredi ortalamalarının ortalamasıdır. u_{0j} : j. kredi müşterisine ait 0 ortalamalı τ_{00} varyanslı tesadüfî etkiyi ifade etmektedir.

Verimize model parametrelerini uyarladığımızda;

1. Aşama:

$$Y_i = \beta_0 + \beta_1(yil) + \beta_2(ay) + \beta_3(evsahipligi) + \beta_4(gelir) + \beta_5(kredimiktari) + \beta_6(geriodeme(ay)) + \beta_7(kredikullanmasebebi) + \beta_8(faiztipi) + \beta_9(kredinotu) + \beta_{10}(toplamgeriodeme) + \beta_{11}(aylikgeriodeme) \tag{4.1}$$

2.Aşama

$$\begin{aligned} \beta_0 &= \gamma_{00} + u_0 \\ \beta_1 &= \gamma_{10} + u_1 \\ \beta_2 &= \gamma_{20} + u_2 \\ \beta_3 &= \gamma_{30} + u_3 \\ \beta_4 &= \gamma_{40} + u_4 \\ \beta_5 &= \gamma_{50} + u_5 \\ \beta_6 &= \gamma_{60} + u_6 \\ \beta_7 &= \gamma_{70} + u_7 \\ \beta_8 &= \gamma_{80} + u_8 \\ \beta_9 &= \gamma_{90} + u_9 \\ \beta_{10} &= \gamma_{100} + u_{10} \\ \beta_{11} &= \gamma_{110} + u_{11} \end{aligned} \tag{4.2}$$

Buradaki;

γ_{00} : genel ortalamadır.

Krediye başvuran müşterinin kredi durumunun "iyi" olmasındaki;

γ_{10} : yıl,

γ_{20} : ay,

γ_{30} : ev sahipliği,

γ_{40} : gelir,

γ_{50} : kredi miktarı,

γ_{60} : geri ödeme (ay),

γ_{70} : kredi kullanma sebebi,

γ_{80} : faiz tipi,

γ_{90} : kredi notu,

γ_{100} : toplam geri ödeme,

γ_{110} : aylık geri ödeme etkisini ifade etmektedir.

$u_0, u_1, ..., u_{11}$ ise tesadüf etkiyi ifade etmektedir.

MODEL 1.

Çizelge 4.5: Model 1 Sonuçları

Model	**Tahmin**	**Standart hata**	**p değeri**
Genelleştirilmiş Doğrusal Model	Sabit	0.0040	< 0.0001

Fisher Skorlama iterasyon sayısı : 5

Çizelge 4.3'e göre tüm verinin kredi müşterilerinin kredi durumlarına göre odds oranı 0.08 çıkmıştır. Bu orana karşılık gelen olasılık değeri de 1/(1+0.08) = 0.93 olarak hesaplanır. Bu orana modelin anlamlılık oranı da denebilir. Çizelge 4.3.'teki p olasılığı $H_0 : \tau_\beta = 0$ hipotezinin red edilmesi demektir. Yani, kredi müşterileri arasında kredinin *iyi* olarak sonuçlanması açısından anlamlı farklılıklar söz konusudur. Başka bir açıdan yapının hiyerarşik olduğu ve bu sonuçlara göre analize devam edilmesi uygundur.

Müşteriler arası bu farklılığı yaratan etkileri açıklayabilmek adına bu adımdan sonra koşullu modellerin kurulumuna geçilmiştir.

Koşullu model için bir diğer aşamada, rastgele etkinin etkisi altındaki model sonuçlarıdır. Bu aşamada rastgele etki olarak banka bünyesindeki değişkenler olan *faiz tipi, kredi notu, geri ödeme süresi* gibi değişkenler bu aşamanın açıklayıcı değişkenleri olarak atanmıştır. Ayrı ayrı veya birlikte modele rastgele etki olarak dahil ederek kredi başvurusu *iyi* sonuçlanan kredi müşterilerinin en çok hangi banka değişkeninden rastgele olarak ne kadar etkilendiğini söylemek mümkün olacaktır.

MODEL 2.

Çizelge 4.4.'e göre 2. modelimize sırasıyla *kredi notu*, *faiz tipi*, *geri ödeme süresi (ay)* tek rastgele değişken olarak, ardından sırasıyla *kredi notu, faiz tipi*, *kredi notu,*

Çizelge 4.6: Model 2 Sonuçları

Rastgele Etki	Tahmin	Standart hata	p değeri	AIC	BIC
kredi notu	sabit	0.2999	< 0.0001	456880	456903.4
faiz tipi	sabit	0.3986	< 0.0001	458314.3	4583337.7
geri ödeme süresi (ay)	sabit	0.1111	< 0.0001	474919.6	474943
kredi notu faiz tipi	sabit	0.4	< 0.0001	453841	453876.5
kredi notu geri ödeme süresi (ay)	sabit	0.3494	< 0.0001	456205.9	456241
geri ödeme süresi (ay) faiz tipi	sabit	0.4155	< 0.0001	458241.3	458276.4
kredi notu geri ödeme süresi (ay) faiz tipi	**sabit**	**0.4411**	**< 0.001**	**453111.9**	**453158.7**

geri ödeme süresi(ay), *geri ödeme süresi (ay), faiz tipi* iki rastgele değişken olarak ve nihayetinde *kredi notu, geri ödeme süresi (ay), faiz tipi* üç rastgele değişken olarak modele atanmıştır. Model analizinden sonra AIC ve BIC sonuçları karşılaştırıldığında %95 önem seviyesinde kredi başvurusunda bulunan müşterilerin kredi durumlarının *iyi* sonuçlanması *kredi notu, geri ödeme süresi (ay), faiz tipi* rastgele etkisi altındadır diyebiliriz.

