



MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE KREDİ RİSKİ TAHMİNİ

Hamdi OKUR

**YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANA BİLİM DALI**

**GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

MAYIS 2020

Hamdi OKUR tarafından hazırlanan “MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE KREDİ RİSKİ TAHMİNİ” adlı tez çalışması aşağıdaki jüri tarafından OY BİRLİĞİ ile Gazi Üniversitesi Bilgisayar Mühendisliği Ana Bilim Dalında YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Danışman: Doç. Dr. Aydın ÇETİN

Bilgisayar Mühendisliği Ana Bilim Dalı, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum.

Başkan: Prof. Dr. Necaattin BARIŞCI

Bilgisayar Mühendisliği Ana Bilim Dalı, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum.

Üye: Dr. Öğr. Üyesi Abdullah ORMAN

Bilgisayar Programcılığı Bölümü, Ankara Yıldırım Beyazıt Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum.

Tez Savunma Tarihi: 07/05/2020

Jüri tarafından kabul edilen bu çalışmanın Yüksek Lisans Tezi olması için gerekli şartları yerine getirdiğini onaylıyorum

.....

Prof. Dr. Sena YAŞYERLİ

Fen Bilimleri Enstitüsü Müdürü

ETİK BEYAN

Gazi Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmada;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmada yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

.....
Hamdi OKUR
07/05/2020

MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE KREDİ RİSKİ TAHMİNİ

(Yüksek Lisans Tezi)

Hamdi OKUR

GAZİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

Mayıs 2020

ÖZET

Son günlerde kişiden kişiye kredi sistemleri yaygınlaşmakta ve geleneksel bankacılık sisteminin yerine alternatif olarak ortaya çıkmaktadır. Bu sistemde kişiler sisteme kayıt olup kaynaklarını istedikleri faiz oranı ve vade ile kredi olarak sunmakta ve müşterilerini kendileri seçebilmektedir. Seçilen müşteriye hızlı bir biçimde doğru miktarda kredi verilmesi amacıyla müşteri özelliklerine göre kredi risklerinin belirlenerek risk değerine göre her müşteri grubuna uygun kredi miktarının hesaplanması gerekmektedir. Lending Club kişiden kişiye finans sistemleri içerisinde en yaygın olanıdır. Bu çalışmada, kredi riskini makine öğrenmesi yöntemleriyle hesaplamak amacıyla regresyon ve ikili sınıflama yöntemlerinin başarımları test edilerek kredi riskine etki eden etmenler araştırılmıştır. Tezde, regresyon ve ikili sınıflama yöntemleri Lending Club veri setiyle eğitilmiş ve test edilmiştir. Test sonuçlarına göre regresyon yöntemleri içerisinde en başarılı sonucu artırılmış karar ağacı regresyon algoritması elde ederken, ikili sınıflandırma türünde ise iki sınıflı karar ağacı algoritması en başarılı sonuçları vermiştir. Yapılan genel veri analizlerinde, kredi riskine etki eden etmenler öncelik derecesine göre gelir grubu, faiz oranı, bölge ve kredi kullanım türü olarak tespit edilmiştir.

Bilim Kodu : 92431
Anahtar Kelimeler : Kişiden Kişiye Kredi, Lending Club, Makine Öğrenmesi, Regresyon, İkili Sınıflı Sınıflandırma
Sayfa Adedi : 79
Danışman : Doç. Dr. Aydın ÇETİN

CREDIT RISK PREDICTION WITH MACHINE LEARNING METHODS

(M. Sc. Thesis)

Hamdi OKUR

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

May 2020

ABSTRACT

Recently, peer-to-peer credit systems have become widespread and emerge as an alternative to the traditional banking system. In order to give the right amount of credit to the selected customer quickly, the credit risks should be determined according to the customer characteristics and the amount of loan suitable for each customer group should be calculated according to the risk value. Lending Club is the most common peer-to-peer financial system. In this study, the factors affecting credit risk were investigated by testing the regression and binary classification methods in order to calculate credit risk with machine learning methods. In the thesis, regression and binary classification methods are trained and tested with the Lending Club data set. According to the test results, the most successful result among the regression methods was the decision tree regression algorithm, while in the binary classification type, the two-class decision tree algorithm gave the most successful results. In the general data analysis, the factors affecting credit risk were determined as income group, interest rate, region and credit use type according to priority level.

Science Code : 92431

Key Words : Peer To-Peer Credit, Lending Club, Machine Learning, Regression, Two Class Classification

Page Number : 79

Supervisor : Assoc. Prof. Dr. Aydın ÇETİN

TEŞEKKÜR

[Tezimi hazırlamamda bana yardımlarını esirgemeyen değerli hocam Doç. Dr. Aydın ÇETİN'e, değerli önerilerinden dolayı sayın Prof. Dr. Necaattin BARIŞÇI ve Dr. Öğretim Üyesi Abdullah ORMAN'a teşekkürü bir borç bilirim.]

İÇİNDEKİLER

	Sayfa
ÖZET	iv
ABSTRACT	v
TEŞEKKÜR.....	vi
İÇİNDEKİLER	vii
ÇİZELGELERİN LİSTESİ.....	x
ŞEKİLLERİN LİSTESİ.....	xi
SİMGELER VE KISALTMALAR.....	xiv
1. GİRİŞ.....	1
2. KAVRAMSAL ÇERÇEVE	5
2.1. Kişiden Kişiye Finans Sistemi	5
2.2. Makine Öğrenmesi	8
2.3. Makine Öğrenmesi Türleri	9
2.4. Gözetimli Öğrenme	9
2.4.1. Sınıflandırma yöntemleri	10
2.4.2. Sınıflandırma yönteminde kullanılan algoritmalar	10
2.4.2. Regresyon yöntemi.....	16
2.5. Gözetimsiz Öğrenme.....	16
2.5.1. Kümeleme	17
2.5.2. Birliktelik kuralı	17
2.5.3. Boyut azaltma.....	18
2.6. Takviyeli Öğrenme.....	19

2.7. Makine Öğrenmesi Öğrenme Stratejileri	20
2.7.1. R dili.....	20
3. ARAŞTIRMA YÖNTEMİ.....	21
3.1. Akış Şeması ve Proje Mimarisi.....	21
3.2. Veri Seti	25
4. BULGULAR VE DEĞERLENDİRME.....	39
4.1. Veri Değerlendirme Ölçütleri.....	39
4.2. Bulgular	41
4.2.1. Doğrusal regresyon	41
4.2.2. Karar ormanı regresyonu.....	42
4.2.3. Yapay sinir ağı regresyonu.....	43
4.2.4. Artırılmış karar ağacı regresyonu.....	44
4.2.5. İki sınıflı yerel derin destek vektör makinesi	45
4.2.6. İki sınıflı karar ormanı	47
4.2.7. İki sınıflı artırılmış karar ağacı.....	48
4.2.8. İkili sınıflı yapay sinir ağları	50
4.3. Değerlendirme.....	51
4.4. Modelin Web Servis Entegrasyonu.....	53
4.5. Kredi Riskine Etki Eden Faktörlerin Analizi	55
5. SONUÇ VE ÖNERİLER	69
KAYNAKLAR	71

	Sayfa
EKLER.....	73
EK-1. R diliyle Normalizasyon Kodları	74
ÖZGEÇMİŞ	79

ÇİZELGELERİN LİSTESİ

Çizelge	Sayfa
Çizelge 3.1. FICO değerlendirmeleri.....	33
Çizelge 3.2. Yüksek FICO değeri niteliği veri dağılımı	36
Çizelge 4.1. Karışıklık matrisi	40
Çizelge 4.3. İkili sınıflandırma yöntemlerinin kıyaslamalı sonuçları.....	52

ŞEKİLLERİN LİSTESİ

Şekil	Sayfa
Şekil 2.1. Lending club gelirleri	7
Şekil 2.2. Lending club çalışma sistemi.....	7
Şekil 2.3. Lending club çalışma sistemi.....	9
Şekil 2.4. Karar ağacı algoritması şeması.....	11
Şekil 2.5. Rasgele orman algoritması şeması.....	13
Şekil 2.6. Destek vektör makinesi algoritması şeması.....	14
Şekil 3.1. İşlem akış şeması	22
Şekil 3.2. Proje detay mimarisi üst kısım	24
Şekil 3.3. Proje detay mimarisi orta kısım	24
Şekil 3.4. Proje detay mimarisi alt kısım	25
Şekil 3.5. Ev sahipliği niteliği.....	26
Şekil 3.6. Çalışma süresi niteliği.....	27
Şekil 3.7. Meslek adı niteliği	27
Şekil 3.8. Eyalet adresleri niteliği	28
Şekil 3.9. Dönem niteliği	29
Şekil 3.10. Borçlanma durumu niteliği.....	30
Şekil 3.11. Borçlanma amacı niteliği.....	31
Şekil 3.12. Faiz oranı niteliği.....	32
Şekil 3.13. Düşük FICO değerlendirme niteliği	34
Şekil 3.14. Yüksek FICO değerlendirme niteliği.....	35
Şekil 3.15. Borç miktarı niteliği.....	36
Şekil 3.16. Doğrulama durumu niteliği.....	37

Şekil	Sayfa
Şekil 4.1. Doğrusal regresyon hata değerleri	42
Şekil 4.2. Doğrusal regresyon hata değerleri dağılımı.....	42
Şekil 4.3. Karar ormanı regresyonu hata değerleri dağılımı	43
Şekil 4.4. Yapay sinir ağı regresyonu hata değerleri	44
Şekil 4.5. Yapay sinir ağı regresyonu hata değerleri dağılımı	44
Şekil 4.6. Artırılmış karar ağacı regresyonu yönteminin hata değerleri	45
Şekil 4.7. Artırılmış karar ağacı regresyonu yönteminin hata değerleri grafiği	45
Şekil 4.8. İkili sınıflı yerel derin destek vektör makinesi ROC grafiği	46
Şekil 4.9. İkili sınıflı yerel derin destek vektör makinesi kesinlik grafiği	46
Şekil 4.10. İkili sınıflı yerel derin destek vektör makinesi kesinlik detay grafiği	47
Şekil 4.11. İkili sınıflı karar ormanı ROC grafiği.....	48
Şekil 4.12. İkili sınıflı karar ormanı kesinlik grafiği	48
Şekil 4.13. İkili sınıflı artırılmış karar ağacı ROC grafiği	49
Şekil 4.14. İkili sınıflı artırılmış karar ağacı kesinlik grafiği.....	49
Şekil 4.15. İkili sınıflı artırılmış karar ağacı kesinlik detay grafiği	49
Şekil 4.16. İkili sınıflı yapay sinir ağları ROC grafiği.....	50
Şekil 4.17. İkili sınıflı yapay sinir ağları kesinlik grafiği	51
Şekil 4.18. İkili sınıflı yapay sinir ağları kesinlik detay grafiği.....	51
Şekil 4.19. Azure ML tahmin şeması.....	54
Şekil 4.20. Azure ML web servis çalışma ekranı	55
Şekil 4.21. Plotly ana menüsü.....	56
Şekil 4.22. Plotly çalışma gridi	56
Şekil 4.23. Bölgelere göre kötü kredi oranları	57
Şekil 4.24. Bölgelere göre kredi talepleri	57

Şekil	Sayfa
Şekil 4.25. Bölgelere göre faiz oranları	58
Şekil 4.26. Bölgelere göre devlet tarafından tahsil oranı.....	58
Şekil 4.27. Yıllar içinde kötü kredi dağılımı.....	59
Şekil 4.28. Yıllar içinde kredilerin bölgesel dağılımı	60
Şekil 4.29. İyi ve kötü kredi faiz oranları	60
Şekil 4.30. İyi ve kötü kredilerin kredi türlerine göre oranları	61
Şekil 4.31. Gelir gruplarının faiz oranları.....	62
Şekil 4.32. Gelir gruplarında iyi kredilerin kredi türlerine göre oranları.....	62
Şekil 4.33. Gelir gruplarında kötü kredilerin kredi türlerine göre oranları.....	63
Şekil 4.34. İyi ve kötü kredilerin toplam kredi türleri arasındaki oranları	64
Şekil 4.35. Gelir gruplarında iyi ve kötü kredilerin kredi türlerine göre oranları.....	64
Şekil 4.36. Kredi doğrulama durumu ve ödenme durumu ilişki grafiği	65
Şekil 4.37. Kredilerin yıllar ve aylar içerisindeki dağılımı.....	65
Şekil 4.38. Kredi vadelerinin yıllara göre dağılımı.....	66
Şekil 4.39. Temerrüte düşen kredilerin vadelerine göre aylık dağılımı.....	66

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Kısaltmalar	Açıklamalar
ABD	Amerika Birleşik Devletleri
API	Application Programming Interface
AUC	Area Under Curves
COD	Coefficient of Determination
FCA	Financial Conduct Authority
FICO	Fair Isaac Corporation
MAE	Mean Absolute Error
ML	Machine Learning
P2P	Peer to Peer
P2PFA	Peer2Peer Finance Association
PCA	Principal Component Analysis
RAE	Relative Absolute Error
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristic Curve
RSE	Relative Squared Error
SVM	Support Vector Machine

1. GİRİŞ

Kişiden kişiye finans sistemleri geleneksel bankacılığın alternatifi olarak ortaya çıkmıştır. Müşteriler, internet sayesinde banka şubesine gitmeden mevduatlarını kredi olarak vermekte ya da kredi bulmaktadır. Bu durum hem rahatlık hem de hız kazandırmaktadır. Kişiden kişiye finans sisteminde, kredi verenler istedikleri risk gruplarına kredilerini ulaştırabilecek ve banka mevduat faizinden daha fazla miktarda gelir elde edebileceklerdir. Geleneksel bankacılıkta kredi ihtiyacı için bankalara başvurmak gerekirken, kişiden kişiye finans sistemlerinde kredi ihtiyacı olan kişiler ve şirketler finans fazlası olanlardan transfer yapabilmektedir. Bu finans sisteminde kredi ihtiyacı olan kişiler ve şirketler kendilerine en uygun vade ve faiz orandan kredi seçebilmektedir.

Kredi verecek olan kişilerde alacakları risk türüne göre faiz isteyebilmektedirler. Mevduatın geri alınabilmesi ve faizin belirlenebilmesi için kredi risklerinin hesaplanmasına yönelik uygulamalar önemlidir. Kredi riskini hesaplayan çeşitli modeller büyük finans kuruluşları tarafından oluşturulmuştur. Bu modellere alternatif olarak makine öğrenmesi ile oluşturulan modeller web servisler aracılığıyla geliştirilen kredi riski yazılımlarına entegre edilebilir.

Kişiden kişiye finans sistemleri arasında en güçlüleri Prosper ve Lending Club olarak tanımlanmaktadır. Lending Club, Prosperdan farklı olarak kredilendirmeye dair verileri açık kaynak olarak kullanıcı ve araştırmacılarla paylaşmaktadır. Bu da, bu alanda araştırma yapılmasının önünü açmaktadır. Literatürde, Lending Club veri setini kullanan çalışmalardan farklı analizler içeren çeşitli çalışmalar yapıldığı görülmüştür. P.Li ve G.Han [1] Doğrusal Regresyon, Yapay Sinir Ağları ve Rasgele Ağaç makine öğrenmesi yöntemleri kullanılarak Lending Club verileri üzerinden kar tahmini gerçekleştirmişlerdir. S. Chang, S.Kim, ve S.Kondo [2], Lojistik Regresyon, Naive Bayes Ağları ve SVM (Destek Vektör Makinesi) yöntemleri kullanılarak risk tahmini çalışması yapmışlardır.

Makine öğrenmesi ile ilgili çalışmalarıyla bilinen Stanford Üniversitesi'nce gerçekleştirilen bir diğer çalışmada ise yine aynı veri seti kullanılarak daha farklı bir problem üzerinde durulmuştur. "P2P Loan Selection" çalışmada Lending Club üzerinde hangi kredi tipinin seçileceği makine öğrenmesi yöntemleri kullanılarak belirlenmiştir. İlgili makalede Lojistik Regresyon, SVM (Destek Vektör Makinesi) ve Doğrusal Discriminant Analizi yöntemleri

kullanılarak en başarılı kredi borç modeli belirlenmeye çalışılmıştır [3]. İncelenen bu çalışmaların ilki kar, ikincisi risk üçüncüsü ise en uygun borç modelini bulmayı amaçlamaktadır. Lending Club veri seti dışında kullanılan daha farklı bir çalışma çalışmada [4] ülkemizdeki bir kamu bankasının verileri yapay sinir ağları ve lojistik regresyon ile değerlendirilmiş ve sonuçları elde edilmiştir. Yöntem olarak, yapay sinir ağının lojistik regresyona göre daha başarılı sonuçlar elde ettiği bu çalışmada tespit edilmiştir. Risk analizi yapılan çalışmada en başarılı sonucu Gaussian uygulanan Naive Bayes algoritması elde ettiği gözlemlenmektedir. Naive Bayes yöntem olarak olasılığa dayalı kestirimlerde bulunmaktadır.

Bu tezde ise, Lending Club veri setini oluşturan veriler göz önüne alındığında istatistiksel ve bilgiye dayalı yaklaşımların da başarılı sonuçlar üretebileceği öngörülmüştür. Bu amaçla yapılan literatür taramasında Lending Club veri seti ile iki sınıflı sınıflandırma algoritmaların olduğu bir çalışmaya ulaşılamamıştır. Ayrıca, “Predicting Default Risk of Lending Club Loans” çalışmasındaki veri setinde daha farklı kolonlar eğitim ve test için kullanılırken bu tezde “kredi durumu” kolonu verilerinin kullanımı kredi riski durumunun diğer parametrelere göre özet bir sonucu olduğu için bu veri seti açısından en doğru sonuçları vereceğini öngörülmüştür.

Bu tezde kredi riski analizi için en başarılı sınıflandırma ve regresyon yöntemi hangisidir ve kredi riskini belirleyen faktörler nelerdir?” sorularına cevap aranmıştır. Tezde, açık kaynaklı ve her yıl güncellenen bir veri seti olması nedeniyle Lending Club veri seti kullanılmıştır.

Lending Club veri tabanında kayıtlı olan ABD’de yerleşik müşterilere ait 2007- 2019 yılları arasında kredi talebinde bulunmuş toplam 111107 müşteriye ait veriler alınmış olup verilerin normalizasyonu ile tüm kayıtların belli bir kısmı olan 88910 tanesi kullanılmıştır. Bu verilerin 71128 tanesi eğitim için 17782 tanesi test için kullanılmıştır. Eğitim ve test oranları değiştirilerek daha farklı başarı sonuçları elde edilebilir.

Bu tez beş bölümden oluşmaktadır. Tezin İkinci Bölüm'ünde kavramsal çerçeve oluşturularak kişiden kişiye finans sistemi hakkında bilgiler sunulmakta ve tezde kullanılan makine öğrenmesi yöntemleri hakkında genel bilgiler verilmektedir. Üçüncü Bölüm'de tezde kullanılan araştırma yöntemi kapsamlı olarak sunulmakta, Dördüncü Bölüm'de ise teze ait bulgular değerlendirmekte ve Beşinci Bölüm'de teze ait sonuçlar sunulmaktadır.

2. KAVRAMSAL ÇERÇEVE

Bu bölümde, Kişiden Kişiyne Finans Sistemi kavramı açıklanarak, Lending Club finans şirketi hakkında bilgiler verilmiş, veri seti yapısı aktarılmış ve tezde kullanılan makine öğrenmesi algoritmaları detaylı olarak incelenmiştir.

2.1. Kişiden Kişiyne Finans Sistemi

İnternet milyonlarca bilgisayar, cep telefonunu ve tabletin birbirleriyle iletişim kurduğu, veri paylaştığı çalışma, bilgi paylaşımı, eğlence ve ticari amaçlar için de kullanılan dünya geneline yayılmış bir ağıdır.

Toplumların bilgi çağına ayak uydurmak için bilgisayar yazılım ve donanımlarını kullanması interneti yaygın bir teknoloji haline getirmiştir. Bilgisayar fiyatlarının düşmesi, yazılımların yaygınlaşması internete daha çok kişinin adapte olmasını sağlamıştır. İnternet sayesinde dünya genelinde hızlı ve kolay bilgi paylaşımı ve iletişim sağlandığı gibi dünyanın küreselleşme kavramına geçişi hızlanmıştır. Ayrıca, internet kamu yönetimini şeffaflaştıran, bilimsel bilginin paylaşımını artıran ve insanların medyaya daha rahat ulaşmasını sağlayan bir teknolojidir. İnternetin yaygın kullanımı bilimsel çalışmaların paylaşılması, uluslararası sosyal ve siyasi ilişkilerin kurulması sağlandığı gibi ticaretin gelişimine de artarak katkı sağlamaktadır. Ticaret günümüzde geleneksel yolların dışında internetin sağladığı kolaylıklar sayesinde farklı boyutlar kazanmıştır.

Geleneksel finans sisteminde fon ihtiyacı için bankalara başvuru yapılarak kredi talep edebilirken, kişiden kişiyne (Peer-To-Peer-P2P) finans sisteminde fon ihtiyacı olan kişiler ve şirketler fon fazlası olanlardan transfer yapabilmektedir. P2P Finans sisteminde fon ihtiyacı olan kişiler ve şirketler kendilerine en uygun fonları seçebilmektedir. Bu yapıyla geleneksel bankacılık sistemi devre dışı bırakılmaktadır. Fon talep edenler uygun koşullarda kredi bulurken fon sağlayıcıları da mevduat faizi gibi çeşitli gelirlerden daha fazlasını elde edebilmektedirler. Kredi verecek olan kişiler de alacakları risk türüne göre faiz isteyebilmektedir. Benzer faiz oranları olduğunda ise örneğin Lending Club finans sisteminde krediyi ilk tanımlayan kişinin fonu kullanılmaktadır.

P2P Finans Sisteminin dünyada ilk kez 2005 yılında İngiltere’de kullanılmıştır. P2P platformlar aracılığıyla sağlanan fon miktarı 2014 yılının 2. çeyreği itibariyle 1.5 milyar sterline ulaşmış durumdadır. Birçok ülkede P2P finans sistemi için ilke ve standartlar bulunmamasına rağmen İngiltere’de finansal tüketici hizmetlerinin gözetiminden sorumlu olan Financial Conduct Authority (FCA) [5] tarafından P2P finansman faaliyetlerini düzenleyen kuralların yanı sıra standartlar ve ilkeler de belirlenmiştir. Hatta İngiltere’de faaliyet gösteren P2P finans hizmeti veren şirketleri temsil eden “Peer2Peer Finance Association (P2PFA)” isimli Bankalar Birliği’ne benzer bir birlik de kurulmuştur.

Dünyaca ünlü Forbes dergisinde yayınlanan yazıda “The Disappearance Of Peer-To- Peer Lending” 2005 ve 2006 yıllarında kurulan iki büyük P2P borç verme sistemi olan Prosper ve Lending Club’ın makalenin yazıldığı tarih itibariyle 6 Milyar Dolardan daha fazla kaynak elde ettiği, buna karşın P2P sisteminin zamanla Şekil değiştirdiği, daha önceden küçük kredi veren şirketlerin yerini büyük yatırımcıların ve finans kuruluşlarının aldığını ve bu durumun küçük kredi sağlayıcılarını zor durumda bıraktığı belirtilmektedir. Yazının devamında P2P kredi sisteminin iş yapış şeklinin değişebileceği, bu tür kuruluşlara bankaların kredi vermek istemedikleri müşterileri yönlendirebilecekleri ifade edilmiştir [6].

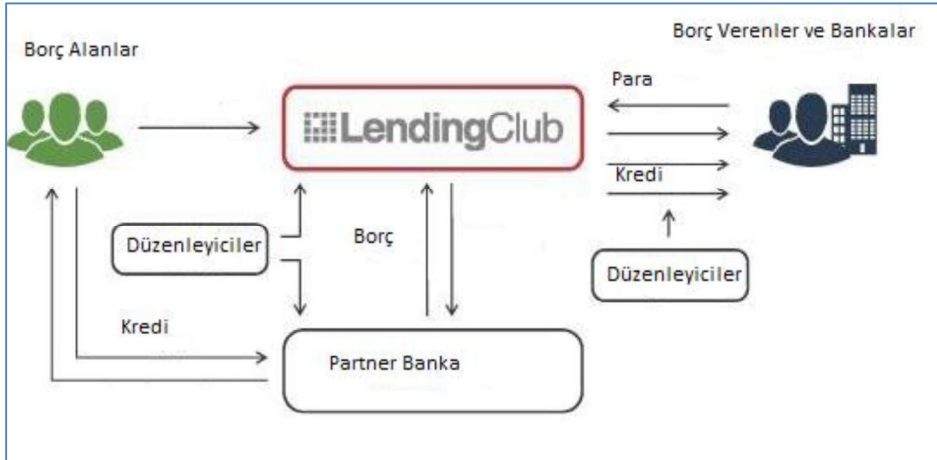
Lending Club’ın kendi web sitesinde Google ile anlaşma yapıldığı açıklanmıştır. Bu anlaşmaya göre Google kendisinde iş yapan küçük şirketlere hızlı ve ucuz fon sağlayabilmek için Lending Club ile protokol imzalamıştır. Lending Club ABD merkezli bir P2P şirkettir ve hızlı bir büyüme göstermektedir [7].

Şekil 2.1’de yer alan grafikte 2016 yılı ilk çeyreğinde Lending Club şirketinin mevcut dönemler içerisinde en yüksek geliri elde ettiği görülmüştür. Bu dönemden sonra gelirin her çeyrekte hemen hemen korunduğu görülebilir [8].



Şekil 2.1. Lending club gelirleri

P2P finans sisteminde Lending Club dışında farklı ülkelerde bulunan sistemler de bulunmaktadır. Bu sistemlerin içinde en büyük paya sahip olanlardan biri Lending Club finans sistemidir. “p2pmoney” isimli web sitesinde hangi P2P sisteminin hangi koşullarda kredi verdiği Şekil 2.2’de görülebilmektedir [9]. Şekil 2.2’de ise Lending Club finans sisteminin çalışma düzeni özetlenmiştir.



