
GAZİ ÜNİVERSİTESİ
Gazi
1926

BÜYÜK VERİDE ÇİZGE TEORİSİYLE TEMERRÜT TAHMİNİ VE MAKİNE ÖĞRENMESİ MODELLERİNİN YORUMLANMASI

Mustafa YILDIRIM

DOKTORA TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANA BİLİM DALI

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

EYLÜL 2021

ETİK BEYAN

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

- 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 çalışmasında 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.

Mustafa YILDIRIM

23/09/2021

BÜYÜK VERİDE ÇİZGE TEORİSİYLE TEMERRÜT TAHMİNİ VE MAKİNE ÖĞRENMESİ MODELLERİNİN YORUMLANMASI

(Doktora Tezi)

Mustafa YILDIRIM

GAZİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

Eylül 2021

ÖZET

Son yıllarda, artan veri kaynağı sayısı, veri toplama, depolama ve işleme maliyetlerinin düşmesi ve veri analizi için yeni yöntemlerin geliştirilmesi büyük veri olarak adlandırılan yeni bir dönemin başlamasına sebep olmuştur. Büyük veri teknolojileriyle daha önce yönetilemeyecek ve işlenemeyecek boyuttaki veriler işlenerek veri içinde saklı olan kıymetli bilgiler keşfedilebilir hale gelmiştir. Bu çalışmada literatürde uzun yıllardır çalışılan ve önemi her geçen gün artan şirketlerin temerrüde düşme tahmini için büyük veri teknolojinden faydalanılmıştır. Bu kapsamda büyük veri platformu üzerinde Makine Öğrenmesi (Machine Learning / ML) ve çizge (graf) teorisinden faydalanarak iki farklı temerrüt tahmin modeli önerilmiştir. Çalışmada, Türkiye'de 2010 ve 2018 yılları arasında faaliyet gösteren 1 milyondan fazla reel sektör şirketinin kredi, bilanço ve fatura veri seti kullanılmıştır. İlk modelde istatistik ve ML algoritmaları kullanılarak kredi ve bilanço veri setleri için iki alt model oluşturulmuş ve bu alt modellerden elde edilen olasılık skorları nihai modelde birleştirilerek en iyi tahmine ulaşılmıştır. İkinci önerilen modelde ise çizge teorisinden faydalanılmıştır. Temerrüde düşmede şirketlerin iç dinamiklerinin yanı sıra ticari ilişki içinde oldukları tedarikçi ve müşterilerinin de önemli olduğu temel varsayımdan yola çıkılmıştır. Bu nedenle, şirketlerin fatura verisi üzerinden ticari ilişkiyi gösteren çizge oluşturulmuştur. Çizge üzerinden temerrüt tahminine fayda sağlayacak yeni değişkenler üretilmiştir. İkinci modelde bu değişkenler kullanılmıştır. Sonuçlara bakıldığında her iki modelin sırasıyla 0,81 ve 0,82 Eğri Altında Kalan Alan (Area Under Curve /AUC) skor elde ettiğini ortaya koymuştur. İkinci modelin daha yüksek tahmin başarısı sağlaması çizge üzerinden elde edilen yeni değişkenlerin temerrüt tahminine katkı sağladığını göstermiştir. Tez kapsamında son olarak temerrüt tahminde karmaşık ML algoritmalarının kullanılmasına getirilen en önemli eleştiri olan sonuçların açıklanabilir olmamasına Yorumlanabilir Makine Öğrenmesi (Interpretable Machine Learning / IML) algoritmalarıyla çözüm aranmıştır. Sonuçlar IML'nin karmaşık ML algoritmalarının açıklanmasında tutarlı ve güvenilir çıktılar verdiğini göstermektedir.

Bilim Kodu	:	92431
Anahtar Kelimeler	:	Temerrüt tahmini, makine öğrenmesi, büyük veri, çizge (graf) teorisi, yorumlanabilir makine öğrenmesi
Sayfa Adedi	:	73
Danışman	:	Prof. Dr. Suat ÖZDEMİR

DEFAULT PREDICTION WITH GRAPH THEORY IN BIG DATA AND INTERPRETATION OF MACHINE LEARNING MODELS

(Ph. D. Thesis)

Mustafa YILDIRIM

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

September 2021

ABSTRACT

In recent years, the increase in the number of data sources, the decrease in the data collection, storage and processing costs, and the development of new methods for data analysis have led to the beginning of a new era called big data. Big data technologies have enabled the process of the data that could not be managed and processed before and explore valuable information hidden in the data. In this study, big data technology is used for the default prediction of companies. Default prediction has been studied for many years in the literature and its importance is increasing day by day. Two different default prediction models are proposed using Machine Learning (ML) and graph theory on a big data platform. In the study, credit, balance sheet and invoice datasets of more than 1 million real sector companies operated in Turkey between 2010 and 2018 are used. In the first model, two sub-models are created for credit and balance sheet datasets by using statistics and ML algorithms, and the probability scores obtained from these sub-models are combined to reach the best estimate in the final model. In the second model, graph theory is employed. It is based on the basic assumption that the internal dynamics of the companies, as well as the suppliers and customers, with whom they have commercial relations, are also important in the default of the companies. Therefore, a graph showing the commercial relationship is created using the invoice data of the companies. New variables that help explore default prediction are generated on the graph. These variables are further used in the second model. The results showed that both models achieved 0.81 and 0.82 Area Under Curve (AUC) scores, respectively. The higher prediction success of the second model showed that the new variables obtained from the graph contributed to the default prediction. Within the scope of the thesis, finally, a solution has been sought with Interpretable Machine Learning (IML) algorithms for the interpretability of the results, which is the most important criticism regarding the use of complex ML algorithms in default prediction. The interpretability results also indicated that IML gives consistent and reliable outcomes in explaining complex ML models.

Science Code	:	92431
Key Words	:	Default prediction, machine learning, big data, graph theory, interpretable machine learning
Page Number	:	73
Supervisor	:	Prof. Dr. Suat ÖZDEMİR

TEŞEKKÜR

Çalışmalarım sırasında her türlü destek, yardım ve katkılarıyla beni yönlendiren değerli hocam Prof. Dr. Suat ÖZDEMİR'e en içten teşekkürlerimi sunarım. Tez izleme komitemde yer alan Prof. Dr. Şeref SAĞIROĞLU ve Dr. Öğr. Üyesi Adnan ÖZSOY hocalarıma verdikleri ufuk açıcı geribildirimlerden dolayı çok teşekkür ederim. Doktora çalışmalarım süresince beraber çalışma fırsatı bulduğum değerli arkadaşım Dr. Feyza YILDIRIM OKAY'a teşekkür ederim.

Çalışmaktan her zaman gurur duyduğum Türkiye Cumhuriyet Merkez Bankası'na, tezim için sağladığı veri ve altyapı desteğinden dolayı teşekkür ederim. Ayrıca tez çalışmama katkılarından dolayı başta Murat TOPKAYA, Merve ARTMAN ve Berkay AKIŞOĞLU olmak üzere tüm çalışma arkadaşlarıma teşekkürlerimi sunarım.

Son olarak hayatım boyunca desteklerini esirgemeyen aileme gönülden sonsuz teşekkür ederim. Doktora sürecinin yoğun ve tempolu çalışma süresince sabır ve anlayış gösteren sevgili eşim ve canım oğluma teşekkürü borç bilirim.

İÇİNDEKİLER

Sayfa

Sayfa

Sayfa

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

SİMGELER VE KISALTMALAR

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

Kısaltmalar	Açıklamalar
ANN	Artificial Neural Network / Yapay Sinir Ağı
AUC	Area Under Curve / Eğri Altında Kalan Alan
BVP	Büyük Veri Platformu
CNN	Convolutional Neural Networks/ Evrişimli Sinir Ağları
CV	Cross Validation / Çapraz Doğrulama
DBN	Deep Belief Network / Derin İnanç Ağları
DL	Deep Learning / Derin Öğrenmesi
DT	Decision Tree / Karar Ağacı
ELI5	Explain Like I'm Five
GB	Gradient Boosting / Dereceli Artırım
GİB	Gelir İdaresi Başkanlığı
GRU	Gated Recurrent Units / Kapılı Yinelemeli Üniteler
HDFS	Hadoop Distributed File System / Hadoop Dağıtık Dosya Sistemi
IML	Interpretable Machine Learning / Yorumlanabilir Makine Öğrenmesi
LIME	Local Interpretable Model-agnostic Explanations / Lokal Vekil Model
IDC	International Data Corporation / Uluslararası Veri Birliği
IoT	Internet of Things / Nesnelerin İnterneti
JSON	JavaScript Object Notation / JavaScript Nesne Notasyonu
KDV	Katma Değer Vergisi
KOBİ	Küçük ve Orta Büyüklükteki İşletmeler
LASSO	Least Absolute Shrinkage And Selection Operator / En Az Mutlak Büzülme Seçici Operatörü
LR	Logistic Regression / Lojistik Regresyon

Kısaltmalar	Açıklamalar
LSTM	Long Short Term Memory / Uzun Kısa Süreli Bellek
MDA	Multiple Discriminant Analysis / Çoklu Diskriminant Analizi
MI	Mutual Information / Karşılıklı Bilgi
ML	Machine Learning / Makine Öğrenmesi
MLP	Multilayer Perceptron / Çok Katmanlı Algılayıcı
NACE	Statistical Classification of Economic Activities in The European Community/ Avrupa Topluluğunda Ekonomik Faaliyetlerin Genel Sınıflandırılması
RBM	Restricted Boltzmann Machine/ Kısıtlı Boltzmann Makineleri
RDD	Resilient Distributed Dataset / Esnek Dağıtık Veri Seti
RF	Random Forest / Rastgele Orman
RMSE	Root Mean Squared Error / Ortalama Karesel Hata
RNN	Recurrent Neural Network / Yinelenen Sinir Ağları
ROC	Receiver Operating Characteristic/ Alıcı İşlem Karakteristikleri
SDA	Single Discriminant Analysis / Tekil Diskriminant Analizi
SFS	Sequential Forward Selection / Ardışık İleri Yönde Seçim
SHAP	SHapley Additive exPlanations / Shapley Katkı Açıklamalar Modeli
SQL	Structured Query Language / Yapılandırılmış Sorgu Dili
SVM	Support Vector Machine / Destek Vektör Makinesi
TCMB	Türkiye Cumhuriyet Merkez Bankası
TDHP	Tek Düzen Hesap Planı
XGBOOST	Extreme Gradient Boosting / Ekstrem Dereceli Artırım
XML	Extensible Markup Language / Genişletilebilir Biçimlendirme Dili
ZPP	Zero Price Probability / Sıfır Fiyat Olasılığı

1. GİRİŞ

Temerrüt, en temel tanımıyla bir şirketin borçlarını ödeme güçlüğüne düşmesi ve taahhütlerini yerine getirememesidir. Bu durum iflasın öncü bir göstergesi olarak kabul edilmektedir. Şirketlerin temerrüde düşme olasılığının tahmin edilmesi çok uzun yıllardır hem akademisyenlerin hem de finans uzmanlarının üzerinde çalıştığı önemli konulardandır. Temerrüdün doğru tahmin edilmesi sigorta şirketleri, bankalar, finansal kurum ve kuruluşlar, yatırımcılar, fon yöneticileri gibi birçok kişi ve kurumun risk yönetiminde önemli bir yere sahiptir.

Temerrüt ve iflas tahmininde ilk yıllarda daha az değişkenle istatistik temelli doğrusal ayrımlar yapan ekonometrik yöntemler kullanılmıştır. 1966 yılında Beaver'ın [1] önerdiği Tekil Diskriminant Analizi (Single Discriminant Analysis / SDA) ve 1968 yılında Altman'ın [2] önerdiği Çoklu Diskriminant Analizi (Multiple Discriminant Analysis / MDA) konuyla ilgili ilk çalışmalardır. Altman'ın çalışması özellikle bu alandaki öncü çalışmalardandır ve hala birçok çalışmada şirketlerin bilanço kalemlerinden oluşturulan değişkenler kullanılmaktadır. Temerrüt tahminiyle ilgili bir diğer önemli çalışma ise 1980 yılında Ohlson'ın [3] önerdiği Lojistik Regresyonla (Logistic Regression / LR) yapılan tahmindir. LR, bu çalışmada da olduğu gibi, hala birçok çalışmada kullanılmaktadır. 1990'lardan itibaren temerrüt ve iflas tahmini için SDA, MDA veya LR'na göre daha yüksek doğruluk oranına sahip birçok farklı Makine Öğrenmesi (Machine Learning / ML) yöntemi önerilmiştir. Bunlarda en sık kullanılanlara örnek olarak Karar Ağaçları (Decision Tree / DT), Destek Vektör Makineleri (Support Vector Machine / SVM), Yapay Sinir Ağları (Artificial Neural Network / ANN), Dereceli Artırım (Gradient Boosting / GB) ve Rastgele Orman (Random Forest / RF) algoritmaları verilebilir. Son yıllarda ise Derin Öğrenme (Deep Learning / DL) algoritmaları temerrüt tahmininde sıklıkla kullanılmaktadır [4-6].

Yukarıda bahsedilen tahmin yöntemlerindeki gelişmelerin yanında bilişim alanında donanım maliyetlerinin ucuzlaması, veri toplama, depolama, analizi gibi işlemlerin kolay hale gelmesi ve verinin önemli bir ekonomik değer olduğunun fark edilmesi Büyük Veri (Big Data) olarak adlandırılan yeni bir dönemi başlatmıştır. Bu dönemin başlamasındaki önemli bir diğer etki ise veri çeşitliğinin artmasıdır. Günümüzde Nesnelerin İnterneti (Internet of Things /IoT), akan veriler, mobil cihazlar, sosyal medya gibi birçok yeni veri

kaynağı oluşmuştur. Büyük veri teknolojisiyle geleneksel yöntemlerle yönetilemeyen ve analiz edilemeyen boyut ve karmaşıklıktaki veriler işlenerek içindeki kıymetli bilgiler çıkarılabilmektedir. Bu durum birçok kurum, kuruluş ve şirketin stratejik kararlar alma sırasında veri güdümlü (data driven) davranmasına ve kararlarını daha analitik çerçevede vermesine imkan sağlamaktadır.

Bu tez kapsamında büyük veri çağının getirdiği, teknolojik imkânlar, artan veri kaynakları ve yeni analiz yöntemlerinden faydalanarak uzun yıllardır çalışılan temerrüt tahmini çalışmalarının geliştirilmesi amaçlanmıştır. Çalışmanın deneysel bölümleri Büyük Veri Platformu (BVP) üzerinde yapılmıştır. Yapılan çalışmalar üç ana bölüme ayrılabilir. İlk bölümde istatistik ve ML yöntemleri kullanılarak oluşturulan temerrüt tahmin modeli açıklanmıştır. İkinci bölümde çizge teorisi yöntemleriyle oluşturulan temerrüt tahmin modeli önerilmiştir. Son bölümde ise temerrüt tahmininde ML yöntemlerinin kullanımına getirilen en önemli eleştiri olan sonuçların yorumlanabilir olmaması problemine çözüm aranmıştır.

Çalışmada Türkiye'de faaliyet gösteren 1 milyondan fazla reel sektör şirketinin 2010-2018 yılları arasındaki bilanço, kredi ve fatura veri setlerinden faydalanılmıştır. Bu veri setleriyle şirketlerin gelecekte temerrüde düşme tahmini yapılmıştır. Temerrüde düşmenin çok farklı kriterleri olmakla birlikte bu çalışmada temerrüt, ödemesi 90 gün geçmiş kredi borcu, çek veya senedin ödenmemesi ve hukuki iflas veya konkordato olarak tanımlanmıştır.

İlk bölümde bilanço ve kredi verilerinden iki alt model oluşturulmuştur. Bu alt modellerde istatistik ve ML yöntemleri kullanılarak karşılaştırmalı olarak değerlendirilmiştir. Alt modellerden elde edilen olasılık skorları farklı bir modelde birleştirilerek nihai sonuçlar elde edilmiştir. Sonuçlar farklı metrikler kullanılarak ayrıntılı olarak değerlendirilmiştir. Ayrıca son yıllardaki tahmin problemlerinde yüksek başarı performansı gösteren DL algoritmalarıyla, kredi veri setinden örneklem seçilerek temerrüt tahmininde DL yöntemlerinin başarısı incelenmiştir.

İkinci bölümde bilanço ve kredi verilerinin yanı sıra şirketlerin fatura verisi temerrüt tahmininde kullanılmıştır. Faturadaki alıcı ve satıcı şirket bilgisi kullanılarak çizge yapısı modellenmiştir. Çizge teorisi yöntemleri kullanılarak temerrüt tahmininde fayda sağlayacak

yeni değişkenler önerilmiştir. Önerilen değişkenlerin ilk bölümde elde edilen temerrüt tahminine katkısı incelenmiştir.

Son bölümde ise temerrüt tahmininde ML yöntemlerine getirilen en önemli eleştiri olan sonuçların yorumlanabilir olmamasına çözüm aranmıştır. ML yöntemlerinde genel olarak karmaşıklık düzeyi arttıkça başarı artmaktadır. Ancak, bununla birlikte modelin yorumlanabilirliği azalmaktadır. Bu durum yüksek başarıya rağmen ML yöntemlerinin etkin kullanılamaması problemine yol açmaktadır. Bu problem için literatürde giderek daha popüler olan Yorumlanabilir Makine Öğrenmesi (Interpretable Machine Learning / IML) yöntemleri önerilmektedir. Ayrıca bu kapsamda IML algoritmalarının güvenilirliği için öznitelik seçimi ve yorumlanabilir modellerle karşılaştırmalı analizi yapılmıştır.

Temerrüt tahmini için tez kapsamında yapılan çalışmanın ana katkıları şunlardır:

- BVP'nin yüksek performansından da yararlanarak literatürdeki en yüksek veri ve değişken sayısıyla yapılan temerrüt tahmini çalışması durumundadır.
- Çalışmada şirketler için sektör, ölçek, aktif büyüklük gibi herhangi bir kısıtlama uygulanmamıştır. Özellikle küçük ölçekli şirketlerin sayısının çok ve veri kalitesinin düşük olmasından dolayı bu şirketlere ait bilgiler birçok temerrüt tahmini çalışmasında dışlanmakta ve bu durum tahmin başarısını yükseltmektedir. Bu çalışma ise hiçbir şirket dışlamamış olmasına rağmen hala yüksek tahmin başarısına sahiptir.
- Bilgimiz dahilinde bu çalışma temerrüt tahmininde fatura verisiyle çizge teorisinden yararlanan ilk çalışmadır. Böylelikle temerrüt tahmininde şirketlerin bilanço, kredi gibi verilerinin yanında fatura verisinden elde edilen, iş ilişkisi içinde oldukları diğer şirketlerin de etkisinin olduğu ortaya konmuştur.
- Karmaşık ML algoritmalarının doğruluk oranının yüksek olmasına rağmen yorumlanabilir olmamasından dolayı temerrüt tahmininde kullanılmaması problemine IML algoritmalarıyla çözüm getirilmiştir.

Çalışmanın geri kalan kısmı ise şu şekilde organize edilmiştir.

Bölüm 2, temerrüt tahmini için önerilen ilk model ayrıntılarıyla açıklanmıştır. Literatürdeki benzeri çalışmalar incelenmiş, büyük veri hakkında genel bilgi verilmiş ve çalışmada kullanılan büyük veri teknolojileri özetlenmiştir. Analizlerde kullanılan sınıflandırma

algoritmaları açıklanmıştır. Kullanılan veri seti, deneysel ortam ve performans metrikleri hakkında detaylı bilgi verilmiştir. ML kullanılarak önerilen temerrüt tahmin modeli açıklanmış ve deneysel sonuçlar analiz edilerek değerlendirmeleri verilmiştir. Ayrıca temerrüt tahmininde DL algoritmalarının gelecek vadettiğini göstermek için ana veri setinden seçilen örneklem şirketler üzerinden analiz yapılmıştır.

Bölüm 3, çizge teorisiyle temerrüt tahmin modelinin ayrıntıları açıklanmıştır. Literatürde çizge teorisi ve ağ analizinin finans alanında kullanımıyla ilgili çalışmalar incelenmiştir. Çalışmada kullanılan çizge teorisi yöntemleri açıklanmış, merkezilik değişkenleri hakkında bilgi verilmiştir. Kullanılan veri setlerinin detayları verilmiştir. Çizge teorisiyle temerrüt tahmin modelinin ayrıntıları, deneysel çalışmalardan çıkan sonuçlar ve değerlendirmeler sunulmuştur.