MODEL 3.

Koşullu model için sonraki aşama sabit etki eklendiği model sonuçlarıdır. Bu aşamada sabit etki olarak zaman kavramını içeren değişkenler olan *yıl, ay* gibi değişkenleri bu aşamanın açıklayıcı değişkenleri olarak atanmıştır. Ayrı ayrı veya birlikte modele sabit etki olarak atanarak kredi başvurusu "iyi" sonuçlanan kredi müşterilerinin *yıl* veya *ay* değişkeninden sabit etki olarak ne kadar etkilendiğini söylemek mümkün olacaktır.

Çizelge 4.5.'e göre Model 3'e sırasıyla *yıl*, *ay* değişkenleri ayrı ayrı sabit değişken olarak modele atanmıştır. Model analizinden sonra AIC ve BIC sonuçları karşılaştırıldığında %95 önem seviyesinde kredi başvurusunda bulunan müşterilerin kredi durumlarının *iyi* sonuçlanması *ay* sabit etkisi altındadır diyebiliriz.

Aynı çizelgede *ay* sabit etkisi altında olan verimizin etki seviyelerini açıklanırken Ocak, Şubat, Mart, Ağustos ve Aralık aylarının sabit etkilerinin anlamlı olduğu söylenebilir.

MODEL 4.

Çizelge 4.7: Model 3 Sonuçları

Sabit Etki	Tahmin	Standart hata	p değeri	AIC	BIC
yıl2008	sabit	0.32261	0.0001*		
yıl2009	sabit	0.065548	0.0001*		
yıl2010	sabit	0.05898	0.0001*		
yıl2011	sabit	0.05637	0.0206*		
yıl2012	sabit	0.05437	0.0071*		
yıl2013	sabit	0.05358	0.0001*		
yıl2014	sabit	0.05333	0.0001*		
yıl2015	sabit	0.05357	0.0001*	425046.3	425046.3
Ocak	sabit	0.2661	0.0005*		
Şubat	sabit	0.1959	0.0164*		
Mart	sabit	0.1789	0.04304*		
Nisan	sabit	0.1972	0.2377		
Mayıs	sabit	0.2652	0.3329		
Haziran	sabit	0.2723	0.0643		
Temmuz	sabit	0.2468	0.2000		
Ağustos	sabit	0.3367	0.0011*		
Eylül	sabit	0.3769	0.0830		
Ekim	sabit	0.2766	0.0576		
Kasım	sabit	0.2235	0.0686		
Aralık	sabit	0.2193	0.0088*	**2391.6**	**2478.3**

Koşullu model için oluşturacağımız son model aşaması rastgele etki altında olan modele bir de sabit etki eklendiği model sonuçlarıdır. Buna karışık etki denilmektedir. Bu aşamada rastgele etki olan *kredi notu, geri ödeme süresi (ay), faiz tipi* değişkenleri ile sabit etkisi olan *ay* değişkeni bu aşamanın açıklayıcı değişkenleri olarak atanmıştır. İç içe etki modeli adı verilen modele kredi başvurusu *iyi* sonuçlanan kredi müşterilerinin belirlenen rastgele ve sabit etki değişkenlerinden ne kadar etkilendiğini söylemek mümkün olacaktır.

Çizelge 4.6.'ya göre Model 4'e sırasıyla rastgele etki olan *kredi notu, geri ödeme süresi(ay), faiz tipi* değişkenleri ile sabit etkisi olan *ay* değişkeninin bu aşamanın açıklayıcı değişkenleri olarak atanmıştır. Model analizinden sonra AIC ve BIC sonuçları karşılaştırıldığında %95 önem seviyesinde kredi başvurusunda bulunan müşterilerin kredi durumlarının *iyi* sonuçlanması iç içe etki modellemesinde rastgele etki *kredi notu, geri ödeme süresi (ay), faiz tipi* ile *ay* sabit etkisi altında modellendiğinde Ocak, Şubat, Ağustos ve Aralık aylarının sabit etkilerinin anlamlı olduğu söylenebilir.