Şekil 2.2. Lending club çalışma sistemi

Şekil 2.2’de görüldüğü gibi sisteme yatırım yapan yatırımcı ve bankalar bulunmaktadır. Bu kişi ve gruplar sisteme faiz karşılığı para aktarmaktadır. Aynı Şekilde ortaklık yapılan

bankalar da müşteri ve kredi desteği sağlamaktadır. Kredi almak isteyen kişiler sistem başvurularını yapmaktadır ve belirli yasal düzenlemeler ile bu finans sistemi denetlenip düzenlenmektedir [10]. Ülkemizde ise P2P finans sisteminin altyapısı henüz kurulmamıştır. Bu alanda faaliyet göstermek için çeşitli şirketler kurulmuştur fakat henüz P2P finansal sisteminin ilkeleri ve standartları belirlenmemiştir.

Kullanılan risk tahmini araçları fon sağlayıcılarının kredi geri ödemelerini almalarında ve tedarikçilerin sermayesinin korunması ve devamlılığı açısından önemlidir. Kredi riski ile ilgili problemler için makine öğrenmesi yöntemleri sıklıkla kullanılmaktadır.

2.2. Makine Öğrenmesi

Makine öğrenmesini farklı yöntemler ile veriyi inceleyip çeşitli şablonları arayan ve bu şablonlara ait olan etiketleri bularak öğrenme işlemi yapan ve devamında öğrenilenlere benzer tahminler yapan sistemler olarak ifade edilebilir. Makine öğrenmesi bu eğitim ve tahmin süreçlerini çeşitli bilimsel yöntemler aracılığıyla oluşturan modeller üretir. Elde edilen modeller veriyi asgari hata ve azami tahmin-kesinlik oranıyla değerlendirmelidir.

Makine öğrenmesi günümüzde büyük verinin analizinde ve bu verinin değerlendirilmesinde kullanılmaktadır. Farklı alanlarda kullanılmakta olan makine öğrenmesi bilhassa tıp, fen bilimleri ve finans alanında kullanılmaktadır. Makine öğrenmesi, kendi kendine öğrenme gerçekleşmesi için çalışma yapılan bir konudur.

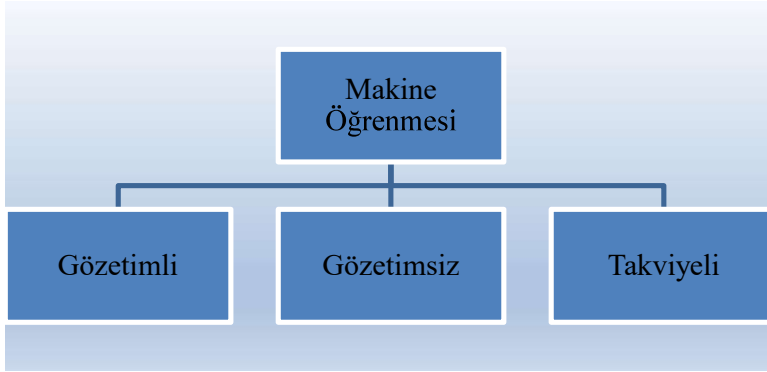
Veriler sürekli büyümektedir ve tüm verileri daha hızlı ve daha yüksek doğrulukla kullanmak gittikçe zorlaşmaktadır. Sahip olduğumuz verilerin büyük çoğunluğu müzik, video, fotoğraf ve dokümanlardan oluşmaktadır. Bu verilere kalıp bulmak insan beyni için çok zordur ve büyük verinin, hesaplaması için çok fazla zamana ihtiyaç duyulmaktadır. Makine öğrenmesinin kullanılmasıyla büyük veri fazla zaman kullanmadan işlenebilir.

Makine öğrenimi yapay zekanın bir alanıdır. Yapay zeka kullanarak akıllı makineler oluşturulabilir, sayısal mantık yürütme, hareket ve konuşma gibi insan benzeri davranışlar elde edilebilir. Yapay zeka insan zekasına benzer makineler tasarlamayı amaçlamaktadır.

Yapay zeka, makine öğrenmesi ve derin öğrenmeyi kapsar. Yapay zeka 1950'lerden beri çalışılan bir alandır ve daha geniş bir alana işaret eder. Makine öğrenmesine 1980'lerde başlanırken derin öğrenme 2000'li yılların ortalarında ortaya çıkmış daha yeni bir teknolojidir.

2.3. Makine Öğrenmesi Türleri

Makine öğrenmesi algoritmaları 3 sınıfa ayrılır. Bunlar 1. Gözetimli Öğrenme (Supervised Learning), 2. Gözetimsiz Öğrenme (Unsupervised Learning), 3. Takviyeli Öğrenme (Reinforcement Learning) olarak ifade edilebilir. Şekil 2.3'te bu türler gösterilmiştir.



Şekil 2.3. Lending club çalışma sistemi

2.4. Gözetimli Öğrenme

Gözetimli öğrenme algoritmaları etiketli verilerden öğrendiklerini kullanarak tahminler yapar. Veriler ve veriye ait sınıflar daha önceden mevcuttur. Bu veriler bilinerek eğitim yapılır. Veriler eğitildikten sonra sistem öğrenir ve sistem yeni gelen verileri yorumlar. Eğitim amaçlı kullanılan veri doğru bir biçimde oluşturulmazsa mevcut modeller kötü tahminler yapabilir. Gözetimli öğrenme algoritmalarının çalışma sistemini açıklamak için çiçek verileri örnek olarak verilebilir. Çiçek verileri eğitim verisi olarak kullanılıp çiçek şekilleri etiketlediğinde model artık çiçek şekillerini tanıyacaktır. Devamında çiçek ve ağaç şekilleri karışık olarak test verisi olarak kullanıldığında ise daha önce eğitilen veriler şekilleri ayırt edebileceklerdir.

Gözetimli öğrenme algoritmaları genelde iki grup Sınıflandırma Yöntemi (Classification Method) ve Regresyon Yöntemi (Regression Method) olarak ayrılmaktadırlar. Tezde, gözetimli algoritmaların her iki grubu da kullanılmıştır.

2.4.1. Sınıflandırma yöntemleri

Bu yöntemde veriler belirli özelliklerine göre sınıflandırılırlar. Sınıflandırma yöntemi yapısal (structure) veya yapısal olmayan (unstructure) verilere uygulanabilmektedir. Sistem hangi verinin hangi durumda, hangi sınıfa ait olacağı bilgisi ile eğitilirse, bu veri eğitildikten sonra yeni gelen veriyi de sınıflandırabilir. Bu yöntemi açıklamak için şöyle bir örnek verilebilir. Bir sınıfta sınav sonrası 20 puan ve altını alanları kötü, 20-50 arasını orta ve 50 üstü puan alanları ise iyi olarak sınıflandırıldıktan sonra bu verilerle model eğitilirse ve sonrasında karışık olan ve sınıflandırılmamış veriler ile test yapılırsa yeni verilerin eğitim verilerine göre sınıflandırıldığı elde edebilir. Sınıflandırma algoritmasının 2 türü bulunmaktadır.

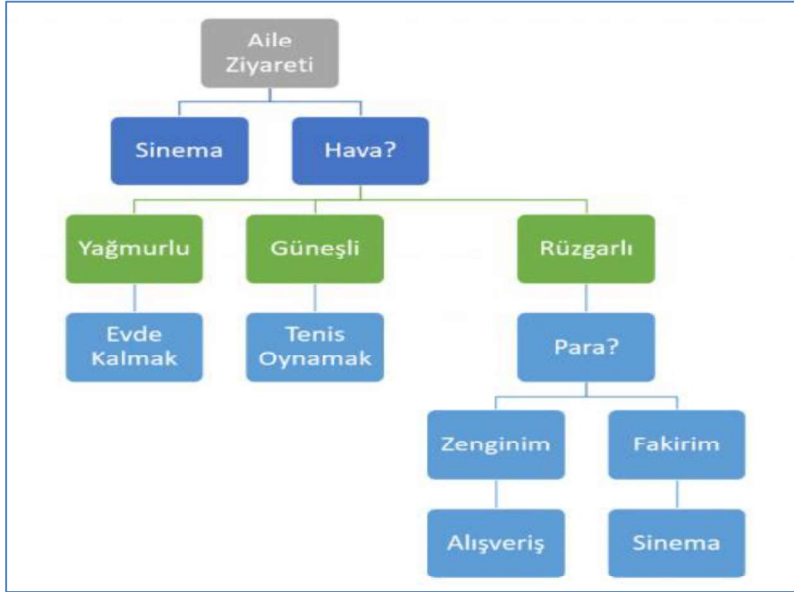
İkili Sınıflandırma (Binary Classification): Bu türde veri iki sınıfa ayrılır. Çoklu Sınıf Sınıflandırma (Multiclass Classification): Bu türde ise ikiden fazla sınıfla veriler sınıflandırılmaktadır. Tezde, ikili sınıflandırma yöntemi kullanılmıştır.

2.4.2. Sınıflandırma yönteminde kullanılan algoritmalar

Bu bölümde tezde kullanılan sınıflandırma yöntemleri incelenerek çalışma sistemleri verilmiştir.

Karar Ağacı (Decision Tree): Veriler sınıfları ile birlikte model içinde değerlendirilerek kurallar üretilir. Karar ağacı karar düğümleri (decision node) ve yaprak düğümleri (leaf node) olan bir ağaç yapısına sahiptir. İlk düğüme kök düğüm (root node) denir. Karar ağaçları bir veya daha fazla dallanma içerebilir. Bu algoritma hem regresyon hem sınıflandırma yönteminde kullanılabilir [11].

Örnek olarak Şekil 2.4'te karar ağacı verilebilir. Karar ağacının en sık kullandığı öğrenme algoritmalarından biri ID3 algoritmasıdır, ağaç üzerinde özyinelemeli olarak bilgi kazanımı prensibine uygun olarak yaprak kalmayınca kadar ID3 algoritması devam eder.



Şekil 2.4. Karar ağacı algoritması şeması

ID3 Algoritması: Bu algoritma sabit bir örnek kümesinden bir karar ağacı oluşturur. Elde edilen ağaç gelecekteki örnekleri sınıflandırmak için kullanılır. ID3 Algoritması Entropi ve Bilgi Kazanımı üzerine inşa edilmiştir.

Entropi (Entropy): Rastgele ve belirsiz durumların ortaya çıkma olasılığıdır [12]. Örneklerin tamamı düzenliyse, entropi sıfırdır. Entropi'nin formülü Eş. 2.1 ile ifade edilmiştir.

$$\text{Entropi}(S) = S - p(I) \log_2 p(I) \quad (2.1)$$

Bilgi Kazanımı (Information Gain): Veri setini bir özellik üzerinden böldükten sonra tüm entropiden çıkarma işlemidir [13]. Eş. 2.2 ile formülü verilmiştir.

$$\text{Gain}(S, A) = \text{Entropy}(S) - \text{Entropy}(S, A) \quad (2.2)$$

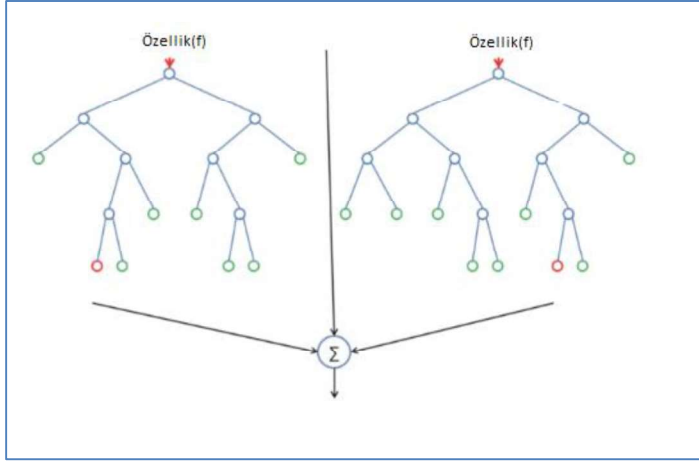
Aşırı Uyum (Overfitting): Eğitim setindeki hataları azaltma sürecinde aşırı uyum ortaya

çıkar. Aşırı uyumdan kaçınmak için öncelikle ağacın büyümesi sınırlandırılır daha sonra, ağaç oluşturulup gereksiz kısımlar çıkartılır.

Artırılmış Karar Ağaçları (Boosted Decision Trees): Karar Ağaçlarının performansını artırmak için Boosting yaklaşımlarından faydalanılması sonucu ortaya çıkan bir algoritmadır. Boosting algoritmalarının temelinde zayıf öğrenenlerin bir araya getirilerek güçlü öğrenicilerin oluşturulması vardır. Bu fikri algoritmik olarak ifade etmek için tek bir tahminsel model formunda modeller serisi oluşturmak gerekir [14]. Eş. 2.3 ile ifade edebiliriz. Boosting yaklaşımı için örnek olarak AdaBoost, Gradient Boosting, XGboost verilebilir. $e_i < p$ şartını sağlayan her i için Eş. 2.3 ile ifade edilebilir.

$$g = \text{Majority Vote } (f_1, f_2, f_3) \quad (2.3)$$

Rastgele Ormanlar (Random Forest): Bu algoritma sınıflandırma yapabilmek için çok fazla karar ağacı kullanılarak sınıflandırma değerinin yükseltilmesi hedefleyen, birden fazla ağaç yetiştiren ve tüm ağaçların oylarına göre nesneleri sınıflandıran bir topluluk modelidir. Yani bir nesne tüm ağaçlardan en fazla oyu alan bir sınıfa atanır. Rastgele orman sınıflandırıcısı, çeşitli veri kümeleri alt örneklerinde bir dizi karar ağacına uyan ve modelin tahmin doğruluğunu iyileştirmek ve aşırı uyumu kontrol etmek için ortalama kullanan bir meta tahmin edicidir. Alt örneklem büyüklüğü her zaman orijinal girdi örneklem büyüklüğüyle aynıdır, ancak örnekler değiştirilerek çizilir. Bu algorithmada sınıflama veya regresyon yönteminde kullanılabilir. Ayrıca çalışma mantığı açısından karmaşık algoritmalar [15]. Örnek olarak, Şekil 2.5 verilebilir. Bu şekilde bir rastgele orman örneği verilmiştir.



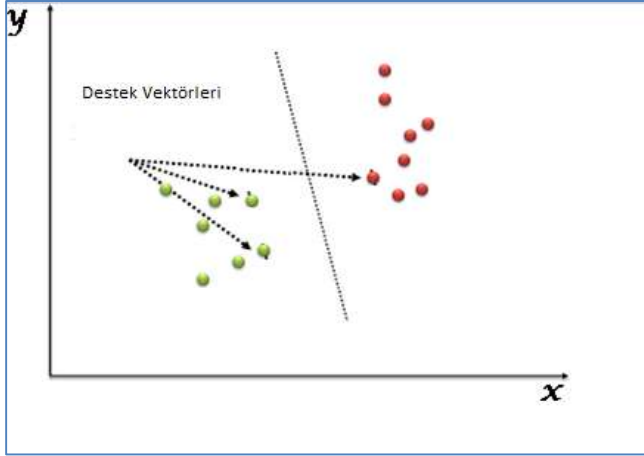
Şekil 2.5. Rasgele orman algoritması şeması

Rastgele Karar Ormanları, Karar Ağaçlarının aşırı uyum sorununu çözerler. Rastgele Ormanlar, aşırı uyum sorununu çözmek için Torbalama Algoritmasını kullanırlar. Torbalama (Bagging): Rastgele Ormanlarda sıklıkla kullanılan bir algoritmadır. Yeniden örneklem yöntemiyle, elde edilen ağaçların bir araya getirilmesi ile gerçekleştirilir. Öncelikle veri seti içinden sürekli örnekler çekilerek yeni ağaçlar oluşturulur ve bu örnekler veri setine yeniden konur. Elde edilen ağaçlar bir ağaç topluluğu oluşturur. Ortaya çıkan ağaçların tamamı modelleme için kullanılarak oylama yapılır. Tahmin işleminin sonucu tek bir ifade olarak ortaya sunulur [16]. Eş. 2.4'te Regresyon için Eş. 2.5'te Sınıflandırma için formülleri verilmiştir.

$$\hat{f}(x) = \sum_{i=1}^T \hat{f}_i(x) \quad (2.4)$$

$$\hat{f}_x = \text{sign}\left(\sum_{i=1}^T \hat{f}_i(x)\right) \quad (2.5)$$

Destek Vektör Makinesi (Support Vector Machine): Destek Vektör Makinesi, örüntü tanıma ve sınıflandırma problemlerinin çözümü için Vapnik tarafından geliştirilmiştir. Bu algoritmanın temelleri istatistiksel öğrenme teorisine diğer bir ifadeyle Vapnik-Chervonenkis (VC) teorisine dayanır. Veri kümesinde benzer gruplar arasına birbirinden en uzak olan noktalardan sınırlar çizen algoritmadır. Destek vektör makinesi, eğitim verilerinin, kategorilere ayrılan alanlardaki geniş bir açıklıkta noktalar ile gösterilmesidir. Yeni örnekler daha sonra aynı alana eşleştirilir ve aralarındaki boşluğun hangi tarafına düştüklerini temel alan bir kategoriye ait olduğu tahmin edilir [17]. Destek vektör makinesi grafik üzerinde Şekil 2.6'da verilmiştir.



Şekil 2.6. Destek vektör makinesi algoritması şeması

SVM sınıflandırıcısı oluşturmak için kullanılan parametreler 3 adettir. Bunlar: çekirdek türü, gama ve C değeridir. Doğrusal ve doğrusal olmayan vektör makinesi türleri vardır. Çalışmada doğrusal vektör makineleri kullanılmıştır.

Doğrusal Vektör Makineleri: Her x_i örneği için, p adet niteliğe sahip girdi, $y_i \in \{1, -1\}$ örneklerin ait olduğu sınıfı temsil eden çıktı ve $x \in p$ yüksek boyutlu girdi vektörü olmak üzere; (x_i, y_i) ikililerinden oluşan n hacimli bir eğitim kümesi S verildiğinde, farklı sınıflara ait örnekleri birbirinden en iyi şekilde ayıracak, doğrusal hiper düzleminin bulunmasına yardımcı olan denetimli öğrenme algoritmaları sınıfına ait makine öğrenimi algoritmasıdır. Burada w , hiper düzlemin normali aynı zamanda ağırlık vektörü ve b sabit olarak tanımlanmıştır. Destek Vektör Makinesi, $w \cdot x + b$ ile ifade edilmiştir.

Farklı sınıflara ait örnekleri ayıran birçok doğrusal düzlem mevcuttur. Fakat, Destek Vektör Makinesi, farklı sınıflara ait destek vektörleri arasındaki uzaklığı maksimize eden ayırma hiper düzleminin bulunmasını amaçlar.

Yapay Sinir Ağları: Bir sinir ağı, bir giriş vektörünü çıktılara çeviren çeşitli katmanlardan oluşan nöronlardan oluşmaktadır. Her katman bir input olarak değerlendirir ve output değerini diğer katmana gönderir. Çoğunlukla ağırlar ileriye dönük olarak ifade edilir ve geri beslemeye sahip değildir. Ağırlık verme işlemi bir üniteden ötekine geçişte kullanılır ve sinir ağının probleme uygulanması için eğitim aşamasında kullanılır.

Bir yapay sinir hücresi beş bölümden oluşmaktadır; Bunlar: Girdiler, Ağırlıklar, Toplama, Fonksiyonu (Birleştirme Fonksiyonu), Aktivasyon fonksiyonu ve Çıktılar'dır.

Girdiler: Girdiler nöronlara gelen verilerdir. Girdiler yapay sinir hücresine bir diğer hücreden gelebileceği gibi direk olarak dış dünyadan da gelebilir. Bu girdilerden gelen veriler toplanır ve nöron çekirdeğine gönderilir.

Ağırlıklar: Yapay sinir hücresine gelen bilgiler girdiler üzerinden çekirdeğe ulaşmadan önce geldikleri bağlantıların ağırlığıyla çarpılarak çekirdeğe iletilir. Bu sayede girdilerin üretilcek çıktı üzerindeki etkisi ayarlanabilmektedir. Bu ağırlıkların değerleri pozitif, negatif veya sıfır olabilir. Ağırlığı sıfır olan girdiler çıktıyı etkilememektedir.

Toplama Fonksiyonu (Birleştirme Fonksiyonu): Toplama fonksiyonu bir yapay sinir hücresine ağırlıklarla çarpılarak gelen girdileri toplayarak o hücrenin net girdisini hesaplayan bir fonksiyondur.

Aktivasyon Fonksiyonu: Bu fonksiyon hücreye gelen net girdiyi işleyerek hücrenin bu girdiye karşılık üreteceği çıktıyı belirler. Aktivasyon fonksiyonu için genellikle doğrusal olmayan bir fonksiyon seçilir. Günümüzde en yaygın olarak kullanılan “Çok katmanlı algılayıcı” modelinde genel olarak aktivasyon fonksiyonu olarak “Sigmoid fonksiyonu” kullanılır [18].

Yapay sinir ağları yapay sinir hücrelerinin birbirine bağlanmasıyla oluşan yapılardır. Yapay sinir ağları üç ana katmanda incelenir; Giriş Katmanı, Ara (Gizli) Katmanlar ve Çıkış Katmanı. Giriş Katmanı, yapay sinir ağına dış dünyadan girdilerin geldiği katmandır. Bu katmanda genelde girdiler herhangi bir işleme uğramadan alt katmanlara iletilmektedir. Ara (Gizli) Katman(lar), giriş katmanından çıkan bilgiler bu katmana gelir. Ara katman sayısı ağdan ağa farklılık gösterir. Ara katman hiç bulunmadığı gibi birden fazla da olabilir. Çıkış Katmanı ise ara katmanlardan gelen bilgileri kullanarak ağın çıktıları üreten katmandır. Yapay Sinir Ağlarının eğitilmesi için, öncelikle örnekler toplanır, ağın topolojik yapısı belirlenir, öğrenme parametreleri belirlenir, ağırlıkların başlangıç değerleri atanır, öğrenme setinden örnekler seçilir, öğrenme sırasında ileri hesaplamaların yapılması sağlanır, gerçekleşen çıktının beklenen çıktı ile karşılaştırılması sağlanır, ağırlıkların değiştirilmesi sağlandıktan sonra öğrenme işlemi tamamlanmış olur [19].

2.4.2. Regresyon yöntemi

Gözetimli öğrenme yöntemleri içerisinde en çok kullanılan yöntemlerden birisi de Regresyon algoritmalarıdır. Regresyon algoritmaları ile çözülecek konular, veri setinin düzenli değerlerden meydana geldiği durumlardan oluşmaktadır. Örneğin, bir makinenin çalıştığı saat, kullanıldığı saatler içinde yaptığı üretim sayısına göre ona bir verimlilik puanı hesaplanmak istendiğinde regresyon algoritmalarından yararlanılabilir. Regresyon algoritmaları gözetimli öğrenme kategorisinde olmasının dışında anormal durum tespiti (anomaly detection) gibi durumlarda gözetimsiz öğrenme algoritmaları ile birlikte kullanılabilir. En sık kullanılan Regresyon algoritmaları, Doğrusal Regresyon, Lojistik Regresyon, Çoklu Doğrusal Regresyon, Polinomal Regresyon, Destek Vektör Makinesi Regresyonu ve Karar Ağacı Regresyonu olarak tanımlanabilir.

Doğrusal Regresyon, sayısal verilerin birbirleriyle ilişkisini doğrusal bağlantıyı kullanarak düzlemdeki verinin modelini en verimli bir biçimde elde etmeye çalışan algoritmadır. Lojistik Regresyon, sonucu belirleyen bir veya daha fazla ilişkisiz değişken bulunan veri setini analiz etmek için düzlemde en iyi eğriyi bulmaya çalışan yöntemdir. Sonuçta sadece iki olası sonuç ortaya çıkmaktadır. Çoklu Doğrusal Regresyon, birden fazla tahminci değişken kullanarak tahmin yapan doğrusal regresyon türüdür. Polinomal Regresyon, verilerin birbirleriyle ilişkisi genellikle doğrusal olsa da bazen doğrusal olmadığı için en doğru ilişkiyi bulmak için bir eğri gerekebilir. Polinom fonksiyonları gibi bir terimin herhangi bir üssü alınarak doğrusal olmayan bir model oluşturulmak istendiğinde bu yöntem kullanılabilir [20]. Destek Vektör Makinesi Regresyonu, Destek Vektör Makinesi ile aynı kurallardan yararlanan, hatanın en üst düzeye çıkarıldığı hiper düzlemi bireyselleştirerek hatayı en aza indirgeyen, hatanın bir kısmının göz ardı edildiği algoritmadır. Örnek olarak bir çalışanın yaşına göre maaşını tahmin eden bir model geliştirmek için bu algoritma kullanılabilir. Karar Ağacı Regresyon ise, Sınıflandırma yönteminde karar ağaçları kullanıldığı gibi, regresyon yönteminde de bu yöntemden yararlanılabilmektedir. Bu algoritma, kök düğümden başlayarak, üstten alta olacak biçimde düğümlerin, kendi aralarında benzer değerlerde olanlarının alt kümelerine ayrılmasını sağlayan bir algoritmadır.

2.5. Gözetimsiz Öğrenme

Bu yöntemde eğitim verisi kullanılarak eğitim yapılmaz ve eğitim yapılmayan veriler

arasında ilişkiler bulup kendi aralarında en yakın değerde olan verilerin kendi içinde kümelenmesi mantığıyla çalışmaktadır. Verilerin hangi sınıfta olduğu bilinmeden sınıflandırma yapmak mümkündür. Veriler eğitilir ve daha sonraki veriler de algoritmanın belirlediği alt gruplarla en bağlantılı gruba gönderilerek sınıflandırma yapılır. Gözetimsiz algoritmalar eğitim verisinin hazırlanması ve anlaşılması zor olan durumlarda sıklıkla kullanılmaktadır. Gözetimsiz algoritmalar 3'e ayrılmaktadır. Bunlar: Kümele (Clustering), Birliktelik Kuralı (Association Rule Mining) ve Boyut Azaltma (Dimensionality Reduction) olarak verilebilir.