Bölüm 4, bölüm 2 ve bölüm 3'te önerilen temerrüt tahmini modellerinin yorumlanabilirlik problemine IML ile çözüm getirilmiştir. IML ile ilgili literatür taraması yapılmış, çalışmada kullanılan Shapley Katkı Açıklamalar Modeli (SHapley Additive exPlanations / SHAP) ve Lokal Vekil Model (Local Interpretable Model-agnostic Explanations / LIME) algoritmaları hakkında bilgi verilmiştir. Tezde önerilen her iki temerrüt tahmini modeli için IML algoritmalarından alınan global ve lokal sonuçlar değerlendirilmiştir. Ayrıca IML ile öznitelik seçimi ve yorumlanabilir modellerin sonuçları karşılaştırılarak IML algoritmalarının güvenilirliği ortaya konulmuştur.

Bölüm 5, tez çalışmasında elde edilen sonuçlar ve değerlendirmeler özetlenmiştir. Ayrıca geleceğe yönelik çalışma önerileri sunulmuştur.

2. MAKİNE ÖĞRENMESİYLE TEMERRÜT TAHMİNİ

Reel sektör şirketlerinin temerrüde düşmesi tahmini çok uzun yıllardır hem akademisyenler hem de profesyonel iş analistleri tarafından çalışılan önemli bir konudur. Temerrüt, kısaca bir şirketin borçlarını ödeme güçlüğüne düşmesidir. Temerrüdün tahmin edilmesi reel sektör şirketleri için ticari ilişki kuracakları şirketleri belirlemede, bankalar için şirketlere kredi sağlarken risklerini ölçmede, Merkez bankaları gibi ekonomi yönetiminden sorumlu kurumlar içinse reel sektörün durumunu anlamada öncü bir gösterge niteliğinde olmaktadır. Bu bölümde Türkiye'de faaliyet gösteren reel sektör şirketlerinin kredi ve bilançoları üzerinden temerrüt tahmini yapılmıştır. Tahmin için BVP üzerinde çalışılmış ve istatistik ve ML yöntemleri kullanılmıştır. Yöntemlerden elde edilen tahmin sonuçları detaylı ve karşılaştırmalı olarak analiz edilmiştir.

Temerrüt tahmini uzun yıllardır çalışılan bir konu olmasına rağmen genellikle çalışmalar kısıtlı şirket sayısıyla ve bu şirketlerin kredi veya bilanço veri setlerinden biriyle yapılmıştır. Bunun genellikle iki önemli sebebi bulunmaktadır. Bunlardan ilki kredi, bilanço gibi veri setlerinin şirketlerin hassas bilgilerini içerdiği için erişiminin zorluğundandır. İkincisiyse veri setlerine erişim sağlansa bile bu büyüklükte veri setlerinin analizinin yapılması için gerekli teknik altyapının olmamasıdır. Bu çalışmada her iki problemde aşılmıştır. Toplam 1 milyonun üzerinde reel sektör şirketinin kredi ve bilanço veri seti kullanılarak temerrüt tahmini yapılmıştır. Bu büyüklükteki veri setlerinin analizi için büyük veri teknolojisinden faydalanılmıştır. Böylelikle sektör, ölçek gibi herhangi bir filtrelemeye maruz bırakılmadan tüm şirketler için temerrüt tahmini yapılabilmiştir. Tahmini güçlendirebilmek için kredi ve bilanço veri setlerinden elde edilen tahminler birleştirilmiştir. Ayrıca istatistiksel ve ML yöntemleri kullanılarak yöntemlerin başarı karşılaştırması yapılması amaçlanmıştır. Görebildiğimiz kadarıyla literatürdeki en çok şirket ve veri sayısıyla yapılan en kapsamlı temerrüt tahmini çalışmasıdır.

2.1. İlgili Çalışmalar

Şirketlerin temerrüt veya iflas tahmini literatürde uzun yıllardır çalışılan bir konudur. Konu ilk yıllarda istatistik temelli doğrusal ayrım yapan ekonometrik analiz yöntemleriyle incelenmiştir. Bu yöntemlerde en çok bilinen ve temel karşılaştırma noktası olarak kabul edilen Altman [2] ve Ohlson'a [3] ait çalışmalardır. Altman çalışmasında dönen varlıkların

toplam varlıklara oranı, toplam karın varlıklara oranı gibi çeşitli bilanço kalemlerinden oluşturduğu değişkenlerle bir şirketin finansal durumunu değerlendirilmiş ve z ayrım skoru olarak adlandırdığı bir nihai değer elde etmiştir. Ohlson ise bir şirketin temerrüde düşme olasılığını doğrusal regresyon kullanarak tahmin etmiştir. Doğrusal regresyon özellikle finans uzmanları tarafından hala sıklıkla kullanılmaktadır. Bunun sebebi görece yüksek tahmin başarısı ve yorumlanabilir olmasıdır.

Son yıllarda ise DT, RF, SVM, Çok Katmanlı Algılayıcı (Multilayer Perceptron / MLP), Ekstrem Dereceli Artırım (Extreme Gradient Boosting / XGBOOST) gibi birçok farklı ML algoritması temerrüt ve iflas tahmininde sıklıkla kullanılmaktadır [7-13]. Bunun en önemli sebebi onların gösterdiği yüksek tahmin başarısıdır. Kim ve arkadaşları [14] temerrüt tahmini için kullanılan ML algoritmalarıyla ilgili çalışmalarından karşılaştırmalı bir değerlendirme sunmuşlar ve ML algoritmalarının temerrüt tahmininde gelecek vaat ettiğini göstermişlerdir. Kim ve Sohn [15] Küçük ve Orta Büyüklükteki İşletmeler (KOBİ) için destek fonu dağılımı amacıyla temerrüt tahminine bakmıştır. Tahminde SVM, LR ve geri beslemeli sinir ağları yöntemlerini kullanmıştır. Deneysel sonuçlar SVM'nin daha yüksek başarı gösterdiğini ortaya koymuştur. Zhou ve arkadaşları [16] iflas tahminindeki dengesiz veri setinden kaynaklanan hassasiyet (precision) probleminin üstesinden gelmek için SVM yöntemini yönlü arama ve öznitelik sıralama yöntemleriyle birleştirmişlerdir. Deneylerde hibrit yöntemle ML yöntemlerini karşılaştırmışlardır. Sonuçlar hibrit yöntemin hassasiyet probleminde ML yöntemlerine göre daha başarılı olduğunu göstermiştir. Ancak bununla birlikte hesaplama maliyetinin arttığı görülmüştür. Wang, Ma ve Yang [17] şirketlerin iflas tahmini için geleneksel boosting (artırma) yöntemini öznitelik seçim yöntemi ile geliştirerek FS-boosting yöntemini önermişlerdir. Artırmalı yöntemin başarısını olumsuz etkileyen veri seti içindeki gürültü, filtrelemeyle öznitelik seçimi yöntemlerinden olan bilgi kazanımı yöntemi kullanılarak temizlenmiştir. Sonuçlar FS-boosting yönteminin tahmini artırdığını göstermiştir. Tsai ve Hsu [18] iflas tahmininde tekil ML algoritmaları ile topluluk tabanlı bagging (torbalama) ve boosting algoritmalarıyla birleştirilen metotları karşılaştırmıştır. Çalışmalarında Avustralya, Almanya ve Japonya kredi veri setlerini kullanmışlardır. Veri setleri sırasıyla 690, 1000, 690 adet şirket verisi içermektedir. Yapılan deneysel çalışmalar topluluk tabanlı yöntemlerin tekil ML algoritmalarına göre daha yüksek tahmin gücüne sahip olduğunu göstermiştir. En iyi tahmin boosting tabanlı DT algoritmasından elde edilmiştir. Barboza ve arkadaşları [19] iflas tahmini için istatistiksel ve ML tabanlı yöntemler için karşılaştırmalı sonuçlar sunmuşlardır. Çalışma 1985 ile 2013 yılları arasından

Kuzey Amerika'da faaliyet gösteren 10 binin üzerinde şirketin verisi kullanılarak yapılmıştır. Altman z skor değişkenlerinin yanında 6 adet yeni değişken önerilmiştir. İlk deneysel çalışma istatistiksel tabanlı diskriminant analiz ve LR ile ML tabanlı SVM, bagging, boosting ve RF algoritmalarının karşılaştırmasıdır. Sonuçlar, ML algoritmalarının istatistiksel algoritmalardan yüksek performans gösterdiğini ortaya koymuştur. Ayrıca çalışan sayısı, satış miktarı gibi yeni değişkenlerin kullanılması başarıyı artırdığını göstermiştir. Literatürde sıkça çalışılan bir diğer konu ise kredi skorlamadır. Kredi skorlama, iflas ve temerrüt tahminine benzer şekilde yapılmaktadır. Nehrebecka [20] LR ve SVM'le karşılaştırmalı olarak kredi skorlama yapmaktadır. Çalışmasında Polonya'da 2007-2016 yılları arasında faaliyet gösteren 100 binin üzerinde şirket verisini kullanmıştır. Bu şirketlerin yıllara göre temerrüt etme oranı değişmekle birlikte ortalama %5 temerrüt oranı bulunmaktadır. Çalışmanın değerlendirme metrikleri Gini indeksi, Kolmogorv-Smirnov ve Eğri Altında Kalan Alan (Area Under Curve / AUC)'dir. Deneysel sonuçlar LR, SVM'in farklı çeşitlerine kıyasla daha doğru sonuçlar verdiğini göstermektedir.

DL algoritmaları son yıllarda birçok tahmin probleminde olduğu gibi temerrüt tahmini probleminde de kullanılmaya başlanmıştır. Bu durum DL algoritmalarının zaman serisi problemlerinde sağladığı yenilikçi ve güçlü çözümlerden kaynaklanmaktadır.

Hosaka [21] Japonya borsalarında işlem gören firmalar için DL ve ML algoritmalarıyla iflas tahmini yapmıştır. Veri setinde 2002 ile 2016 yılları arasındaki 2164 şirket verisi kullanılmış ve bu şirketlerden 102 adedi iflas etmiştir. Çalışmada finansal oranlar üzerinde CNN ile görüntü tanıma yöntemi, doğrusal diskrimant analizi, SVM, MLP, AdaBoost kullanılmıştır. Sonuçlar CNN ile görüntü tanıma yönteminin tahmin başarısının diğerlerinden yüksek olduğunu göstermiştir.

Jing ve arkadaşları [22] temerrüt tahmininin olasılık skoru için günlük hisse senedi getirilerini kullanarak Sıfır Fiyat Olasılığı (Zero-Price Probability / ZPP) ve Uzun Kısa Süreli Bellek (Long Short Term Memory / LSTM) ile oluşturulan hibrit bir yöntem önermiştir. Çin'deki inşaat ve emlak sektörü firmaları veri setini kullanarak yapılan çalışmada önerilen hibrit yöntemi Kealhofer, McQuown ve Vasicek (KMW) ve sabit varyanslı ZPP ile karşılaştırmışlardır. Deneysel sonuçlar hibrit yöntemin diğer modellerden daha yüksek başarı sağladığını göstermektedir. Hisse senedi getirisiyle ilgili bir diğer çalışma Yeh [23] tarafından yapılmıştır. Temerrüde düşen ve düşmeyen firmaların hisse

senedi getirilerinden yararlanarak Derin İnanç Ağları (Deep Belief Network / DBN) ve SVM ile temerrüt tahmini yapılmıştır. Sonuçlar DBN'nin tahmin başarısının yüksek olduğunu göstermektedir. DBN kullanılarak yapılan bir diğer çalışma ise Luo ve arkadaşlarına aittir [24]. Bu çalışmada kredi temerrüt takası verisi üzerinden kredi skorlama için Kısıtlı Boltzmann Makineleri (Restricted Boltzmann Machines / RBM) ile DBN birleştirilmiştir. Yapılan deneysel çalışmalarda önerilen DBN yönteminin LR, MLP ve SVM algoritmalarına göre doğruluk ve AUC metrikleriyle değerlendirildiğinde daha yüksek başarı sağladığı görülmektedir. Kredi derecelendirmeyle ilgili yapılan bir diğer çalışma ise Addo ve arkadaşlarının [25] bir bankanın 2016 ve 2017 yılları içinde 115 288 adet sağlıklı 1731 adet temerrüde düşmüş şirket veri seti üzerinden kredi skorlaması çalışmasıdır. Bu veri setinde finansal durum, gelir tablosu, bilanço, nakit akış tablosu gibi şirket raporlarından oluşturulan toplam 235 adet değişken bulunmaktadır. Çalışma kredi derecelendirme DL tabanlı 4 yöntem ile LR, RF ve GB algoritmaları kullanılarak yapılmıştır. AUC ve Ortalama Karesel Hata (Root Mean Squared Error / RMSE) ile yapılan değerlendirmede DL tabanlı yöntemlerin ML yöntemlerine göre yüksek başarılı olduğunu göstermiştir. Bu tez kapsamında DL ile temerrüt tahmini yapılmıştır. Ancak veri setimizin tümünü DL algoritmalarıyla analiz edebilecek donanıma sahip olmadığımız için sadece veri setinden seçilen örneklemler üzerinden DL ile temerrüt tahmini yapılmıştır.

2.2. Büyük Veri

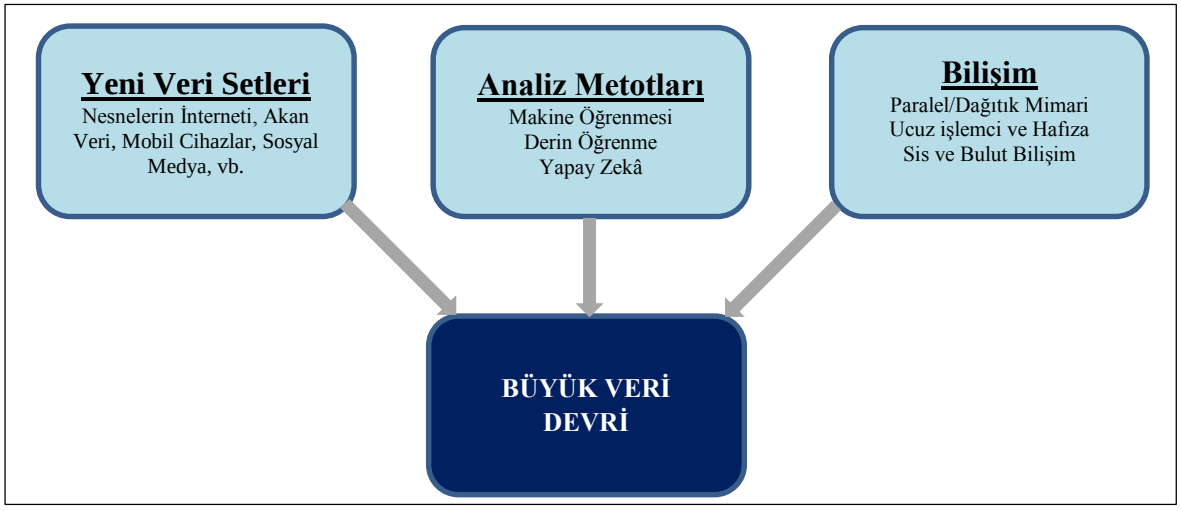
Veri oluşturma, depolama ve bu verilerden faydalanmanın kökeni milattan önceki dönemlere kadar dayanmaktadır. Bilgisayarların kullanılmaya başlanması veri oluşturma, depolama gibi işlemleri daha kolay hale getirmiş ve oluşturulan veri miktarı giderek artmıştır. Günümüzde hemen herkesin sahip olduğu bilgisayar, akıllı telefon, tablet gibi cihazlarla insanların oluşturduğu verilerin yanında son yıllarda hayatımıza IoT ile giren internete bağlı nesnelerde çok önemli boyutlarda veri üretmektedir. Dünyadaki toplam veri miktarının geometrik olarak büyümesi ve verinin önemli bir ekonomik değer olduğunun fark edilmesiyle birlikte yeni döneme "büyük veri dönemi" denilmeye başlanmıştır.

Büyük hacimli ham veri içerisinden veri analiz yöntemleri kullanılarak yorumlanabilir kıymetli bilginin çıkarılması, kurum ve kuruluşların stratejik kararlarında, olası risklerin yönetiminde ve geleceği tahmin etmede önemli bir yere sahiptir. Birçok kurum, kuruluş ve şirket stratejik kararlarını almada veri güdümlü sonuç çıkarma yöntemlerini kullanmaya

başlamıştır. Büyük veriye yatırım yapan sadece teknoloji şirketleri olmayıp, sağlık, tarım, eğitim, eğlence gibi birçok farklı sektörde faaliyet gösteren şirketler başarıyla sonuçlanan büyük veri projeleri yürütmektedir. Büyük veri sadece büyük şirketler için olmayıp, küçük şirketler, devlet kurumları, finans kuruluşları, üniversite ve araştırma kuruluşları gibi birçok alanda kullanım bulabilmektedir.

2.2.1. Büyük veri kavramı

Büyük veri, İngilizce adıyla "Big Data" karmaşıklık ve veri boyutunun büyüklüğü nedeniyle geleneksel veri yöntemleriyle yönetilemeyen ve işlenemeyen verilerin işlenebilir ve anlamlı biçime sokulmuş halidir. Büyük veri kavramı ilk olarak 2005 yılında Roger Magoulas tarafından bilgisayar dünyasına kazandırılmıştır. Büyük veri bağımsız (stand-alone) bir teknoloji değildir. Büyük veri bir olgu olmanın yanı sıra aynı zamanda bir süreci ifade etmektedir. Büyük veri, geleneksel yöntemlerle saklanamayacak ve işlenemeyecek kadar büyük ve çeşitli veriyi ifade eden bir olgu, büyük boyutlu verilerin analizini ve bilgi kazanımını sağlayan bir süreçtir [26]. Büyük verinin sağladığı depolama ve işleme imkânlarıyla yeni veri kaynakları daha değerli hale gelmiştir.