Çizelge 4.8: Model 4 Sonuçları

Rastgele Etki	Sabit Etki	Tahmin	Standart hata	p değeri	AIC	BIC
	Ocak	sabit	0.2826	0.0015*		
	Şubat	sabit	0.2742	0.0429*		
	Mart	sabit	0.3229	0.2437		
	Nisan	sabit	0.2209	0.1652		
kredi notu	**Mayıs**	sabit	0.3104	0.2245		
geri ödeme süresi (ay)	**Haziran**	sabit	0.4081	0.1002		
faiz tipi	**Temmuz**	sabit	0.3544	0.2078		
	Ağustos	sabit	0.4286	0.0042*		
	Eylül	sabit	0.7002	0.1211		
	Ekim	sabit	0.3139	0.0755		
	Kasım	sabit	0.2413	0.0507		
	Aralık	sabit	0.2811	0.0297*	**2837.8**	**4259.7**

4.4 MLM ile VERİYİ AÇIKLAMA SÜRECİ

Yapılan çok aşamalı modelleme sonucunda, kredi çekmek isteyen kredi müşterilerinin aralarında anlamlı farklılıklar olduğu tespit edilmiştir. Kredi müşterilerinin kredi kullanmalarında etkili olan faktörler ve bu faktör etkilerinin araştırılmak istenen değişkenlere göre farklılık gösterip göstermediği koşullu modeller oluşturularak araştırılmıştır. Oluşturulan ilk modelde hiyerarşik bir yapının oluşturulup , araştırma konusu senaryosu altında bir yapı oluşturulabilineceğine karar verilmiştir. İkinci modelde, kredi başvurusunda bulunan müşterilerin kredi durumlarının *iyi* sonuçlanması *kredi notu, geri ödeme süresi (ay), faiz tipi* rastgele etkisi olduğu tespit edilmiştir. Üçüncü modelde, yine kredi başvurusunda bulunan müşterilerin kredi durumlarının *iyi* sonuçlanması *ay* sabit etkisi altında olduğu ve özellikle bu modelde Ocak, Şubat, Mart, Ağustos ve Aralık aylarının sabit etkilerinin anlamlı olduğu gösterilmiştir. Dördüncü ve son modelde ise kredi başvurusunda bulunan müşterilerin kredi durumlarının *iyi* sonuçlanması iç içe etki modellemesinde rastgele etki *kredi notu, geri ödeme süresi (ay), faiz tipi* ile *ay* sabit etkisi altında modellendiğinde Ocak, Şubat, Ağustos ve Aralık aylarının sabit etkilerinin anlamlı olduğuna karar verilmiştir.

5 SONUÇLAR ve TARTIŞMA

Kredi skorlama, kredi değerinin tespiti amacıyla, bankalar veya diğer finansal kuruluşlar tarafından firmaların ahlakî ve malî durumlarını doğru olarak tespit etmek için yapılan değerlendirme ve sınıflandırma faaliyetidir. Kredi derecelendirmeleri yapan kuruluşlar, kredi risklerini iyi bilmeli ve yönetmelidirler. Belirsizlik ve korunmasızlık finansal risklerin temelini oluşturmaktadır. Bu çerçevede, herhangi bir finans kuruluşunun kredi kullandırdıkları müşterilerinin mali durumlarını incelemenin yanında, tüm finans sisteminden kullandıkları toplam kredi miktarını görmeleri ve bu kişilerle ilgili güncel bilgileri edinmeleri, derecelendirme kararları için yararlı olmaktadır. Finans kuruluşları, güvenilir ve açıkça tanımlanmış kredi derecelendirme kriterlerine uygun olarak çalışmalıdırlar. Bu kuruluşlar, derecelendirme hatalarını minimize edilmesine yönelik bağımsız bir sistem kurduklarında, kredi notu verme sürecindeki zararların oluşmasını öngörüp olası hataları engelleyecek önlemleri alabilirler. Kuruluşlar, uluslararası kredi faaliyetlerinde standart kredi risklerine ek olarak, kredi alan kişi veya kurumların ülkesine ait ekonomik koşullarla ilişkili riskleri de dikkate almaları gerekmektedir. Çünkü, kredi kurumlarının finansal başarısızlığı aynı zamanda sosyo-ekonomik sonuçlar açısından önem arz etmektedir. Ülke riski, bir ülkedeki kredi yükümlüleri ya da o ülkeye yapılan yatırımlar açısından önemli etkiler yaratabilecek şekilde ülkenin ekonomik, siyasî ve sosyal koşullarıyla bağlantılı tüm risklerini içermektedir.

Bu tarz sorunlara çözüm olabilmek adına, önerilen yaklaşımda kredi skorlamada YZT ile MLM temel alan hibrit bir yaklaşım elde edilmiş ve finansal açıdan veri detaylandırılmıştır. Burada geliştirilen modeller literatüre katkı sağlayacağı gibi, finans ve sosyal alanlara uygulama potansiyeli de bulunmaktadır. Literatür incelendiğinde, kredi skorlama da çoğunlukla YSA ve DVM alışılagelmiş kullanımlarıyla karşılaşılmaktadır (Chen ve diğ., 2009; Ionescu, 2018; Roy ve Urolagin, 2019). Önerilen yaklaşımla geliştirilen yaklaşımlar ile hem YZT'nin hem de MLM'nin hibritleşen model kestirimi

elde edilmeye çalışılmıştır. Elde edilen sonuç yapısıyla veri detaylı açıklanarak kredi veren finans kurumlarına yol gösterici nitelikte bir yaklaşım olması sağlanmıştır.

MLM için literatürde yapılan araştırmalarda kredi skorlama verileri üzerinde böyle bir uygulamaya rastlanmamıştır. Sigorta sektöründe yapılan çalışmaların yanı sıra, yapılan çalışmaların, güvenlik, sağlık, eğitim, öğrenci veya öğretmen başarısını ve performansını etkileyen faktörleri modellemede kullanıldıkları görülmüştür. Önerilen yaklaşımda vurgulanmak istenen, etkin değişken modellemeyle birlikte hiyerarşik bir havuzlanmış MLM ile ekonomi sektörüne daha detaylı ve yüksek güvenilirlikli bir çalışma bırakmaktır.