2.5.1. Kümeleme

Kümeleme (Clustering), veri setinin tümündeki verilerin kendi aralarındaki benzerliklerine göre gruplara ayrılması işlemidir. Her bir grup küme olarak tanımlanabilir. Tüm kümelerde birbirine en yakın veriler olmalı ve aralarında benzerlik olmayan veriler de mümkün olduğunca farklı kümelerde olmalıdır. Sonuç olarak kümelerin arasındaki benzerliğin az olması önemlidir.

2.5.2. Birliktelik kuralı

Birliktelik kuralı (Association rule mining), kümeleme yönteminden farklı olarak değişkenler arası bağlantıları bulmak için kurallar arayan bir yöntemdir. Veriler içindeki önceki hareketleri inceleyip birlikte gerçekleşme durumlarını tespit eden bir veri madenciliği yöntemidir. Elde edilen örüntüler daha sonra yapılacak veri analizlerinde tahmin yapılmasını sağlayabilir. Birliktelik Kuralı ile özellikle e-ticaret sitelerindeki sepet yönetimi ve analizi yapılabilir [21].

Örnek olarak, bu yöntemle e-ticaret sitesinde müşterilerin hangi ürünleri tercih ettiğini tespit edebiliriz. Sepet analizi ile sitenin kazancını ve performansını yükseltmek için her kullanıcı için kişisel ürün tavsiyelerinde bulunulabilir. Kullanıcılar için indirimler ve promosyonlar yapılabilir. Stok güncel tutulup benzer ürünler birbirine yakın gruplanabilir.

Birliktelik Kuralı ile ilgili en sıklıkla kullanılan algoritmalarından bazıları şunlardır; Apriori algoritması, Eclat algoritması ve FP-growth algoritması. İsimleri verilen bu algoritmaların hepsinin amacı veriler arasındaki ilişkiyi ortaya çıkartmaktır. Bu işlemi gerçekleştirmek için

tümü farklı matematiksel yöntemler kullanır. Bu kural ve kümeleme yöntemleri arasındaki en önemli fark şudur. Kümeleme veri noktalarıyla ilgilenirken birliktelik kuralı bu veri noktalarının nitelikleri arasındaki bağlantıları bulmakla ilgilenir.

2.5.3. Boyut azaltma

Boyut azaltma (Dimensionality reduction) işlemi, sahip olunan hedef için en başarılı sonucu verecek olan öznitelikler ile çalışmayı amaçlayan ve kullanılmayan öznitelikleri eksilterek, veri boyutunu azaltan işlemdir. Öznitelik (Feature), Veri setindeki tüm kolonlar bir özniteliktir. Örnek olarak birçok veri setinde kullandığımız niteliklerden bazıları: isim, cinsiyet, yaş, doğum tarihidir.

Daha detaylı olarak boyut azaltma işlemi için şu örnek verilebilir. Bir şirketin personelinin maaş ve özlük bilgilerinin olduğu bir veri seti için boyut azaltma işlemi yapılabilir. Veri seti aşağıdaki gibi alındığında:

- İsim,
- Soy isim,
- Kan grubu,
- Doğum Yeri,
- Doğum Tarihi,
- Yaşı,
- Maaşı,
- Sicil Numarası,
- Derece ve Kademesi

Personelin yaşına ve kademesine göre maaş tespiti yapmak için kademe ve derece bilgileri kullanılacaktır. Fakat personelin yaşına ve kan grubuna ihtiyaç duyulmadığı için bu öznitelikleri veri setinden çıkartarak daha rahat analiz yapılabilir. Bu yöntemin diğer adı “Feature Selection” olarak ifade edilmektedir. Veri boyutu kontrol edilmesi zor bir duruma gelerek büyüdüğünde, gereksiz öznitelikleri tespit etmek gerekmektedir. Kullanılmayan öznitelikleri çıkartmak için PCA (Principal Component Analysis) adı verilen bir algoritma kullanılmaktadır. Boyut azaltma işlemi için kullanılan 3 kavram aşağıda tanımlar halinde verilmiştir.

Öznitelik Çıkarma ve Öznitelik Mühendisliği (Feature Extraction and Feature Engineering): Ham verilerin modelleme yapılabilmesi için gerekli özniteliklere dönüştürülmesi, ham veri içinden kullanılabilen verinin işlenerek kullanılmasıdır. Örnek olarak biyometrik verilerden oluşan parmak izi ve göz retinası bilgileri mevcutsa ve sadece göz retinası verileri kullanılacaksa parmak izi verilerinin kullanılmasına gerek bulunmamaktadır. Veri ham haldeyken parmak izi verileri görmezden gelinebilir [22].

Öznitelik Dönüşümü (Feature Transformation): Veri setini analiz ederken kullanılan algoritmanın kesinlik ve doğruluk oranını artırmak için verilerin dönüştürülmesi işlemidir. Öznitelikler birleştirilerek elde edilebilir ya da mevcut olan öznitelige yeni bir nesne eklenerek yeni öznitelik oluşturulabilir. Bu yöntemlere öznitelik dönüşümü denilmektedir. Dönüştürülen yeni öznitelik, eski özniteliklerin özelliklerini barındırabilir. Fakat dönüşüm sonrası eski özellikleri barındırmak zorunda değildir.

Öznitelik Seçimi (Feature Selection): Gereksiz özelliklerin kaldırılması işlemidir. Tezde R kodları ile öznitelik seçimi yapılmış gereksiz nitelikler silinmiş ve Kredi Durumu (Loan Status) özniteliğine ikili sınıflandırmanın kullanılabilmesi için özellik dönüşümü uygulanmıştır.

2.6. Takviyeli Öğrenme

Bu öğrenme yöntemi daha gözetimli ve gözetimsiz makine öğrenmesi yöntemlerinden daha farklıdır. Algoritmanın temelinde insan bilişsel altyapısı vardır. Takviyeli öğrenme işlemi ile öğrenme süreci dış çevreden alınan geri bildirim ile yapılmaktadır. Bu yöntemle olası durumların, amaç olarak alınıp alınmadığının kontrolü yapılır. Denemeler yapıldığında hedefe ulaşamazsa ceza, ulaşırsa ödül sinyali alınır ve model ceza sinyali aldığı hareketi daha sonra tekrar etmez. Ödül sinyali aldığı deneyimden yararlanıp öğrenmeye devam eder ve hep azami ödülü amaçlayarak düzenli öğrenmeye işlevini sürdürür. Bu yüzden modelde en doğruya ulaşamaz, öğrenme işlemi durdurulmaz ve sürekli devam eder.

2017’de Google’ın geliştirdiği yapay zeka uygulaması AlphaGo, zekaya dayanan Go oyununda dünya şampiyonu Ke Jie’yi mağlup etmişti. Bu uygulamanın öğrenme mantığında takviyeli öğrenme yöntemi bulunmaktadır. Bu yöntemde amaçlanan insan beyninin çalışma sistemine en yakın algoritmalar geliştirmektir. Özellikle, Deep Learning algoritmaları takviyeli öğrenme yöntemleri kullanılarak geliştirilmiştir.

2.7. Makine Öğrenmesi Öğrenme Stratejileri

Makine öğrenmesini öğrenmek için ilk olarak bir programlama dili ya da platformu belirlenmelidir. Programlama dili bilmeyen kişiler için tezde kullanılan Microsoft Windows Azure Machine Learning uygulaması mevcuttur. Bu uygulamada geniş bir veri seti havuzu bulunmaktadır. Programlama yapmadan sürükle bırak ile veri setlerine modeller geliştirilebilir. Geniş veri setlerinin yanında çok sayıda makine öğrenmesi algoritması havuzu da bulunmaktadır. Ayrıca R ve Python dilleri de bu uygulama da kullanılabilir. Bizim çalışmamızda R dili ile verilerin normalizasyonu yapılmış öznitelik dönüşümü ve çıkarımı yapılmıştır.

Azure ML dışında doğrudan R ve Python dillerini de kullanarak makine öğrenmesi yapılabilir. R ve Python dilleri için çeşitli geliştirme arayüzleri mevcuttur. Bu arayüzlerle programlama yapılabilir. Python dilinin kullanılabilmesi için bu dilin kütüphanelerinin kullanım amaçları çok iyi öğrenilmelidir.

Eğer Python dili seçilecekse kullanılacak araç öncelikli olarak seçilmelidir. Bu araçlardan bazılarını şöyle verebiliriz: Jupyter/IPython, PyCharm, Spyder ve Rodeo. Eğer R dili seçilecekse R Studio kullanılabilir. Python daha çok programcılar tarafından seçilirken R dili istatistik uzmanları tarafından seçilmektedir.

2.7.1. R dili

R dili istatistiksel ve matematiksel hesaplamalar yapılan ve grafiksel sonuçlar elde edilebilen bir programlama dilidir. Gün geçtikçe kullanımı artan R özellikle istatistik bilimi ile ilgilenen uzmanlar ve sosyal bilimciler arasında daha yaygın olarak kullanılmaktadır. R dili S dilinin gelişmiş bir türüdür. R Yeni Zelanda'nın Auckland Üniversitesinde 1993 yılında Ross Ihaka ve Robert Gentleman tarafından geliştirilmiştir. Bu kişilerin baş harfleri dilin adına verilmiştir. R dili açık kaynak kodlu bir veri programlama dili olup R Foundation tarafından desteklenmektedir.

3. ARAŞTIRMA YÖNTEMİ

Bu bölümde araştırma yöntemi ve veri seti üzerinde durulmuştur. Araştırmanın kapsamı içerisindeki müşterilerin kredi risklerinin belirlenmesi ve kredi riskine neden olan faktörlerin belirlenmesi amacıyla bu çalışma betimsel tarama modeli niteliğindedir. Betimsel çalışma ve tarama modelinde, betimsel araştırmalar genelde verilen bir durumu açıklamak, değerlendirmeler yapmak ve olaylar arasındaki olası ilişkileri ortaya çıkarmak için yürütülür. Tarama modeli ise çok sayıda elemandan oluşan bir evrende, evren hakkında genel bir yargıya varmak amacıyla, evrenin tümü ya da ondan alınacak bir grup, örnek ya da örneklem üzerinde yapılan tarama düzenlemeleridir. Tezde hem Lending Club veri setinin içinden bir kesit alınarak tarama modeli uygulanmış hem de belirli bir durumu açıklamak ve değerlendirmeler yapmak için veri analizleri yapılarak betimsel çalışma uygulanmıştır.

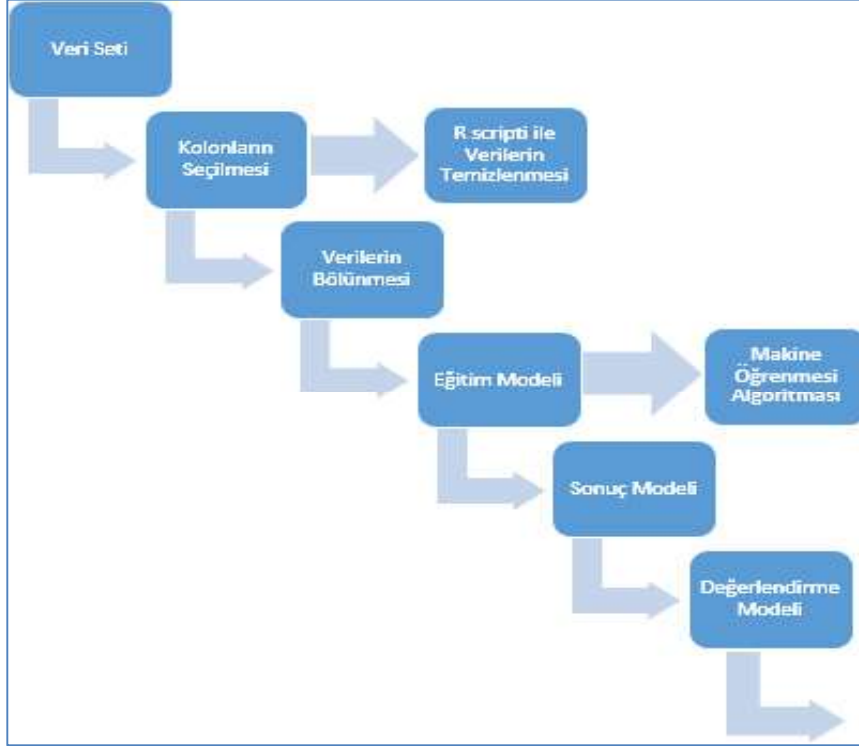
Tezde Lending Club Kredi verileri kullanılmıştır. Bu veriler için Python dili ve ekstra araçlar kullanılarak makine öğrenmesi teknikleri uygulanabilir. Fakat tezde bu yöntemden daha kolay olan aynı zamanda Python ve R dilleri de uygulama üzerinde yazılabilen, makine öğrenmesi algoritmaları sürükleyip bırakılabilen bir platform olan Microsoft Azure Machine Learning platformu tercih edilmiştir. Bu uygulama üzerinde Regresyon ve İkili Sınıflandırma yöntemleri ile veriler değerlendirilmiştir.

3.1. Akış Şeması ve Proje Mimarisi

Öncelikle, Microsoft Azure Machine Learning platformu üzerinde hesap açılmıştır. Hesap açıldıktan sonra Lending Club veri seti indirilerek kullanıma hazır hale getirilmiştir. Uygulama sürükleyip bırak mantığı ile çalışmaktadır. Uygulamanın sol bölümünde kullanılacak olan araçlar mevcuttur. Buradan arama yapılarak istenilen araçlar sürükleyip bırak yöntemiyle çalışma bölümüne getirilmelidir.

Azure ML’de kullanılacak tüm makine öğrenmesi yöntemlerinde ana adımlar şunlardır: 1. Veri setinin çalışmaya alınması, 2. Kolonların seçilmesi (Select columns in dataset), 3. Verilerin bölünmesi (Split Data), 4. Eğitim modeli (Train model), 5. Çalışma modeli (Train Model), 6. Seçilecek Makine Öğrenmesi Modeli (Örn: Doğrusal Regresyon, SVM gibi), 7. Skor modeli (Score Model), 8. Değerlendirme modeli (Evaluate Model), 9. Web Servis

Çıktıları (Web Service Output). Bu adımların özet olarak verildiği bir akış şeması Şekil 4.4'te verilmiştir.



Şekil 3.1. İşlem akış şeması

Bu adımlara verilerin normalizasyonu gibi süreçlerde eklenebilir. Normalizasyon işlemi öznitelikler içinde gereksiz olan ya da örtüşmeyen kolonların temizlenmesi işlemidir ve sıklıkla makine öğrenmesi işlemlerinde ham verinin düzenlenmesi için kullanılmaktadır. Öncelikle veri seti projeye sürükleyip bırak işlemi ile getirilmiştir. Çıkış bölümüyle kolon seçim işleminin yapıldığı bölümün girişi bağlanmıştır.

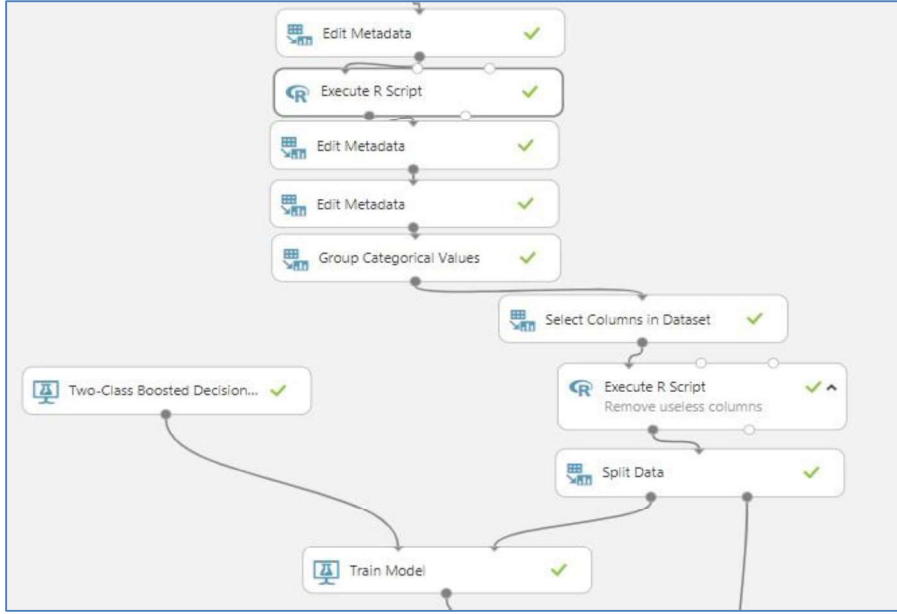
Veri seçimlerinin yapıldığı bölümde kullanılacak olan kolonlar eklenip kullanılmayacak olan kolonlar çıkarılmaktadır. Kolon seçimleri modülünün çıkışı ise veri bölme modülünün girişine bağlanmalıdır. Veri bölme bölümünde mevcut verilerin ne kadarının test için ne kadarının eğitme işlemi için kullanılacağı belirlenmektedir. Bu modül açıldığında Split Data Modülü (Şekil 4.5) ekranı gelecektir. Bu ekrandaki “Splitting mode” seçeneğinde veri bölmesi için Azure ML’de kullanılan teknikler gelmektedir. Tezde Split Rows tekniğini kullanılmıştır. Bu teknikte verinin ne kadarının bölüneceği belirlenebilmektedir. Şekil 4.5’te

“Fraction of rows in the first output dataset“ seçeneğinde satırların %50’sinin bölündüğü örnek olarak gösterilmiştir. Fakat biz projemizde bu oranı %80 (eğitim) ve %20 (test) olarak kullanılmıştır. Rastgele tohum (Random seed) işleminde ise rastgele sayı üretme işlemi kullanılacaksa buraya bir değer girilip rastgele sayı üretimi kullanılabilir. Ayrıca, şekil 4.5’te çalışmaya başlama, bitiş tarihleri, çalışma süresi ve durumları hakkında bilgi verilmektedir. Verilerin içinde bulunan boş değerler, mükerrer kayıtlar ve çok büyük rakamlardan oluşan hatalı girilmiş değerler R kodlarıyla temizlenmiştir. Borç durumu özniteliği makine öğrenmesi için doğrudan kullanılmıştır. Bu kolona regresyon ve ikili sınıflı sınıflandırma yöntemleri uygulanıp yöntemlerin başarımları karşılaştırmalı olarak gözlemlenmiştir.

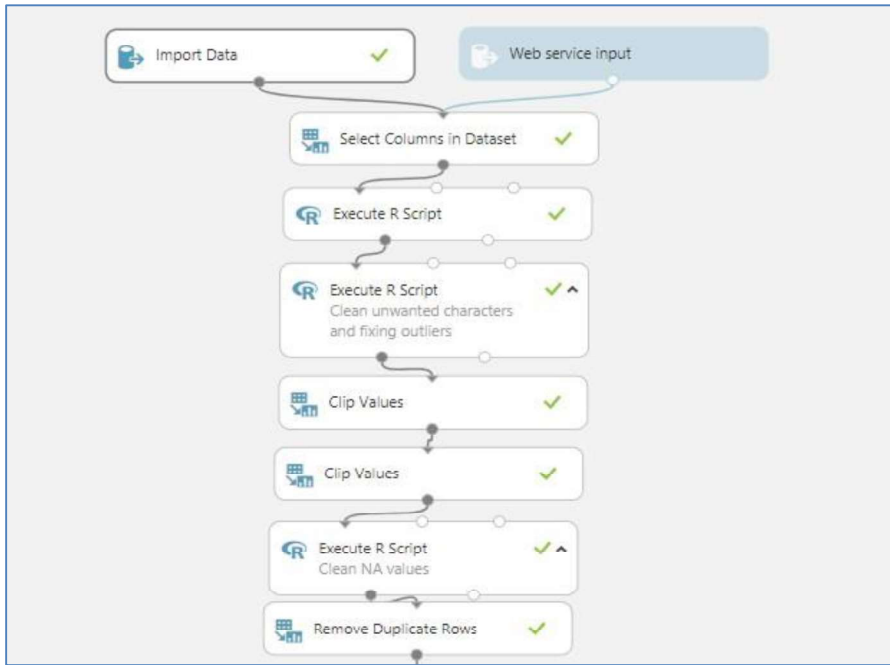
Borç durumu kolonunda 2’den fazla değer bulunmaktadır. Tamamen ödenen krediler (iyi krediler) “0” olarak tanımlanırken geç ödenen ve temerrüte düşen krediler (kötü krediler) “1” olarak tanımlanmıştır. Bu işlemin yapılmasının nedeni ikili sınıflı sınıflandırma yöntemlerinin ikili(binary) veri yapısıyla kullanılmasıdır. İkili veri yapısını oluşturmak için R scriptleri kullanılmıştır. R kodları ile Borç Durumu kolonundaki değerler 0 ve 1’e dönüştürülmüştür.

Eğitim seti tüm algoritmalarda toplam verinin %80’dir. Test verisi ise %20’ye karşılık gelmektedir. Sonuç modelinde gerçek değerlerin bulunduğu 0 ve 1’lerden oluşan borç durumu kolonu ile makine öğrenmesi yöntemleriyle elde edilen sonuçların bulunduğu tahmin kolonu yan yana getirilmiştir. Değerlendirme kısmında ise görselleştirme yapıp makine öğrenmesi algoritmalarının başarı oranları ölçülmüştür.

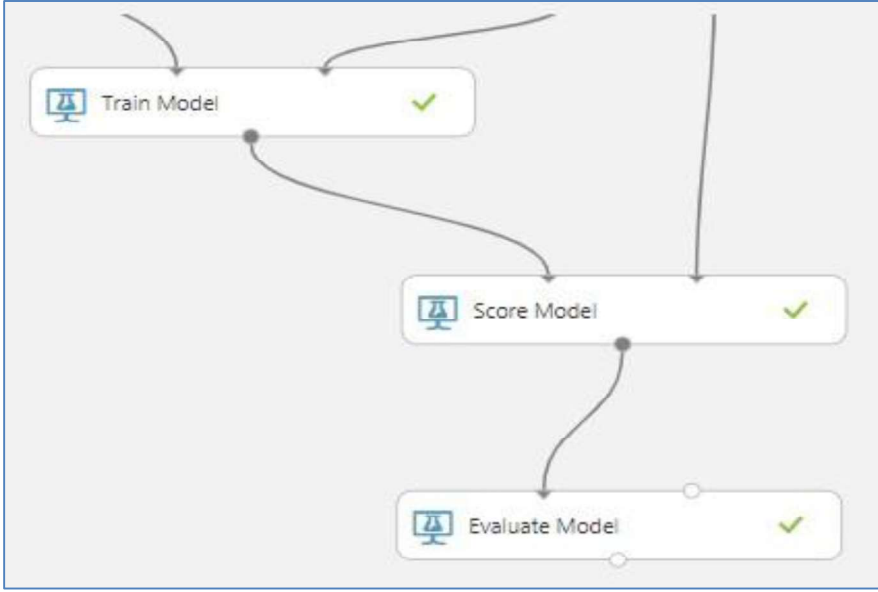
Tüm adımların gösterildiği mimari Şekil 3.2, 3.3 ve 3.4'deki şekillerde gösterilmiştir.



Şekil 3.2. Proje detay mimarisi üst kısım



Şekil 3.3. Proje detay mimarisi orta kısım



Şekil 3.4. Proje detay mimarisi alt kısım

Kullanılan yöntemler Regresyon ve İki Sınıflı Sınıflandırma yöntemleridir. Sınıflandırma ve Regresyon yöntemlerinin içinden alt algoritmalar seçilmiştir. Tezin makine öğrenmesi kısmında regresyon ve ikili sınıflandırma yöntemleri tercih edilmiştir. Tezin ek kısmında verilerin normalize edilmesi için kullanılan R kodları verilmiştir. Bu kodlar ile verilerin içinde bulunan boş değerler, mükerrer kayıtlar ve çok büyük rakamlardan oluşan hatalı girilmiş değerler R dili kullanılarak temizlenmiştir.

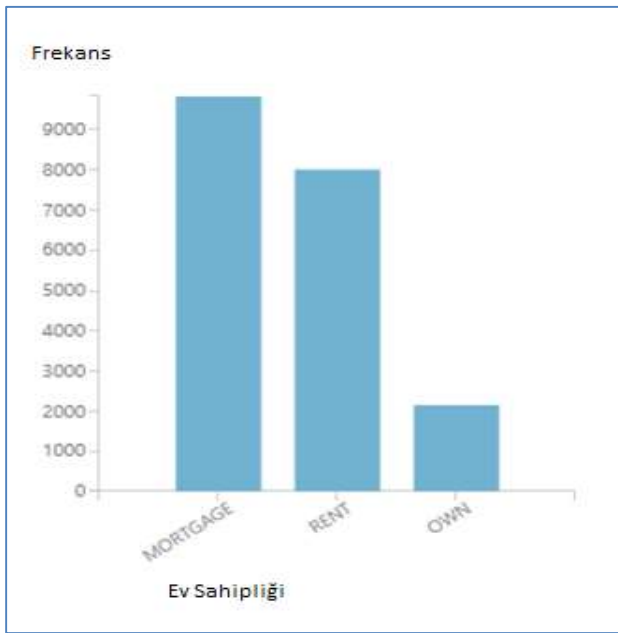
3.2. Veri Seti

Betimsel araştırmaların temel amacı mevcut bir duruma açıklık getirmek, değerlendirmelerde bulunmak ve olaylar arasında var olan ilişkileri tespit etmektir. Evren hakkında genel bir yargıya ulaşmak için bir evren ya da evrenden alınacak bir örneklem üzerinde yapılan çalışmalar tarama modeli kapsamında değerlendirilmektedir. Bu nedenle tüm Lending Club veri setinden bir kesit alınarak çalışmada kullanılmıştır.