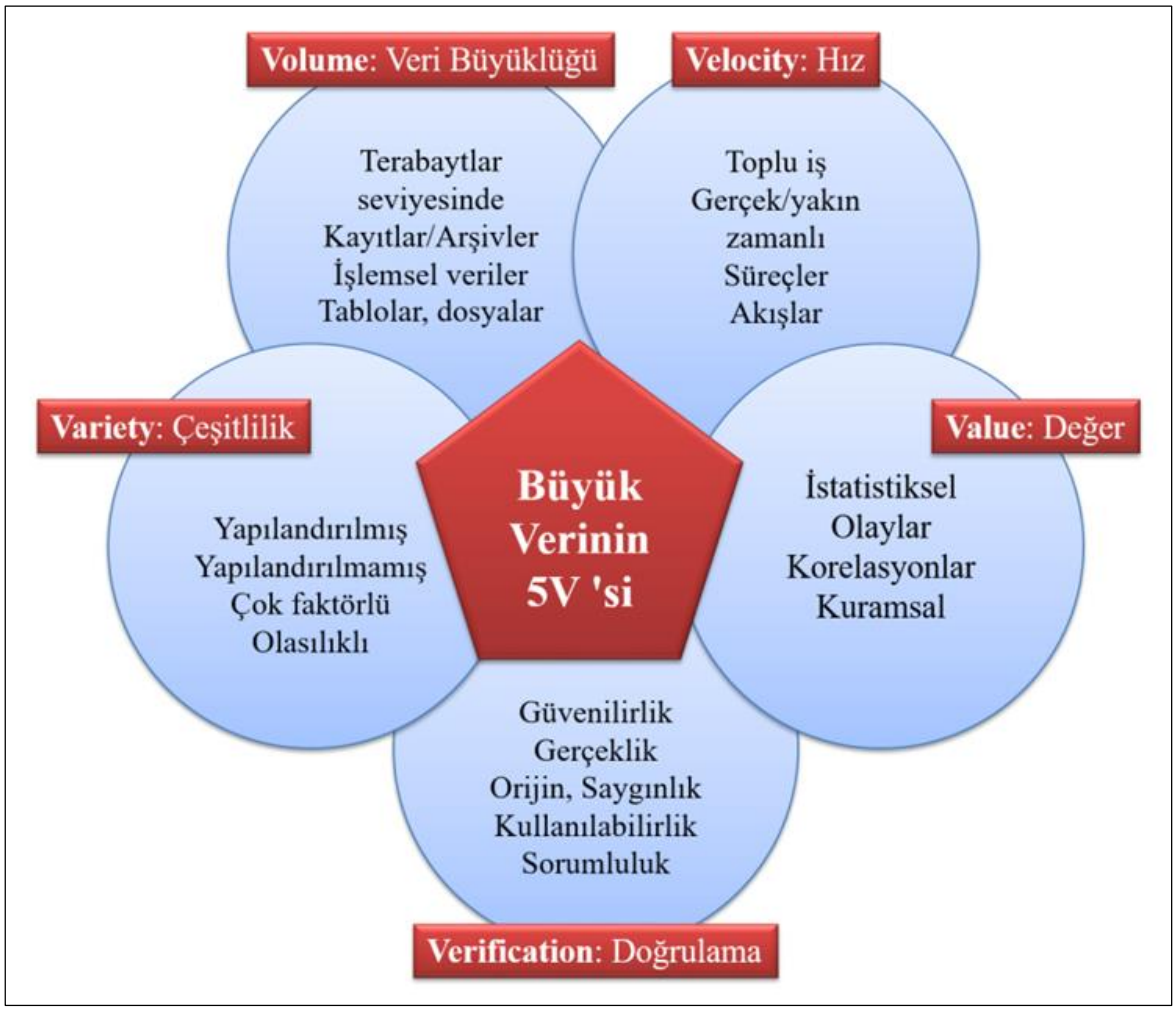
Şekil 2.1. Büyük veri devri

Şekil 2.1'de görüldüğü gibi büyük veri devrinin arkasında yeni veri kaynaklarının oluşması, karmaşık problemlerin çözümü için yeni analiz metotlarının geliştirilmesi ve bununla birlikte ucuzlayan donanım maliyeti yatmaktadır. Depolama maliyetlerinin düşmesi ve artan teknolojik cihazlar birçok yeni veri kaynağı oluşmuştur. Bunlardan en

önemlisi ise nesnelerin internetidir. Önceki yıllarda dijital verilerin birçoğunu insanlar üretmekteyken gelişen ve ucuzlayan sensör teknolojisiyle birlikte makineler insanların ürettikleriyle kıyaslanamayacak boyutlarda veri üretmeye başlamışlardır. Yeni veri kaynakları, astronomik büyüyen dijital verilerin analizi için geleneksel istatistiki analiz yöntemlerini yetersiz kılmış, bu nedenle yeni analiz yöntemleri geliştirilmiştir. Bu analiz yöntemleri büyük miktardaki veri üzerinde işlem yaparak veri içerisinde gizli kalmış bilgileri çıkarmakta geleneksel yöntemlere göre çok başarılıdır. Veri miktarının üstel şekilde büyümesiyle verinin tekil bilgisayarda analizi mümkün olmamaktadır. Bu derece büyük veriler, dağıtık ve paralel bilgisayar mimarileriyle birden fazla bilgisayarı aynı anda kullanarak analiz edilmeye başlanmıştır. Bu gelişmeler sonucunda büyük verinin önemi giderek artmıştır [27].

2.2.2. Büyük veri birleşenleri

Büyük verinin karakteristik özellikleri hakkında literatürde çok sayıda çalışma mevcuttur. Bu çalışmalarda ortak olan 3V (Volume, Velocity, Variety) ile simgelenen temel karakteristik özellikler olan hacim, hız ve çeşitlilikle birlikte son yıllarda doğrulama (Veracity) ve değer (Value) eklenerek Şekil 2.2'de görüldüğü gibi 5V olarak simgelenmektedir [28].



Şekil 2.2. Büyük veri birleşenleri [28]

Hacim: Verinin boyutunu açıklamaktadır. Uluslararası Veri Birliği (International Data Corporation / IDC) istatistiklerine göre 44 farklı veri üreticisi araç vardır. Sensörlerden süper bilgisayarlara, kişisel bilgisayarlardan sunuculara, arabalardan uçaklara kadar çok farklı türde veri üreticisi, metin, e-posta, video, ses, sensör verisi gibi birçok farklı formata veri üretmektedir. Veri boyutları ifade edilirken gigabayt ve terabaytlardan zetabaytlara geçilmiştir.

Hız: Verinin üretim hızını ifade etmektedir. Büyük hacimli veri ambarlarında bulunan statik olarak duran veri dışında hızı giderek artan ve gerçek zamanlı olarak yönetilip analiz edilmesi gereken akışkan veriler büyük veri teknolojisinin çözmesi gereken problemlerinden biridir. Borsadaki hisse senedi değeri verileri, mobil ve IoT cihazlarından üretilen sensör verileri hızlı ve akan veriye örnek oluşturmaktadır.

Çeşitlilik: Üretilen verinin format farklılıklarını açıklamaktadır. Genel olarak veri, yapısal, yarı yapısal ve yapısal olmayan şeklinde ayrılmaktadır. Yapısal veriye ilişkisel veri tabanlarında tablolar şeklinde tutulan veri, yarı yapısala XML, JSON formatlarında özellikle dağıtık sistemlerden toplanan veri, yapısal olmayana ise metin, ses, video türündeki veriler örnek olarak gösterilebilir.

Doğrulama: Veri doğrulama, verinin güvenilirliğini kapsamaktadır. Gerçek dünyada veri içerisine dâhil olan gürültü ve aykırı durumlar verinin doğruluğunu azaltmaktadır. Doğruluğundan emin olunmayan veriden yapılacak analizlere güvenerek alınacak kritik kararlar çok büyük problemlere neden olabilmektedir. Veriden büyük değer üretebilmek için verinin doğrulanması ve analize uygunluğundan emin olunması gerekmektedir.

Değer: Veriden kıymetli bilgi üretmektir. Büyük veri çalışmalarının amacını oluşturan en önemli birleşendir. Yukarıda bahsedilen tüm birleşenler veri üretim, işleme ve analiz katmanları veri içerisinden artı değer çıkarmak için yapılmaktadır. Üretilen değer verinin içeriğine toplanma amacına, uygulama alanına bağlı olarak değişkenlik göstermektedir. Değerin doğru ve zamanlı olarak üretilmesi karar alıcılar için çok önemli olmaktadır.

2.2.3. Büyük veri teknolojileri

Veri miktarındaki artış, üretim hızı, saklanma, işleme, analiz ve değerlendirme tekniklerindeki değişiklikler teknolojik olarak yeni ihtiyaçlar doğurmuştur. Büyük veriyle çalışırken karşımıza farklı veri kaynaklarından verinin toplanması, veri hacminin yönetilmesi, yapısal, yarı yapısal veya yapısal olmayan gibi farklı yapı ve formattaki verilerin depolanması, dağıtık/paralel hesaplama ve gerçek zamanlı akan verinin analizi gibi birçok konuda teknik ve teknolojik ihtiyaçlar doğmuştur. Bu nedenle birçok teknolojik çözüm üretilmiştir. Aşağıda bu çalışmada kullanılan önemli büyük veri teknolojilerinden kısaca bahsedilmiştir.

Apache spark

Büyük ölçekli verileri paralel olarak işleyebilen açık kaynak kodlu bir veri analizi aracıdır. Büyük veri kullanan veri analistleri tarafından sıklıkla kullanılmaktadır. Bunun en önemli sebebi hızlı işlem yapabilmesidir. Hızlı işlem yapabilmesini ise veri yapısı

olarak Esnek Dağıtık Veri Setine (Resilient Distributed Dataset / RDD) borçludur. Hata ve değişime karşı toleransının yüksek olması ve dağıtık olarak çalışabilmesi RDD'leri veri analizinde güçlü kılmaktadır. Spark 2 sürümüyle RDD ile birlikte veri çerçevesi (dataframe) yapısı eklenmiştir. Kolay kullanımı nedeniyle veri çerçevesi daha çok tercih edilmektedir. Ancak veri çerçevesi kullanılsa bile Spark arka planda işlemlerini yaparken veri çerçevesini RDD'lere dönüştürmektedir.

Spark'ın çekirdeği Spark SQL, Spark Streaming, MLlib ve GraphX olmak üzere 4 temel birleşenden oluşmaktadır. Bunlardan Spark SQL yaygın bilenen SQL sorgu dilini kullanarak verileri sorgulamamızı, Spark Streaming akan verilerde gerçek zamanlı veri analizi, MLlib ML ile veri analizi ve GraphX ise çizge analiz işlemlerini gerçekleştirmektedir.

Python, Scala, Java veya R programlama dilleriyle Spark geliştirmesi yapılabilmektedir. Bu çalışmada Spark üzerindeki uygulama geliştirmeleri Python PySpark kütüphanesiyle yapılmıştır.

HDFS

Hadoop, verileri tek bir bilgisayarda tutmayıp, farklı sunucularda denormalize olarak saklayan ve Map-Reduce (haritala-indirge) veri işleme modeline göre verileri işleyen açık kaynak kodlu Apache kütüphanesidir. HDFS (Hadoop Distributed File System), Hadoop dağıtık dosya sistemidir. Java tabanlı olup hata toleransı yüksek, güvenilir, ölçeklenebilir bir veri depolama çözümü sunmaktadır. Map-Reduce veri işleme modeline destek vermesiyle veri işleme ve analizinden yüksek performans sağlamaktadır.

Hive

HDFS üzerinde depolanan verilere SQL benzeri sorgulama yapabilmemizi sağlayan araçtır. Kullanıcılardan aldığı sorguyu Map-Reduce mimarisine uygun hale getirerek bir soyutlama sağlar. Böylelikle kullanıcı tıpkı ilişkisel veri tabanı sorgulaması yapar gibi dağıtık dosya sistemi üzerindeki verilerde sorgulama yapabilir.

Zeppelin

Büyük veri teknolojileriyle çalışabilen web tabanlı bir editördür. Python, SQL, Scala, R gibi birçok programla diliyle çalışabilir. Veri analizi, görselleştirme, notlar alma (Markdown), veri çıkarma gibi birçok işlemi kolay hale getirmektedir. Bir diğer önemli özelliği aynı not defteri (notebook) üzerinden farklı dillerle çalışma imkanı sunmasıdır.

Hue

Büyük veri üzerinde çalışabilen açık kaynak kodlu bir SQL sorgulama asistanıdır. Apacha Hive, Impala Presto gibi birçok kaynağa kolayca sorgu gönderebilmektedir. Sorgu paylaşımı, sonuçların grafikle sunulması, veri çıkarma gibi birçok işlemde kolaylık sağlamaktadır.

2.3. Sınıflandırma Algoritmaları

Şirketlerin temerrüt tahmini ikili sınıflandırma (binary classification) problemidir. Bu nedenle burada çalışmada sınıflandırma için kullanılan istatistik ve ML temelli algoritmalar açıklanmaktadır.

2.3.1. Lojistik regresyon

Lojistik regresyon gözetimli öğrenme yöntemlerindendir. LR, bir veya birden fazla sınıftan oluşan gözlemleri sınıflandırabilen istatistik temelli bir algoritmadır. Bağımsız değişkenler (giriş değerleri) ile bağımlı değişken (sonuç değerleri) arasındaki ilişkiyi açıklamaktadır. Farklı birçok sınıflandırma probleminde olduğu gibi temerrüt ve iflas tahmininde de sıklıkla kullanılmaktadır [20,29]. Yaygın kullanımının en önemli sebebi kurulan modelin basit ve yorumlanabilir olmasıdır. ML ile sınıflandırma çalışmalarında genellikle referans algoritma olarak kullanılmaktadır [30].

LR maliyet fonksiyonu olarak lojistik veya sigmoid fonksiyonu kullanmaktadır. Bu fonksiyon 0 ile 1 arasında olasılık değeri almaktadır. LR sonuç değerleri Y, aşağıdaki denklemde görüldüğü gibi giriş değerleri ve katsayılarla hesaplanmaktadır:

$$Y=ln(\frac{P}{1-P})=b_0+b_1*x_1+b_2*x_2+...+b_k*x_k \quad (2.1)$$

P sigmoid fonksiyonu, b_0 yanlılık (bias), *k* girdilerin sayısı, x_k her bir girdi değeri, b_k *k* girdisine ait katsayıdır. Sigmoid *P* aşağıdaki denklem ile hesaplanmaktadır:

$$P=\frac{1}{1+e^{-Y}} \quad (2.2)$$

LR'de bağımsız değişkenlerin artması modelin karmaşıklığını artırmaktadır. Modelin karmaşıklığı artığında ise yüksek varyans ve düşük yanlılığa sebep olmaktadır. Bu durumda modelde aşırı öğrenme (overfitting) problemi oluşmaktadır. Böyle bir modelde eğitim veri setinde yüksek başarı görülürken modelin hiç görmediği test veri setinde başarı düşmektedir. Bu durumdan kaçınmak için En Az Mutlak Büzülme Seçici Operatörü (Least Absolute Shrinkage And Selection Operator / LASSO), RIDGE, ElasticNet gibi düzenleyiciler kullanılmaktadır. Bu düzenleyiciler modelin karmaşıklığını azaltarak aşırı öğrenme problemini çözmektedir. RIDGE en küçük kareler yöntemini kullanarak katsayılar arasında en düşük varyanslı katsayıları bulmaya çalışmaktadır. LASSO, RIDGE benzemekte olup aralarındaki en temel fark LASSO'nun en küçük kareler yerine mutlak değeri kullanmasıdır. Bu sebep LASSO değişkenleri tamamıyla ihmal edebilmektedir. ElasticNet ise LASSO ve RIDGE'in birleşiminden oluşan bir düzenleyicidir. Bu algoritmaların çalışmaları sırasında ayar parametresi olarak kullandıkları λ değeri 0 ile 1 arasında değiştirilerek LASSO ve RIDGE düzenleyicileri bir arada kullanılmaktadır.

2.3.2. Karar ağacı

Gözetimli ML yöntemlerinden biri olan karar ağacı, sınıflandırma ve regresyon problemleri için çözüm aramaktadır. Düğüm ve yapraklardan oluşan hiyerarşik bir ağaç yapısı oluşturmaktadır. Verilen bir veri setini kök düğümden başlayarak her bir düğümdeki karar kurallarına göre bölerek sonucu belirlemektedir. Bu nedenle düğümlerdeki bölme işleminin nasıl yapılacağı başarı etkileyen en önemli faktördür. Karar ağaçları en iyi bölen özniteliği seçmek için bilgi kazancı (information gain) ve Gini gibi metotları kullanmaktadır. Bu çalışmada bilgi kazancı metoduyla karar ağacı oluşturulmuştur [31]. Bilgi kazancı en iyi özniteliği belirlerken entropy ölçütünden

yararlanır [32]. Entropy, rastgeleliğin bir ölçütüdür ve aşağıdaki formülle hesaplanır:

$$Entropy(S)=\sum_{i=1}^{c}-p_i log_2 p_i \tag{2.3}$$

Burada *S* eğitim veri setini temsil etmektedir. *c*, hedef değişkenin alabileceği farklı değer sayısını göstermektedir. p_i ise hedef değişkenin *i* sınıfının eğitim verisindeki olasılığını göstermektedir. Aşağıdaki formülle bilgi kazancı hesaplanmaktadır.

$$Gain(S,A)=Entropy(S)-\sum_{v\in Values(A)}(|S_V|/|S|)Entropy(S_V) \tag{2.4}$$

Burada *A* her bir özniteliği, *Values(A)*, *A* özniteliğinin alabileceği olası değerleri S_V ise *A* özniteliğinin *S* eğitim veri setinde *v* değerini almış alt kümesini göstermektedir. *A* özniteliğine ait bilgi kazancı hedef değişenin entropy değerinden *A* özniteliğinin entropy değeri çıkarılarak hesaplanmaktadır. Bilgi kazancı veri setindeki her bir öznitelik için ayrı olarak hesaplandıktan sonra en yüksek bilgi kazancına sahip olan öznitelik karar ağacı için düğüm olarak belirlenmektedir.

Karar ağacı algoritması kolaylıkla görselleştirilebildiği için anlaması ve yorumlanması kolaydır. Hem sayısal hem de kategorik verilerle çalışabilmektedir. Bu gibi avantajlarının yanı sıra eğitim verisi ezberleyerek aşırı öğrenme problemine açıktır. Bu problemin çözümü için budama veya parametre kısıtlamaları (maksimum ağaç derinliği, bir düğümün alabileceği minimum örnek sayısı gibi) uygulanmaktadır. Bu çalışmada aşırı öğrenme problemi için parametre kısıtlamaları kullanılmıştır.

2.3.3. Rastgele orman

Gözetimli ML yöntemlerinden olan rastgele orman, temelde karar ağaçlarını kullanan kolektif öğrenme (ensemble learning) algoritmasıdır. Regresyon ve sınıflandırma problemlerinde uygulanabilmektedir. Karar ağaçlarının en önemli sorunu olan aşırı öğrenme problemini çözmek için veri setinde ve özniteliklerde rastgele seçimler yaparak alt setler oluşturmaktadır. Bu alt veri setlerini kullanarak birçok ağaç eğitilmektedir. Eğitilen bu ağaçların her biri bireysel olarak tahmin üretmektedir. Nihayetinde bu

tahminler arasındaki en çok oy alan sonuç tahmin olarak seçilmektedir [33].

Algoritma eğitim sırasında farklı veri setleri kullandığı için varyans düşmektedir. Böylelikle aşırı öğrenmeden kaçınmaktadır. Ayrıca rastgele seçtiği alt veri kümelerinde aykırı değer (outlier) bulunma şansı da azalmaktır. En önemli dezavantajı yorumlanabilirliğinin olmamasıdır [34].

2.3.4. Dereceli artırma

Sınıflandırma ve regresyon problemleri için kullanılan gözetimli ML algoritmasıdır. Kolektif öğrenme algoritmalarındandır. Karar ağacı gibi zayıf öğrenicileri bir araya getirerek güçlü öğreniciler yapmayı hedeflemektedir. Dereceli artırma yönteminin en önemli farkı ağaçlar sıralı olarak oluşturulmakta ve her bir ağaç önceki ağaçların sonuçlarından elde ettikleri bilgileri kullanarak gelişmektedir. Başka bir ifadeyle ağaçlar oluşurken önceki ağaçların yaptıkları hatalardan ders alarak büyümektedir [35].

GB algoritmasındaki temel amaç seçilen kayıp fonksiyonunu minimize etmektir. GB farklı kayıp fonksiyonları kullanmaktadır. Aşağıda bu çalışmada kullanılan kayıp fonksiyonu denklemi vardır. Burada N örneklem sayısını, y_i i'inci örneklemdeki özniteliği, $F(x_i)$ ise i'inci örneklem için tahmin değerini göstermektedir.

$$Loss=2\sum_{i=1}^{N} log(1+exp(-2y_i F(x_i))) \tag{2.5}$$

GB algoritması aşağıdaki adımları kullanarak çalışmaktadır.

1. Hedef değişkenini ortalama değerini hesapla
2. Her bir örneklem için hedef değer ve tahmin arasındaki farkı alarak artık değerleri (residual) hesapla
3. Artık değerlerin tahmini için karar ağacı oluştur.
4. Aynı giriş değişkenleri ile yeni artık değerleri hesapla
5. Önceki artık değerler ile yeni artık değerleri topla
6. Aşırı öğrenme veya artık değerler sabit kalana kadar 2'den 5'e kadar adımlar tekrarla

2.4. Kullanılan Veri Setleri ve Temerrüt Tanımı

Çalışmada kullanılan tüm veri setleri Türkiye Cumhuriyet Merkez Bankası (TCMB) veri tabanlarından elde edilmiştir. Veri setleri son derece hassas verileri içermesi nedeniyle çalışmalar TCMB sistemleri üzerinde maskeli verilerle yapılmıştır. Çalışmada şirketlerin bilanço, kredi ve temerrüt veri seti kullanılmıştır.

2.4.1. Bilanço veri seti

Bilanço şirketlerin belirli bir dönem sonundaki varlık, yükümlülük ve öz kaynaklarını gösteren finansal tablodur. Ülkemizde şirket bilançoları Tek Düzen Hesap Planına (TDHP) uygun olarak Gelir İdaresi Başkanlığı (GİB) tarafından yıllık frekansta toplanmaktadır. Bilanço kullanılarak yapılan birçok temerrüt tahmini çalışması mevcuttur [36-38].