Bilgiyi farklı bakış açılarıyla ele almak veya bu bilgileri önemli verilere çevirmek uzun bir süreçtir. Oluşturulan MLM modellemeleriyle çalışma zaman alıcı görünse de önerilen yaklaşımda tüm risk oluşturabilecek detaylar dikkate alınıp, finansal açıklar üzerindeki problemlere yoğunlaşarak, finans kurumlarından sağlanan gerçek yaşam verileri üzerinde YZT modellerini temel alan bir hibrit sistemi ile çözüm getirilmeye çalışılmıştır. Önerilen yaklaşım, YZT yerine zaman serileri modelleri temel alınması sağlanarak özellikle finansal veriler için etkisi olan "zaman" kavramını detaylı karşılaştırabilmek amaçlı yine bir hibrit modellemeye için izleyen çalışmalara olanak sağlamaktadır. Önerilen yaklaşımda oluşturulan senaryo yerine, araştırmacı tarafından kurulacak başka hiyerarşik yapılar ve farklı değişkenler ile araştırılan etkilerin incelenmesi planlanmaktadır.

KAYNAKLAR

Abdou, H., Pointon, J., ve El-Masry, A. (2018). Neural nets versus conventional techniques in credit scoring in egyptian banking. *Science Direct*, 35:1275–1292.

Abellan, J. ve Castellano, J. G. A. (2017). A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Systems with Applications*, 73:1–10.

Aitkin, M., Anderson, D., ve Hinde, J. (1981). Statistical modelling of data on taeching styles(with discussion). *Journal of Royal Statistical Society, Series*, 144:419–461.

Alaraj, M. ve Abbod, M. F. (2016). Classifiers consensus system approach for credit scoring. knowledge-based systems. *Knowledge-Based Systems*, 104:89–105.

Backwith, N. (1972). Multivariate analysis of sales response of competing brands to advertising. *J. MArketing Res.*, (9):168–176.

Belotti, T. ve Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. pages 3302–3308.

Bhatia, S., Sharma, P., Burman, R., Hazari, S., ve Hande, R. (2017). Credit scoring using machine learning techniques. *Knowledge-Based Systems*, 161(11):975–8887.

Breiman, L. (2001). Random forest. *Machine Learning*, 45:5–32.

Bunker, R. P., Naeem, M. A., ve Zhang, W. (2017). Improving a credit scoring model by incorporating bank statement derived features. *Auckland University of Technology, Expert Systems with Applications*.

Carlson, R. (1978). Seemingly unrelated regression and the demand for automobiles and their use for forecasting in an energy crisis. *J. Business*, (51):243–262.

Chen, W., Ma, C., ve Ma, L. (2009). Mining customer credit using hybrid support vector machine technique. *Expert Systems with Applications*, 36(4):7611–7616.

Ciarleglio, M. ve Makuch, R. W. (2007). Hierarchical linear modeling: An overview. *Yale University School of Medicine, New Haven, CT, USA*, pages 91–92.

Dempster, A. P., Laird, N. M., ve Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Dielman, T. E. (1979). Pooled cross-sectional and time series data analysis. *Marcel Dekker,Inc.*

Durand, D. (1941). *Risk Elements in Consumer Instalment Financing*. National Bureu of Economic Research, New York.

Ferguson, I. ve Leech, J. (1978). Generalized least squares estimation of yield functions. *Forest Sci.*, (24):27–42.

Ghodselahi, A. ve Amirmadhi, A. (2011). Application of artificial intelligence techniques for credit risk evaluation. *International Journal of Modelling and Optimization*, 1(3):243–249.

Goh, R. Y. ve Lee, L. S. (2019). Credit scoring: A review on support vector machines and metaheuristic approaches. *Advances in Operations Research*, 30.

Granger, C., Engle, R., Ramathan, R., ve Andersen, A. (1979). Residental load curves and time -of-day pricing: An econometric analysis. *J. Econometrics*, (9):13–32.

Henley, W. E. ve Hand, D. J. (1996). A k-nn classifer for assessing consumer credit risk. *Statistician*, 65:77–95.

Hothorn, T., Hornik, K., ve Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674.

Ilter, D. ve Kocadağlı, O. (2019). Credit scoring by artificial neural networks based cross-entropy and fuzzy relations. *Sigma Journal of Engineering and Natural Sciences*.

Ionescu, M. S. (2018). Use of neural networks in the business process modeling.

Johnson, R. W. (1992). *Legal, social and economic issues implementing scoring in the U.S.*

Kaggle (2018). Ireland dummy bank, `https://www.kaggle.com/mrferozi/loan-data-for-dummy-bank`.

Keramati, A. ve Yousefi, N. (2011). A proposed classification of data mining techniques in credit scoring. *Proceeding of the 2011 International Conference on Industrial Engineering and Operations Management*, pages 22–224.

Khasman, A. (2010). Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. *Expert Systems with Applications*, 37(9):6233–6239.

Kuznetsov, S. O. (2020). Are you a good barrower? mining interpretable pattern structures in credit scoring. *Asian Journal of Economics and Banking*.