Veri setinde kullanılan kolonlar loan status (kredinin ödenme durumu), issue date (kredi yayın tarihi), loan amount (kredi miktarı), term (borç vadesi), int rate (faiz oranı), installment (ilk ödeme tutarı), last payment amount (son ödenen tutar), loan purpose (borç amacı), loan title (borç başlığı), emp title (borç alan kişinin mesleği), home ownership (ev durumu), zip code (bölge veya ülke id'si) gibidir. Özellikle borç durumu kolonu makine öğrenmesi için

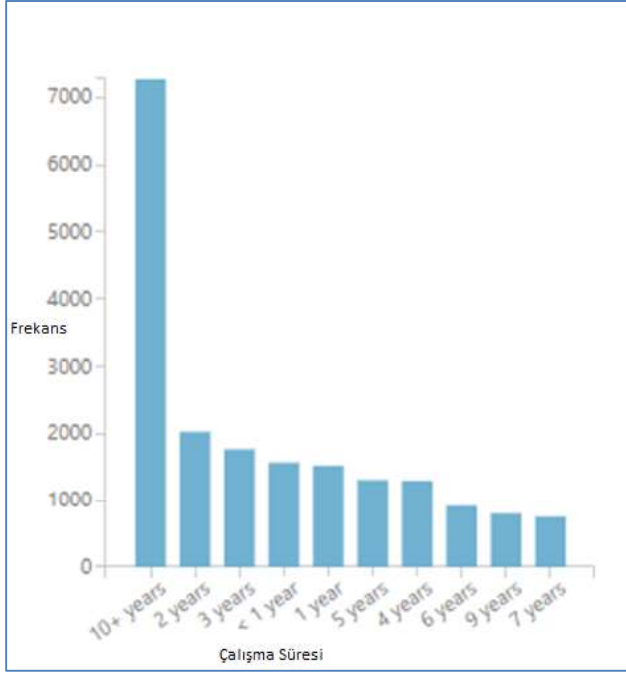
doğrudan kullanılmıştır. Bu bölümün devamında veri setindeki niteliklerin tanımları ve kapsamaları verilmiştir.

Ev Sahibi Olma (Home Ownership): Veri setindeki bu nitelik kayıtlı kullanıcıların sahip oldukları evlerin durumunu göstermektedir. Kişilerin çoğunluğu Mortgage kredisi ile ev almışken daha az olan kısmının kiralık evde oturduğu görülmektedir. En az olan kısım ise kendi evine sahip olanlardır. Bu verilerden birçok kişinin ev kredisi için sistemden yararlandığı ya da yararlanmak isteyeceği çıkarımı yapılabilir.



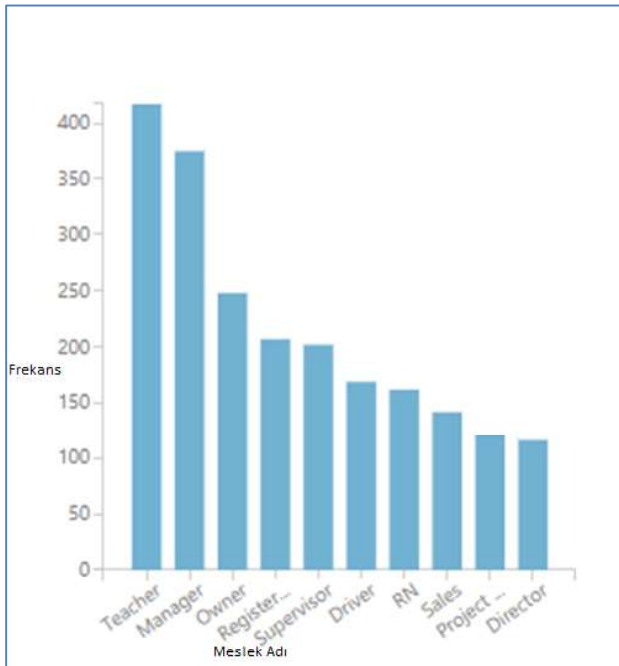
Şekil 3.5. Ev sahipliği niteliği

Çalışma süresi (Employee Length): Veri setindeki bu nitelik kayıtlı kullanıcıların çalışma sürelerini göstermektedir. Sistemdeki kişilerin büyük çoğunluğunun 10 yıldan daha fazladır iş hayatında olduğu görülmektedir. 1 yıl ve daha az süre çalışanların veri setinde azınlık oldukları görülebilir.



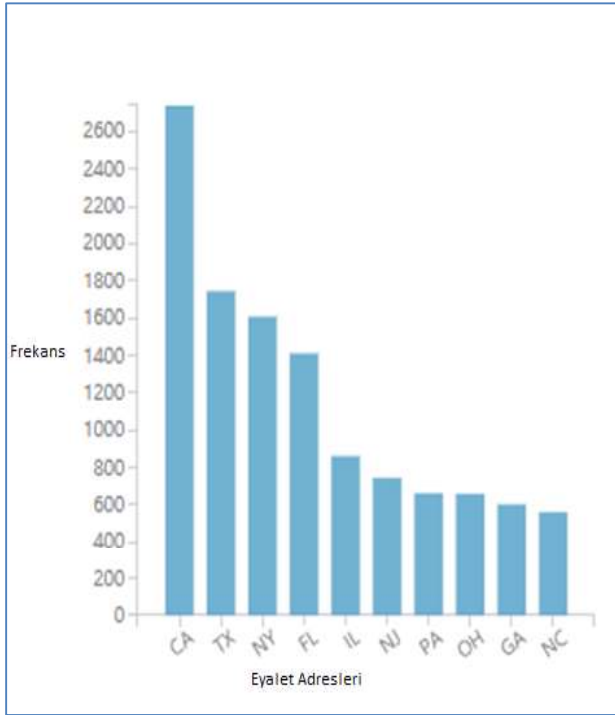
Şekil 3.6. Çalışma süresi niteliği

Meslek Türü (Employee Title): Veri setindeki bu nitelik kayıtlı kullanıcıların mesleklerini göstermektedir. Sistemdeki kişilerin en fazlası öğretmen olduğu en azının ise yönetici olduğu görülmektedir.



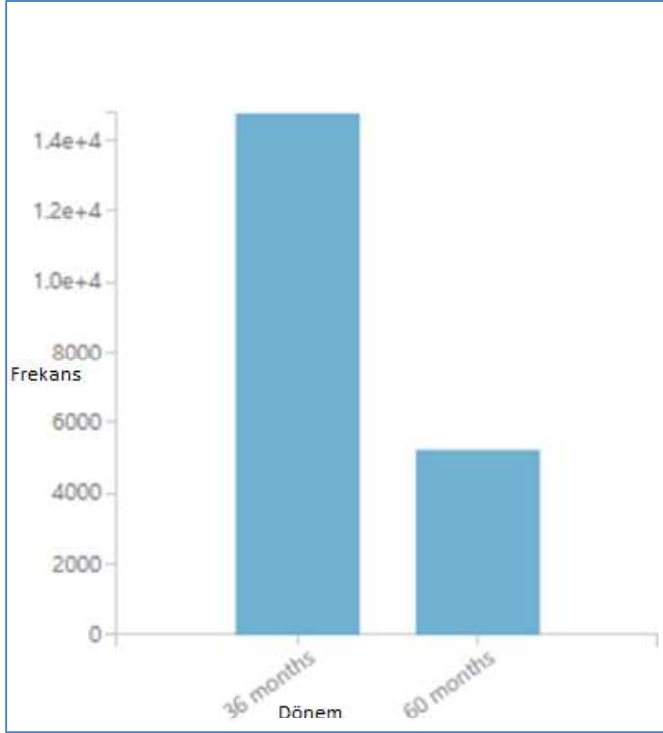
Şekil 3.7. Meslek adı niteliği

Eyalet Adresleri (State Addresses) : Veri setindeki bu nitelik kayıtlı kullanıcıların yaşadıkları eyaletleri göstermektedir. Sistemdeki kişilerin en fazlasının California daha sonra sırasıyla Texas ve New York'ta yaşadığı en azının ise North Carolina'da yaşadığı görülmektedir. Tezdeki genel veri analizlerinde de elde ettiğimiz gibi ABD'nin batı ve güney eyaletlerinin ekonomik gücünün yüksek olduğu özellikle kuzey eyaletlerinin ekonomik gücünün daha zayıf olduğu sisteme kayıtlı kullanıcıların yaşadığı yerlerden de görülebilir.



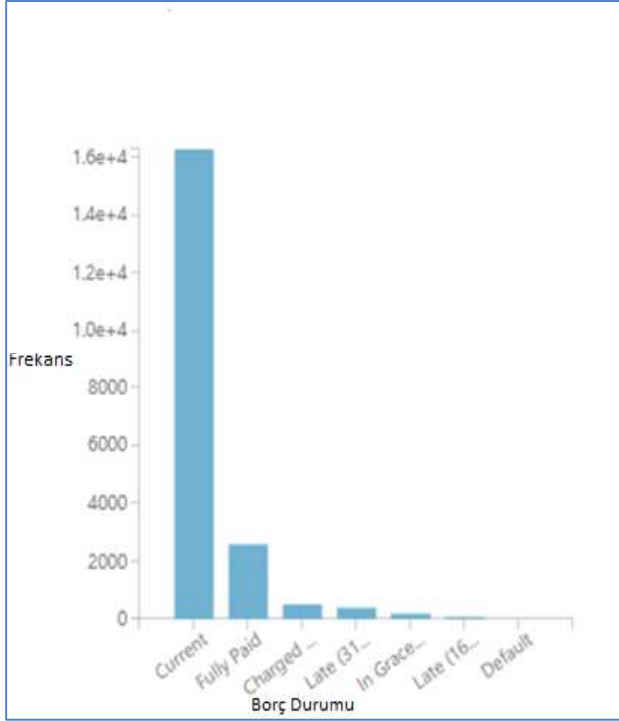
Şekil 3.8. Eyalet adresleri niteliği

Dönem (Term): Veri setindeki bu nitelik kayıtlı kullanıcıların kredi ödeme vadelerini göstermektedir. Sistemdeki kişilerin çoğunluğunun 36 ay vade ile diğer kısmının da 60 ay vade ile borçlandığı görülmektedir.



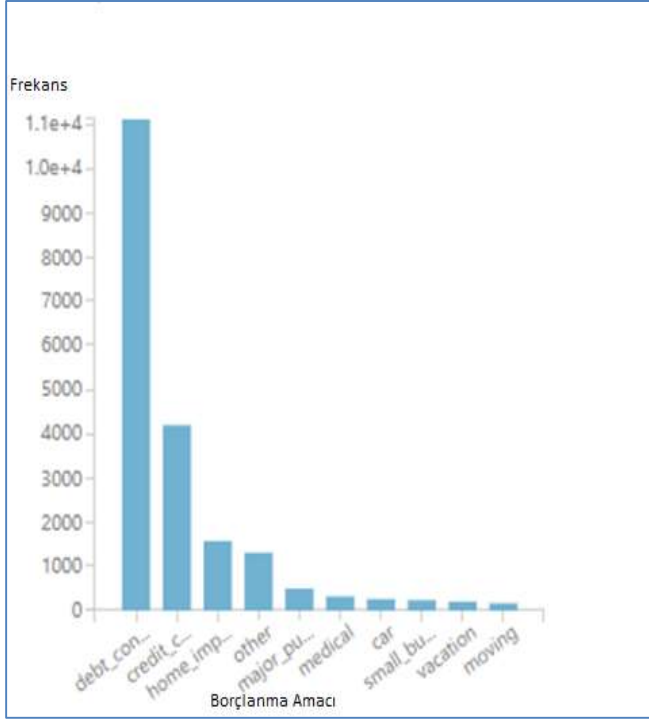
Şekil 3.9. Dönem niteliği

Borçlanma Durumu (Loan Status): Veri setindeki bu nitelik kayıtlı kullanıcıların kredi ödeme durumlarını göstermektedir. Sistemdeki kişilerin çoğunluğunun mevcut kredi olduğu sonra sırasıyla tam ödeme ve geç ödemelerin geldiği görülebilir. Bu nitelik tezimiz için önem arz etmektedir. Makine öğrenmesi algoritmaları bu niteliğe uygulanmıştır. İkili sınıflandırma algoritmalarının uygulanması için bu nitelik iki kısma ayrılmıştır. Geç ödenen ve ödenmeyen krediler kötü kredi olarak alınırken zamanında ödenen krediler ise iyi kredi olarak alınmıştır. Kötü ve iyi kredilerin farklı niteliklerle ilişkileri verilerin genel analizinde incelenmiştir.



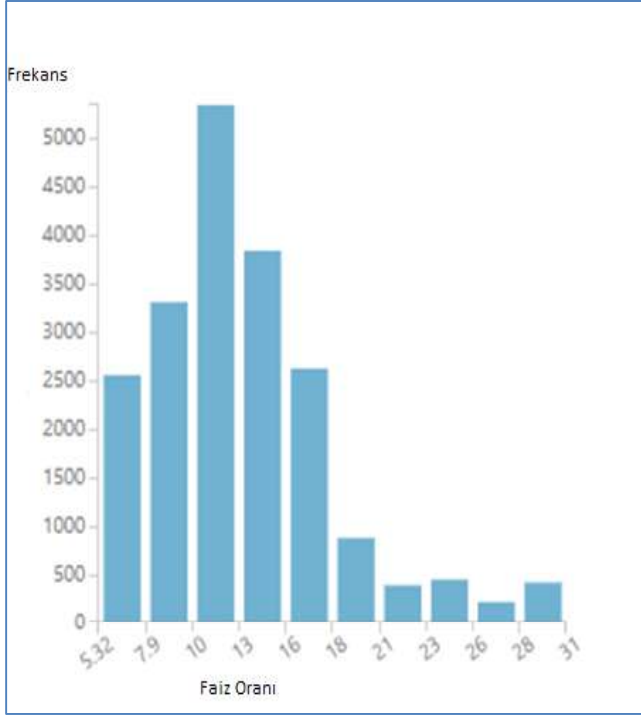
Şekil 3.10. Borçlanma durumu niteliği

Borçlanma Amacı (Loan Purpose): Veri setindeki bu nitelik kayıtlı kullanıcıların kredi amaçlarını göstermektedir. Sistemdeki kişilerin çoğunluğunun debt consoloditaion (destek- ihtiyaç kredisi) olduğu görülmektedir. Bu borç tipini kredi kartı ve ev tadilatı amacıyla çekilen kredilerin takip ettiği görülmektedir. En az değerin ise taşınma ve seyahat için alınan krediler olduğu ifade edilebilir.



Şekil 3.11. Borçlanma amacı niteliği

Faiz Oranı (Interest Rate): Şekil 3.12’de görülebileceği gibi veri setindeki bu nitelik faiz oranlarını göstermektedir. Sistemdeki faiz oranlarının büyük çoğunluğunun 10 ve 13 arasında olduğu görülmektedir. En az faiz oranının ise 20 üzerindeki faiz oranları olduğu görülmektedir.



Şekil 3.12. Faiz oranı niteliği

Düşük FICO Değeri (FICO Range Low): Şekil 3.13'te görülebileceği gibi veri setindeki bu nitelik FICO puanlarını göstermektedir. FICO, firmalara ve tüketicilere ürün ve hizmet sağlayan yazılım şirkettir. Fair Isaac Corporation olarak bilinen şirket, 2009 yılında adını FICO olarak değiştirmiştir. Dünyada en çok kullanılan ve tanınan tüketici kredisi puanlarını üretmesiyle tanınır. Tüm dünyadaki bankalar bu puanı kullanmaktadır ve birçok banka tüketicilere risk durumlarını talep etmeleri takdirde vermektedir.

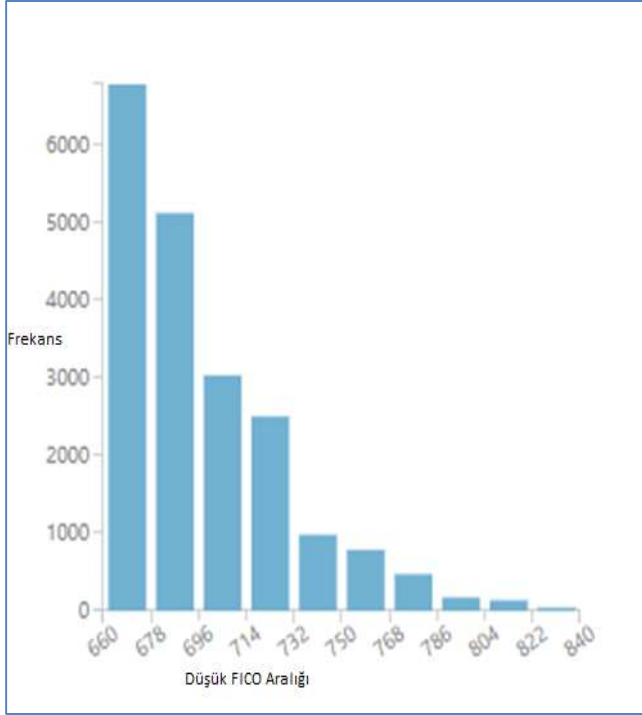
FICO puanları 300 ile 850 arasında bulunmaktadır [23]. Genellikle, 650'nin üzerindeki puanlar çok iyi bir kredi geçmişini göstermektedir. Diğer yandan, 620'nin altındaki puanları olan kişiler, istedikleri faiz oranlarında kredileri genellikle zor bulmaktadır. Bankalar kredi verecekleri zaman bir tüketicinin FICO puanını dikkate alırlar, fakat aynı anda borçlunun çalışma süresi ve istenen kredi türü gibi ve tüketicinin geliri gibi diğer detayları da göz önünde bulundururlar.

Çizelge 3.1. FICO değerlendirmeleri

FICO Değerlendirmeleri	Derece
<580	Kötü
580-669	Orta
670-739	İyi
740-799	Çok İyi
800+	Mükemmel

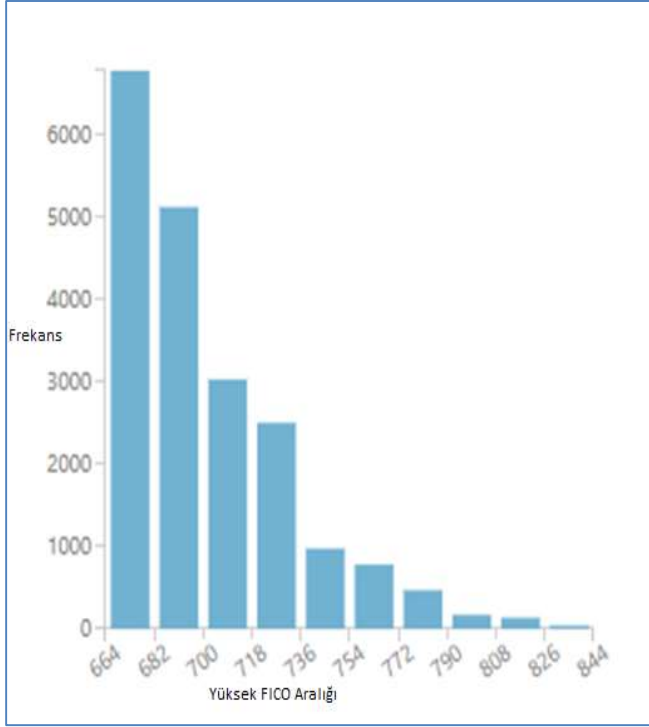
Bu Şekilde FICO puanlarının hangi aralıklarda ne anlama geldikleri verilmiştir. 580 puan altının riskli bir kullanıcı olduğu görülmektedir. Bu kullanıcılar bankalardan kredi alacakları zaman yüksek faizle borçlanmaktadır. 670 üzerindeki puanlar ise iyi olarak belirtilen ve kredi bulma açısından kolay olan puanlardır. Daha sonra da vereceğimiz üzere sistemimizdeki FICO puanlarının çoğunluğu 660 puanın üzerinde olup çoğunluğu iyi FICO puanına sahip kullanıcılardan oluşmaktadır.

FICO puanları hesaplanırken kişinin ödeme geçmişi, borç miktarı, kredi geçmişi uzunluğu, kredi karması (kredi kartı, taşıt kredileri ve ipotekler gibi çeşitli ödemeler) ve son olarak da yeni krediler (yeni krediler ve yeni açılan hesaplar riski artırır ve puanı düşürür) kullanılır. Bu değerler belirli katsayılarla çarpılarak FICO puanı hesaplanır. FICO puanlarının çeşitli türleri bulunmaktadır. FICO'nun genel tanımından sonra veri yapımızdaki düşük FICO puanlarını incelediğimizde, sistemdeki en düşük FICO puanlarının büyük çoğunluğunun 660 ve 696 arasında olduğu görülmektedir. 800 puan üzerinde ise az sayıda kişi yer almaktadır. 600 puanın altında puan bulunmamaktadır.



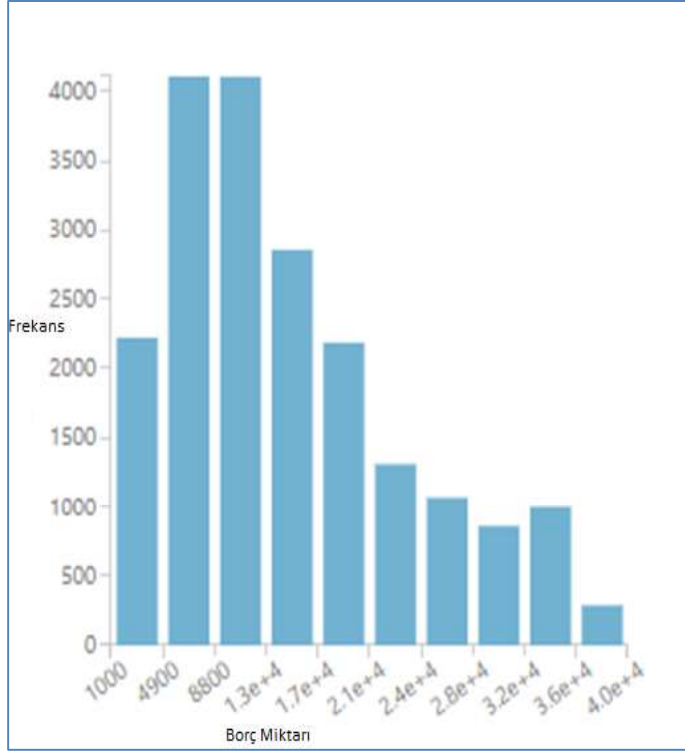
Şekil 3.13. Düşük FICO değerlendirme niteliği

Yüksek FICO Değeri (FICO Range High): Şekil 3.14'te görülebileceği gibi sistemdeki en yüksek FICO puanlarının büyük çoğunluğunun 664 ve 700 arasında olduğu görülmektedir. 800 puan üzerinde ise az sayıda kişi yer almaktadır. 600 puanın altında puan bulunmamaktadır.



Şekil 3.14. Yüksek FICO değerlendirme niteliği

Borç Miktarı (Loan Amount): Veri setindeki bu nitelik borç olarak verilen kredi miktarlarını göstermektedir. Sistemdeki kredi tutarlarının büyük çoğunluğunun 5000 ve 13000 arasında olduğu görülmektedir. En az kredi miktarının ise 20000 üzerindeki rakamlar olduğu görülmektedir.



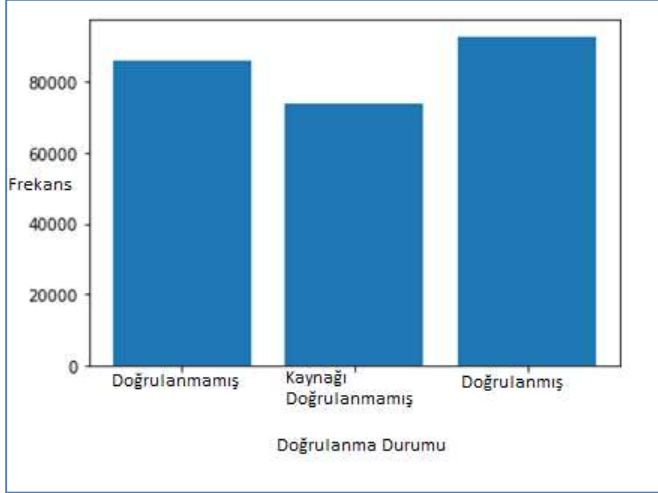
Şekil 3.15. Borç miktarı niteliği

Çizelge 3.2’de verilen kredi tutarlarının ortalamasının 14597 olduğu görülmektedir. Medyan’ın 12000, en az değerin 1000, en yüksek değerin 40000 ve standart sapma değerinin 9254,3418 olduğu verilmiştir. Aynı zamanda 1094 adet farklı kredi tutarının olduğunu ve kredi tutarı olmayan kayıp değerin 0 adet olduğunu ifade edebiliriz.

Çizelge 3.2. Yüksek FICO değeri niteliği veri dağılımı

İstatistikler	
Ortalama	14597.7726
Median	12000
Min	1000
Max	40000
Standart Sapma	9254.3418
Tek Değerler	1094

Verification Status (Doğrulama Durumu): Veri setindeki bu nitelik doğrulanan ve kaynakları doğrulanan kredilerin sayıca doğrulanamayan kredilerden daha fazla olduğu görülmektedir. Bu nitelik çalışılan veri setinin çoğunluğunun kaynaklı ve doğru verilerden oluştuğunu göstermiştir.



Şekil 3.16. Doğrulama durumu niteliği

Bahsedilen bu nitelikler Lending Club veri setini özetlemektedir. Bu niteliklerin neyi ifade ettiği tez amacında belirlenen kredi riski modeli oluşturma ve veri analizi yapma konusunda önem arz etmektedir. Özellikle, borç durumu kolonu makine öğrenmesi ve verilerin genel analizinde kullanılmıştır. Diğer niteliklerden ise veriler üzerinde yapılan analizlerde yararlanılmıştır.

4. BULGULAR VE DEĞERLENDİRME

Bu bölümde tezde elde edilen bulgular değerlendirilmiştir. Bulguların değerlendirilmesinde veri değerlendirme ölçütleri belirlenmiş ve bu ölçütlerle ilişkili olarak her bir yöntemin başarımı değerlendirilmiştir. Değerlendirme sonuçları özet olarak bir tablo halinde sunulmuştur.

4.1. Veri Değerlendirme Ölçütleri

Bu bölümde regresyon ve sınıflandırma yöntemlerin başarı oranlarını belirleyen metrik değerleri verilmiştir.