Temerrüt tahmini için bilanço kalemleri üzerinde literatürde sıklıkla kullanılan 61 adet bilanço oranı hesaplanmıştır. Hesaplanan bu oranlar temerrüt tahmininde bilanço değişkenleri olarak kullanılmıştır. Bu değişkenler 4 temel kategoriye ayrılabilmektedir. Bunlar likidite, finansal pozisyon, devir hızı ve karlılık kategorileridir. Bu çalışmada Türkiye'de 2010-2018 yılları arasında faaliyet gösteren şirketlere ait bilançolardan oluşturulan değişkenler kullanılmıştır. Toplam 1 016 315 farklı şirkete ait bilanço üzerinde çalışılmıştır.

2.4.2. Kredi veri seti

Kredi veri seti şirketlerin ülkemizde faaliyet gösteren bankalardan kullandıkları tüm kredileri içermektedir. Kredi veri seti bankalardan aylık frekansta Kredi Kayıt Bürosu tarafından derlenmektedir. Veri seti üzerinden temerrüt tahmini için sıklıkla kullanılan 57 adet değişken oluşturulmuştur. Bu değişkenler 6 temel kategoriye ayrılabilmektedir. Bunlar kredi/limit, limit yeterlilik, faktöring, ödeme/ödememe, yeniden yapılandırma ve bankacılık ilişkileridir. Kredi veri seti aylık frekansta olması nedeniyle oluşturulan değişkenlerin yıllık bazda ortalamaları kullanılmıştır.

2.4.3. Temerrüt veri seti

Temerrüt, gerçek veya tüzel kişinin borcunu zamanında ve usulüne uygun olarak ödeyememesi, ödeme güçlüğüne düşmesi durumudur. Tezin temel amacı temerrüde düşecek şirketlerin önceden tahmin edilmesidir. Literatürde farklı temerrüt tanımları olmakla birlikte bu çalışmada şirketlerin temerrüde düşmesi aşağıdaki kriterlerle belirlenmiştir.

- Ödemesi 90 gün gecikmiş krediler
- Çek ve senet ödenmemesi
- Hukuki iflas veya konkordato

Çalışmada, 2010 ile 2018 yılları arasındaki veriler kullanılarak yapılmıştır. Yıllara göre şirket sayıları ve temerrüde düşme oranları tablo Çizelge 2.1'de verilmiştir.

Çizelge 2.1. Yıllara göre şirket sayıları ve temerrüde düşme oranları

Yıl	Şirket Sayısı	Temerrüt Oranı
2010	561 040	0,066325
2011	588 928	0,065607
2012	597 359	0,052610
2013	610 013	0,064018
2014	627 342	0,058156
2015	654 201	0,047626
2016	681 753	0,045043
2017	708 775	0,047369
2018	761 498	0,033091

2.5. Makine Öğrenmesi ile Temerrüt Tahmin Modeli (DPModel-1)

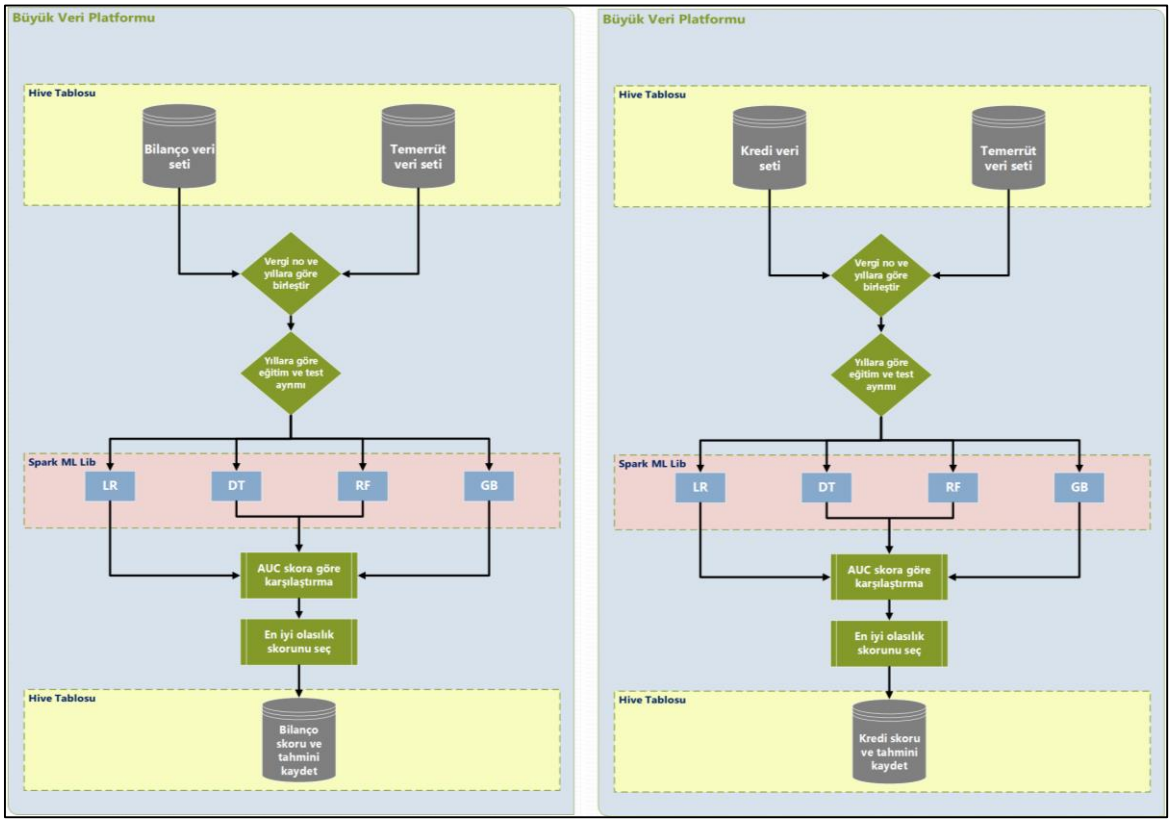
Bu bölümde hem literatürde hem de gerçek dünyada sıkça çalışılan şirketlerin temerrüde düşme tahmini için DPModel-1 önerilmiştir. Literatürdeki çalışmalara bakıldığında problemi araştıranların çoğunlukla finans uzmanı olduğu ve LR, RIDGE, LASSO gibi geleneksel istatistiksel yöntemlerle tahminler yapıldığı görülmektedir [39-43]. Son

yıllarda ise tahmin gücü daha yüksek olan ML modelleriyle tahminler yapılmaya başlanmıştır [44-46]. Tezin bu bölümünde temerrüt tahmini için istatistiksel yöntemlerinden LR, ML yöntemlerinden ise DT, RF ve GB algoritmaları seçilmiştir. Önceki çalışmalara bakıldığında birçok farklı yöntemin temerrüt tahmini için kullanıldığı görülmektedir. Çalışmada kullanılan yöntemlerin seçilmesindeki en önemli 3 gerekçe aşağıda açıklanmaktadır;

- LR algoritması literatürde temerrüt tahmini için kullanılan temel ölçüt (benchmark) algoritmasıdır. Uzun yıllardır bu algoritma temerrüt tahmini için sıklıkla kullanılmıştır. Bunun en önemli sebebi başarı oranının istatistiksel yöntemler içinde yüksek olması ve sonuçlarının yorumlanabilir olmasıdır. Seçilen ML yöntemleri de literatürde sıklıkla kullanılmakta ve yüksek başarı oranı göstermektedir. Özellikle GB ve RF algoritmaları birçok çalışmada en yüksek başarıyı göstermiştir [8,47,48].
- Toplam 115 değişken kullanılmaktadır. Literatürdeki birçok çalışmada modelleme yapılmadan önce öznitelik seçimi yapılmaktadır. Öznitelik seçimi yapıldığında bilgi kaybı olması kaçınılmaz hale gelmektedir. Ayrıca ekstra bir çalışma maliyeti oluşmaktadır. Bu nedenlerle çalışmada öznitelik seçimi yapılmamış, bunun yerine ağaç tabanlı ML yöntemleri seçilmiştir. Ağaç tabanlı algoritmaların doğası gereği öznitelik seçimine ihtiyaç duymadan çalışabilmektedir. Ayrıca çalışmada kullanılan LR algoritmasında ızgara arama (grid search) ile ElasticNet parametresi değiştirilerek (ElasticNet=1 ise LASSO, ElasticNet=0 ise RIDGE) başarıyı düşüren değişkenler cezalandırılarak önceden öznitelik seçimine ihtiyaç duyulmamaktadır.
- Çalışma BVP üzerinde yapılmıştır. BVP dağıtık bir mimaridir. Birçok ML algoritmasının dağıtık mimari üzerinde çalışacak şekilde henüz uygulaması yapılmamıştır. Örneğin son yıllarda sıklıkla kullanılan ve yüksek başarı gösteren XGBOOST algoritması henüz Spark ML kütüphanesinde mevcut değildir. Bu nedenle literatürde yüksek başarı gösteren tüm ML algoritmaları kullanılamamıştır.

Tahmin için bilanço ve kredi veri setinden yararlanılmıştır. İlk olarak şirketlerin bilanço ve kredi veri setinde tahminde kullanılacak değişkenler yıllık frekansta oluşturulmuştur. Veri setleri 2010-2018 yılları arasını içermektedir. Bilanço veri setinden 61, kredi veri setinden 57 adet değişken üretilmiştir. Tüm değişkenler için boş değerler sıfır ile doldurulmuştur.

Kredi ve bilanço veri setlerinden oluşturulan değişkenleri tek bir tabloda birleştirmek yerine ayrı alt modeller oluşturulmuştur. Bunun en önemli sebebi her şirketin bilançosu varken şirketlerin sadece bir kısmının kredi kullanmasından kaynaklanmaktadır. Aksi halde tüm değişkenler bir araya getirildiğinde kredi kullanmayan şirketlerin kredi değişkenlerinin hepsi 0 olarak doldurulması gerekecektir. Bu durum modelin tahmin gücünü düşürecektir. Dolayısıyla bilanço ve kredi değişkenleriyle ayrı alt modeller oluşturulmuştur. Alt modellere ait akış şeması Şekil 2.3'de görülmektedir.



(a) (b)

Şekil 2.3. Bilanço (a) ve kredi (b) verisiyle oluşturulan alt modellerin akış şeması

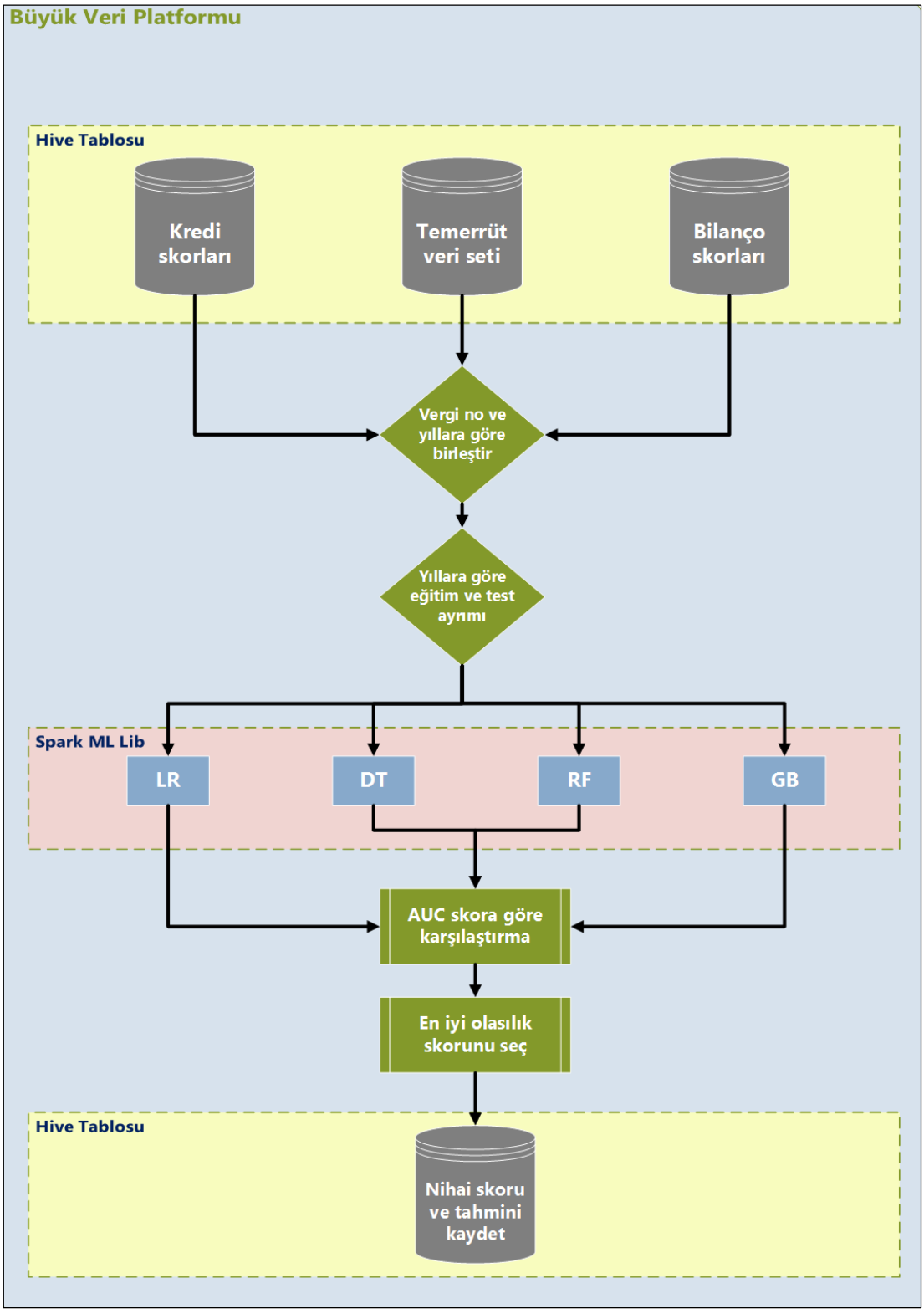
Bilanço değişkenleriyle yapılan alt modelin akış şeması Şekil 2.3 (a) görülmektedir. 2010-2018 yılları arasındaki bilançolardan oluşturulan 61 adet değişken yıl ve vergi numarası kullanılarak yine yıllık frekanstaki temerrüt tablosuyla birleştirilmiştir. Zaman bilgisi içeren bu veri seti standart ML çalışmalarında yapıldığı gibi rastgele olarak eğitim ve test verisi olarak ayrılmayıp gerçek hayatı simule edebilmesi için bir önceki yıl verisi ile sonraki yıl tahmini yapılmıştır. Bu nedenle verimiz 2010 yılında başladığı halde tahminlerimiz 2011 yılından başlamaktadır.

Modelde tahmin algoritmaları için Apache Spark kütüphanesinden faydalanılmıştır. Bunlar LR, DT, RF ve GB algoritmalarıdır. Bu algoritmaların başarı oranları başlangıçta aldıkları parametrelere göre değişkenlik göstermesi nedeniyle en iyi başarıyı bulabilmek için her birine ayrı olarak ızgara arama parametre optimizasyon algoritması uygulanmıştır. Böylelikle parametreler manuel olarak değil başarıyı maksimize eden en uygun değerlerle optimize edilmiş olarak verilmiştir.

Kullanılan ML algoritmalarının en önemli parametreleri optimize edilmiştir. LR algoritması için maksimum yineleme (maxIter), öznitelik cezalandırma (elacticNet) ve düzenleme parametresi (regParam) optimize edilmiştir. DT algoritması için maksimum yineleme sayısı, maksimum ağaç derinliği (maxDepth) kullanılmıştır. RF algoritması için maksimum ağaç sayısı (numTrees) ve maksimum ağaç derinliği kullanılmıştır. GB algoritması içinse maksimum yineleme sayısı, öğrenme oranı (learningRate), maksimum ağaç derinliği, yinelemelerdeki adım büyüklüğü (stepSize) optimize edilmiştir.

Her bir model için AUC, doğruluk, kesinlik, duyarlılık, f1 skoru metrikleri hesaplanmıştır. Şirketlerdeki temerrüt oranı yıllara göre %3-6 arasındadır. Bir diğer deyişle veri setimiz dengesiz olarak dağılmış bir veri setidir. Bu nedenle modellerin başarısı kıyaslanırken doğruluk yerine AUC metriğine göre seçim yapılmıştır. Çalışmanın sonunda her bir yıla ait en yüksek başarıyı gösteren modelin sonuçları ve 0 ile 1 arasındaki olasılık skorları (probability score) Hive tablosuna yazılmıştır.

Şekil 2.3 (b) kredi değişkenleriyle yapılan çalışmanın akış şeması görülmektedir. Bu çalışma bilanço değişkenleriyle yapılan çalışmaya benzemektedir. Bilanço değişkenleri için gerçekleştirilen tüm adımlar kredi değişkenleri içinde gerçekleştirilmiştir. Çıkan sonuçlar ve olasılık skorları benzer şekilde Hive tablosuna yazılmıştır.



Şekil 2.4. DPModel-1'in nihai tahmin akış şeması

Bilanço ve kredi değişkenleri üzerinden oluşturulan alt modellerden elde edilen en iyi olasılık skorları Şekil 2.4'deki akış şemasında görüldüğü gibi birleştirilerek DPModel-1'in nihai tahminleri elde edilmiştir. Bu birleştirme sırasında şirketin kredi bilgisi yoksa

doğrudan bilanço alt modelinden alınan tahminler DBModel-1'in nihai tahminleri olarak belirlenmiştir. Kredi bilgisi varsa daha önce oluşturulan bilanço ve kredi alt modellerinden elde edilen skorlar akışta görüldüğü gibi tekrar modellenerek DBModel-1'in nihai tahminini geliştirilmiştir.

2.6. Deneysel Ortam ve Performans Metrikleri

Bu bölüm çalışmanın yapıldığı ortam ve sonuçların değerlendirilmesinde kullanılan performans metrikleri açıklanmaktadır.

2.6.1. Deneysel ortam

Tüm çalışmada büyük veri platformu (BVP) kullanılmıştır. Bu platformda Hortonworks Data Platform 3.1.5 kullanılmaktadır. Platform 1 kenar sunucusu (edge node), 3 ana sunucu (master node) ve 4 veri sunucusu (data node) bulunmaktadır. Her bir sunucuda 256 GB RAM, 2x18 core Intel Xeon (R) Gold 6154 CPU 3.00GHz işlemci bulunmaktadır. Dağıtık dosya sistemi olarak HDFS'in 3.1.1 sürümü kullanılmaktadır. Platform üzerindeki veriler SQL sorgusu yapılabilen ve geniş veri setlerini dağıtık olarak yönetebilen Hive tabloları olarak tutulmaktadır. Platformda bir SQL asistanı olan Hue mevcuttur. Hue, kolaylıkla Hive tablolarına SQL sorguları gönderebilmektedir. ML algoritmaları için Spark 2.3.2 sürümü kullanılmıştır. Spark'ı kullanmak için PySpark kütüphanesinden yararlanılmıştır. Ayrıca platform üzerinde Python 3.8 yorumlayıcı ve editör olarak ise Zeppelin Notebook 0.8.0 kuruludur.

2.6.2. Performans metrikleri

Modellerinin sonuçlarının değerlendirilmesinde farklı metrikler kullanılabilmektedir. Bu çalışmadaki gibi dengesiz veri setleri ile oluşturulan modellerin karşılaştırılmasında sıklıkla kullanılan ROC yönteminden faydalanılmaktadır. ROC bir olasılık eğrisi olup x ekseninde yanlış pozitif oranı (FPR) ve y ekseninde doğru pozitif oranı (TPR)'i göstermektedir. FPR ve TPR aşağıdaki gibi hesaplanmaktadır.

$$FPR = \frac{YP}{YP + DN} \tag{2.6}$$

$$TPR = \frac{DP}{DP + YN} \tag{2.7}$$

ROC eğrisinin altında kalan alana AUC denilmektedir. AUC model performansının özeti olarak kabul edilmektedir. AUC değeri ne kadar büyükse sınıfları ayırt etmede model o kadar başarılıdır. İdeal bir modelde AUC değeri 1'dir.