Lewis, E. M. (1992). An introduction to credit scoring. *Athena Press,San Rafael, CA*.

Lindley, D. V. ve Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, 34(1):1–41.

Longford, N. (1995). Random coefficient models. pages 519–570.

MacKey, D. J. C. (1992). Bayesian methods for adaptive models.

Mason, W. M., Wong, G., ve Entwisle, B. (1983). Contextual analysis through the multilevel linear model. *Sociological Methodology*, pages 72–103.

Mays, E. (1998). Credit risk modeling. *Glenlake Publishing, Chicago*.

Mehta, J. S., Narasimham, G. V. L., ve Swamy, P. A. V. B. (1978). Estimation of a dynamic demand function for gasoline with different schemes of parameter variation. *J. Econometrics*, (7):263–279.

Moerbeek, M., Breukelen, G. J. P., ve Berger, M. F. (2003). A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies. *Journal of Clinical Epidemiology*, 56:341–350.

Myers, J. H. ve Forgy, E. W. (1963). The development of numerical credit evaluation systems. *J. Amer Statist. Assoc.*, 58:799–806.

Neal, R. M. (1996). Bayesian learning for neural networks.

Olaniyan, R. ve Maheswan, M. (2017). Recent developments in resource management in cloud computing and large computing clusters.

Ong, C. S. (2005). "building credit scoring models using genetic programming. *Expert systems with Applications*, 29:41–47.

Oreski, S., Oreski, D., ve Oreski, G. (2012). Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. *Expert systems with Applications*, 39:12650–12617.

Osborne, J. (2000). Advantages of hierarchical linear modeling, practical assessment. *Research Evaluation*, 7(1).

R-Studio (2019). The r project for statistical computing, `https://cran.r-project.org/bin/windows/base/`. ®R package version 4.0.3.

Raudenbush, S. W. ve Bryk, A. S. (1992). Hierarchical linear models. *Newbury Park, CA: Sage*.

Rosenberg, E. ve Gleit, A. (1994). Quantative methods in credit management: A survey. *Oper.Res.*, 42:589–613.

Roy, A. G. ve Urolagin, S. (2019). Credit risk assesment using decision tree and support vector machine based data analytics. *Springer*.

Sheiner, L., Rosenberg, B., ve Melmon, K. (1972). Modeling of individual pharmacokinetics for computer-aided drug dosage. *Comput. Biomedical Res.*, (5):441–459.

Soui, E. (2019). Rule-based credit risk assessment model using multi-objective evolutionary algorithms. 126:144–157.

Sullivan, L. M., Dukes, K. A., ve Losina, E. (2004). Hierarchical modelling: An introduction to hierarchical linear modelling. *Tutorials in Biostatistics: Statistical Modelling of Complex Medical Data,Wiley Online Library*, 2(1).

Tezergıl, S. A. ve Bülbül, S. (2018). Çok aşamalı loıstık regresyon ve fınans sektörüne bır uygulama.

Thomas, L. C. (1998). *Methodologies for classifying applicants for credit*.

Thomas, L. C., Edelman, D. B., ve Crook, J. N. (2002). Credit scoring and its applications.

Wonderlic, E. F. (1952). An analysis of factors in granting credit. *Indiana Univ.Bull.*, 50:163–176.

Xiao, H., Xiao, Z., ve Wang, Y. (2016). Ensemble classification based on supervised clustering for credit scoring. *Applied Soft Computing*, 43:73–76.

Yufei, X. ve Liu, N. (2017). A boosted decision tree approach using bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, 78:225–241.

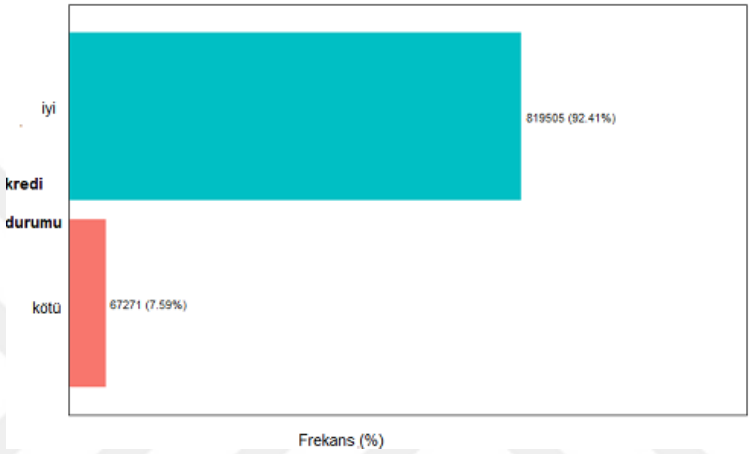
Zeng, J., Erzurumluoglu, M. A., Elsworth, L. B., Kemp, P. J., Haycock, C. P., Hemani, G., Tansey, K., ve Laurin, C. (2017). Ld hub: a centralized database and web interface to perform ld score regression that maximizes the potential of summary level gwas data for snp heritability and genetic correlation analysis. *Bioinformatics*, 33(2):272–279.