Ortalama Mutlak Hata (Mean Absolute Error): Regresyon algoritmalarından elde edilen hata (error) değerleri modelin ne kadar yanıldığını yani gerçek değerlerle tahminlenen değerler arasındaki farkları gösterir. MAE, bu değerler arası farkların mutlak değerlerinin ortalamasını göstermektedir, bu değer ne kadar 0'a yakın olursa modelin o kadar başarılı olduğunu söylenebilir.

Karesel Ortalama Hata (Root Mean Squared Error): Karesel Ortalama Hata tahmin edilen değerlerle gerçek değerler arasındaki farkların karelerinin ortalamasının kareköküdür. Bu hata değerlerinin sıfıra yaklaşması modelin daha az hata yaptığını gösterir.

Determinasyon Katsayısı (Coefficient of Determination): Determinasyon katsayısı ise tahmin edilen değerlerle gerçek değerler arasındaki korelasyonun karesidir. 0'a yaklaşması aradaki korelasyonun zayıf olduğunu, 1'e yaklaşması ise güçlü olduğunu gösterir. Regresyon yöntemlerinden sonra ikili sınıflı sınıflandırma yöntemleri veri setine uygulandığında elde edilen sonuçlar incelenmiştir. Regresyon yöntemlerinde başarı oranını hata değerleri gösterirken ikili sınıflı sınıflandırma yöntemlerinde ise doğruluk ve kesinlik değerleri başarı oranını göstermektedir. İkili sınıflı sınıflandırma algoritmalarının ölçütleri Karışıklık Matrisindeki “True Positive”, “False Positive”, “False Negative” ve “True Negative” değerleri kullanılarak hesaplanır.

Karışıklık Matrisi (Confusion Matrix): Karışıklık matrisi sınıflandırma makine öğrenmesi yöntemlerinde elde edilen doğruluk ve yanlışlık sonuçlarının gösterildiği çizelgedir.

Çizelge 4.1’de gösterilmiştir. Bu çizelgedeki değerler sırasıyla “True Positive”, “True Negative”, “False Pozitif”, “False Negative” şeklindedir.

Çizelge 4.1. Karışıklık matrisi

Gerçek/Tahmin Edilen	Negatif	Pozitif
Negatif	998	0
Pozitif	1	1

Gerçek Pozitifler (TP): Bu değer doğru tahmin edilen pozitif değerlerdir, gerçek sınıfın değerinin olumlu olduğu ve tahmin edilen sınıfın değerinin de olumlu olduğu anlamına gelir.

Gerçek Negatifler (TN): Bu değer doğru tahmin edilen negatif değerlerdir, yani gerçek sınıfın değeri olumsuz ve tahmin edilen sınıfın değeri de olumsuzdur. Yanlış pozitifler ve yanlış negatifler, bu değerler gerçek sınıfınız öngörülen sınıfa aykırı olduğunda ortaya çıkar.

Yanlış Pozitifler (FP): Gerçek sınıfın olumsuz olması durumunda ve tahmin edilen sınıfın olumlu olması durumunda bu değer elde edilir.

Yanlış Negatifler (FN): Gerçek sınıf olumlu olduğunda ve olumsuz olarak tahmin edilen sınıf olduğunda bu değer elde edilmektedir.

Bu dört parametre anlaşıldıktan sonra bu değerlerin kullanıldığı Doğruluk, Hassaslık, Geri Çağırma ve F1 puanlarının hesaplanması formülleri ile verilmiştir.

Doğruluk (Accuracy): Doğruluk en sezgisel performans ölçütüdür. Doğru bir Şekilde gözlemlenmiş gözlemin toplam gözlemlere oranı olup Eş. 4.1 ile ifade edilmektedir.

$$\text{Doğruluk} = \frac{GP + GN}{GP + YP + YN + GN} \quad (4.1)$$

Kesinlik (Precision): Doğru tahmin edilmiş olumlu gözlemlerin, tahmin edilen toplam pozitif gözlemlere oranıdır. Doğru tahmin edilmiş pozitif gözlemlerin, toplam pozitif gözlemlere oranıdır. Eş. 4.2 ile ifade edilmektedir.

$$\text{Kesinlik} = \text{GP} / \text{GP} + \text{YP} \quad (4.2)$$

Duyarlılık (Recall): Hatırlama, doğru tahmin edilmiş pozitif gözlemlerin gerçek sınıftaki tüm gözlemlere oranıdır. Duyarlılık Eş. 5.3'te ifade edilmektedir.

$$\text{Duyarlılık} = \text{GP} / \text{GP} + \text{YN} \quad (4.3)$$

AUC (Area Under Curves): Eğrinin altında kalan alanı ifade eder. 1'e yaklaşması başarı oranını artırır.

F1 Score: Kesinlik ve Duyarlılığın ağırlıklı ortalamasıdır. Bu nedenle, bu puan hem yanlış pozitifleri hem de yanlış negatifleri hesaba katmaktadır.

4.2. Bulgular

Tezde 8 farklı makine öğrenmesi (Regresyon yöntemlerinden; Doğrusal, Karar Ormanı, Yapay Sinir Ağı ve Artırılmış Karar Ağacı kullanılırken İkili Sınıflı Sınıflandırma yöntemlerinden ise İkili Sınıflı Yerel Derin Destek Vektör Makinesi, İkili Sınıflandırma Karar Ormanı, İkili Sınıflı Yapay Sinir Ağları ve İkili Sınıflandırma Artırılmış Karar Ağacı) algoritması kullanılmıştır. Bu bölümde bu algoritmaların sonuçları incelenmiş ve hangilerinin daha başarılı olduğu değerlendirilmiştir.

4.2.1. Doğrusal regresyon

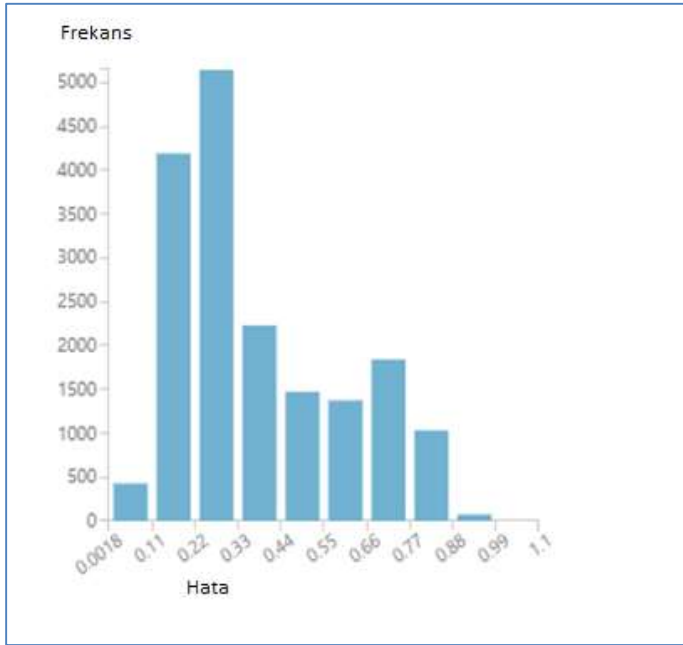
Doğrusal Regresyon (Lineer Regression), sayısal girdi ve çıktılar arasındaki doğrusal ilişkiyi tespit etmeyi amaçlayan ve verinin modelini doğrusal olarak çıkartmaya çalışan yöntemdir.

Şekil 4.1 Azure ML'den alınmıştır. Şekil 4.1'deki sonuçlara bakıldığında, şekildeki hata değerleri modelin ne kadar yanlış olduğunu yani gerçek değerlerle tahminlenen değerler arasındaki farkların ne olduğunu göstermektedir. MAE ve RMSE değerleri sıfırdan uzak olarak görülebilir. Dolayısıyla doğrusal regresyondan daha başarılı yöntemleri aramalıyız. Azure ML parametre seçenekleri üzerinde Doğrusal Regresyon için kullanılan parametreler

için Çözüm metodu olarak Sıralı En Küçük Kareler Yöntemi tercih edilmiş, L2 düzenleme ağırlığı olarak 0.001 seçilmiş ve rastgele sayı üretme verisi seçilmemiştir.

Metrikler	
Mean Absolute Error	0.382989
Root Mean Squared Error	0.437661
Relative Absolute Error	0.943449
Relative Squared Error	0.943708
Coefficient of Determination	0.056292

Şekil 4.1. Doğrusal regresyon hata değerleri



Şekil 4.2. Doğrusal regresyon hata değerleri dağılımı

4.2.2. Karar ormanı regresyonu

Kök düğümden başlayan düğümler ile verilerin kendi içlerinde benzer değerlere sahip olanlarını karar ağaçlarını kullanarak alt kümelerine ayıran regresyon yöntemine Karar Ormanı Regresyonu(Decision Forest Regression) denir.

Azure ML parametre seçenekleri üzerinde Karar Ormanı Regresyonu için kullanılan parametreler: Yeniden Örneklemeye metodu olarak Torbalama(Bagging) yöntemi tercih edilmiş, Eğitim modu olarak tek parametre, karar ağacı sayısı olarak 8, karar ağaçlarının maksimum derinliği 32, düğüm başına rastgele bölünme sayısı 128, yaprak düğümü başına asgari örnek sayısı 1 olarak seçilmiştir.

rows	columns
1	6
	Negative Log Likelihood
	Mean Absolute Error
	Root Mean Squared Error
	Relative Absolute Error
	Relative Squared Error
	Coefficient of Determination
view as	
	1603289899.071394
	0.300879
	0.409639
	0.741165
	0.826716
	0.173284

Şekil 4.3. Karar ormanı regresyonu hata değerleri dağılımı

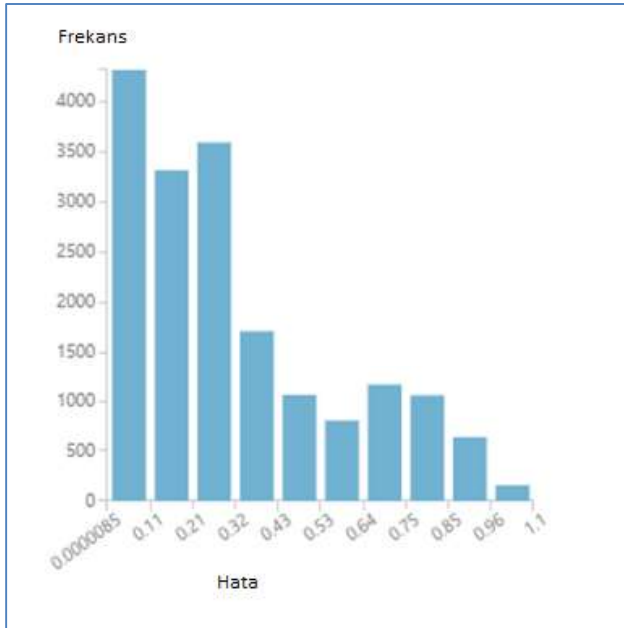
Şekil 4.3'ü incelediğimizde MAE ve RMSE hata değerleri doğrusal regresyon yöntemine göre daha başarılı sonuç vermiştir.

4.2.3. Yapay sinir ağı regresyonu

Yapay sinir ağları insan beyninden esinlenerek geliştirilmiş, ağırlıklı bağlantıları kullanarak birbirine bağlanan ve her biri kendi belleğine sahip olarak dağıtık bir biçimde çalışan yapılardır. Azure ML parametre seçenekleri üzerinde Yapay Sinir Ağları Regresyonu için kullanılan parametreler: Eğitim modu olarak tek parametre, gizli katman tanımlaması olarak tamamı bağlantılı, gizli düğüm sayısı olarak 100, öğrenme oranı 0.005, öğrenme yineleme sayısı 100, başlangıç öğrenme ağırlıklarının çapı 0.1, momentum değeri 0, normalleştirme türü Min-Max normalleştirme olarak seçilmiştir. Azure ML parametre ekranında üzerinde normalleştirme türleri arasında Binning(Gruplama), Gaussian ve normalleşme türü olmayan seçenekleri de mevcuttur. Projenin sonuç modülüne Yapay Sinir Ağı Regresyonu (Neural Network Regression) algoritması bağlayarak elde ettiğimiz sonuçları inceleyelim. Değerlendirme modülünün çıktısındaki görselleştirme sekmesinden Şekil 4.4 ve 4.5'deki sonuçları alabiliriz. Şekil 4.4'te metrik değerleri gösterilmiştir.

Metrikler	
Mean Absolute Error	0.319787
Root Mean Squared Error	0.411761
Relative Absolute Error	0.787758
Relative Squared Error	0.835319
Coefficient of Determination	0.164681

Şekil 4.4. Yapay sinir ağı regresyonu hata değerleri



Şekil 4.5. Yapay sinir ağı regresyonu hata değerleri dağılımı

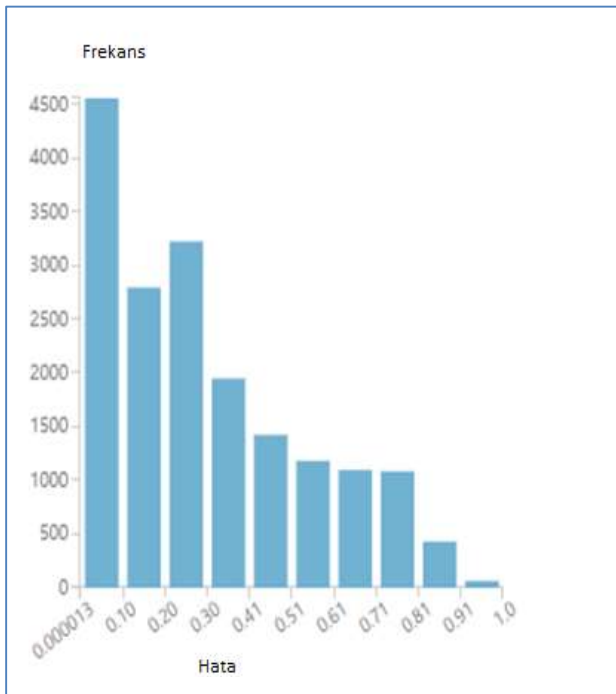
4.2.4. Artırılmış karar ağacı regresyonu

Ağaç içindeki zayıf düğümleri belirli kurallar çerçevesinde Artırılmış(Boosting) algoritmalarıyla birleştirerek güçlü ağaçlar oluşturan karar ağacı türüdür. Azure ML parametre seçenekleri üzerinde Artırılmış Karar Ağacı Regresyonu için kullanılan parametreler: Eğitim modu olarak tek parametre, ağaç başına maksimum yaprak sayısı 20, yaprak düğümü başına minimum örnek sayısı 10, öğrenme oranı 0.2, toplam inşa edilen ağaç sayısı 100 olarak seçilmiş, rastgele sayı üretimi ise seçilmemiştir. Veri setinde Artırılmış Karar Ağaçları Regresyonu (Boosted Decision Tree Regression) denenmiştir. Şekil 4.7’de

MAE ve RMSE değerleri daha önceden denenilen tüm regresyonlara göre 0'a daha yakın olduğu için daha başarılı sonuç vermiştir. CoD değeri diğer yöntemlere göre 1'e yakın bir sonuç vermiştir. Bu değer 1'e yakın olması modelin başarısını göstermektedir.

Metrikler	
Mean Absolute Error	0.301801
Root Mean Squared Error	0.38679
Relative Absolute Error	0.743451
Relative Squared Error	0.737076
Coefficient of Determination	0.262924

Şekil 4.6. Artırılmış karar ağacı regresyonu yönteminin hata değerleri

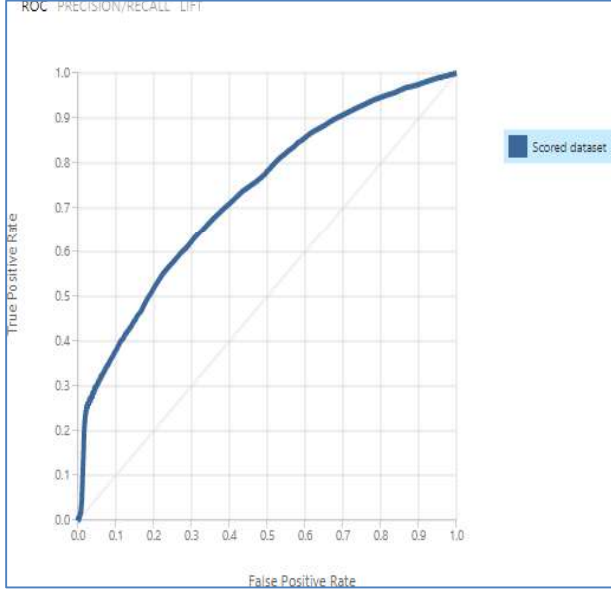


Şekil 4.7. Artırılmış karar ağacı regresyonu yönteminin hata değerleri grafiği

4.2.5. İki sınıflı yerel derin destek vektör makinesi

İkili sınıfa ayrılmış verilerde veri setinde birbirine benzeyen gruplar arasına birbirinden en uzak olan noktalardan sınırlar çizen algoritmadır. Azure ML parametre seçenekleri üzerinde

İkili Sınıflı Yerel Derin Destek Vektör Makinesi için kullanılan parametreler: Eğitim modu olarak tek parametre, ağaç derinliği 3, Lambda W 0.1, Lambda Theta 0.01, birincil Lambda Theta 0.01, Sigmoid keskinliği 1, yineleme sayısı 15000, normalizasyon türü olarak Min-Max normalizasyonu seçilmiş, rastgele sayı üretimi ise seçilmemiştir. Projenin sonuç modülüne İkili Sınıflı Lokal Derin Destek Vektör Makinesi algoritmasını bağlayarak elde ettiğimiz sonuçlar incelendiğinde şekil 4.8, 4.9 ve 4.10'daki sonuçlar elde edilmiştir.



Şekil 4.8. İkili sınıflı yerel derin destek vektör makinesi ROC grafiği

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
1439	3595	0.769	0.737	0.5	0.729
False Positive	True Negative	Recall	F1 Score		
513	12232	0.286	0.412		
Positive Label	Negative Label				
1	0				

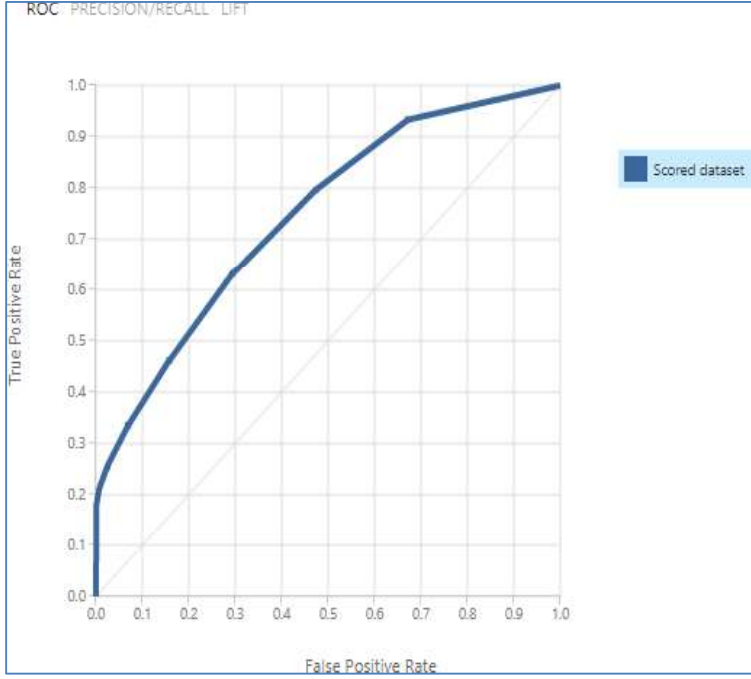
Şekil 4.9. İkili sınıflı yerel derin destek vektör makinesi kesinlik grafiği

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	67	73	0.008	0.717	0.026	0.479	0.013	0.718	0.994	0.000
(0.800,0.900]	599	98	0.047	0.745	0.227	0.796	0.132	0.742	0.987	0.000
(0.700,0.800]	541	85	0.082	0.770	0.372	0.825	0.240	0.765	0.980	0.002
(0.600,0.700]	121	103	0.095	0.771	0.395	0.787	0.264	0.770	0.972	0.004
(0.500,0.600]	111	154	0.110	0.769	0.412	0.737	0.286	0.773	0.960	0.007
(0.400,0.500]	145	217	0.130	0.765	0.431	0.685	0.315	0.777	0.943	0.012
(0.300,0.400]	279	461	0.172	0.755	0.461	0.610	0.370	0.785	0.907	0.025
(0.200,0.300]	2531	6802	0.697	0.514	0.504	0.355	0.873	0.881	0.373	0.377
(0.100,0.200]	640	4752	1.000	0.283	0.441	0.283	1.000	1.000	0.000	0.729
(0.000,0.100]	0	0	1.000	0.283	0.441	0.283	1.000	1.000	0.000	0.729

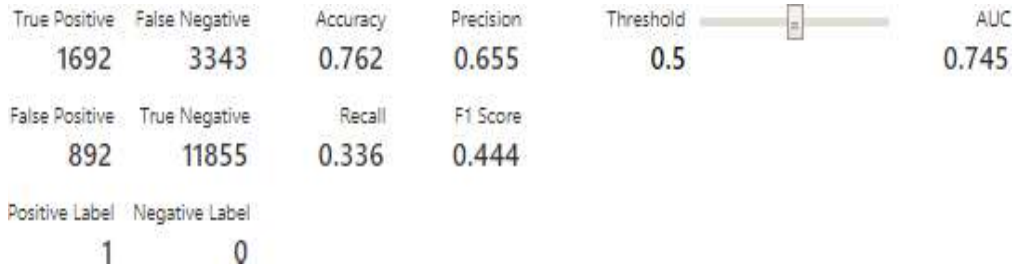
Şekil 4.10. İkili sınıflı yerel derin destek vektör makinesi kesinlik detay grafiği

4.2.6. İki sınıflı karar ormanı

İkili Sınıflı Karar Ormanı (Two Class Decision Forest) ikili sınıfa ayrılmış verilerde karar ağaçlarının kullanıldığı algoritmadır. Azure ML parametre seçenekleri üzerinde İkili Sınıflı Karar Ormanı için kullanılan parametreler: Yeniden Örnekleme metodu olarak Torbalama(Bagging) yöntemi tercih edilmiş, Eğitim modu olarak tek parametre, karar ağacı sayısı olarak 8, karar ağaçlarının maksimum derinliği 32, düğüm başına rastgele bölünme sayısı 128, yaprak düğümü başına asgari örnek sayısı 1 olarak seçilmiştir. Bu yöntemin sonuçları elde edilmiştir. Doğruluk ve Kesinlik değerleri 1'e yakın olduğu için başarılı bir tahmin yapıldığı ifade edilebilir. Sonuçlar, Şekil 4.11 ve 4.12'de verilmiştir.



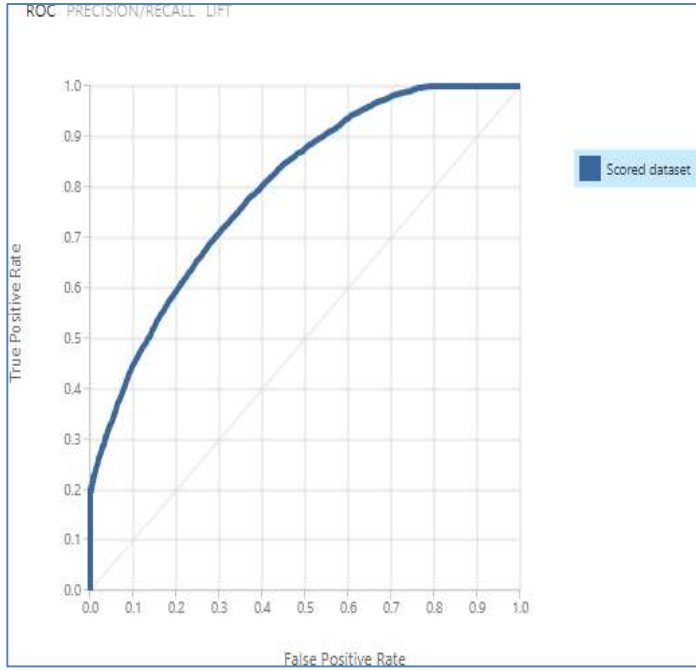
Şekil 4.11. İkili sınıflı karar ormanı ROC grafiği



Şekil 4.12. İkili sınıflı karar ormanı kesinlik grafiği

4.2.7. İkili sınıflı artırılmış karar ağacı

İkili Sınıflı Artırılmış Karar Ağacı (Two Class Boosted Decision Tree), ikili sınıfa ayrılmış verilerde artırılmış karar ağaçlarını kullanan algoritmadır. Azure ML parametre seçenekleri üzerinde İkili Sınıflı Artırılmış Karar Ağacı için kullanılan parametreler: Eğitim modu olarak tek parametre, ağaç başına maksimum yaprak sayısı 20, yaprak düğümü başına minimum örnek sayısı 10, öğrenme oranı 0.2, toplam inşa edilen ağaç sayısı 100 olarak seçilmiş, rastgele sayı üretimi ise seçilmemiştir. Şekil 4.13, 4.14 ve 4.15’de verilen sonuçlara göre Doğruluk ve Kesinlik değerleri 1’e yakın olduğu için başarılı bir tahmin yapıldığı ifade edilebilir.