Ayrıca model başarısını değerlendirebilmek için kesinlik (precision), duyarlılık (recall), doğruluk (accuracy), ve ağırlıklandırılmış f1 skor (weighted f1 score) kullanılmıştır. Bu değerlendirme kriterleri Çizelge 2.2'de gösterilen karışıklık matrisi (confusion matrix) üzerinden hesaplanmaktadır.

Çizelge 2.2. Karışıklık matrisi

		Gerçek Durum	
		Pozitif	Negatif
Tahmin	Pozitif	DP	YP
	Negatif	YN	DN

Karışıklık matrisi gerçek durum ile modelin tahmin ettiği değerleri kıyaslamada kullanılmaktadır. Kesinlik karışıklık matrisi üzerinden aşağıdaki gibi hesaplanmaktadır. Kesinlik modelin pozitif olarak tahmin ettiklerinden gerçekte kaçının pozitif olduğunun oranını göstermektedir.

$$Kesinlik = \frac{DP}{DP + YP} \tag{2.8}$$

Duyarlılık, modelin pozitif durumları ne kadar başarılı olarak tahmin ettiğinin oranıdır.

$$Duyarlılık = \frac{DP}{DP + YN} \tag{2.9}$$

Doğruluk ise tahminlerin doğruluk oranını göstermektedir ve aşağıdaki gibi hesaplanmaktadır.

$$Doğruluk = \frac{DP+DN}{DP+DN+YP+YN} \tag{2.10}$$

f1 skor ise kesinlik ve duyarlılığın harmonik ortalamasıyla üretilen bir kriterdir.

$$f1\ Skor = 2 * \frac{Duyarlılık * Kesinlik}{Duyarlılık + Kesinlik} \tag{2.11}$$

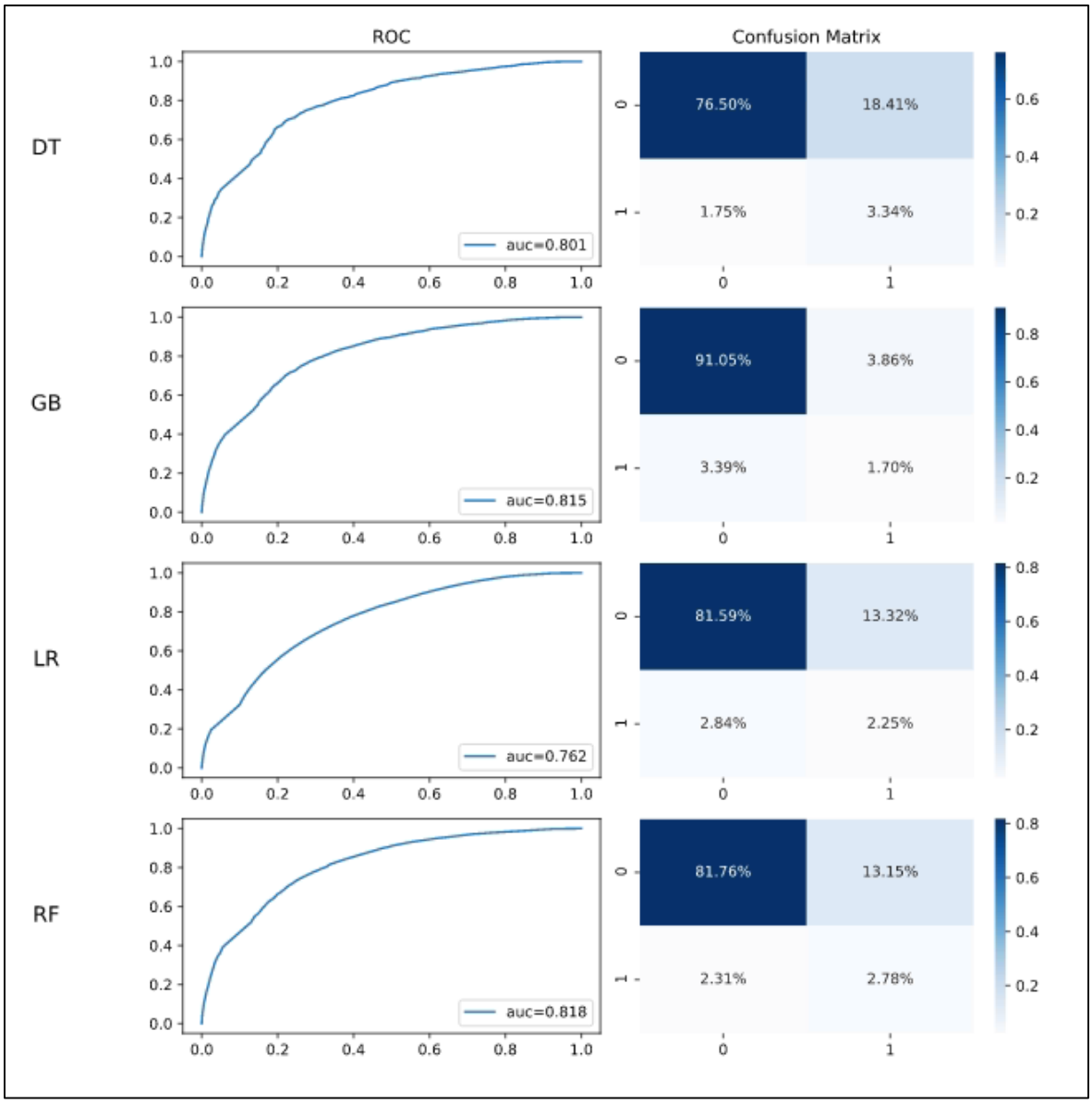
2.7. Deneysel Sonuçlar ve Değerlendirmesi

DPModel-1, Şekil 2.3'de görüldüğü gibi bilanço ve kredi rasyolarıyla oluşturulan iki alt modele sahiptir. Bu alt modellerden elde edilen en iyi tahminler Şekil 2.4'de gösterilen nihai modelle tekrar modellenerek en iyi tahmin değerlerine (olasılık skorları) ulaşılmıştır. Çizelge 2.3 yıl ayrımı olmaksızın tüm tahminlerin sonuçlarından elde edilen metrikleri göstermektedir. Ayrıca Şekil 2.5'de modellerin ROC eğrisi ve karşılaştırma matrisi mevcuttur. Buna göre en iyi tahmini yapan algoritma RF algoritması olmuştur. Geleneksel yöntemlerden olan LR en düşük AUC sonucunu göstermiştir.

Sınıflandırma metrikleri hesaplanırken varsayılan eşik değer 0,5 alınmaktadır. Fakat bu dengesiz veri setlerinde f1 skoru ve doğruluğu düşürmektedir. Bu nedenle çalışmada f1 skoru maksimum yapan eşik değer hesaplanmış ve tablolardaki tüm metrikler bu eşik değer üzerinden hesaplanmıştır. Her bir modele ait eşik değer tablolarda mevcuttur. Buna göre RF algoritması 0,818 AUC değeri ile en iyi tahmini yapmaktadır.

Çizelge 2.3. DPModel-1 ortalama model başarıları

Model	Eşik Değer	AUC	F1 Skor	Doğruluk	Kesinlik	Duyarlılık
DT	0,08	0,801	0,249	0,798	0,153	0,655
GB	0,20	0,815	0,319	0,828	0,306	0,334
LR	0,08	0,762	0,218	0,838	0,145	0,443
RF	0,12	0,818	0,264	0,845	0,174	0,545

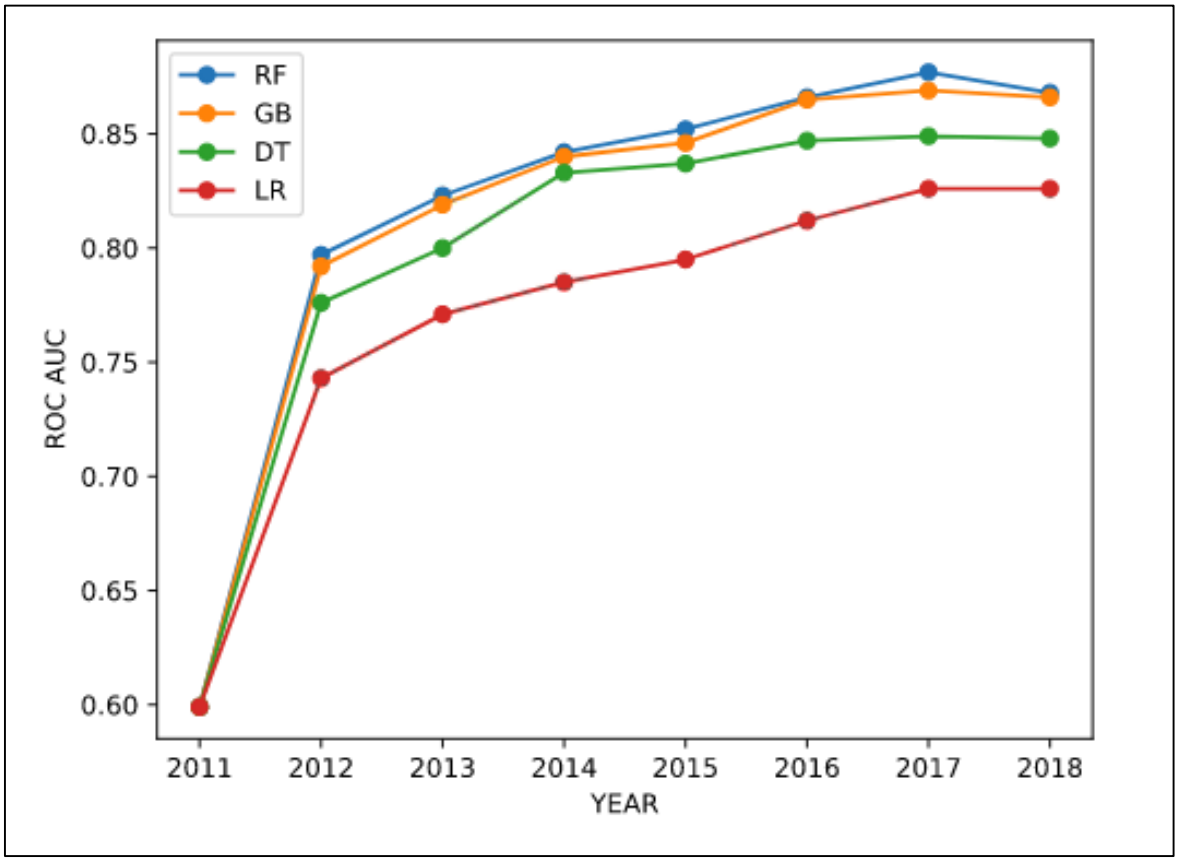


Şekil 2.5. DPModel-1 sonuçlarının ROC eğrisi ve karşılaştırma matrisi

DPModel-1'de eğitim ve test veri setleri yıllara göre ayrılmıştır. Veri seti 2010 yılından başladığından tahminler 2011 yılından başlamıştır. Bir önceki yıl eğitim, sonraki yıl ise test veri seti olarak kullanılmıştır. Çizelge 2.4'de yıllara göre modellerin başarı oranları mevcuttur.

Yıllar itibariyle genel olarak en yüksek tahmini RF algoritması yapmaktadır. Ancak bazı yıllarda GB algoritmasının en iyi tahmini yaptığı görülmüştür. Modellerin başarı oranının yıllar itibariyle arttığı görülmektedir. Bu durumun öncelikle veri kalitesiyle ilişkili olduğu düşünülmektedir. Yıllara göre toplanan verinin kalitesi artığı görülmektedir. Bir diğer sebebi ise yıllara göre artan şirket sayısıdır. Şirket sayısındaki artış modelin daha iyi

eğitildiğini göstermektedir. 2018 yılındaki başarıdaki düşüş kredi verisinin eksikliğinden kaynaklanmaktadır. Şekil 2.6'da yıllara göre AUC skorları görülmektedir.



Şekil 2.6. Yıllara göre AUC skorları

Çizelge 2.4. DPModel-1 yıllara göre model başarıları

Yıl	Model	Eşik Değer	AUC	F1 Skor	Doğruluk	Kesinlik	Duyarlılık
2011	DT	0,13	0,599	0,152	0,363	0,083	0,870
2011	GB	0,17	0,599	0,152	0,363	0,083	0,870
2011	LR	0,12	0,599	0,152	0,363	0,083	0,870
2011	RF	0,13	0,599	0,152	0,363	0,083	0,870
2012	DT	0,11	0,776	0,296	0,927	0,299	0,292
2012	GB	0,14	0,792	0,301	0,904	0,244	0,394
2012	LR	0,07	0,743	0,227	0,876	0,169	0,348
2012	RF	0,09	0,797	0,302	0,912	0,259	0,362
2013	DT	0,08	0,800	0,358	0,894	0,292	0,461
2013	GB	0,14	0,819	0,371	0,918	0,365	0,377

Çizelge 2.4.(devam) DPModel-1 yıllara göre model başarıları

2013	GB	0,14	0,819	0,371	0,918	0,365	0,377
2013	LR	0,06	0,771	0,278	0,875	0,22	0,376
2013	RF	0,08	0,823	0,375	0,916	0,356	0,395
2014	DT	0,33	0,833	0,357	0,910	0,305	0,430
2014	GB	0,20	0,840	0,347	0,878	0,252	0,557
2014	LR	0,12	0,785	0,279	0,890	0,225	0,366
2014	RF	0,27	0,842	0,362	0,900	0,288	0,487
2015	DT	0,17	0,837	0,341	0,934	0,326	0,356
2015	GB	0,21	0,846	0,360	0,936	0,341	0,381
2015	LR	0,10	0,795	0,268	0,922	0,243	0,297
2015	RF	0,18	0,852	0,363	0,926	0,309	0,439
2016	DT	0,07	0,847	0,320	0,873	0,211	0,661
2016	GB	0,17	0,865	0,382	0,942	0,369	0,396
2016	LR	0,08	0,812	0,294	0,930	0,271	0,322
2016	RF	0,15	0,866	0,397	0,936	0,347	0,464
2017	DT	0,11	0,849	0,381	0,908	0,280	0,596
2017	GB	0,17	0,869	0,419	0,929	0,341	0,544
2017	LR	0,09	0,826	0,320	0,929	0,292	0,354
2017	RF	0,11	0,877	0,411	0,927	0,332	0,540
2018	DT	0,26	0,848	0,334	0,949	0,293	0,387
2018	GB	0,22	0,866	0,338	0,934	0,253	0,510
2018	LR	0,14	0,826	0,289	0,953	0,288	0,290
2018	RF	0,18	0,868	0,340	0,937	0,261	0,488

DPModel-1'in nihai sonuçlarına bakıldığında AUC skoruna göre en yüksek başarıyı RF algoritması vermiştir. İstatistiksel yöntem olan LR'la modelde kullanılan ML algoritmaları karşılaştırıldığında ML yöntemlerinin önemli ölçüde daha yüksek başarı sağladığı görülmektedir. Algoritmalar arasındaki bu farkın temel sebebi LR algoritmasının doğrusal olmasından (linear), ML algoritmalarının doğrusal olmamasından (non-linear) kaynaklanmaktadır. ML algoritmalarına bakıldığında ise DT algoritması RF ve GB algoritmalarına göre daha zayıf bir tahmin edicidir. Bunun gerekçesi veri miktarı ve değişken sayılarının çokluğundan dolayı tek bir ağacın

öğrenmede yeterli olmamasıdır. Bu problemi RF ağaç miktarını artırarak, GB ise zayıf öğrenicileri iterasyonla güçlü öğrenicilere çevirerek başarmaktadır.

DPModel-1'de kullanılan tüm algoritmalarının sonuçlarına bakıldığında kabul edilebilir seviyede başarılı olduğu görülmektedir. Bu başarının çeşitli gerekçeleri vardır. Bunlardan ilki her bir algoritmanın ızgara aramayla parametrelerinin optimizasyonundan kaynaklanmaktadır. Bir diğer sebep ise K Katmanlı Çapraz Doğrulama (Cross Validation / CV) kullanılarak modellerin birden fazla kez çalıştırılıp sonuçların ortalamasının alınmasındandır. Son olarak algoritmaların sınıflandırma metrikleri hesaplanırken eşik değer olarak varsayılan değeri yerine f1 skoru maksimize eden eşik değeri bulmaktan kaynaklanmaktadır.

DPModel-1'de elde edilen sonuçların seviyesinin kabul edilebilir olmasının bir diğer önemli sebebi ise kullanılan veri setindeki şirketler seçilirken herhangi bir filtre kullanılmamasıdır. Birçok temerrüt tahmini çalışması belirli filtreler kullanılarak seçilen şirketler üzerinde yapılmaktadır. Filtreler genellikle şirket ölçeği, sektörü veya bilanço büyüklüğü belirli bir düzeyin üzerindeki şirketler olmaktadır. Bu filtreler genellikle şahıs veya küçük ölçekli bilanço kalitesi düşük ve şirket dışı etkenlerden kolay etkilenen şirketleri dışlamaktadır. Bu çalışmada herhangi bir filtre olmaksızın tüm şirketler için temerrüt tahmini yapılmış ve kabul edilebilir başarı elde edilmiştir.

2.8. Temerrüt Tahmininde Derin Öğrenmeyle Kritik Analiz

DL algoritmaları literatürde giderek daha fazla kullanılan güçlü tahmin algoritmaları arasında yer almaktadır. Birçok farklı sınıflandırma probleminde önemli tahmin başarısı göstermektedir [4,49-52]. Bu alt çalışma bölümünde temerrüt tahmini için DL algoritmalarının başarısı analiz edilmiştir. Bu bölümün temel amacı temerrüt tahmininde DL algoritmalarının katkısının ölçülmesidir.

DL algoritmalarının yüksek hesaplama maliyetine sahiptir. Bu nedenle grafik kartlarına ve yüksek sistem gereksinimlerine ihtiyaç duymaktadır. Ancak çalışma sırasında kullanılan sistemlerde tüm veriyi DL ile inceleyebilecek gerekli teknik gereksinim olmadığı için çalışma örnek veri seti üzerinden yapılmıştır. Çalışma sırasında 2.9 Ghz i7 işlemci ve 16 GB ram'li bir bilgisayar kullanılmıştır. Örnek veri seti, kredi veri setinden

2010-2017 yılları arasında faaliyet gösteren ve toplam kredi riski 100 bin TL üzeri olan şirketlerden 67 694 adet şirket seçilerek oluşturulmuştur. Bu şirketlerdeki temerrüt oranı %12'dir. Şirketlere ait yıllık frekansta 57 adet kredi değişkeni kullanılmıştır.

Deneysel çalışma sınıflandırma problemi için sıklıkla kullanılan LSTM, Yinelenen Sinir Ağları (Recurrent Neural Network / RNN) ve Kapılı Yinelemeli Üniteler (Gated Recurrent Units / GRU) DL algoritmaları kullanılmıştır. Çalışmada Python Keras kütüphanesinden yararlanılmıştır. Algoritmaların ağ yapısı ve parametreleri manuel olarak belirlenmiştir. Her 3 algoritma için aynı değerler kullanılmıştır. Her biri için 2 gizli katman oluşturulmuş ve bu katmanlardaki nöron sayısı sırasıyla 64 ve 128 olarak belirlenmiştir. Aktivasyon fonksiyonu olarak sigmoid kullanılmıştır. Kayıp fonksiyonu için "binary-crossentropy" seçilmiştir. Başarı metriği olarak AUC belirlenmiştir. Her bir model için yineleme sayısı (epoch) 100 olarak ayarlanmıştır.

2010-2016 yıllarındaki veriler eğitimde 2017 yılı ise test veri seti olarak kullanılmıştır. Ana çalışmaya benzer olarak bir sonraki yılda temerrüde düşme tahmin edilmiştir. Çizelge 2.5'de test veri setinden elde edilen sonuçlar görülmektedir.

Çizelge 2.5. DL algoritmalarının tahmin sonuçları

Model	AUC	F1 Skor	Doğruluk	Kesinlik	Duyarlılık
RNN	0,790	0,140	0,883	0,717	0,077
GRU	0,762	0,164	0,880	0,550	0,096
LSTM	0,803	0,228	0,886	0,676	0,137

Örneklem veri seti üzerinden elde edilen çalışma sonuçlarına göre DL algoritmalarının temerrüt tahmininde başarılı olduğu gözlenmektir. Algoritmalardan elde edilen AUC skora bakıldığında birbirlerine yakın sonuçlar verdiği görülmektedir. En yüksek başarı 0,803 AUC skoru ile LSTM algoritmasından elde edilmiştir. Deneysel çalışma kısıtlı sistemle gerçekleştirildiğinden daha büyük ağ yapıları ve parametre optimizasyonlarıyla geliştirilmeye açıktır. Ancak buna rağmen başarı oranı DL algoritmalarının temerrüt tahmininde umut vaat edici olduğunu göstermektedir.