EKLER

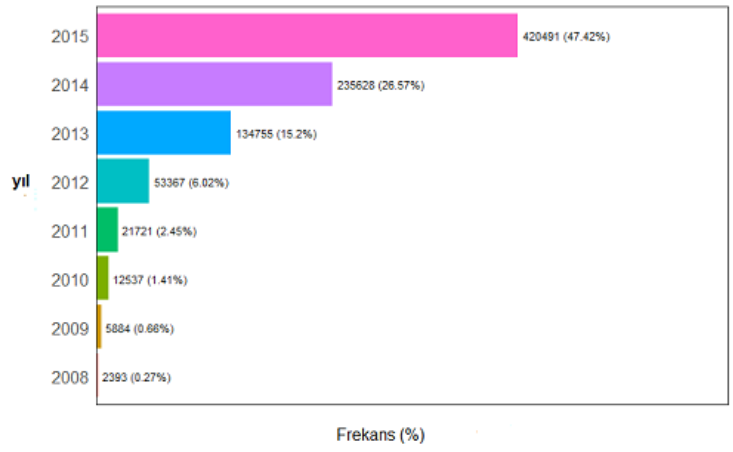
EK A : Veri Hakkında Özet Bilgiler
EK B : YZT Model Sınıflandırması
EK C : Çaprazlama Performans Sonuçları

7 Veri Hakkında Özet Bilgiler

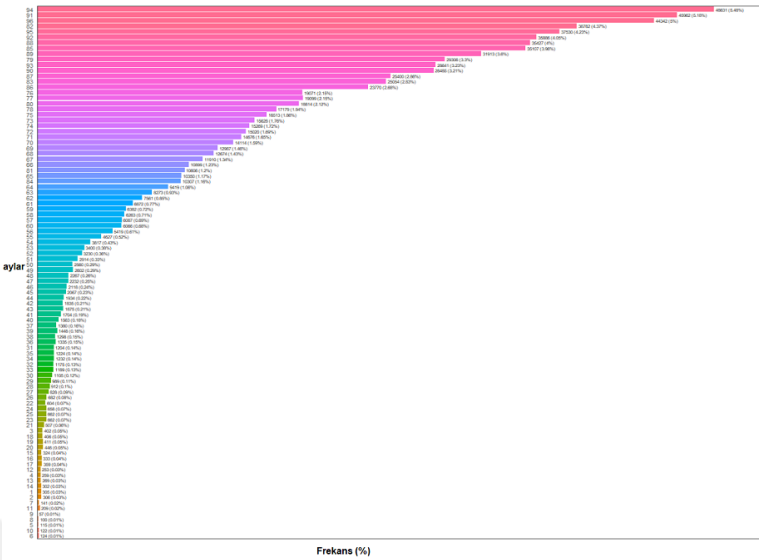
7.1 Kredilerin Kullanım Durum Dağılımı



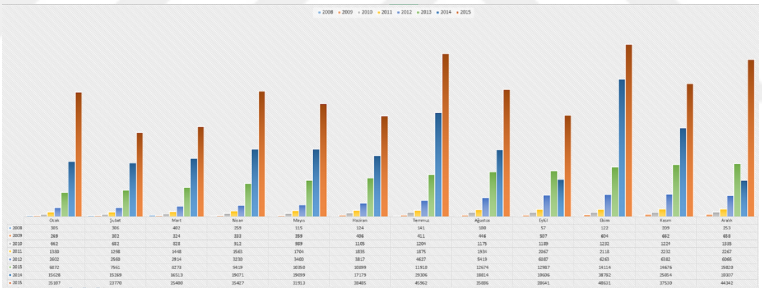
7.2 Kredilerin Kullanımının Yıllara Göre Dağılımı



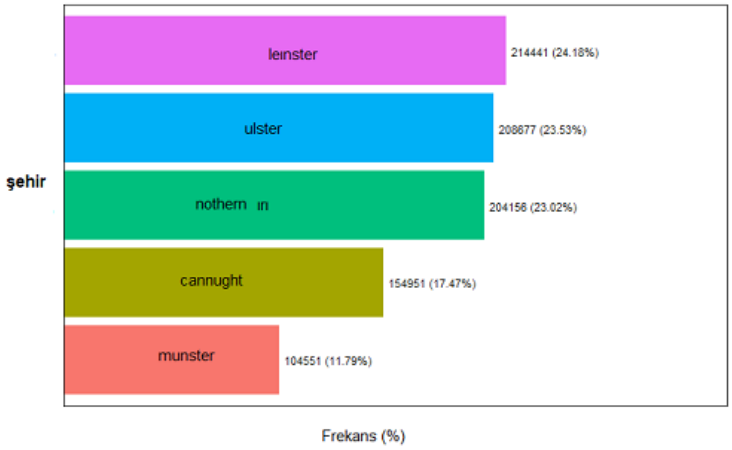
7.3 Kredilerin Kullanımının Aylara Göre Dağılımı