Şekil 4.13. İkili sınıflı artırılmış karar ağacı ROC grafiği

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
2035	2999	0.775	0.669	0.5	0.797
False Positive	True Negative	Recall	F1 Score		
1007	11738	0.404	0.504		
Positive Label	Negative Label				
1	0				

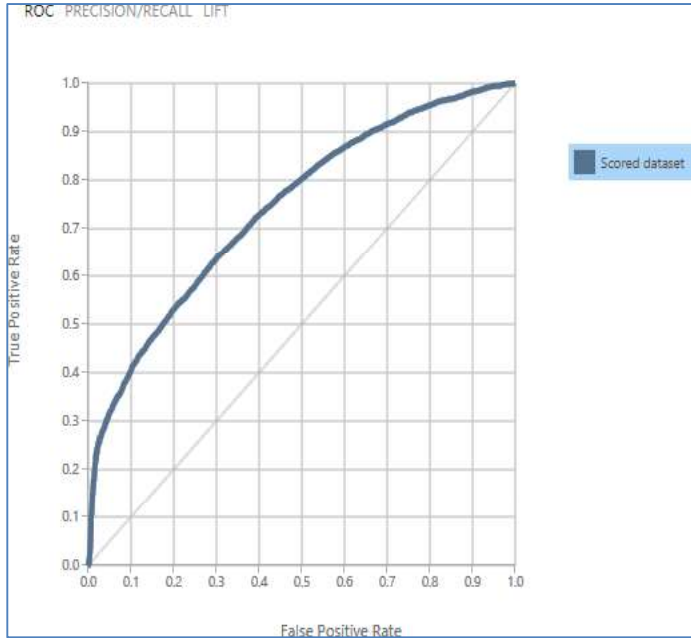
Şekil 4.14. İkili sınıflı artırılmış karar ağacı kesinlik grafiği

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	1003	5	0.057	0.773	0.332	0.995	0.199	0.760	1.000	0.000
(0.800,0.900]	44	41	0.061	0.773	0.342	0.958	0.208	0.761	0.996	0.001
(0.700,0.800]	192	138	0.080	0.776	0.384	0.871	0.246	0.768	0.986	0.003
(0.600,0.700]	324	304	0.115	0.777	0.441	0.762	0.310	0.779	0.962	0.010
(0.500,0.600]	472	519	0.171	0.775	0.504	0.669	0.404	0.796	0.921	0.024
(0.400,0.500]	544	809	0.247	0.760	0.547	0.587	0.512	0.817	0.858	0.054
(0.300,0.400]	680	1287	0.358	0.726	0.572	0.512	0.647	0.845	0.757	0.113
(0.200,0.300]	833	2144	0.525	0.652	0.569	0.438	0.813	0.888	0.588	0.236
(0.100,0.200]	708	2830	0.724	0.533	0.536	0.373	0.954	0.952	0.366	0.434
(0.000,0.100]	234	4668	1.000	0.283	0.441	0.283	1.000	1.000	0.000	0.797

Şekil 4.15. İkili sınıflı artırılmış karar ağacı kesinlik detay grafiği

4.2.8. İkili sınıflı yapay sinir ağları

İkili Sınıflı Yapay Sinir Ağları (Two Class Neural Network), ikili sınıfa ayrılmış verilerde yapay sinir ağlarını kullanan algoritmadır. Eğitim seti tüm çalışmada kullanıldığı gibi %80 olarak alınmıştır. Azure ML parametre seçenekleri üzerinde İkili Sınıflı Yapay Sinir Ağları için kullanılan parametreler: Eğitim modu olarak tek parametre, gizli katman tanımlaması olarak tamamı bağlantılı, gizli düğüm sayısı olarak 100, öğrenme oranı 0.1, öğrenme yineleme sayısı 100, başlangıç öğrenme ağırlıklarının çapı 0.1, momentum değeri 0, normalleştirme türü Min-Max normalleştirilmesi olarak seçilmiştir. Azure ML parametre ekranında üzerinde normalleştirme türleri arasında Binning(Gruplama), Gaussian ve normalleşme türü olmayan seçenekleri de mevcuttur. Şekil 4.16, 4.17 ve 4.18'de görüleceği gibi doğruluk değeri 0.772 ve kesinlik değeri ise 0.782 olarak elde edilmiştir.



Şekil 4.16. İkili sınıflı yapay sinir ağları ROC grafiği

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
1353	3681	0.772	0.782	0.5	0.742
False Positive	True Negative	Recall	F1 Score		
378	12367	0.269	0.400		
Positive Label	Negative Label				
1	0				

Şekil 4.17. İkili sınıflı yapay sinir ağları kesinlik grafiği

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	58	26	0.005	0.719	0.023	0.690	0.012	0.719	0.998	0.000
(0.800,0.900]	452	59	0.033	0.741	0.181	0.857	0.101	0.737	0.993	0.000
(0.700,0.800]	443	92	0.064	0.761	0.309	0.843	0.189	0.755	0.986	0.001
(0.600,0.700]	233	69	0.081	0.770	0.367	0.828	0.236	0.765	0.981	0.002
(0.500,0.600]	167	132	0.097	0.772	0.400	0.782	0.269	0.771	0.970	0.005
(0.400,0.500]	349	406	0.140	0.768	0.453	0.685	0.338	0.782	0.938	0.015
(0.300,0.400]	673	1120	0.241	0.743	0.510	0.555	0.472	0.803	0.851	0.051
(0.200,0.300]	1172	2872	0.468	0.648	0.531	0.426	0.705	0.843	0.625	0.184
(0.100,0.200]	1219	5107	0.824	0.429	0.484	0.325	0.947	0.914	0.225	0.522
(0.000,0.100]	268	2862	1.000	0.283	0.441	0.283	1.000	1.000	0.000	0.742

Şekil 4.18. İkili sınıflı yapay sinir ağları kesinlik detay grafiği

4.3. Değerlendirme

Regresyon yöntemleri içerisinde en başarılı yöntemin Artırılmış Karar Ağacı Regresyonu olduğu görülmüştür. MAE, RMSE, RAE ve RSE değerlerinin tamamının 0'a yakın olması o modelin başarısını göstermektedir. Bu algorithmada bu değerlerin tamamı diğer algoritmalara göre daha küçüktür ve sıfıra daha yakındır. Regresyon yöntemleri içerisinde en başarısız sonuç elde edilen algoritma Doğrusal Regresyon olmuştur. Doğrusal Regresyon hata değerlerinde 0'a en uzak sonuçlar vermiştir. CoD değeri için 1'e en yakın sonucu da yine Artırılmış Karar Ağacı Regresyonu verdiği için regresyon yöntemleri içerisinde en başarılı sonucu bu algoritmanın verdiği ifade edilebilir. Sınıflandırma yöntemlerinde ise en başarılı sonucu veren algoritmanın İkili Sınıflı Artırılmış Karar Ağacı yöntemi olduğu tespit edilmiştir. Çünkü; Doğruluk, Kesinlik, F1 Skor, Duyarlılık değerleri 1'e en yakın değerler bu algorithmadan elde edilmiştir. Elde edilen bulguların sonuçlarını daha iyi incelemek için tüm sonuçlar kendi aralarında Çizelge 4.2 ve 4.3'te kıyaslanmıştır.

Regresyonlar kendi aralarında, sınıflandırma yöntemleri kendi aralarında değerlendirilip hangi yöntemin daha başarılı olduğu tespit edilmeye çalışılmıştır. Öncelikle kullanılan regresyon yöntemlerin sonuçları Çizelge 4.2’de incelenmiştir.

Çizelge 4.2. Regresyon yöntemlerinin kıyaslamalı sonuçları

	Doğrusal Regresyon	Karar Ormanı Regresyonu	Yapay Sinir Ağları Regresyonu	Artırılmış Karar Ağacı Regresyonu
Ortalama Mutlak Hata	0.38 2	0.30 0	0.319	0.301
Karesel Ortalama Hata	0.43 7	0.40 9	0.411	0.386
Göreceli Mutlak Hata	0.94 3	0.74 1	0.787	0.743
Göreceli Karesel Hata	0.94 3	0.82 6	0.835	0.737
Determinasyon Katsayısı	0.05 6	0.17 3	0.164	0.262

Çizelge 4.3. İkili sınıflandırma yöntemlerinin kıyaslamalı sonuçları

	İkili Sınıflı Yerel Derin Destek Vektör Makinesi	İkili Sınıflı Karar Ormanı	İkili Sınıflı Artırılmış Karar Ağacı	İkili Sınıflı Yapay Sinir Ağı
Doğruluk	0.769	0.762	0.775	0.772
Kesinlik	0.737	0.655	0.789	0.782
Duyarlılık	0.286	0.336	0.404	0.269
F1 Skoru	0.412	0.444	0.504	0.400
AUC	0.729	0.745	0.797	0.742

Sınıflandırma yöntemlerinin sonuçları ise Çizelge 4.3’te verilmiştir. Bu sonuçlara göre en başarılı sonucu veren algoritmanın İkili Sınıflı Artırılmış Karar Ağacı olduğu ifade edilebilir. Çünkü; Doğruluk, Kesinlik, Duyarlılık 1’e en yakın değerler bu algoritmadan elde edilmiştir.

Literatürdeki benzer çalışmalarla karşılaştırmalar: Lending Club veri setini literatürde kullanan daha önce bahsedilen “Predicting Default Risk of Lending Club Loans”

çalışmasında ise Lojistik Regresyon, Naive Bayes Ağları ve SVM (Destek Vektör Makinesi) yöntemleri kullanılarak risk tahmini yapılmıştır. Lojistik Regresyonda 30 iterasyon(yineleme) kullanıldığında Doğruluk 92.8, Kesinlik 96.6 Duyarlılık 95.1 bulunmuştur. Naive Bayes algoritmasında Gaussian modelinde Doğruluk 91.1, Kesinlik 96.6, Duyarlılık 92.7 değerleri elde edilmiştir. SVM algoritmasında ise Sigmoid kernel(çekirdek) kullanıldığında Doğruluk 90.8, Kesinlik 91.8, Duyarlılık 98.2 değerleri elde edilmiştir. Çalışmamızdan daha başarılı doğruluk, kesinlik ve duyarlılık sonuçları elde edilmiştir. Özellikle kendi çalışmamızdaki SVM algoritması ile kıyaslandığında ortak olarak Sigmoid fonksiyonu kullanılmasında rağmen bu çalışmada daha başarılı sonuçlar elde edilmiştir. Bunu da kullanılan parametrelere ve verilerin daha başarılı normalize edilmesine bağlıyoruz.

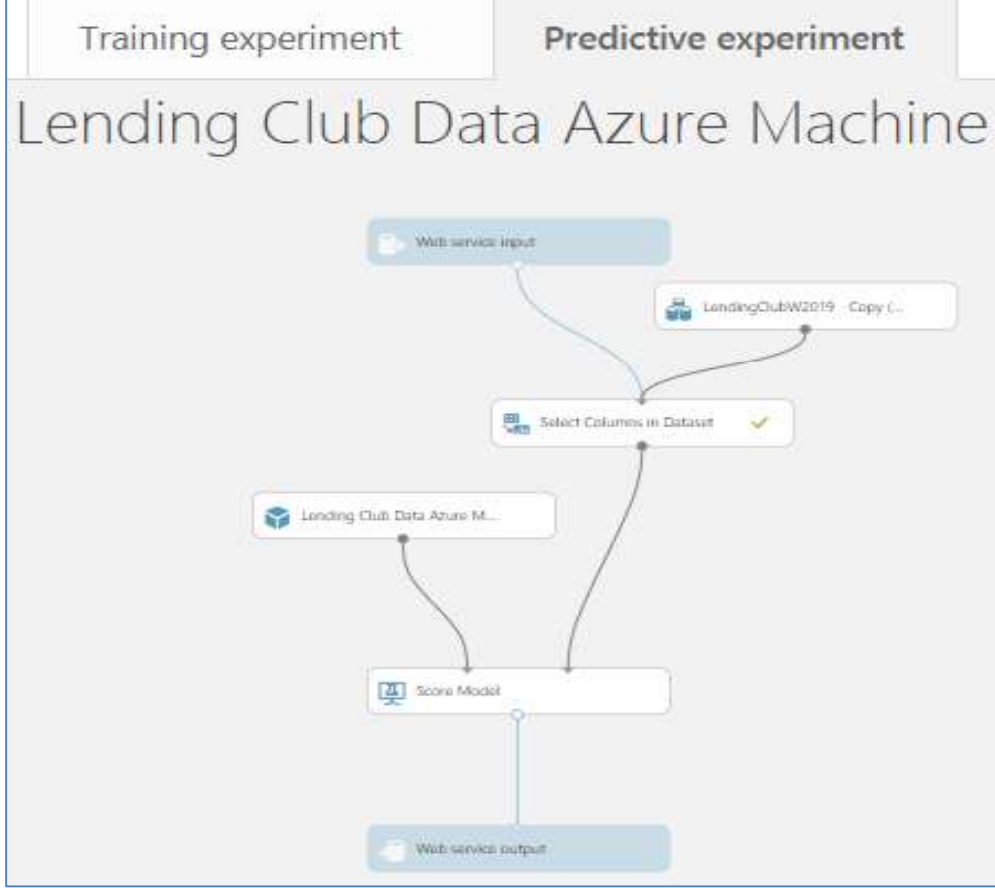
Diğer bir çalışma olan “Credit risk prediction in an imbalanced social lending environment” çalışmasında ise Rastgele Orman Algoritmasında Doğruluk 0.692, Kesinlik 0.69 ve Duyarlılık 0.717 elde edilirken Lojistik Regresyonda ise Doğruluk 0.693, Kesinlik 0.71 ve Duyarlılık 0.723 olarak bulunmuştur. Çalışmamızdan daha kötü performans değerleri elde edildiği görülmektedir [24].

Başka bir çalışma olan “Performance Analysis of Credit Scoring Models on Lending Club Data” çalışmasında ise Lojistik Regresyonda Doğruluk 0.7913 ve AUC değeri 0.6979, SVM Doğruluk 0.7887 ve AUC değeri 0.6967, Rastgele Orman Algoritmasında Doğruluk 0.7883 ve AUC değeri ise 0.6928 elde edilmiştir. Bu çalışma ise kendi çalışmamıza yakın sonuçlar vermiştir [25]. Farklı çalışmalardan elde edilen sonuçlarda genellikle kullanılan model, parametre ve fonksiyonlar sonuçları değiştirmiştir. Bazı çalışmalarda Kesinlik, Doğruluk gibi performans metrikleri yerine daha farklı metrikler kullanılmıştır.

4.4. Modelin Web Servis Entegrasyonu

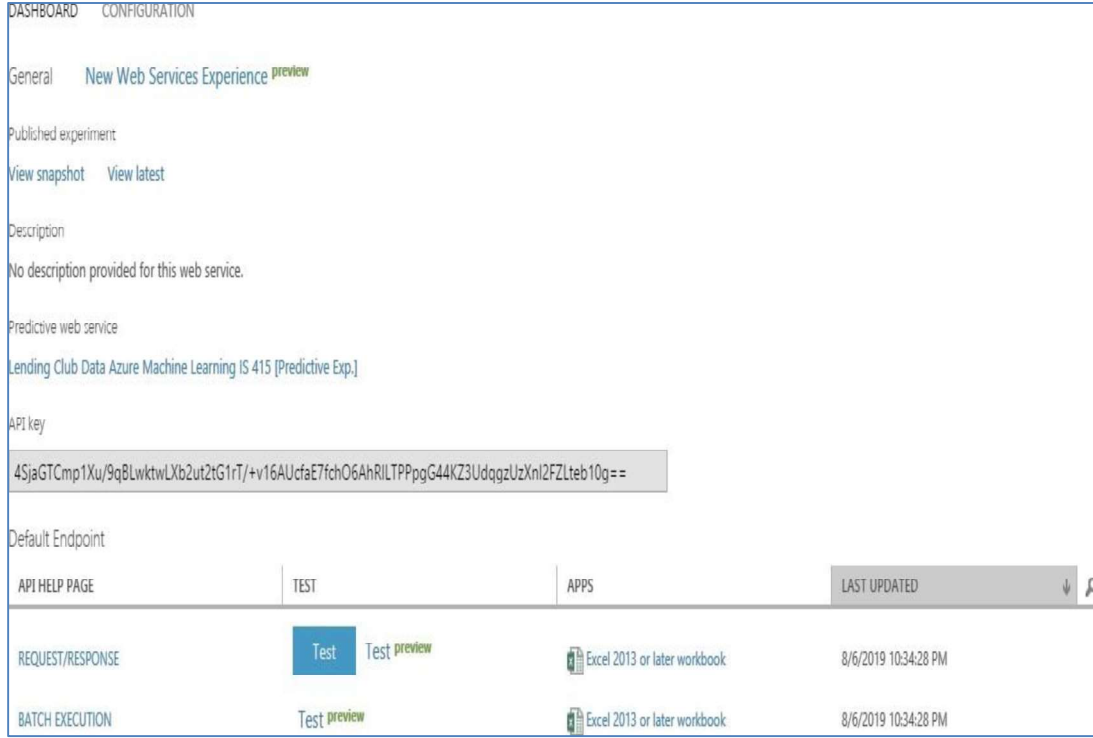
Değerlendirme sonuçları alındıktan sonra elde edilen en başarılı algoritmalar web servis üzerinden bir yazılıma entegre edilmiştir. Azure ML uygulamasının değerlendirme modülüne web servis giriş modülü bağlanarak web servis yapılabilir. Çalışma alanı içerisinde temelde iki sekme bulunmaktadır. Bu sekmelerden birisi şu ana kadar çalışılan alan olan Eğitilen(Training Experiment) sekmesidir. Bu sekme altında makine öğrenmesi yöntemleri kullanılıp veri analizi yapılmaktadır. Diğer sekme ise Tahmin (Predictive Experiment)

sekmesidir. Bu sekmeye geçildiğinde çalışmanın mimarisi değişmektedir. Şekil 4.19’da tahmin mimarisi görülmektedir.



Şekil 4.19. Azure ML tahmin şeması

Tahmin mimarisinde web servis giriş ve çıkış modülleri görülmektedir. Alt menüden web servis çalıştırılınca(Deploy Web Service) seçildiğinde ise web servisin yayınlanma süreci başlatılmış olunur. Web servis çalıştırılınca(deploy) Şekil 4.20’deki ekran elde edilecektir.

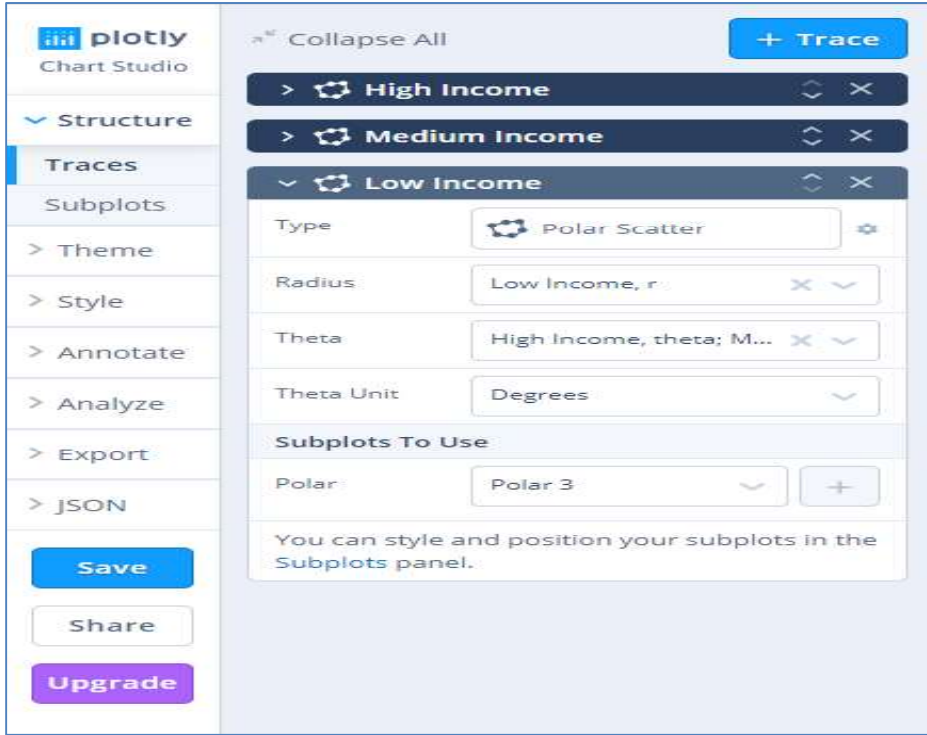


Şekil 4.20. Azure ML web servis çalışma ekranı

Hazırlanan web servis web API formatındadır. Azure ML uygulaması model üretildikten sonra web servis entegrasyonu için API Key değeri oluşturmaktadır. Bu değerler API arayüz uygulamalarında kullanılabilir. Test butonuna basılarak Azure Servis uygulamalarına geçilerek web servis kullanılabilir. İşlenen veriler web servis haline getirildikten sonra bir uygulama arayüzü ile kullanıcıların parametre (yaş, kredi tutarı, kredi puanı vs...) seçebildiği bir kredi risk tahmini uygulaması yapılabilir.

4.5. Kredi Riskine Etki Eden Faktörlerin Analizi

Bu aşamada ise tezin 2. Problemi olan “Kredi riskini belirleyen faktörler nelerdir?” sorusuna cevap aramak için Lending Club verilerinin genel analizi yapılmıştır. Bu analizle riske etki eden faktörler incelenmiş, iyi ve kötü kredilere etki eden etmenler değerlendirilmiştir. Verilerin analizi yapılırken ücretsiz ve online veri analiz ve görselleştirme sitesi olan Plotly kullanılmıştır. Elde edilen sonuçlar grafiklerle gösterilmiştir. Bu grafiklerin sonuçlarına göre veri setindeki niteliklerin iyi ve kötü kredi oranlarına etkileri incelenmiştir.



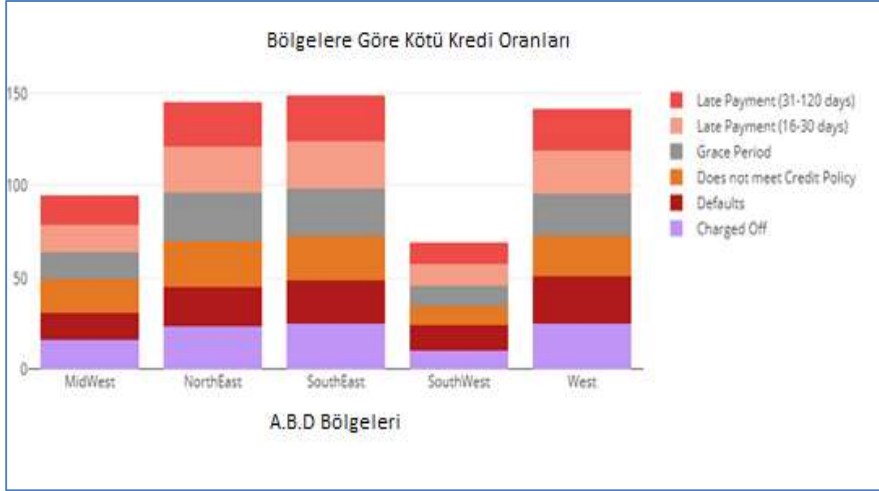
Şekil 4.21. Plotly ana menüsü

Şekil 4.21’de verilen Plot.ly sitesinde sol menüde subplotlar ve yapıların yönetimi yer almaktadır. Bu menüde yapı, subplot, tema, stil gibi sekmeler yer almaktadır. Sağındaki kısımda ise gruplandırıdığımız verilerin hangi şekil ve kolonlarla korelasyona gireceği belirlenmektedir.

Unnamed grid 1 X +													
	Good Loans, r	Good Loans, theta	Bad Loans, r	Bad Loans, theta	E	F	G	H	I	J	K	L	M
1	13.27	Fully Paid	16.14	Default Rate									
2	12.95	Current	16.02	Charged Off									
3	12.44	Issued	14.6	C.P. Charged Off									
4	13.98	No C.P. Fully Paid	15.35	In Grace Period									
5			15.7	Late (16-30 days)									

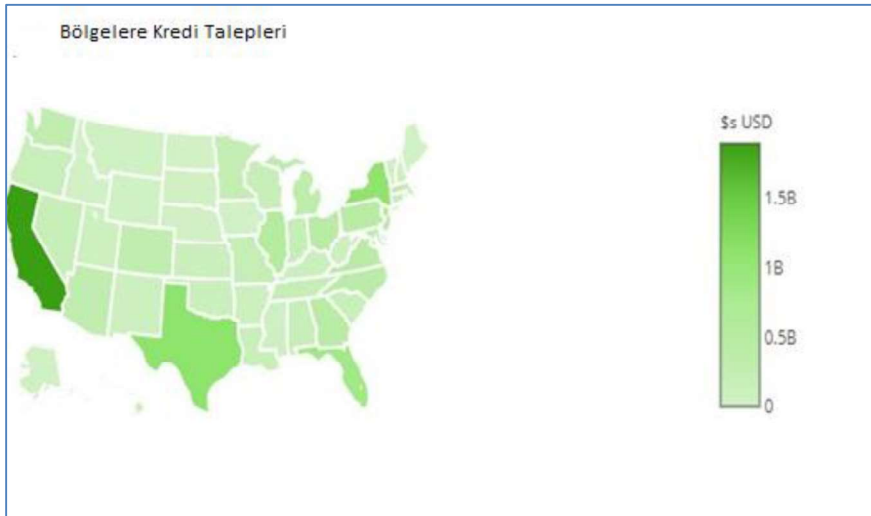
Şekil 4.22. Plotly çalışma gridi

Şekil 4.22’deki gridde ise korelasyona girecek verilerin yönetilmesi gerçekleştirilmiştir. Kötü ve iyi kredilerin gruplandırılarak bölge ve yıl gibi farklı parametrelerle korelasyonları incelenmiştir.