3. ÇİZGE TEORİSİYLE TEMERRÜT TAHMİNİ

Temerrüt tahmininde kredi ve bilanço verileri literatürde sıklıkla kullanılmaktadır. Kredi verileri şirketlerin finansal durumunu, bilançolar ise yıllık olarak şirketlerin varlık ve bu varlıkların elde edildiği kaynağı göstermektedir. Şirketlerin temerrüt tahmininde önemli olan bir diğer etkense şirketlerin faaliyetleri sırasındaki alım satım ilişkisi içinde oldukları tedarikçi ve müşterileridir. Bu bölümde temerrüt tahmininin geliştirilmesi için şirketlerin ticari ilişkileri çizge teorisiyle modellenerek faydalı bilgiler çıkarılması amaçlanmıştır. Şirketler arasındaki ticari ilişki fatura veri seti üzerinden oluşturulmuştur.

Bir şirketin alım satım ilişkisi içinde olduğu müşteri veya tedarikçilerinin temerrüde düşme olasılığı, şirketin temerrüt ihtimalini değiştireceği varsayımı, bu bölümün temel motivasyonunu oluşturmaktadır. Şirketlerin alım satım ilişkisini modellemede kullanılacak en iyi yapı çizgedir. Çizge yapısıyla tedarik zinciri kolaylıkla simule edilebilmektedir. Böylelikle her bir şirketin ilişkili olduğu diğer şirketler belirlenebilir. Ayıca şirketin çizge içindeki önemi belirlenir. Amacımız çizge içinden elde edilecek değişkenlerle temerrüt tahmininin geliştirilmesidir.

3.1. İlgili Çalışmalar

Çizge teorisi veya ağ analizi yöntemleri görebildiğimiz kadarıyla temerrüt tahmini probleminde bu zamana kadar kullanılmamıştır. Bu nedenle literatürde kullanımı incelenirken finans alanındaki diğer çalışmalar taranmıştır.

Park ve Yang [53] ağ analiziyle kriz öncesi ve kriz sonrası ülkeler arasındaki sermaye akışını incelemişlerdir. Uluslararası Para Fonu (International Monetary Fund / IMF) portföy yatırım veri seti üzerinden 2002 ve 2018 yılları arasında 24 adet gelişmiş, 41 adet gelişmekte olan ülkenin verisi kullanılmıştır. Çalışmada ülkeler düğüm, kenarlar sermaye akışı olacak şekilde çizge oluşturulmuş ve iç derece (indegree), dış derece (outdegree) ve özvektör (eigenvektör) merkeziliğine bakılmıştır. Sonuçlar Gayri Safi Yurtiçi Hasıla (Gross Domestic Product / GDP) büyüklüğü ve off-shore piyasası gelişmişliğinin sermaye akışına önemli etkisinin olduğunu göstermiştir.

Leon [54] bankalar arası güvenli ve güvensiz piyasalarda uygulanan faiz oranlarıyla ağ

analizinden elde edilen merkezilik değişkenlerinin arasındaki ilişkiye bakmıştır. Çalışmasında Meksika merkez bankasının 2007-2016 yılları arasındaki günlük işlem kayıtları kullanılmıştır. Ülkelerin finansal istikrarı için önemli olan piyasalarda uygulanan faiz oranı geleneksel ekonometrik yöntemlerden olan RIDGE, LASSO ve ElasticNet kullanılarak analiz edilmiştir. Sonuçlar güvenilir piyasalarda Pagerank, güvensiz piyasalarda Katz merkeziliği faiz oranıyla ilişkili olduğunu göstermiştir. Genel olarak bir bankanın merkeziliği arttıkça daha yüksek oranlarda borç verir ve daha düşük oranlarda fon alır.

Benzeri diğer çalışma [55] Temizsoy ve arkadaşlarına aittir. Regresyon analizi kullanarak farklı ağ analizi merkezilik metrikleriyle bankalar arası faiz oranlarını incelemişlerdir. 2006-2009 yılları arası İtalya veri setini kullanmışlardır. Sonuçlar Leon'un çalışmasına benzerdir. Borçlanan ne kadar merkeziyse faiz oranı o derece düşük olmaktadır. Her iki çalışmada batmayacak kadar iyi bağlantılı (too interconnected to fail) teorisini doğrulamaktadır.

Bankacılık sektörünün performansında kredi önemli bir rol oynamaktadır. Bhattachayra [56] ve arkadaşları bankaların kredi riskiyle finansal entegrasyon arasındaki ilişkiyi araştırmışlardır. Bunun için sendikasyon kredileriyle oluşturdukları çizge üzerinden finansal entegrasyonu ölçmüşlerdir. 1988 ile 2014 yılları arasında araç değişken teknikleri kullanılarak 39 ülkenin panel verisi üzerinde analizler yapılmıştır. Sonuçlar borçlanan bankaların kredi riskinin azalmadığını göstermektedir. Ayrıca borçlanan bankaların yakınlık merkeziliği yüksekse kredi riskinin önemli ölçüde azaldığı gözlemlenmiştir.

Ticari krediler en basit haliyle bir tedarikçinin müşterisine sağladığı borç olarak tanımlanmaktadır. Birçok şirketin bilançosunda önemli bir yere sahiptir. Gofman ve Wu [57] üretim ağlarında ticari krediler ile karlılık arasındaki ilişkiyi araştırmıştır. 2003-2018 yılları arasında faaliyet gösteren 5000 firmayla yapılan çalışmada tedarik zinciri ağ analiz yöntemleriyle incelenmiştir. Sonuçlar üzerinden birçok farklı gerçek çıkarılmıştır. Bunlardan başlıcası bir firmanın tedarik zincirinin üst kısmında yer alması nette ticari kredi miktarında artışa sebep olur. Ağ üzerindeki merkezi firmalar fazla kredi verir, fazla kredi alır ve nette ticari kredi tutarlar.

Kuzubaş ve arkadaşları [58] ağ analizinde merkezilik ölçüleriyle finansal sistemdeki sistemik risk oluşturabilecek kuruluşları belirlemeye çalışmışlardır. Çalışmalarını Türkiye'de 2000 yılında gerçekleşen finansal kriz sırasında bankalar arası piyasadaki

borçlanma verisi üzerinden yapmışlardır. Özellikle bu kriz sırasında bankacılık sisteminden çıkmak zorunda kalan Demirbank'ın kriz öncesi ve kriz esnasındaki ağdaki merkezilik ölçülerine bakılmıştır. Krizden önce ağ yapısında birçok borç alan ve veren banka olurken, kriz sırasında borç alanlar azalmış olduğu gözlemlenmiştir. Bir diğer bulgu kriz öncesinde bankalar arası işlem hacmi küçük ve işlem sayısı çokken. Kriz sırasında işlem hacimleri büyümüş ve işlem sayısı azalmıştır. Bu durum ağ üzerinde Demirbank'ın daha merkezi bir konum almasına sebep olmuş ve ağ yıldız topolojisine dönüşmüştür.

3.2. Çizge Teorisi Analizi

Matematiğin bir dalı olan çizge teorisi temelde küme teorisine dayanmaktadır. Çizge teorisi nesnelerin arasındaki ilişkileri modellemek için kullanılan matematiksel yapıları incelemektedir. Çizge, düğümler ve kenarlardan oluşan bir tür ağdır. Çizgeler genellikle G notasyonuyla gösterilmektedir. Çizgelerin genel matematiksel gösterimi G(V,E) şeklindedir. Burada V düğümleri, E ise kenarları temsil eden kümelerdir. Çizgeler kenarları üzerinde yön bilgisine göre yönlü ve yönsüz olarak ayrılmaktadır. Yönsüz çizgelere örnek olarak sosyal ağlardaki arkadaşlık ilişkisi verilebilir. Örneğin A kişisi B ile arkadaş ise B kişisi de A ile arkadaş olmak zorundadır. Yönlü çizgelere örnek ise yine sosyal ağlarda sıklıkla olan takipçi mekanizması gösterilebilir. Örneğin A kişisi B'nin takipçisi olması B'nin de A kişisinin takipçisi olması anlamına gelmemektedir.

Çizgelerde bir diğer önemli özellik kenarlara ağırlık verildiği durumdur. Kenar ağırlığı düğümler arasındaki ilişkileri ifade eden sayısal bir değerdir. Çizge grafiklerinde kenar ağırlığı genellikle çizgi kalınlığı şeklinde gösterilmektedir. Ağırlıklı çizgenin matematiksel notasyonu G(V, E, W) şeklinde olup W ağılıkları ifade etmektedir.

Bir düğümün çizge içinde ne kadar önemli olduğu merkezilik metrikleriyle hesaplanır. Başka bir ifadeyle merkezilik düğümün çizge içindeki önemini göstermektedir. Merkezilik hesaplamak için farklı metrikler mevcuttur. Bunlardan en popülerleri ve bu çalışmada kullanılanları derece merkeziliği, yakınlık merkeziliği, arasındalık merkeziliği ve özvektör merkeziliğidir.

3.2.1. Derece merkeziliği

Derece merkeziliği bir düğüme bağlı kenarların sayısıdır. Yönlü çizgelerde iki farklı derece merkeziliği hesaplanabilmektedir. Bunlar iç derece ve dış derece olarak isimlendirilmektedir. İç derece düğüme giren kenarların sayısını, dış derece ise düğümlerden çıkan kenarların sayısını göstermektedir.

Derece merkeziliği C_D aşağıdaki formül ile hesaplanmaktadır:

$$C_D(x)=\frac{\sum_{y=1}^{N} a_{xy}}{N-1} \tag{3.1}$$

Burada N çizge içindeki düğümlerin sayısını göstermektedir a_{xy} x ve y arasında kenar varsa 1 yoksa 0 değerini almaktadır.

3.2.2. Yakınlık merkeziliği

Yakınlık merkeziliği, bir düğümün çizge içindeki diğer düğümlere olan yakınlığını göstermektedir. Bu metrik bir düğümün diğer düğümlere ne kadar hızlı ulaşabildiğini göstermektedir. Yakınlık merkeziliği, bir düğümün diğer düğümlere en kısa yollarının ortalamasının tersi ile hesaplanmaktadır.

Yakınlık merkeziliği C_c aşağıdaki formül ile hesaplanmaktadır:

$$C_C(x)=\frac{N-1}{\sum_{y} d(y,x)} \tag{3.2}$$

Burada N çizge içindeki düğümlerin sayısını göstermektedir. $d(y,x)$ ise x ile y düğümü arasındaki en kısa yolun uzunluğunu göstermektedir.

3.2.3. Arasındalık merkeziliği

Arasındalık merkeziliği bir düğümün diğer düğümler arasındaki geçiş/köprü konumunda olmasını ölçmektedir. Bir düğümün arasındalık merkeziliği yüksek olması bu düğümün

doğrudan bağlantısı olmayan düğümler arasında köprü görevi görerek çizgenin iletişimi için önemli olduğunu göstermektedir.

Arasındalık merkeziliği C_B aşağıdaki formül ile hesaplanmaktadır:

$$C_B(x)=\sum_{u\neq v\neq x}\frac{\sigma_{uv}(x)}{\sigma_{uv}} \tag{3.3}$$

Burada σ_{uv} düğüm *u* ile düğüm *v* arasında kalan en kısa yolların sayısını, $\sigma_{uv}(x)$ ise düğüm *u* ile düğüm *v* arasındaki en kısa yollardan kaç tanesinin *x* düğümü üzerinden geçtiğinin sayısını göstermektedir.

3.2.4. Özvektör merkeziliği

Özvektör merkeziliği sadece bir düğümle bağlantılı düğümlerin sayısını göz önünde bulundurmayıp ayrıca bağlantılı olan düğümlerin önemini hesaba katmaktadır. Sosyal ağlar üzerinde bir örnek vermek gerekirse bir kişinin sadece kaç adet arkadaşının olduğuna bakarak önem derecesi belirlemek yerine arkadaşlarının popülerliği başka bir ifade ile arkadaşlarının arkadaş sayısını da göz önünde bulundurarak hesaplama yapmaktadır. Bu metriğe prestij metriği denmektedir. Gerçek hayatı simule etmekte çok yararlı kriterlerden biridir.

Özvektör merkeziliği C_E aşağıdaki formül ile hesaplanmaktadır:

$$C_E(x)=\frac{1}{\lambda}\sum_{y\in M(x)}C_E(y)=\frac{1}{\lambda}\sum_{y\in G}a_{x,y}C_E(y) \tag{3.4}$$

Burada λ özdeğeri göstermektedir. $M(x)$, *x* düğümün komşularının kümesini göstermektedir. *y* komşu düğümleri gösterir ve $a_{x,y}$ *x* ve *y*'nin komşu olup olmamasına göre 0 ile 1 değerini almaktadır.

3.3. Kullanılan Veri Setleri ve Temerrüt Tanımı

Çizge teorisiyle temerrüt tahmini için DPModel-1'de kredi ve bilanço veri setlerinden elde edilen olasılık skorları, fatura veri seti ve temerrüt veri seti kullanılmıştır. Tüm veri setleri TCMB veri tabanlarından elde edilmiş ve TCMB sistemleri üzerinde çalışılmıştır.

3.3.1. Bilanço ve kredi olasılık skorları

DPModel-1'deki bilanço ve kredi veri setlerinden oluşturulan alt modellerden elde edilen en iyi tahminleri yapan algoritmaları olasılık skorları şirket ve yıl ayrımında kaydedilmiştir. Bu veri setinde yıl, vergi numarası, kredi skoru ve bilanço skoru değişkenleri bulunmaktadır.

3.3.2. Fatura veri seti

Fatura bir mal/hizmetin satıldığını gösteren, satıcı şirketin düzenleyerek müşterisine ulaştırması gereken belgedir. Türkiye'de faaliyet gösteren şirketler Katma Değer Vergisi (KDV) hariç 5000 TL ve üzeri alış veya satış faturalarını GİB'e beyan etmektedir. Bu beyannameler aylık olarak düzenlenmektedir. Çalışmada 2010-2018 yılları arasında şirketlerin GİB'e gönderdikleri beyanname satış formlarından oluşturulan fatura veri seti kullanılmıştır. Bu veri setinde temelde beyanname yılı, ayı, satıcı ve müşteri vergi numarası ve mal/hizmet bedeli bulunmaktadır. Kredi ve bilanço skorları ve temerrüt yıllık frekansta veri setleri olduğu için fatura veri seti yıllık frekansa çevrilmiştir. Bu çevrimle veri seti, yıl, satıcı vergi numarası, müşteri vergi numarası ve satıcı ile müşteri arasında gerçekleşen mal/hizmet bedeli olarak düzenlenmiştir. Yıllara göre tekil satıcı şirket sayısı, tekil müşteri sayısı ve toplulaştırılmış fatura miktarları Çizelge 3.1'de verilmiştir.

Çizelge 3.1. Yıllara göre şirket, müşteri ve fatura sayısı

Yıl	Satıcı sayısı	Müşteri sayısı	Fatura sayısı
2010	619 926	1 682 087	8 691 608
2011	674 054	1 938 182	10 568 381
2012	719 875	2 084 469	11 751 065
2013	765 376	2 261 197	12 933 493
2014	818 371	2 390 387	14 277 163

Çizelge 3.1. (devam) Yıllara göre şirket, müşteri ve fatura sayısı

2015	878 920	2 628 272	15 709 458
2016	929 370	2 855 222	16 785 048
2017	988 131	3 138 366	18 996 823
2018	828 442	1 715 799	8 434 067

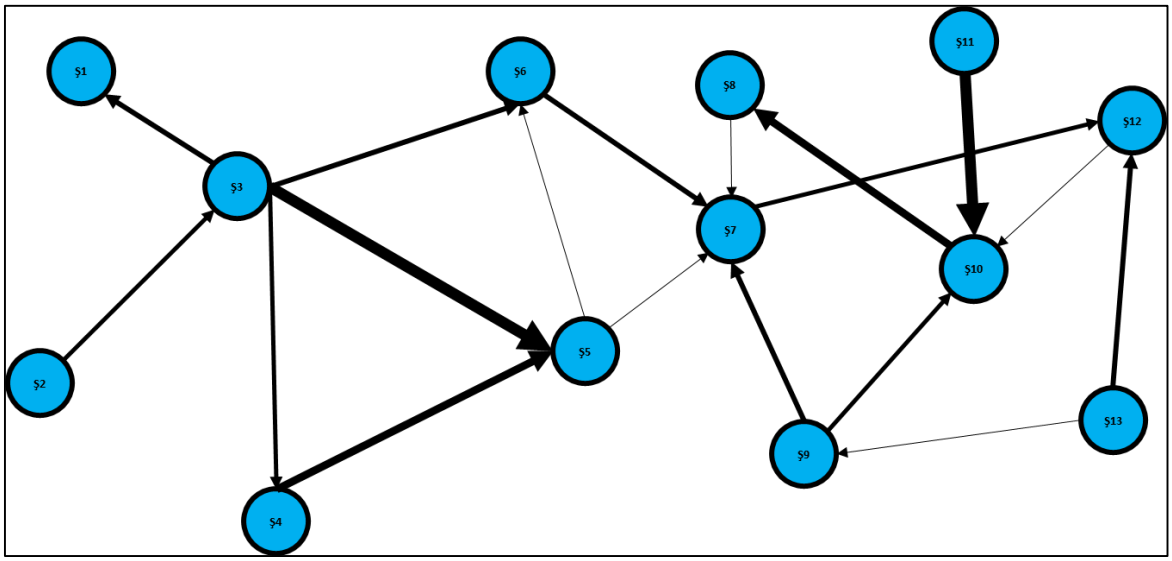
Çizelge 3.1'e bakıldığında müşteri sayısı satıcı sayısından oldukça büyüktür. Bunun sebebi müşterinin şahıs veya aylık tek bir müşteriye 5000 TL üzerinde satış yapamayan küçük işletmelerin olmasıdır. 2018 yılındaki veri miktarındaki azalışın sebebi kullandığımız veri setinin tüm yılı içermemesinden kaynaklanmaktadır.

3.3.3. Temerrüt veri seti

Kullanılan temerrüt veri seti DPModel-1 ile aynı veri setidir. Şirketlerin temerrüde düşmesini, kredi ödemesi 90 gün geçmiş veya çek veya senet ödemesini yapmamış veya hukuki olarak iflas veya konkordato ilan etmiş şirketler olarak belirlenmiştir. Çizelge 2.1'de yıllara göre şirketlerin sayısı ve temerrüt oranı verilmiştir.

3.4. Çizge Teorisiyle Temerrüt Tahmin Modeli (DPModel-2)

Çizge teorisiyle temerrüt tahmini için fatura veri seti üzerinden çizge oluşturulmuştur. Fatura veri seti, GİB'in aylık frekansta beyanname satış formundan elde edilmiştir. Bu veri setinde aylık toplam KDV hariç 5000 TL üzerinde faturaların bilgisi mevcuttur. Tedarik zinciri gereği fatura verisi çok karmaşık bir çizge yapısı oluşturmaktadır. Bununla birlikte içerisinde temerrüt tahmini için kıymetli bilgi barındırmaktadır. Çünkü temerrüde düşen bir şirket ilk olarak tedarik zinciri içerisindeki kendisine daha yakın olan diğer şirketlere zarar vermektedir. Bu karmaşık çizge yapısı Apache Spark'ın GraphX kütüphanesi yardımıyla simule edilmiştir. Oluşturulan çizgede düğümler (node) şirketleri, kenarlar (edge) ise iki şirket arasındaki toplam fatura miktarını göstermektedir.



Şekil 3.1. Ağırlıklandırılmış yönlü çizge yapısı

Şekil 3.1'de ağırlıklandırılmış yönlü çizgenin temsili bir gösterimi oluşturulmuştur. Şekilde düğümler şirketi, kenarlar yıllık toplam fatura miktarını göstermektedir. Kenarların yönü satan şirketten alan şirkete doğrudur. Kenarın kalınlığı ise fatura miktarına göre belirlenmektedir. Temsili çizgeye benzer şekilde tüm veri seti için ağırlıklandırılmış yönlü çizge oluşturup üzerinde daha sonra temerrüt tahmininde kullanılacak aşağıdaki gibi yeni değişkenler hesaplanmıştır.