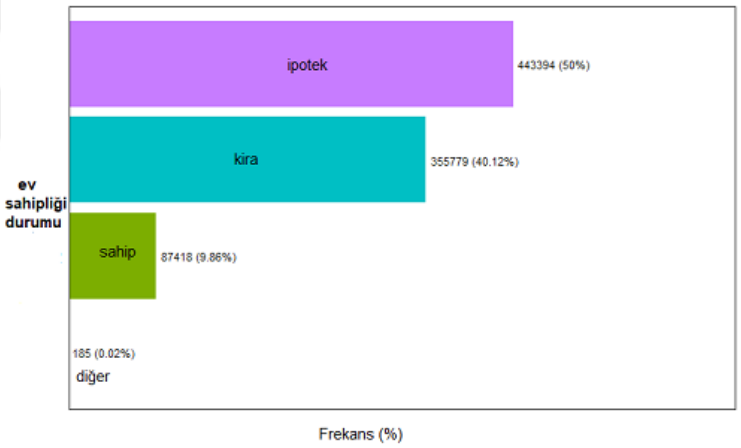
7.4 Kredilerin Kullanımının Yıl & Ay Göre Dağılımı



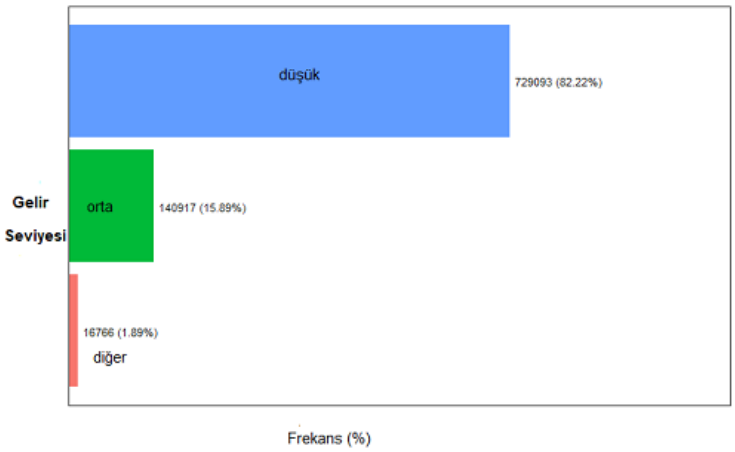
7.5 Kredilerin Kullanımının Şehir Göre Dağılımı



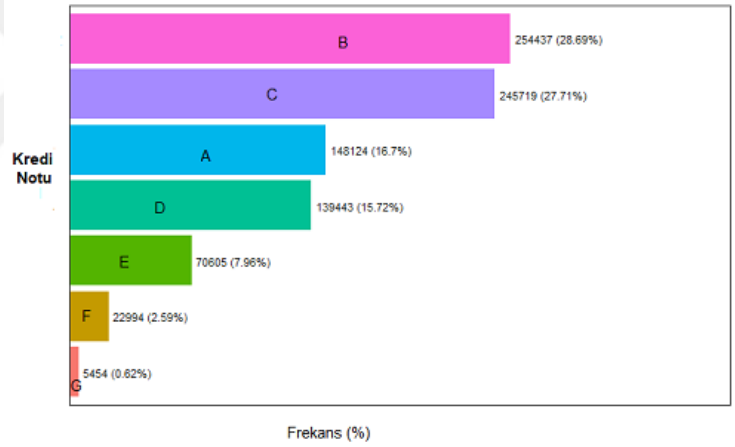
7.6 Kredi Müşterilerinin Ev Sahipliği Durum Dağılımı



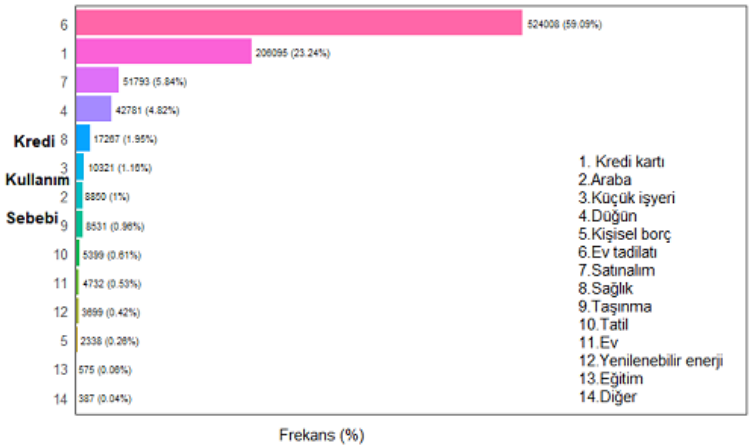
7.7 Kredi Müşterilerinin Gelir Seviyelerine Göre Dağılımı



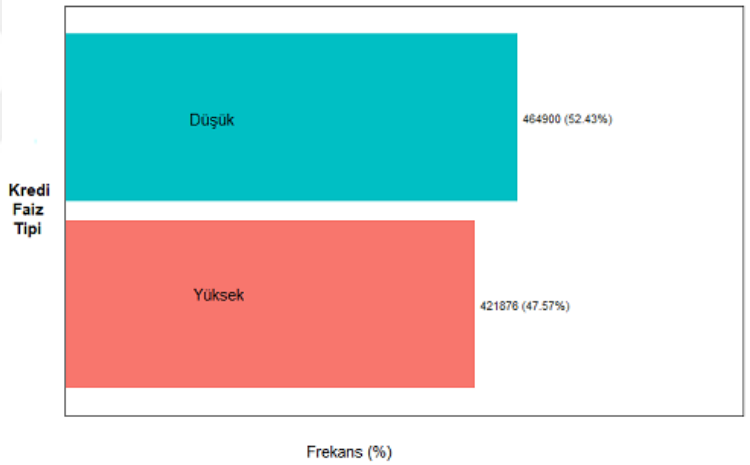
7.8 Kredi Müşterilerinin Belirlenen Kredi Notuna Göre Dağılımı