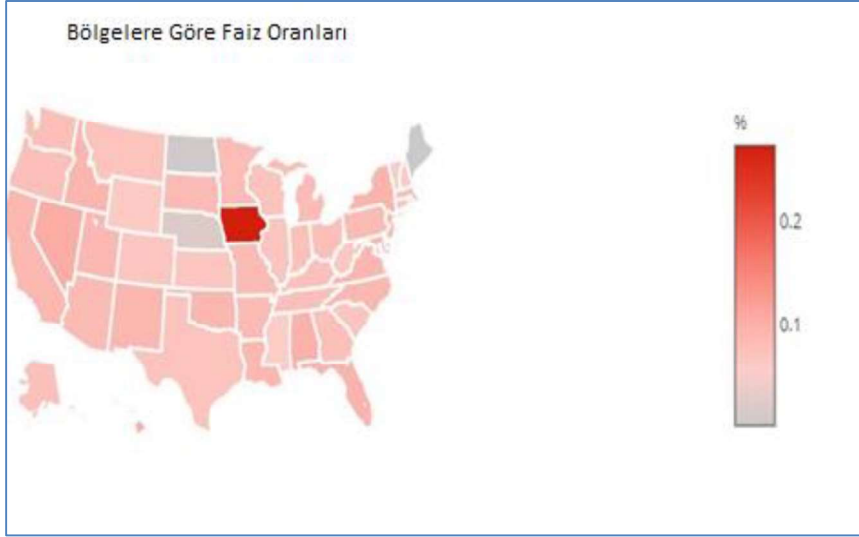
Şekil 4.23. Bölgelere göre kötü kredi oranları

Şekil 4.23'deki grafikte kötü ve iyi kredi oranlarının ABD'nin hangi bölgelerinde nasıl dağıldığı görülmektedir. Batı ve güneybatı eyaletlerinin kötü kredi oranlarının düşük olduğu görülürken kuzey ve doğu eyaletlerinin kötü kredi oranlarının daha yüksek olduğu tespit edilmiştir.



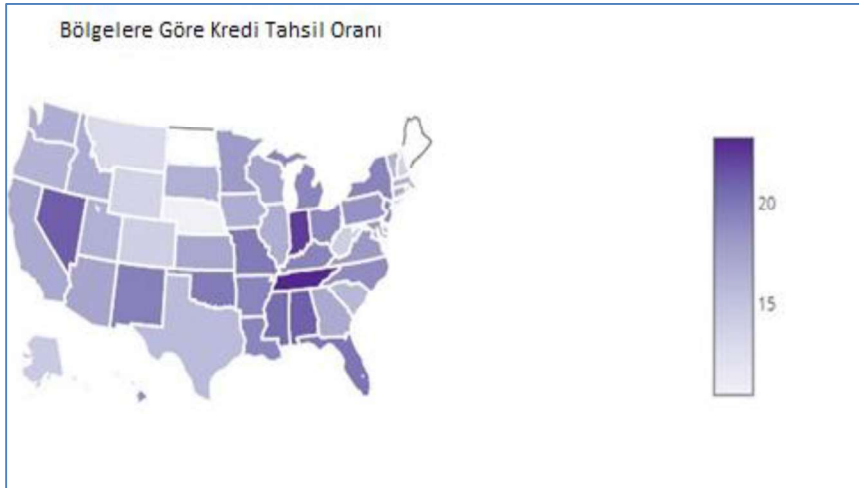
Şekil 4.24. Bölgelere göre kredi talepleri

Şekil 4.24'de ABD'nin hangi eyaletlerinin daha çok kredi talep ettiği görülmektedir. Batı ve güneybatı eyaletlerinin kredi taleplerinin daha fazla olduğu görülmektedir. Kuzey eyaletlerinin ise kredi taleplerinin daha az olduğu sonucuna varılabilir. ABD'nin maddi refahı en yüksek eyaletinin şeklimizde elde ettiğimiz sonuçta olduğu gibi California olduğu bilinmektedir.



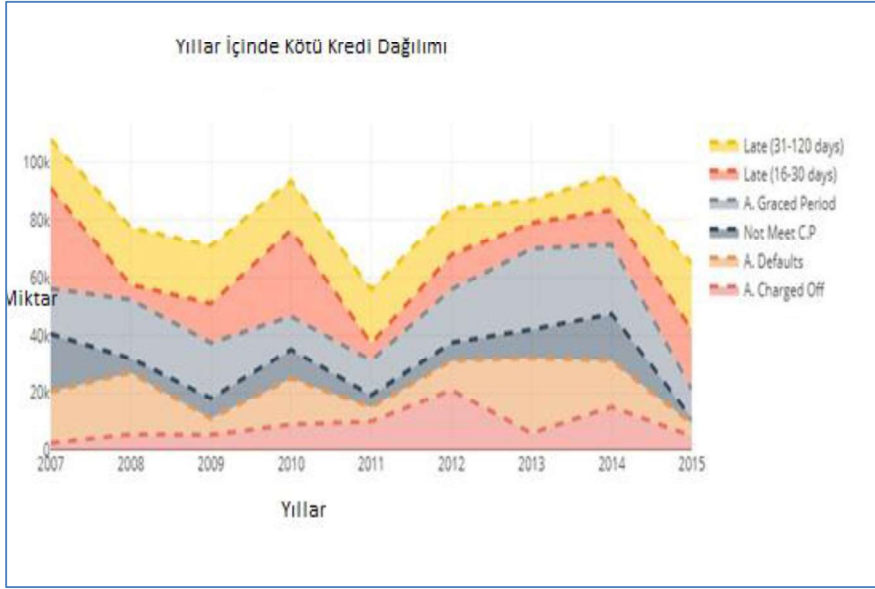
Şekil 4.25. Bölgelere göre faiz oranları

Şekil 4.25'te ABD'nin hangi eyaletlerinde faiz oranının daha yüksek olduğu görülmektedir. Kuzey eyaletlerinde kırmızı renk oranının artmıştır. Daha önceki grafiklerimizde kuzey eyaletlerinin kredi talep oranlarının düşük ve kötü kredi oranlarının daha yüksek olduğu görülmüştür. Şekilde görüldüğü gibi Iowa eyaleti kredi riski açısından en yüksek risk grubundadır. Bu sonuçtan kötü kredi oranlarının yüksek olduğu kuzey eyaletlerinde faiz oranlarının riske uygun bir biçimde arttığı sonucunu çıkarabiliriz.



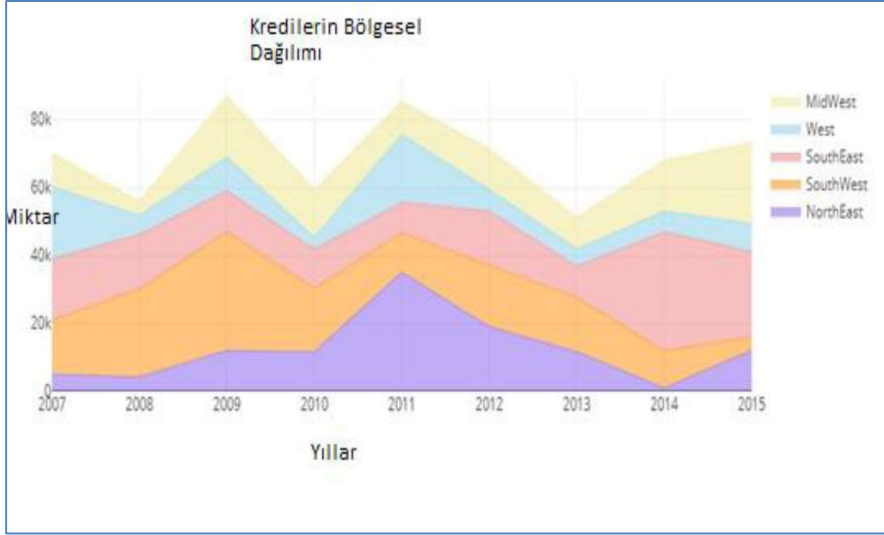
Şekil 4.26. Bölgelere göre devlet tarafından tahsil oranı

Şekil 4.26’da ABD’nin tüm eyaletlerinde devlet tarafından tahsil edilen kredilerin faiz oranlarının dağılımları görülmektedir. Batı eyaletleri yüksek kredi almalarına rağmen genellikle temerrüte daha az düşmüş ve düşük faiz oranları ile krediler tahsil edilmiştir. Orta bölgelerde ise temerrüte düşen kredilerin faiz oranlarının daha yüksek olduğu görülmektedir.



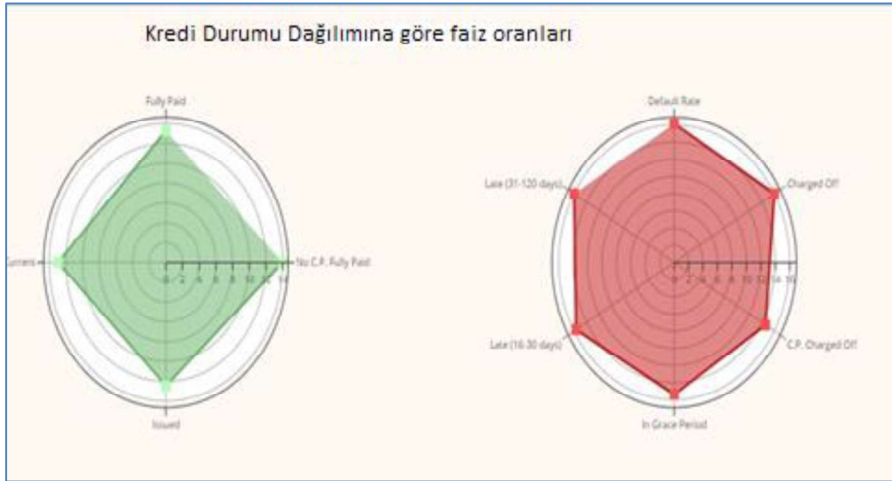
Şekil 4.27. Yıllar içinde kötü kredi dağılımı

Şekil 4.27’de ise yıllar içerisinde kötü kredinin dağılımı görülmektedir. 2011 yılında kötü kredilerin sayısının azaldığı ifade edilebilirken 2014 yılında kötü kredi miktarının çok fazla arttığı görülmektedir.



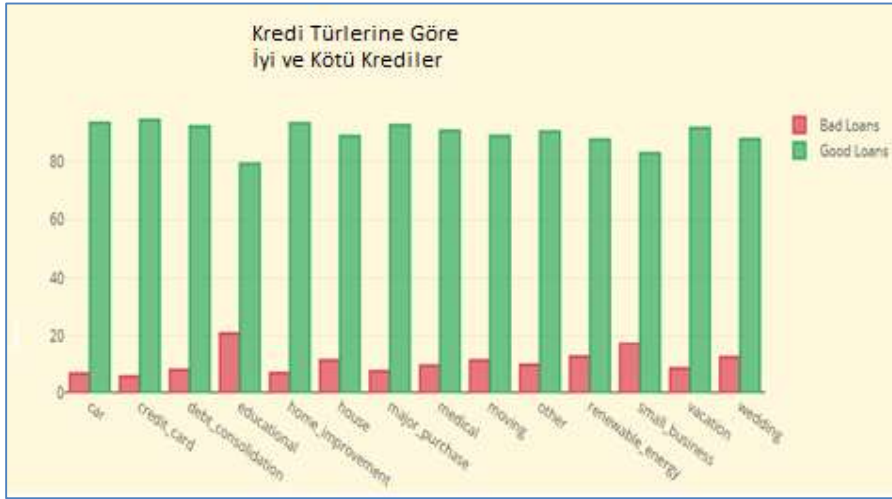
Şekil 4.28. Yıllar içinde kredilerin bölgesel dağılımı

Şekil 4.28’de yıllar içerisinde verilen kredilerin yıllar içerisinde hangi bölgelerde nasıl dağıldığı görülmektedir. 2009 ve 2011 yılında verilen kredilerin sayısının arttığı ifade edilebilir. En yüksek kredi miktarlarının yine batı eyaletlerinde olduğu en düşük kredi miktarlarının ise daha önceden tespit ettiğimiz gibi kuzey eyaletleri olduğu görülmüştür.



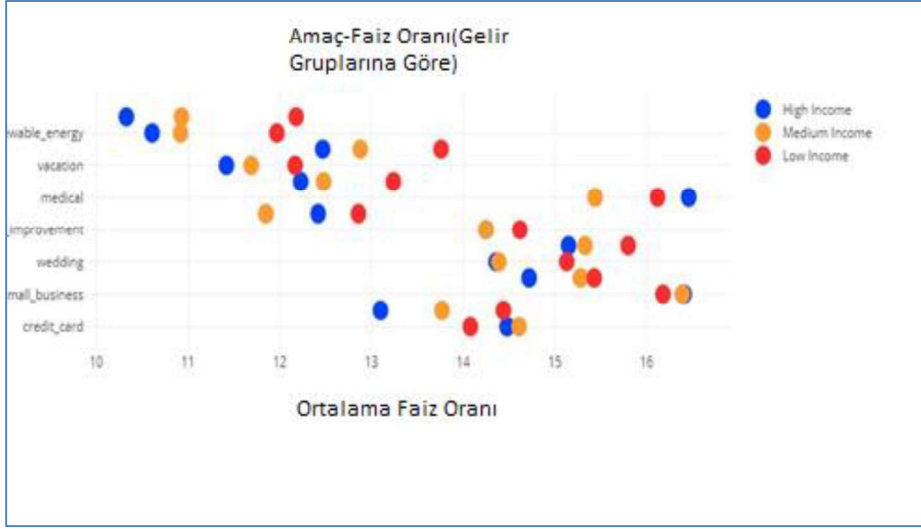
Şekil 4.29. İyi ve kötü kredi faiz oranları

Şekil 4.29’da kötü ve iyi kredilerin faiz oranları ile ilişkileri incelenmiştir. Tamamı ödenen iyi kredilerde faiz oranları genellikle daha düşüktür. Tamamı ödenmeyen kötü kredilerin faiz oranları ise daha yüksektir. Geç ödenen kötü krediler ise tamamen ödenmeyen kredilere göre daha düşük faizde alınmıştır. Bu sonuçlara göre faiz oranı arttıkça ödeme güçlüğü ortaya çıktığı tespit edilebilir.



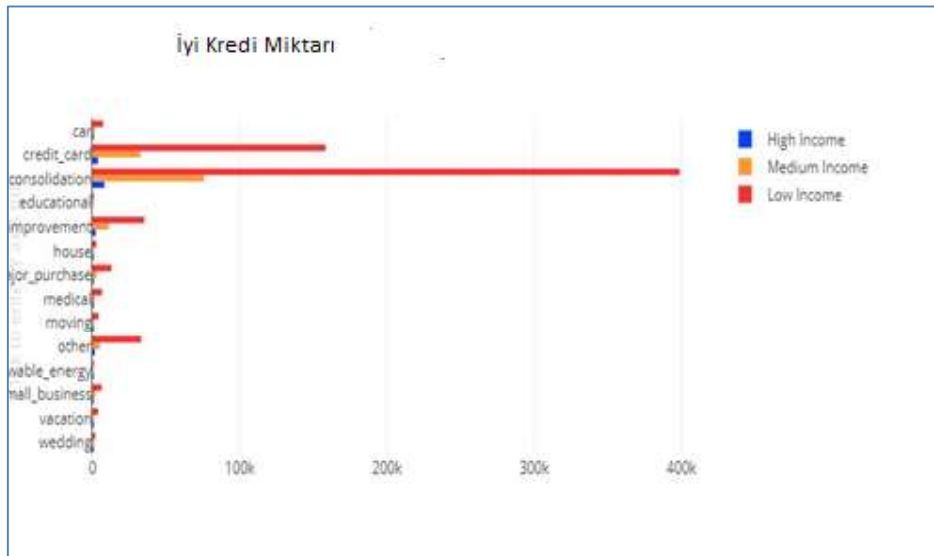
Şekil 4.30. İyi ve kötü kredilerin kredi türlerine göre oranları

Şekil 4.30’da en yüksek kötü kredilerin eğitim ve küçük işler için alınan krediler olduğu görülmüştür. İyi kredilerde ise ev tadilatı ve kredi kartı borcu için alınan kredilerin çok yüksek olduğu görülürken bu borç tiplerinde kötü kredi oranının en düşük olduğu görülmüştür.



Şekil 4.31. Gelir gruplarının faiz oranları

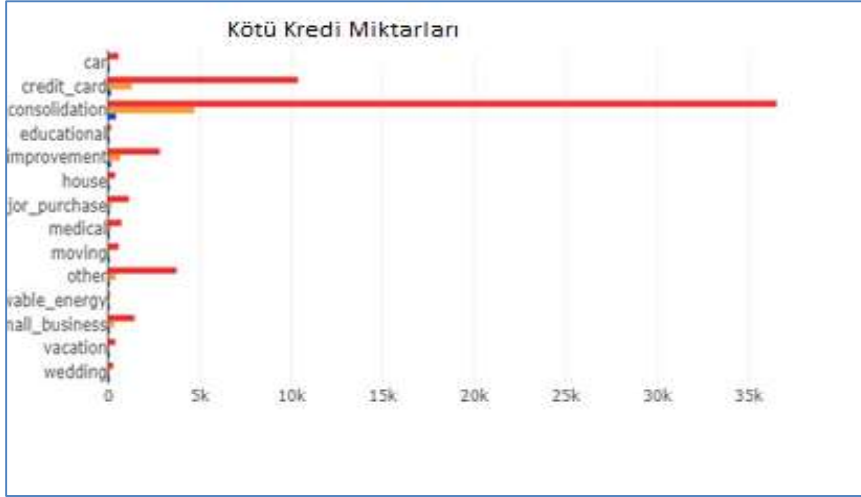
Şekil 4.31’de gelir gruplarının hangi kategorilerde ne oranda faiz ile borçlandıkları görülmektedir. Yüksek gelir gruplarının sol eksene yakınlığı görülmektedir. Yüksek gelir gruplarından daha az oranda faiz alındığı sonucu çıkarılabilir. Düşük ve orta gelir grubunun ise daha yüksek faiz oranından borçlandıkları görülmüştür.



Şekil 4.32. Gelir gruplarında iyi kredilerin kredi türlerine göre oranları

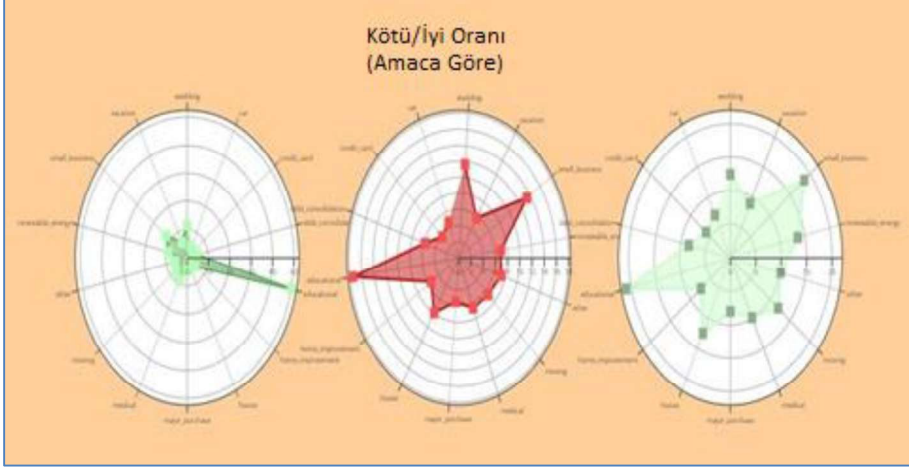
Şekil 4.32’de farklı gelir gruplarında iyi kredilerin hangi kredi türlerine göre dağılımı gösterilmiştir. Borç birleştirme için alınan kredilerin özellikle düşük ve orta gelirli gruplarda daha fazla olduğu görülmektedir.

Bu türden sonra ise kredi kartı için alınan iyi kredilerin düşük gelir gruplarında daha fazla olduğu ortaya çıkmıştır. Yüksek gelir grupları iyi krediler içerisinde de kötü krediler içerisinde de azınlıkta kalmıştır.



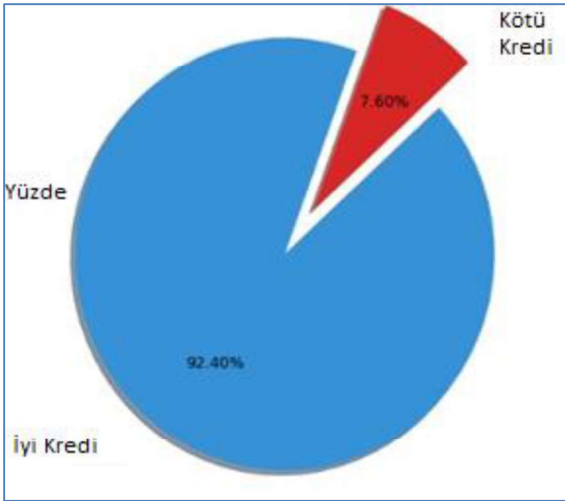
Şekil 4.33. Gelir gruplarında kötü kredilerin kredi türlerine göre oranları

Şekil 4.33'te farklı gelir gruplarında kötü kredilerin hangi kredi türlerine göre dağılımı gösterilmiştir. Borç birleştirme için alınan kredilerin özellikle düşük ve orta gelirli gruplarda daha fazla olduğu görülmektedir. Bu türden sonra ise kredi kartı için alınan kötü kredilerin düşük gelir gruplarında daha fazla olduğu ortaya çıkmıştır. Yüksek gelir grupları iyi krediler içerisinde de kötü krediler içerisinde de azınlıkta kalmıştır.



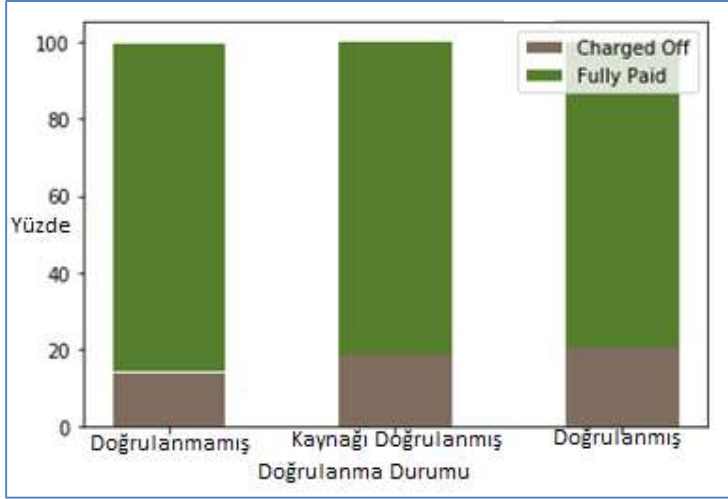
Şekil 4.34. İyi ve kötü kredilerin toplam kredi türleri arasındaki oranları

Şekil 4.34'te kötü ve iyi kredilerin toplam kredi türleri arasındaki oranları verilmiştir. Kötü krediler içerisinde eğitim amaçlı alınan kredilerin oranının diğer kredi türlerinden daha yüksek oranda olduğu görülmüştür. İyi kredide ise ev tadilatı için alınan kredilerin yüksek oranda olduğu ve eğitim için alınan kredilerin ise en düşük oranda olduğu elde edilmiştir.



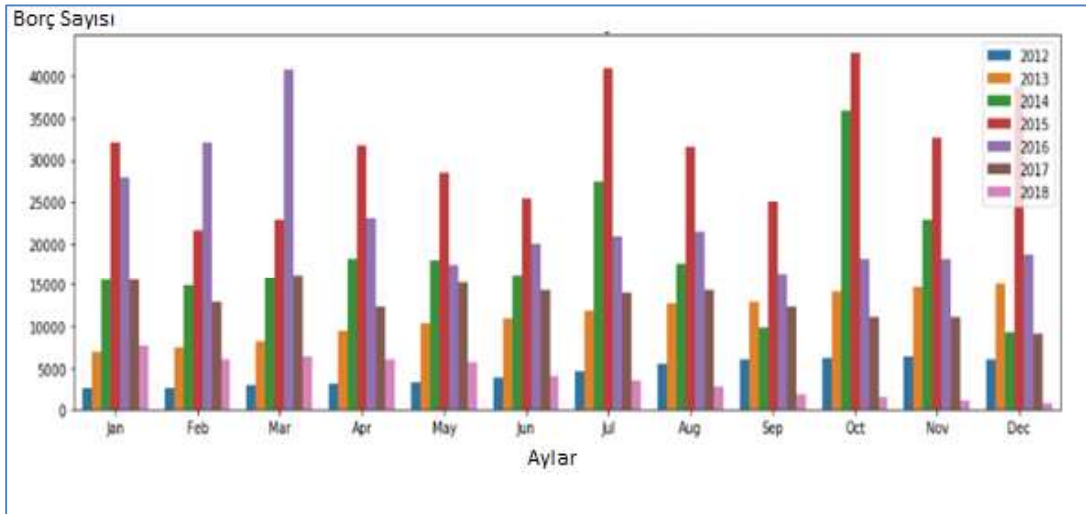
Şekil 4.35. Gelir gruplarında iyi ve kötü kredilerin kredi türlerine göre oranları

Şekil 4.35'de kötü ve iyi kredilerin toplam krediler içindeki dağılımını göstermektedir. İyi krediler %92'lik bir kısmı yaklaşık %8'lik bir oranda kötü kredileri oluşturmaktadır. Veri setinde yüksek çoğunluğun iyi kredilerde olduğu görülmüştür.



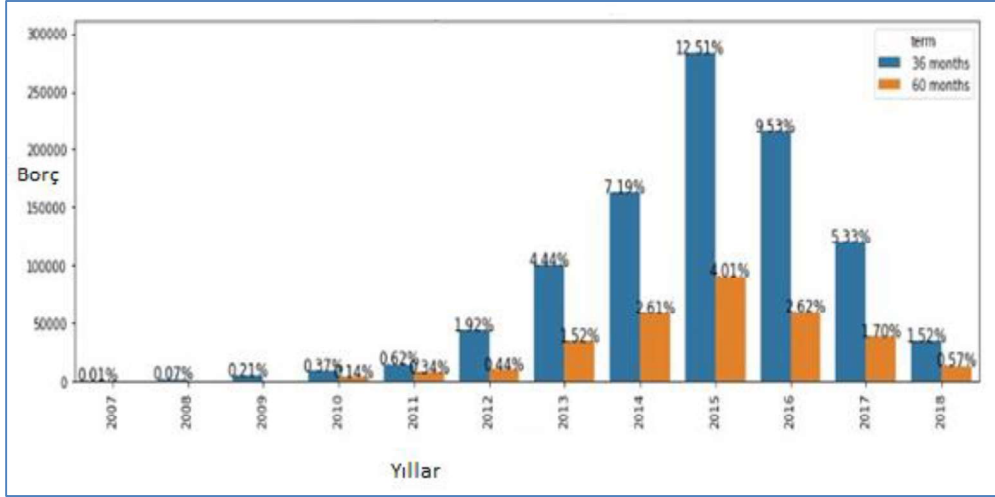
Şekil 4.36. Kredi doğrulama durumu ve ödenme durumu ilişki grafiği

Şekil 4.36’da doğrulanan- doğrulanamayan kredilerin tamamen ödenen ve tahsil edilen krediler arasındaki dağılımı görülmektedir. Veri setinin çoğunluğu doğrulanmış ya da kaynakları doğrulanmıştır. Doğrulanan ve doğrulanamayan verilerin çoğunluğu tamamen ödenen kredilerdir.