3.4.1. Ağırlıklandırılmış müşteri skoru değişkeni

Çizge analizinden hesaplanacak ilk değişken ağırlıklandırılmış müşteri skorudur. Bu değişken görebildiğimiz kadarıyla literatürde bulunmamaktadır. Ancak bu çalışmada temerrüt tahmininde önemli katkı sağlamıştır. Ağırlıklandırılmış müşteri skorunun hesaplanması için çizge içindeki tüm şirketlerin daha önce atanmış bir temerrüt olasılık skorunun bulunması gerekmektedir. Bunun için DPModel-1'de elde ettiğimiz skorlar kullanılmıştır. Ancak çizgedeki tüm müşteriler için skor DPModel-1'de bulunmamaktadır. Bunun en önemli sebebi müşterinin gerçek şahıslar, bilanço hazırlamak zorunda olmayan küçük işletmeler ve ihracat yapılan yabancı şirketlerdir. Ağırlıklandırılmış müşteri skoru oluşturabilmek için bir şirketin satış yaptığı tüm müşterilerinin skorunun olması gerektiğinden skoru olmayan müşterilere skor atanması gerekmektedir. Bu tarz müşteri skorları bilinmeyen şirketlerin müşterilerinin faaliyet gösterdiği sektör biliniyorsa sektörünün ortalama skoru müşteri skoru olarak atanmıştır. Eğer şirket ihracatçı ve

müşterisinin sektörü de belli değilse şirketin kendisinin faaliyet gösterdiği sektörün skoru müşteri skoru olarak kullanılmıştır.

Sektörlerin skorları, DPModel-1'deki o sektörde faaliyet gösteren şirketlerin aldıkları skorların ortalaması olarak belirlenmiştir. Sektör ayrımı Avrupa Topluluğunda Ekonomik Faaliyetlerin Genel Sınıflandırılması (Statistical Classification of Economic Activities in The European Community / NACE) kodu ile yapılmıştır. NACE kodlarından 4 basamakta sektör ayrımı kullanılmıştır. Her sektörün ortalama skoru yıllık olarak hesaplanmıştır. Çizgedeki tüm şirketlerin müşterilerinin skoru aşağıdaki sözde kod ile belirlenmiştir.

- DBModel-1 skorlarını 4 basamakta sektör kodları (NACE) ile birleştir.
- Yıllara göre ortalama sektör skoru: Yıl ve sektöre göre gruplandır ve skorların ağırlıklı ortalamasını kaydet
- Her bir yıl için
  - Her bir şirketin için
    - Müşteri listesi: Ağ üzerinden müşterilerini bul
    - Müşteri listesindeki her bir müşteri için
      - Eğer müşterinin DPModel-1 de skoru varsa
        - Müşteri skoru = DPModel-1 skoru
        - Döngüden çık
      - Eğer müşterinin DPModel-1'de skoru yok ama sektörü belli ise (perakendeciler gibi)
        - Müşteri skoru: Ortalama sektör skoru
        - Döngüden çık
      - Eğer müşterinin DPModel-1'de skoru yok ve sektörü de belli değilse (ihracatçılar gibi)
        - Müşteri skoru: Şirketin kendi sektörünün ortalama skoru
        - Döngüden çık
    - Her müşteri için müşteri skorunu kaydet
  - Her şirket için müşterilerinin skorunu kaydet
- Her yıl için şirketlerin müşteri skorlarını kaydet

Yukarıdaki sözde kodla çizge üzerindeki her bir şirkete ait müşteri skoru belirlenmiştir. Bu durumda çizgedeki her bir şirketin müşterilerine ait ayrı ayrı müşteri skorları elde edilmiştir.

Temerrüt tahmininde kullanılacak olan ağırlıklandırılmış müşteri skorunu hesaplamak için gerekli olan diğer değişken satış oranıdır. Satış oranı, bir şirketin bir yıl içinde müşterisine yaptığı satışın toplam satışına oranıdır. Bu oran her bir şirketin tüm müşterileri için yıllık olarak hesaplanmıştır. Böylelikle bir müşterinin şirket için önemi satış oranı ile belirlenmiştir.

Ağırlıklandırılmış müşteri skoru, sözde kodla elde edilen müşteri skorunun satışa oranı ile ağırlıklandırılmasıyla elde edilmiştir. Her bir şirkete ait müşteri sayısı *n*, müşteri skoru *MS* ve fatura miktarı *FM* ile gösterilirse ağırlıklandırılmış müşteri skoru aşağıdaki formülle hesaplanmaktadır.

$$Ağırlıklandırılmış\ Müşteri\ Skoru = \frac{\sum_{i=1}^{n} MS_i * \frac{FM_i}{\sum FM}}{n} \tag{3.5}$$

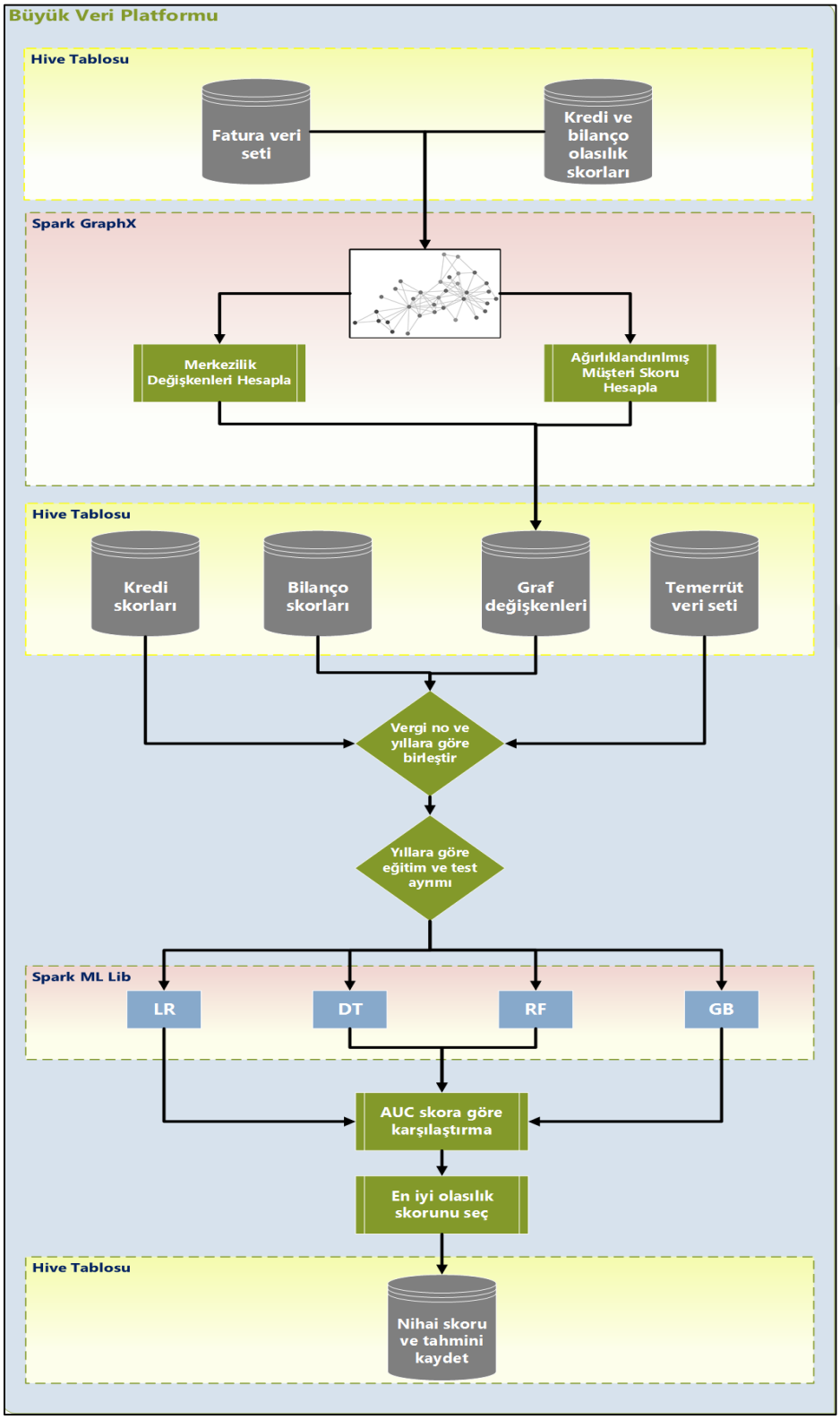
Bu formül çizgedeki tüm şirketler için çalıştırılmıştır. Böylelikle çizge içinde satış yapan tüm şirketlere ait ağırlıklandırılmış müşteri skorları elde edilmiştir. Böylelikle şirketin satışlarında müşterinin önemi ve müşterinin temerrüt olasılığı skoru üzerinde müşterinin şirketin temerrüt etme olasılığına etkisi hesaplanmıştır.

3.4.2. Çizge teorisinde merkezilik değişkenleri

Çizge içindeki önemli şirketlerin belirlenmesinde merkezilik göstergelerinin büyük önemi vardır. Şirketlerin fatura bilgisi üzerinden oluşturduğumuz çizgede şirketin çizge için önemini belirlenmesi olası bir temerrüt durumunda diğer şirketleri nasıl etkileneceğini görmemiz açısından büyük önem arz etmektedir.

Şirketler ve onların fatura ilişkisi üzerinden oluşturduğumuz çizgede derece merkeziliği, yakınlık merkeziliği, arasındalık merkeziliği ve özvektör merkeziliği yıl ve şirket bazında ayrı ayrı hesaplanmış ve her bir şirket için bu değerler Hive tablosuna kaydedilmiştir. Böylelikle tüm yıllar ve tüm şirketlere ait merkezilik değişkenleri oluşturulmuştur. Bir şirket

ne kadar merkezi ise o şirketin olası bir temerrüdü çizgedeki diğer şirketleri o derecede etkilemektedir.



Şekil 3.2. DPModel-2 çizge teorisiyle tahmin akış şeması

DPModel-2'ye ait akış şeması Şekil 3.2'de gösterilmiştir. DPModel-2'yi diğer temerrüt tahmini modellerinden ayıran en önemli kısım olan çizge analizidir. Çizgeyi oluşturmak için fatura veri seti ve DPModel-1'den gelen en iyi tahminler kullanılmıştır. Çizge, ağırlıklı yönlü çizge olarak oluşturulmuştur. Şirketler düğüm, fatura miktarları kenarları oluşturmaktadır. Yön bilgisi satış yönünü göstermektedir. Örneğin yön A şirketinden B şirketine doğru ise A şirketi B şirketine satış yaptığını göstermektedir. Ağırlık ise faturada yazılı olan mal veya hizmet bedelidir.

Oluşturulan çizge üzerinde temerrüt tahmini için kullanılacak olan değişkenler elde edilmiştir. Bunlardan ilki ağırlıklandırılmış müşteri skorudur. Diğerleri ise şirketlerin çizge içindeki konumlarından kaynaklanan önem derecelerini göstermektedir. Tüm değişkenler şirket ve yıl bazında ayrı ayrı oluşturulmuştur.

Çizge değişkenleriyle birlikte DPModel-2'de kullanılacak olan DPModel-1'in alt modellerinden gelen kredi ve bilanço olasılık skorları ve temerrüt veri seti birleştirilmiştir. Birleştirme yıl ve vergi numarası üzerinden yapılmıştır. DPModel-1'de olduğu gibi DPModel-2'de de eğitim ve test verisi yıllara göre bölünmüş ve bir önceki yıl verisi ile sonraki yıl tahmini yapılmıştır. Yine DPModel-1'de kullanılan tahmin algoritmaları ve değerlendirme metrikleri kullanılmıştır. Yıllara göre elde edilen en iyi skorlar Hive tablosuna kaydedilmiştir.

3.5. Deneysel Sonuçlar ve Değerlendirmesi

DPModel-2 temerrüt tahmini için çizge teorisi ve ML'yi bir arada kullanan yenilikçi bir modeldir. DPModel-2, şirketler arasındaki karmaşık ilişkiyi çizge teorisiyle açıklar ve güçlü tahminler için ML kullanır. Temerrüt tahmininde sıklıkla kullanılan bilanço ve kredi verileriyle birlikte fatura verisinden faydalanmaktadır.

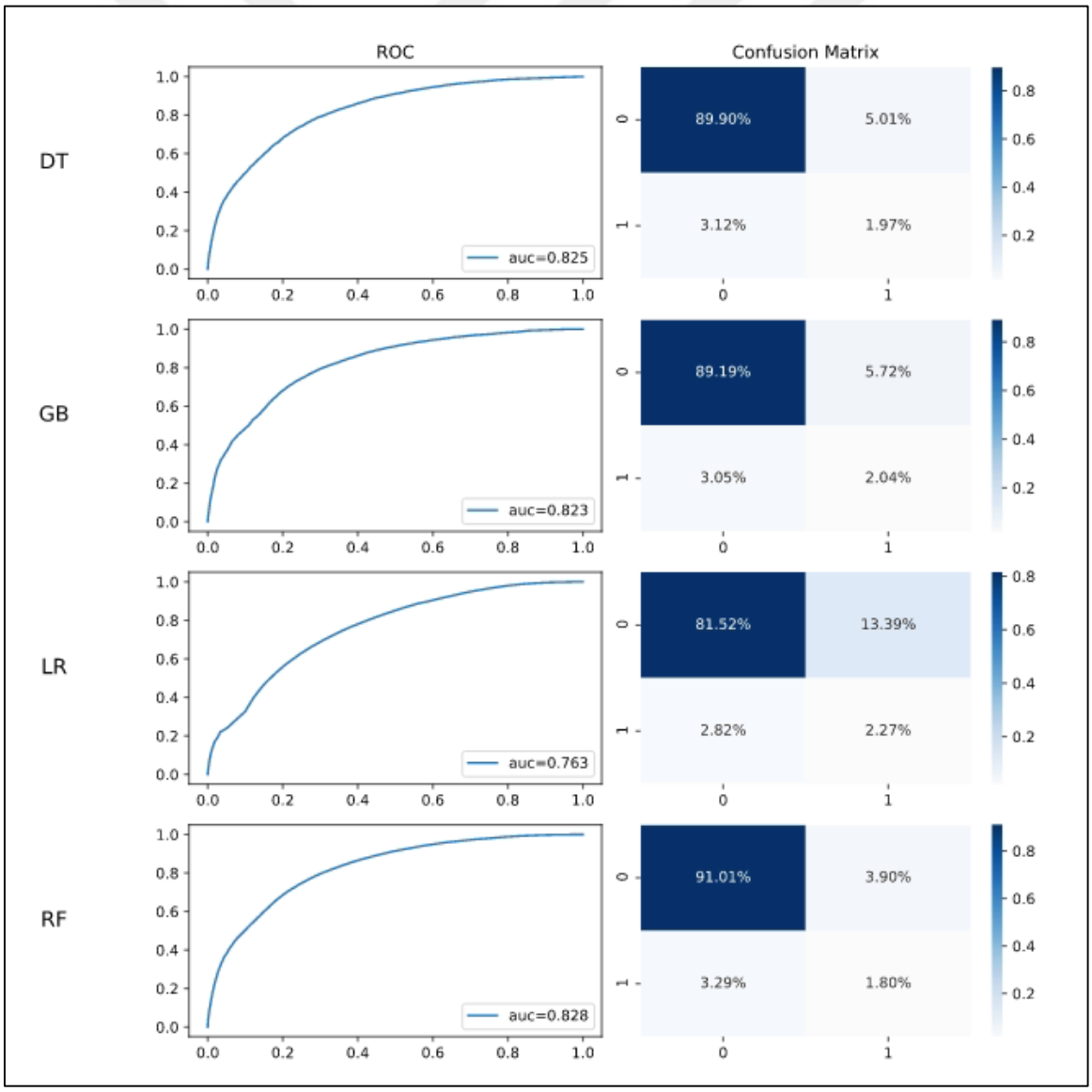
Şekil 3.2'de akış şemasını gördüğümüz DPModel-2 tüm veri setimiz için çalışılmıştır. Çizelge 3.2 yıl ayrımı olmaksızın tüm tahminlerden elde edilen sonuçları göstermektedir. Ayrıca Şekil 3.3'de modelin ROC eğrisi ve karşılaştırma matrisi mevcuttur.

DPModel-1'de olduğu gibi DPModel-2'de f1 skoru maksimum yapan eşik değer hesaplanmış ve tüm tablolar bu değer üzerinden hesaplanmıştır. Ayrıca tahmin

algoritmalarına ızgara aramayla parametre optimizasyonu uygulanmıştır. ML algoritmalarının optimizasyonu için seçilen parametreler DPModel-1'de kullanılanlarla aynıdır.

Çizelge 3.2. DPModel-2 ortalama model başarıları

Model	Eşik Değer	AUC	F1 Skor	Doğruluk	Kesinlik	Duyarlılık
DT	0,18	0,825	0,326	0,919	0,282	0,387
GB	0,18	0,823	0,317	0,912	0,263	0,400
LR	0,08	0,763	0,219	0,838	0,145	0,446
RF	0,19	0,828	0,334	0,928	0,316	0,355



Şekil 3.3. DPModel-2 sonuçlarının ROC eğrisi ve karşılaştırma matrisi

Yıl ayrımı olmaksızın sonuçlara bakıldığında DT, GB ve RF birbirine yakın başarı göstermekle birlikte en yüksek başarıyı RF algoritması göstermiştir. LR ise görece daha düşük tahmin başarısı göstermektedir. Bunu en önemli sebebi LR algoritmasının doğrusal yöntemlerden olmasından kaynaklanmaktadır. DPModel-2'nin yıllara göre tahminlerinin sonuçları Çizelge 3.3'de gösterilmektedir.

Çizelge 3.3. DPModel-2 yıllara göre model başarıları

Yıl	Model	Eşik Değer	AUC	F1 Skor	Doğruluk	Kesinlik	Duyarlılık
2011	DT	0,13	0,653	0,168	0,624	0,098	0,581
2011	GB	0,16	0,651	0,167	0,570	0,096	0,657
2011	LR	0,12	0,603	0,152	0,363	0,083	0,870
2011	RF	0,13	0,653	0,168	0,603	0,098	0,612
2012	DT	0,10	0,804	0,306	0,911	0,259	0,375
2012	GB	0,14	0,804	0,309	0,920	0,282	0,340
2012	LR	0,07	0,748	0,229	0,875	0,169	0,354
2012	RF	0,11	0,811	0,312	0,917	0,277	0,359
2013	DT	0,12	0,818	0,375	0,922	0,386	0,365
2013	GB	0,13	0,820	0,372	0,909	0,333	0,422
2013	LR	0,07	0,770	0,278	0,895	0,247	0,317
2013	RF	0,09	0,822	0,381	0,912	0,346	0,424
2014	DT	0,17	0,841	0,340	0,863	0,236	0,609
2014	GB	0,24	0,845	0,374	0,912	0,320	0,450
2014	LR	0,12	0,785	0,279	0,890	0,225	0,365
2014	RF	0,29	0,848	0,372	0,911	0,316	0,451
2015	DT	0,21	0,854	0,367	0,933	0,334	0,407
2015	GB	0,20	0,853	0,369	0,933	0,337	0,409
2015	LR	0,10	0,795	0,267	0,923	0,244	0,296
2015	RF	0,19	0,859	0,371	0,931	0,326	0,430
2016	DT	0,16	0,870	0,406	0,940	0,368	0,453
2016	GB	0,15	0,870	0,394	0,937	0,346	0,458
2016	LR	0,08	0,812	0,294	0,931	0,271	0,321
2016	RF	0,16	0,877	0,411	0,941	0,374	0,457

Çizelge 3.3. (devam) DPModel-2 yıllara göre model başarıları

2017	DT	0,16	0,883	0,439	0,938	0,386	0,509
2017	GB	0,17	0,878	0,439	0,935	0,372	0,537
2017	LR	0,09	0,826	0,320	0,929	0,292	0,355
2017	RF	0,17	0,890	0,444	0,941	0,399	0,500
2018	DT	0,29	0,872	0,359	0,947	0,301	0,445
2018	GB	0,25	0,871	0,360	0,948	0,302	0,446
2018	LR	0,14	0,826	0,289	0,953	0,288	0,290
2018	RF	0,31	0,874	0,362	0,950	0,314	0,426

Çizelge 3.3'de 2011-2018 yılları arasında DPModel-2'nin tahmin sonuçları görülmektedir. En yüksek AUC başarısı 0,890 ile 2017 yılında RF algoritmasından gözlemlenmiştir. DPModel-1'e benzer şekilde yıllara göre artış ve son yıl olan 2018'de bir miktar düşüş gözlenmiştir. Bunun sebebi DPModel-1'de de açıklandığı gibi veri kalitesi ve sayısından kaynaklanmaktadır.