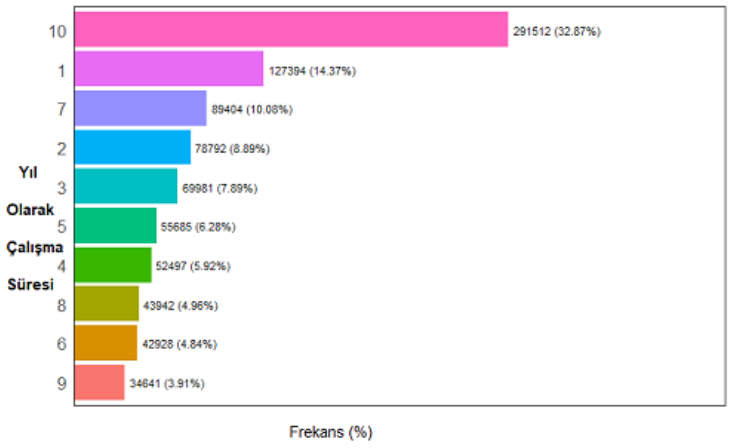
7.9 Kredi Müşterilerinin Kredi Kullanma Sebeplerine Göre Dağılımı



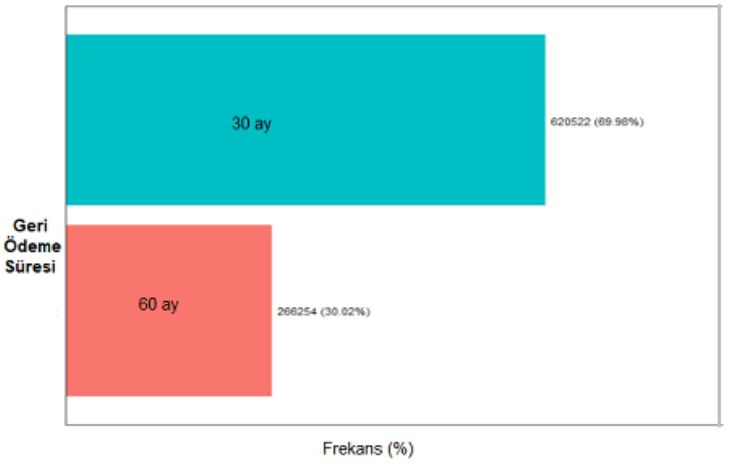
7.10 Kullanılan Kredilerin Faiz Tipine Göre Dağılımı



7.11 Kredi Müşterilerinin Yıl Olarak Çalışma Süreleri Göre Dağılımı



7.12 Kredi Müşterilerinin Ay Olarak Geri Ödeme Süreleri Dağılımı



7.13 Kredi Verisindeki Sürekli Değişkenlerin Normalleştirmeden Önceki Bilgileri

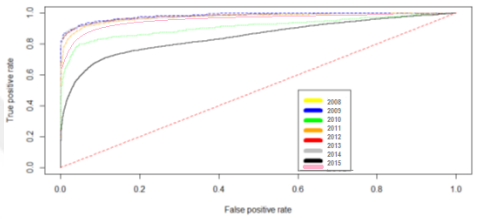
	faiz oranı	toplam ödeme	yıllık gelir	kredi miktarı	aylık ödeme
Mean	13.247706	7558.642	75034.54	14759.685	436.8340
Stdev	4.382623	7672.220	64697.94	8435.092	244.1734
Median	12.990000	4093.660	65000.00	13000.000	382.5500
Minimum	5.320000	0.000	0.00	500.000	15.6700
Maximum	28.990000	57777.580	9500000.00	35000.000	1445.4600

7.14 Kredi Verisindeki Sürekli Değişkenlerin Normalleştirdikten Sonraki Bilgileri

	faiz oranı	toplam ödeme	yıllık gelir	kredi miktarı	aylık ödeme
Mean	0.000000	0.000000	0.000000	0.000000	0.000000
Stdev	1.000000	1.000000	1.000000	1.000000	1.000000
Median	-0.068605	0.142954	-0.372935	-0.208615	-0.590710
Minimum	-1.462035	-6.995273	-1.275410	-1.690519	-1.093655
Maximum	1.936192	0.142954	1.995288	2.399537	1.325210

8 YZT Model Sınıflandırması

8.1 Sınıflandırmada Kullanılan AUC, KS ve Gini Endeksi için Yıllara göre TP-FN Grafiği



9 Çaprazlama Performans Sonuçları

9.1 Çaprazlama Sonuçlarının Yıllara göre Ortalama Performans Değerleri

Bölünme Oranları	2008	2009	2010	2011	2012	2013	2014	2015
%60 - %40	0.76	0.78	0.75	0.71	0.76	0.72	0.70	0.71
%70 - %30	**0.82**	**0.88**	0.81	**0.80**	**0.70**	**0.71**	**0.73**	**0.72**
%80 - %20	0.81	0.87	**0.82**	**0.80**	0.69	0.68	0.70	0.71