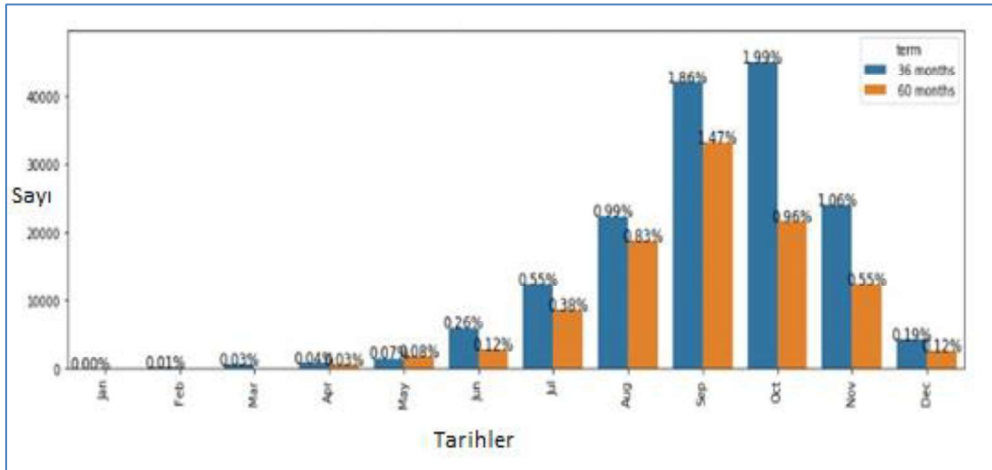
Şekil 4.37. Kredilerin yıllar ve aylar içerisindeki dağılımı

Şekil 4.37’de kredilerin yıllar ve aylar içerisinde dağılımı görülmektedir. Yıl içerisinde temmuz ve kasım aylarında kredi miktarlarının arttığını ifade edebiliriz.



Şekil 4.38. Kredi vadelerinin yıllara göre dağılımı

Şekil 4.38’de 36 aylık vadeli kredilerin 60 aylık kredilerin tamamında her yılda daha yüksek olduğu görülmektedir. 2015 yılında ise 36 aylık krediler tüm yıllardaki 36 aylık kredilerden daha fazladır.



Şekil 4.39. Temerrüte düşen kredilerin vadelerine göre aylık dağılımı

Şekil 4.39’da temerrüte düşen kredilerin ekim ve kasım aylarında diğer aylara göre daha fazla olduğu ayrıca 36 aylık kredilerdeki temerrüt oranının 60 aylık kredilerden daha fazla olduğu ifade edilebilir.

Batı ve güneybatı eyaletlerinin kötü kredi oranları düşük elde edilmişken kuzey ve doğu yıllar içerisinde kötü kredinin dağılımı incelendiğinde 2011 yılında kötü kredilerin miktarının azaldığı elde edilirken 2014 yılında kötü kredi miktarının çok fazla arttığı sonucu ortaya çıkmıştır.

Verilen kredilerin yıllar içerisinde hangi bölgelerde nasıl dağıldığı incelendiğinde 2009 ve 2011 yılında verilen kredilerin sayısının arttığı görülürken en yüksek kredi miktarlarının yine batı eyaletlerinde olduğu en düşük kredi miktarlarının ise daha önceden tespit ettiğimiz gibi kuzey eyaletlerinde olduğu saptanmıştır.

Kötü ve iyi kredilerin faiz oranları ile ilişkileri incelenmiştir. Tamamı ödenen iyi kredilerde faiz oranları genellikle daha düşük olurken tamamı ödenmeyen kötü kredilerin faiz oranları ise daha yüksek elde edilmiştir. Geç ödenen kötü krediler ise tamamen ödenmeyen kredilere göre daha düşük faizde alınmıştır. Bu sonuçlara göre faiz oranı arttıkça ödeme güçlüğü ortaya çıktığı tespit edilebilir.

En yüksek miktarda kötü kredinin eğitim ve küçük işler için alınan krediler olduğu görülmüştür. İyi kredilerde ise ev tadilatı ve kredi kartı borcu için alınan kredilerin çok yüksek olduğu elde edilirken kötü kredi oranının ise en düşük oranda olduğu görülmüştür. Gelir gruplarının hangi kategorilerde ne oranda faiz ile borçlandıkları incelendiğinde, Yüksek gelir gruplarından daha az oranda faiz alındığı sonucu çıkarılabilir. Düşük ve orta gelir grubunun ise daha yüksek faiz oranından borçlandıkları görülmüştür.

Farklı gelir gruplarında iyi kredilerin kredi türlerine göre dağılımı incelendiğinde, borç birleştirme (mevcut borcun başka bir kredi ile kapatılması) için alınan kredilerin özellikle düşük ve orta gelirli gruplarda daha fazla olduğu görülmüştür. Borç birleştirmeden sonra ise kredi kartı için alınan iyi kredilerin düşük gelir gruplarında daha fazla olduğu ortaya çıkmıştır. Yüksek gelir grupları iyi krediler içerisinde de kötü krediler içerisinde de azınlık olarak görülmüştür. Kötü krediler içerisinde eğitim amaçlı alınan kredilerin oranının diğer kredi türlerinden daha yüksek oranda olduğu görülmüştür. İyi krediler arasında ise ev tadilatı için alınan kredilerin yüksek oranda olduğu ve eğitim için alınan kredilerin ise en düşük oranda olduğu elde edilmiştir.

Tüm krediler incelendiğinde, iyi kredilerin toplam kredilerin %92’lik bir kısmını oluşturduğu, kötü kredilerin ise yaklaşık %8’lik bir kısmı meydana getirdiği belirlenmiştir. Bu sonuçla, veri setinde çoğunluğun iyi kredilerde olduğu görülmüştür. Kredilerin doğrulanma durumlarına göre tamamen ödenen ve tahsil edilen krediler incelendiğinde veri setinin çoğunluğunun doğrulanmış ya da kaynakları doğrulanan krediler olduğu belirlenmiştir. Ayrıca, doğrulanan ve doğrulanamayan verilerin çoğunluğu tamamen ödenmiş kredilerdir.

Kredilerin yıllar ve aylar içerisinde genel dağılımı incelendiğinde temmuz ve kasım aylarında kredi miktarlarının arttığı görülmüştür. ABD’nin tüm eyaletlerinde devlet tarafından tahsil edilen kredilerin faiz oranlarının dağılımları incelendiğinde batı eyaletleri yüksek miktarda kredi almalarına rağmen genellikle temerrüte daha az düşmüş ve düşük faiz oranları ile krediler tahsil edilmiştir. Orta bölgelerde ise temerrüte düşen kredilerin faiz oranlarının daha yüksek olduğu görülmektedir.

Veri setindeki kredi vadeleri tüm veri setinde 36 ve 60 aydır. 36 aylık vadeli kredilerin 60 aylık kredilerin tamamından her yılda daha yüksek olduğu görülmüştür. 2015 yılında ise 36 aylık krediler tüm seneler içindeki 36 aylık kredilerden daha çok miktardadır. Temerrüte düşen kredilerin ekim ve kasım aylarında diğer aylara göre daha fazla olduğu ayrıca 36 aylık kredilerdeki temerrüt oranının 60 aylık kredilerden daha fazla olduğu elde edilmiştir.

5. SONUÇ VE ÖNERİLER

Küreselleşen dünyada P2P finans sistemleri önemli bir yer tutmaktadır. Kişiden kişiye finans sistemi sayesinde insanlar kendi mevduatlarını istedikleri faiz ve vadeye kullanıma açacaklar ve birikimlerini değerlendirmiş olacaklardır. Özellikle, bankaya gitmeden internet üzerinden başvuru yapılması P2P sistemini daha hızlı ve kolay bir finans aracı haline getirmektedir. İnternet üzerinden kolaylıkla başvuru yapılan finans sistemleri olan P2P finans sistemleri hızlı bir şekilde yükselmektedir ve yakın gelecekte geleneksel bankacılığa daha fazla entegre olacak, hatta ondan ciddi paylar alacak hale gelecektir.

Lending Club veri seti için yapılacak risk tahmini ile elde edilen başarılı algoritmalar benzer veri setleri için de kredi riski tahmini yapılmasını sağlayabilecektir. Kredi riski tahminleri kullanılarak finansal çıkarımlar yapılabilir ve geleneksel bankacılık sistemleri için de risk analizi yapılabilir. Çalışmada Lending Club verileriyle kredi riski tahmini makine öğrenmesi yöntemleri uygulanarak yapılmaya çalışılmıştır. Elde edilen en başarılı algoritmalar web servis ile yazılımlara entegre edilebilir.

Makine öğrenmesi ile risk analizi dışında çalışmanın sonunda verilerin genel analizi de yapılarak iyi ve kötü kredilere etki eden etmenler araştırılmıştır. Veriler iyi ve kötü kredi olarak gruplandırılmış ve farklı niteliklerle ilişkileri incelenmiştir. Çalışmada elde edilen bulgulara göre problem, (Kredi riski analizi için en başarılı sınıflandırma ve regresyon yönteminin belirlenmesi) ve alt problemleri (Kredi riskini belirleyen faktörlerin tespiti) incelendiğinde; en başarılı sınıflandırma ve regresyon yöntemleri elde edilen bulgulara göre sınıflandırma yöntemleri için Artırılmış İki Sınıflı Karar Ağacı algoritması, regresyon yöntemleri için ise Artırılmış Karar Ağacı algoritmasıdır. Kredi riskini belirleyen faktörleri öncelik sırasına göre gelir dağılımı, faiz oranı, müşterinin bulunduğu bölge ve kredi türü olarak tespit edilmiştir.

ABD’de en çok kredinin California ve Texas eyaletlerinde alındığı tespit edilmiştir. ABD’de kötü kredi oranlarının en düşük güneybatı eyaletlerinde olduğu tespit edilirken kuzey ve doğu eyaletlerinde daha yüksek oranda kötü kredi tespit edilmiştir. ABD’nin güneybatı eyaletlerinin ekonomik açıdan daha güçlü olduğu sonucunu çıkarabiliriz. Kredi durumu kolonu içinde kötü kredi olarak kabul edilen krediler bölgelerine göre gruplanmış ve

sonuçları elde edilmiştir. Kredi kullanım nedenlerinin kötü ve iyi krediyle ilişkisi incelendiğinde eğitim amaçlı kullanılan kredilerin kötü kredi oranı en yüksek ve iyi kredi oranı en düşük krediler olduğu tespit edilmiştir. Kötü kredi oranı en düşük ve iyi kredi oranı en yüksek kredilerin ise ev tadilatı için çekilen krediler olduğu görülmüştür. Bu sonuçlardan öğrencilerin kredi ödeme güçlüğü yaşadığını ifade edebiliriz. Faiz ve gelir grupları arasındaki ilişki incelendiğinde ise yüksek gelir gruplarından daha az oranda faiz alındığı sonucu çıkarılabilir.

Düşük ve orta gelir grubunun ise daha yüksek faiz oranından borçlandıkları görülmüştür. Buna göre veri setinin bulunduğu ABD’de coğrafi bölgelerin, kredi kullanım türlerinin, faiz oranlarının ve gelir gruplarının borç ödeme veya temerrüde düşme oranlarını etkilediği tespit edilmiştir. Elde edilen korelasyonlar incelenip makine öğrenmesi kullanılarak sadece borç durumu niteliği değil diğer niteliklerin de birbirleriyle ilişkileri üzerine modeller oluşturulabilir. Böylece çok farklı niteliklerin eğitilip test edilmesi gerçekleştirilebilir. R modülü üzerinde yazılacak kodlarla ekonomik hesaplamalarla yeni sanal nitelikler oluşturulabilir ve bu nitelikler makine öğrenmesi yöntemleriyle eğitilip test yapılarak yeni risk modelleri elde edilebilir.

Veri setinde bulunmayan yeni sanal niteliğin oluşturulması amacıyla gereken ekonomik ve matematiksel hesaplamalar için özellikle bu alanda çalışan finans uzmanlarıyla çoklu disiplinli bir çalışma yapılması gerekmektedir. Finans uzmanlarının önerileriyle FICO puanına benzer bir risk puanı oluşturabilir ve bu puanın başarısı makine öğrenmesi ile test edilebilir. Lending Club veri setinde üretilen makine öğrenmesi modeli web servis aracılığı ile bir yazılıma bağlanabilir böylece çeşitli parametrelerin kullanılabildiği bir makine öğrenmesi tabanlı bir yazılım elde edilmiş olunur. Üretilecek ekonomik model için piyasada bulunan mevcut risk uygulamaları incelenip bu modellerden fikir alınıp yapılacak olan modelde benzer bir model oluşturabilir.

KAYNAKLAR

1. İnternet: Li P. and Han G., Lending Club Loan Default and Profitability Prediction, URL: <http://cs229.stanford.edu/proj2018/report/69.pdf>, Son Erişim Tarihi: 12.02.2020.
2. İnternet: Chang S., Kim S. and Kondo S., Predicting Default Risk of Lending Club Loans, http://cs229.stanford.edu/proj2015/199_report.pdf, Son Erişim Tarihi: 12.02.2020.
3. İnternet: Feis A., Mehta V., Morris S., Solitario J. and Graaf C., P2P Loan Selection, URL: <http://stanford.edu/class/msande448/2016/final/group4.pdf>, Son Erişim Tarihi: 12.02.2020.
4. İnternet: Budak H. ve Erpolat S., Kredi Riski Tahmininde Yapay Sinir Ağları ve Lojistik Regresyon Analizi Karşılaştırılması, URL: https://www.researchgate.net/publication/276233830_Kredi_Riski_Tahmininde_Yapay_Sinir_Aglari_ve_Lojistik_Regresyon_Analizi_Karsilastirilmesi, Son Erişim Tarihi: 12.02.2020.
5. İnternet: FCA, URL: <https://www.fca.org.uk>, Son Erişim Tarihi: 02.10.2019.
6. İnternet: The disappearance of peer to peerlending, URL: <https://www.forbes.com/sites/groupthink/2014/10/14/the-disappearance-of-peer-to-peerlending>, Son Erişim Tarihi: 14.10.2019.
7. İnternet: Lending club press, URL: <https://www.lendingclub.com/public/lending-club-press-2015-01-15.action>, Son Erişim Tarihi: 02.11.2019.
8. İnternet:Lending-club-makes-a-surprise-comeback, URL:<https://www.businessinsider.com/lending-club-makes-a-surprise-comeback-2017-8>, Son Erişim Tarihi: 20.12.2019.
9. İnternet: Companies, URL: <http://www.p2pmoney.co.uk/companies.htm>, Son Erişim Tarihi: 22.12.2019.
10. İnternet: As filed with the Securities and Exchange Commission, URL: <https://www.sec.gov/Archives/edgar/data/1409970/000119312514323136/d766811ds1.htm>, Son Erişim Tarihi: 02.01.2020.
11. Johnson D., Oles F., Zhang T and Goetz T, 2002, A decision-tree-based symbolic rule induction system for text categorization by Article, *Ibm Systems Journal*, 41(3), 428-437.
12. İnternet: Hbchap9, URL: <http://www.ise.bgu.ac.il/faculty/liorr/hbchap9.pdf>, Son Erişim Tarihi: 19.05.2020.
13. İnternet: The ID3 Algorithm, URL: <https://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm>, Son Erişim Tarihi: 19.05.2020.

14. İnternet: Ağaca Dayalı Yöntemlerde Bagging ve Boosting Arasında ne Fark var? , URL: <https://www.veribilimiokulu.com/agaca-dayali-yontemlerde-bagging-ve-boosting-arasinda-ne-fark-var>, Son Erişim Tarihi: 19.05.2020.
15. İnternet: Electromyographic Patterns during Golf Swing: Activation Sequence Profiling and Prediction of Shot Effectiveness, URL: <https://pubmed.ncbi.nlm.nih.gov/27120604/>, Son Erişim Tarihi: 19.05.2020.
16. İnternet: Bagging and Boosting, URL: <https://www.mit.edu/~9.520/spring06/Classes/class10.pdf>, Son Erişim Tarihi: 19.05.2020.
17. İnternet: Destek Vektör Makineleriyle Sınıflandırma Problemlerinin Çözümü İçin Çekirdek Fonksiyonu Seçimi, URL: https://www.researchgate.net/publication/291347708_Destek_vektor_makineleriyle_sinfla ndirma_problemlerinin_cozumu_icin_cekirdek_fonksiyonu_secimi, Son Erişim Tarihi: 19.05.2020.
18. İnternet: İleri Algoritma Analizi, URL: <http://www.ibrahimcayiroglu.com/dokumanlar/ilerialgoritmaanalizi/ilerialgoritmaanalizi-5.hafta-yapaysiniraglari.pdf>, Son Erişim Tarihi: 19.05.2020.
19. İnternet: Yapay Sinir Ağları ile Borsa Endeksi Tahmini, URL: <http://www.acarindex.com/dosyalar/makale/acarindex-1423905602.pdf>, Son Erişim Tarihi: 18.05.2020.
20. İnternet: Regression-analysis, URL: <https://www.listendata.com/2018/03/regression-analysis.html>, Son Erişim Tarihi: 10.01.2020.
21. İnternet: Association-rule-mining-be, URL: <https://towardsdatascience.com/association-rule-mining-be4122fc1793>, Son Erişim Tarihi: 13.01.2020.
22. İnternet: Getting data ready for modelling feature engineering feature selection dimension reduction, URL: <https://towardsdatascience.com/getting-data-ready-for-modelling-feature-engineering-feature-selection-dimension-reduction-39dfa267b95a>, Son Erişim Tarihi: 10.02.2020.
23. İnternet: Ficoscore, URL: <https://www.investopedia.com/terms/f/ficoscore.asp>, Son Erişim Tarihi: 25.11.2019.
24. İnternet: Credit risk prediction in an imbalanced social lending environment, URL: <https://arxiv.org/ftp/arxiv/papers/1805/1805.00801.pdf>, Son Erişim Tarihi: 25.05.2020.
25. İnternet: Performance Analysis of Credit Scoring Models on Lending Club Data, URL: <https://www.semanticscholar.org/paper/Performance-Analysis-of-Credit-Scoring-Models-on-Polena/020eb2cbf5e818ad2cf240c1d11e934d8fdb581f>, Son Erişim Tarihi: 25.05.2020.

EKLER

EK-1. R diliyle Normalizasyon Kodları

Giriş bağlantı noktalarını (portlara) değişkenlere atama için aşağıdaki kodları kullanıyoruz.

```
loan_df <- maml.mapInputPort (1) # class: data.frame
```

```
loan_df$"Loan Status" <- ifelse (loan_df$"Loan Status" == "Charged Off", "1", "0")
```

```
# Select data.frame to be sent to the output Dataset port
```

```
maml.mapOutputPort ("loan_df");
```

```
library ("stats")
```

Giriş bağlantı noktalarını (portlara) değişkenlere atama için aşağıdaki kodları kullanıyoruz.

```
loan_df <- maml.mapInputPort (1) # class: data.frame
```

```
#Sayısal verileri temizleme ve sayısal olmayan karakterleri silme loan_df$"Maximum
```

```
Open Credit" <- gsub ("^[^0-9]", "", loan_df$"Maximum Open  
Credit")
```

```
loan_df$"Maximum Open Credit" <- as.numeric (loan_df$"Maximum Open Credit")
```

```
loan_df$"Monthly Debt" <- gsub ("^[^0-9]", "", loan_df$"Monthly Debt")
```

```
loan_df$"Monthly Debt" <- as.numeric (loan_df$"Monthly Debt") loan_df$"Months
```

```
since last delinquent" <- gsub ("^[^0-9]", "", loan_df$"Months since  
last delinquent")
```

EK-1. (Devam) R diliyle Normalizasyon Kodları

```
loan_df$"Months since last delinquent" <- as.numeric (loan_df$"Months since last
delinquent")
```

```
loan_df$"Bankruptcies" <- gsub ("^[^0-9]", "", loan_df$"Bankruptcies")
```

```
loan_df$"Bankruptcies" <- as.numeric (loan_df$"Bankruptcies")
```

```
loan_df$"Tax Liens" <- gsub ("^[^0-9]", "", loan_df$"Tax Liens")
```

```
loan_df$"Tax Liens" <- as.numeric (loan_df$"Tax Liens")
```

```
#Mevcut iş değişkenlerinde yıllara kategori ekleme
```

```
loan_df$"Years in current job" <- gsub ("[1-5]\\syyears", "1-5 years", loan_df$"Years
in current job")
```

```
loan_df$"Years in current job" <- ifelse (loan_df$"Years in current job" == "1 year",
"1-5 years", loan_df$"Years in current job")
```

```
loan_df$"Years in current job" <- gsub ("[6-9]\\syyears", "6-9 years", loan_df$"Years
in current job")
```

```
loan_df$"Years in current job" <- gsub ("n/a", NA, loan_df$"Years in current job")
```

```
#1000'in üzerindeki kredi puanlarında kredi puanını sabitleme
```

```
loan_df$"Credit Score" <- ifelse (loan_df$"Credit Score" > 1000, 650,
loan_df$"Credit Score")
```

```
# Çıktı veri kümesi portuna gönderilecek data frame ögesi seçilmektedir.
```

```
maml.mapOutputPort ("loan_df");
```

EK-1. (Devam) R diliyle Normalizasyon Kodları

Giriş bağlantı noktalarını (portlara) değişkenlere atama için aşağıdaki kodları kullanıyoruz.

```
loan_df <- maml.mapInputPort (1) # class: data.frame
```

```
### Nas (Boş) değerleri temizleme
```

```
loan_df$"Bankruptcies" <- ifelse (is.na (loan_df$"Bankruptcies") == TRUE, 0,
loan_df$"Bankruptcies")
```

```
loan_df$"Tax Liens" <- ifelse (is.na (loan_df$"Tax Liens") == TRUE, 0,
loan_df$"Tax Liens")
```

```
loan_df$"Credit Score" <- ifelse (is.na (loan_df$"Credit Score" == TRUE), 0,
loan_df$"Credit Score")
```

```
loan_df$"Annual Income" <- ifelse (is.na (loan_df$"Annual Income") == TRUE, 0,
loan_df$"Annual Income")
```

```
loan_df$"Years in current job" <- ifelse (is.na (loan_df$"Years in current job") ==
TRUE, "1-5 years", loan_df$"Years in current job")
```

```
loan_df$"Months since last delinquent" <- ifelse (is.na (loan_df$"Months since last
delinquent") == TRUE, 0, loan_df$"Months since last delinquent")
```

```
loan_df$"Maximum Open Credit" <- ifelse (is.na (loan_df$"Maximum Open Credit")
== TRUE, 7000, loan_df$"Maximum Open Credit")
```

```
####
```


EK-1. (Devam) R diliyle Normalizasyon Kodları

```
maml.mapOutputPort ("loan_df");
```

Giriş bağlantı noktalarını (portlara) değişkenlere atama için aşağıdaki kodları kullanıyoruz.

```
loan.df <- maml.mapInputPort (1) # class: data.frame
```

```
#d <- subset (loan.df, loan.df$"Loan Status" == "1")
```

```
#n <- 2
```

```
#d <- do.call ("rbind", replicate (n, d, simplify = FALSE))
```

```
#loan.df <- rbind (loan.df, d)
```

```
if (FALSE)
```

```
{
```

```
loan.df$"Annual Income" <- NULL
```

```
loan.df$"Bankruptcies" <- NULL
```

```
loan.df$"Tax Liens" <- NULL
```

```
loan.df$"Current Loan Amount" <- NULL
```

```
loan.df$"Years in current job" <- NULL
```

```
loan.df$"Monthly Debt" <- NULL
```

```
loan.df$"Current Credit Balance" <- NULL
```

```
loan.df$"Home Ownership" <- NULL
```

```
loan.df$"Term" <- NULL
```

```
loan.df$"Number of Open Accounts" <- NULL
```

```
loan.df$"Bankruptcies" <- NULL
```

```
loan.df$"Tax Liens" <- NULL
```

```
loan.df$"Purpose" <- NULL loan.df$"Years  
of Credit History" <- NULL  
loan.df$"Months since last delinquent" <- NULL  
loan.df$"Number of Credit Problems" <- NULL  
}
```

```
loan.df$"Loan ID" <- NULL
```

```
# Çıktı veri kümesi portuna gönderilecek data frame ögesi seçilmektedir.  
maml.mapOutputPort ("loan.df");
```

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı : OKUR, Hamdi
 Uyuğu : T.C.
 Doğum tarihi ve yeri : 11.10.1985, Elazığ
 Medeni hali : Bekar
 Telefon : 0 (539) 726 40 85
 e-mail : hamdiokur1@gmail.com



Eğitim

Derece	Eğitim Birimi	Mezuniyet Tarihi
Y.Lisans	Gazi Üniversitesi / Bilgisayar Mühendisliği	Devam ediyor
Lisans	Fırat Üniversitesi / Bilgisayar Mühendisliği	2009
Lise	Elazığ Anadolu Lisesi	2002

İş Deneyimi

Yıl	Yer	Görev
2015-Halen	KOSGEB	KOBİ Uzmanı
2011-2015	TEİŞ Genel Müdürlüğü	Mühendis

Yabancı Dil

İngilizce

Yayınlar

Okur H. and Cetin A., (2019), "Credit Risk Estimation With Machine Learning," *2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, Ankara, Turkey.

Hobiler

Gezi, Kitap, Fotoğraf



GAZİ GELECEKTİR..