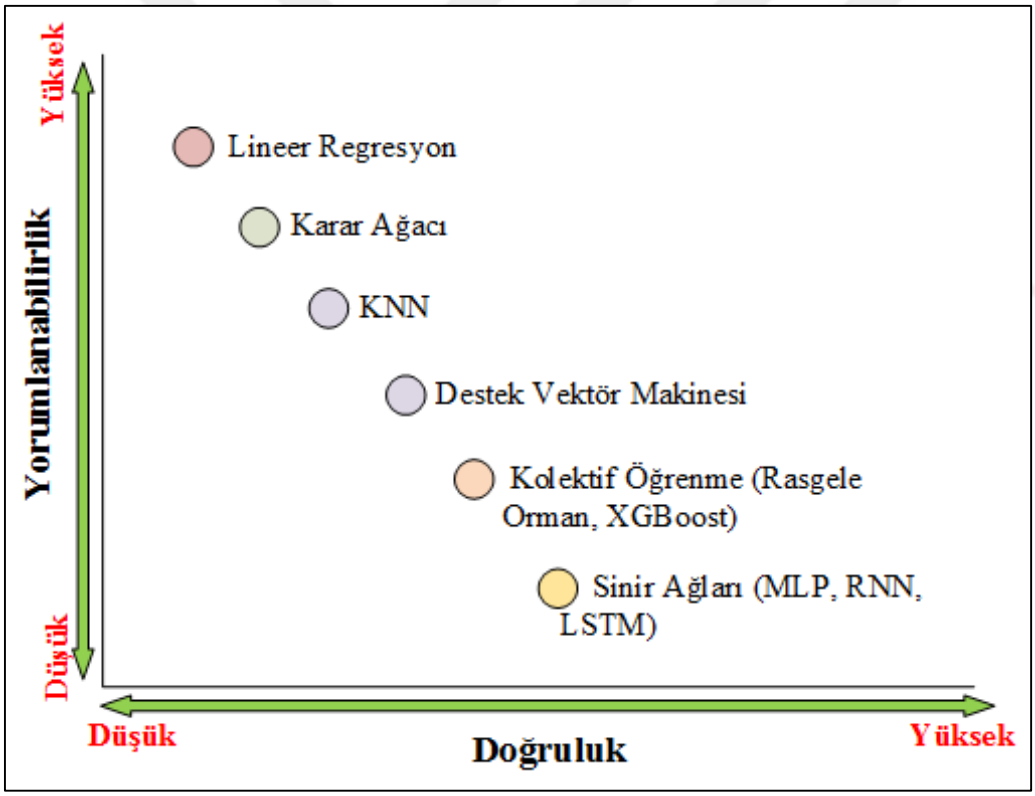
Tezde önerilen DPModel-1 ve DPModel-2'nin yıllara göre sonuçlarına bakıldığında her iki modelinde başarısının giderek artığı gözlemlenmektedir. Bunun en önemli sebebi veri kalitesinin ve miktarının yıllara göre artmasıdır. Özellikle bilanço veri setindeki kalitenin artışı buna sebep olmuştur. DPModel-2'de kullanılan fatura veri setindeki kalite ve miktarındaki artışta önemli bir diğer etkendir. DPModel-2'deki başarının DPModel-1'e göre daha yüksek olması çizge üzerinde oluşturulan değişkenlerin temerrüt tahminine olumlu katkısından kaynaklanmaktadır. İlk kez önermiş olduğumuz ağırlıklandırılmış müşteri skoru ile şirketin temerrüt etme olasılığına müşterilerinin katkısı ölçülmüştür. Çizge teorisi merkezilik değişkenleriyle ise şirketin çizge içindeki öneminin katkısı ölçülmüştür.

Önerilen modellerden elde edilen AUC skorlar literatürdeki diğer çalışmalarla karşılaştırıldığında kabul edilebilir başarıdadır. Literatürdeki bazı çalışmalarda skorların bizim elde ettiğimiz skorlardan daha yüksek olduğu görülmektedir [6,59]. Ancak bu çalışmalarda görülen örneklem seçimi ve filtre uygulanması bizim çalışmamızda uygulanmamıştır. Bu durum başarı oranını aşağı çeken veri kalitesi düşük birçok küçük ölçekli şirketin analize katılması anlamına gelmektedir. Küçük ölçekli şirketlerin temerrüdü sadece şirketin bilanço, kredi, fatura gibi değişkenleriyle ölçülemeyip; şirketin sahibi,

lokasyonu gibi farklı etkenlerden çok daha fazla etkilenmektedir. Bu duruma rağmen küçük ölçekli şirketleri analizde tutmamızın amacı tezde önerilen modellerin gerçek hayattaki tahmin başarısını daha net gözlemlemektir. Literatürdeki çalışmaların skorlarının yüksek olmasının bir diğer önemli sebebi temerrüt oranının daha dengeli olmasından kaynaklanmaktadır [5-7,59].

4. IML İLE ÖZNİTELİKLERİN SONUCA KATKISININ ÖLÇÜLMESİ

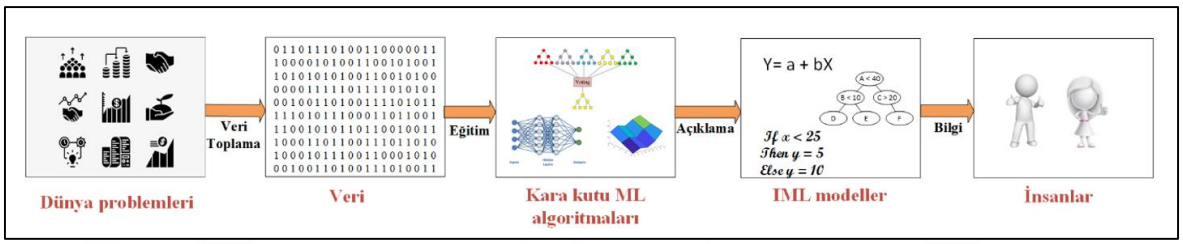
Tahmin yöntemlerinde doğruluk ve yorumlanabilirlik arasında ters bir ilişki vardır. Şekil 4.1'de görüldüğü gibi tahmin doğruluğu düşük olan yöntemlerin yorumlanabilirliği yüksek, tersinin ise yorumlanabilirliği düşüktür. RF, GB, DL algoritmaları gibi genellikle yüksek tahmin başarısına sahip algoritmalara getirilen en önemli eleştiri kara kutu şeklinde çalışmasıdır. Diğer bir deyişle modelin kararlarının nasıl oluştuğunun insan zekâsı ile anlaşılamamasıdır. Bu durum örneğin bir film tavsiye sistemi için çok önemli bir problem teşkil etmezken, bir hastalık teşhis sistemi veya kredi skoru belirleme sisteminde kritik öneme sahiptir.



Şekil 4.1. ML algoritmalarında yorumlanabilirlik ve doğruluk arasındaki ters ilişki

Son yıllarda karmaşık ML yöntemlerinin bu eksikliğinin giderilebilmesi için IML algoritmaları önerilmiştir. Böylelikle sağlık, finans gibi kritik alanlarda sonuçların açıklanabilir olması için doğruluktan vazgeçilmek zorunda kalınmaması amaçlanmıştır. Tezin konusu olan temerrüt tahmini de son derece kritik bir öneme sahiptir. Bu nedenle literatürdeki birçok çalışma karmaşık ML yöntemlerinin yorumlanabilir olmamasından

dolayı eleştirilmiştir. Bu bölümde önerdiğimiz temerrüt modellerin sonuçları IML algoritmalarıyla incelenmiştir. Şekil 4.2.'de IML'nin genel akış şeması görülmektedir. Çeşitli kaynaklardan toplanan veriler karmaşık makine öğrenmesi yöntemleriyle analiz edilmektedir. Analiz sonuçlarının insanlar tarafından yorumlanabilir olması için IML modeller kullanılmaktadır.



Şekil 4.2. IML'nin genel akışı şeması

Önerdiğimiz her iki modelde en yüksek tahmin başarısı RF algoritmasıyla elde edilmiştir. Bu algoritma karmaşık bir ağaç yapısı kullanarak tahmin yaptığından yorumlanabilirliği düşüktür. Temerrüt etme ihtimali olan şirketlerin doğru tahmini, başta şirketlerin kendileri olmak üzere ekonomi yönetimi, bankacılık, sigortacılık gibi birçok alanda kritik öneme sahiptir. Önerdiğimiz modellerin gerçek dünya problemine uygulanabilmesi için sonuçlarının insan zekâsıyla anlaşılabilir olması gerekmektedir. Bu nedenle her iki model içinde IML algoritmaları kullanılarak modellere açıklanabilirlik eklenmesi amaçlanmıştır.

4.1. İlgili Çalışmalar

Yorumlanabilir makine öğrenmesi yeni bir kavram olmakla birlikte literatürde giderek yoğun olarak çalışılan bir konu haline gelmiştir. IML, ML algoritmalarının başarısı kadar sonuçları bulurken nasıl davrandığının bilinmesi gerekli olan sağlık [60], finans [61], otonom sistemler [62] gibi alanlarda sıklıkla kullanılmaktadır.

Demajo ve arkadaşları [63] XGBOOST algoritmasıyla yapılan kredi skorlama için geniş bir perspektiften açıklanabilirlik sunan bir iş çerçevesi önermişlerdir. Bu iş çerçevesinde global, lokal ve öznitelik tabanlı lokal olmak üzere farklı açılardan açıklanabilirlik sunulmuştur. Global açıklanabilirlik için SHAP+Grip, lokal için PromoDash ve öznitelik tabanlı lokal açıklanabilirlik için Anchors kullanılmıştır. Bu modeller HELOC (Home Equity Line of Credit) ve LC (Lending Club) veri setlerinde uygulanmıştır. Önerilen iş çerçevesi problemi

mantıksal kurallarla global açıklanabilirlik sunan BRCG (Boolean Rules via Column Generation) ile karşılaştırılmıştır. Önerilen iş çerçevesi tutarlılık, basitlik, doğruluk, etkililik, kolay anlaşılabilirlik, detay yeterliliği ve güvenilirlik açısından analiz etmek için işlevsel temelli, uygulama temelli ve insan temelli olmak üzere üç farklı değerlendirme yaklaşımı benimsenmiştir. Çalışma sonucunda önerilen iş çerçevesinin kredi skorlama sırasında finans uzmanları için önemli bir yarar sağlayacağını göstermişlerdir.

Bussmann ve arkadaşları [64] Demajo'nun çalışmasına benzer şekilde yorumlanabilirliği düşük XGBOOST algoritmasıyla birebir borçlanmalarda kredi riski tahmininin açıklanabilirliği üzerinde çalışmışlardır. Her bir şirketin kredibilitesini açıklayabilmek için Shapley değeri kullanılmıştır. Çalışma ECAI (External Cedit Assessment Institution) veri seti üzerinde yapılmıştır. Veri setinde 15 binin üzerinde KOBİ ölçeğindeki şirketin 2015 yılı resmi bilanço değişkenleri kullanılmıştır. Çalışma sonuçları LR ile karşılaştırılmıştır. Karşılaştırmada ROC eğrisinden yararlanılmıştır. XGBOOST algoritması LR algoritmasına göre daha yüksek başarı göstermiştir. Shapley değeri hesaplanarak açıklanabilirlik problemi çözülmüştür.

Grath ve arkadaşları [61] kara kutu şeklinde çalışan sınıflandırıcıları karşı olgu açıklama yöntemiyle açıklamayı önermişlerdir. Karşı olgu açıklama yöntemi örneklem bazlı bir yorumlanabilirlik yöntemidir. Kısaca bir örneklemin tahmininde kullanılan özniteliklerin değeri minimal değiştiğinde tahminin nasıl değişeceğini açıklamaya çalışmaktadır. Yazarlar temel karşı olgu açıklanabilirliğinin başarısını artırmak için 2 ağırlıklandırma yöntemi önermiştir. Yöntemler HELOC kredi veri seti üzerinde test edilmiştir. Sonuçlar önerilen yöntemin temel karşı olgu yöntemine göre daha hassas ve daha yorumlanabilir olduğunu göstermiştir.

Futagami ve arkadaşları [65] şirket satın alma tahmini için SHAP değeri ile yorumlanabilirlik aramışlardır. Şirket satın alma tahmininde şirketin kuruluş tarihi, lokasyonu, sektörü, şirket haberleri, şirket yöneticileri gibi finansal olmayan bilgileri ve şirketler arasında oluşturulan alım satım ağ analizinden elde ettikleri değişkenleri kullanmışlardır. Deneysel çalışmada CruchBase 2000-2018 yılı veri setinden yaralanmışlardır. ML algoritması olarak LigthGBM kullanmışlardır. Sonuçları açıklayabilmek için SHAP algoritmasından yararlanmışlardır. Çalışma çizge teorisinden elde ettikleri değişkenlerin tahminde önemli fayda sağladığını ortaya çıkarmıştır.

4.2. Yorumlanabilir Makine Öğrenmesi Metotları

ML algoritmaları için yorumlanabilirlik, metodun tahmin veya kararlarının sebepleri insanların anlayabileceği düzeyde olmasını ifade etmektedir. Yorumlanabilirlik özellikle kritik problemlerde son derece önemli olmaktadır. Çünkü ML algoritmaları gerçek dünya uygulamalarında kullanılmakta olduğundan insanların hayatını doğrudan etkileyebilmektedir. Örneğin son yıllarda ML algoritmalarıyla çalışan sürücüsüz araçlar tasarlanmaktadır. Bu araçların olası bir kazaya karışması durumunda algoritmanın neden bu kazayı yaptığının açıklanması gerekmektedir. Algoritmaların yorumlanabilir olmasının bir diğer önemli gerekçesi algoritmanın eğitim setinde beklenmedik yanlılıklar (bias) yapma ihtimalidir. Örneğin bir kredi başvuru kabul uygulamasında algoritmanın bir bölge, grup veya ırktan olan başvuruları doğrudan reddetmesi gibi yanlılık problemleri doğurabilir. Eğitim veri setinde bölge, grup veya ırkla ilgili değişken olmasa bile algoritma eğitimi sırasında bir gruba olumlu veya olumsuz yanlılık yapma durumu olabilir. ML algoritmalarının yorumlanabilir olmasının önemli bir diğer gerekçesi ise algoritmalardan yeni bilgilerin yorumlanabilirlikle elde edilmesindendir. Başka bir deyişle eğer algoritmayı yorumlayabilirsek daha önce fark edilmemiş önemli ek bilgi ve örüntüleri keşfetme imkânımız olabilir.

IML algoritmaları farklı şekillerde sınıflandırılmaktadır. Bunlardan ilki içsel yorumlanabilir modeller (intrinsic IML) ve agnostik modeller (post hoc models)'dir. İçsel modeller LR, DT gibi algoritmaların doğası gereği sonuçların yorumlanabildiği modellerdir. Agnostik modellerse SHAP, LIME, ELI5 (Explain Like I'm Five) [66] gibi model eğitimi tamamlandıktan sonra yorumlanabilirlik sağlayan modellerdir. Bir diğer ayrım ise modelin bağımlı veya bağımsız olarak sınıflandırılmasıdır. Bu ayrım adından da anlaşılacağı gibi modelin belirli ML algoritmasıyla mı yoksa ML algoritması seçmeksizin mi çalışabildiğine göre yapılmaktadır. Bir diğer sınıflandırma ise yorumlanabilirliğin lokal veya global olmasına göredir. Yorumlanabilirliğin veri içerisindeki bir örneklem için yapılması durumu lokal, veri setinin tamamı için yapılması durumu ise global olarak isimlendirmektedir.

Tez kapsamında agnostik IML modelleri kullanılmıştır. "IML" dendiğinde agnostik IML modellerden bahsedilmektedir. Agnostik IML modellerinden SHAP ve LIME algoritmaları kullanılmıştır. Bu algoritmalar aşağıda açıklanmaktadır.

4.2.1. Lokal vekil model (LIME)

LIME [67] ML algoritmalarının veri seti içindeki tekil örneklerinin yorumlanmasında kullanılan modeldir. LIME tekil örneklerin açıklanmasında yorumlanabilir vekil bir model kullanır. Bu vekil model LR, DT gibi yorumlanabilir bir modeldir. Çalışması sezgiseldir. ML algoritmalarını tamamen bir kara kutu olarak kabul eder. ML algoritmalarının girdilerini ve tahminlerini alır. Amacı ML'nin bir örneklem için belirli bir tahmini neden yaptığını anlamaktır. Bunun için seçilen örnekleme yakın olan yeni örneklemler oluşturur. Yakınlık derecesine göre ağırlıklandırılır. LIME, ağırlıklandırılmış yeni veri setiyle yorumlanabilir vekil modelle eğitilir. Vekil modelin değişkenlere verdiği ağırlıklarla yorumlanabilirlik sağlanmış olur.

LIME matematiksel olarak aşağıdaki gibi formüllenir:

$$explanation(x) = \underset{g \in G}{arg\,min}\, L(f, g, \pi_x) + \Omega(g) \tag{4.1}$$

Burada x açıklanmak istenen örneği, g LR veya DT gibi vekil modeli, f RF gibi tahminleri yapan asıl modeli ve π_x ise x örneğinin yakınlık ölçeğini temsil etmektedir. L en küçük ortalama kare hatası gibi kayıp fonksiyonunu minimum yapan fonksiyondur. $\Omega(g)$ vekil modelin öznitelik sayısını düşük tutmak için kullanılır. Ancak genellikle pratikte kullanıcı öznitelik sayısını kendi belirler.

4.2.2. Shapley katkı açıklamalar modeli (SHAP)

SHAP, Lundberg ve Lee tarafından 2017 yılında önerilmiş agnostik bir IML algoritmasıdır [68]. SHAP oyun teorisi yöntemlerinden Shapley değeri temel alınarak geliştirilmiştir. SHAP'in amacı bir örneklemin her bir özniteliğinin tahmine olan katkısını açıklamaktır. Bunun için temel aldığı Shapley değeri, bir takım oyununda takımdaki her bir oyuncunun skora olan marjinal katkısını ölçmek için geliştirilmiştir. Bunu takımda her bir oyuncunun teker teker takımdan çıkarılıp yerine ortalama bir oyuncunun alınması durumunda skorun nasıl değişeceğini ölçerek yapmaktadır. Böylelikle her bir oyuncunun skora olan marjinal katkısı ölçülebilmektedir. Her bir oyuncu birer birer çıkarılarak ölçüm yapıldığı için

takımdaki oyuncu sayısının artması, Shapley değerinin hesaplanmasını üstel olarak artırmaktadır. Benzer şekilde SHAP modeli her bir oyuncuyu her bir öznitelik olarak kabul etmektedir. Her bir özniteliğin ML algoritmasının tahminine olan katkısını ölçerek yönteme açıklanabilirlik aramaktadır.

SHAP, LIME ve Shapley değerini birleştirmektedir. SHAP aşağıdaki gibi formülize edilir.

$$g(z') = \varphi_0 + \sum_{j=1}^{M} \varphi_j z'_j \tag{4.2}$$

g açıklanabilir model olmak üzere M maksimum öznitelik sayısı, $z' \in [0,1]$ M basitleştirilmiş yeni veri setidir. $\varphi_j \in R$ olmak üzere j özniteliğinin tahmine olan katkısı, yani Shapley değeridir. Shapley değeri φ_i aşağıdaki formülle hesaplanmaktadır.

$$\varphi_i = \sum_{S \subseteq M \setminus i} \frac{|S|!(|M|-|S|-1)!}{|M|!} \left[f(S \cup i) - f(S) \right] \tag{4.3}$$

φ_i değeri, öznitelik i kullanılarak yapılan tahmin ile i olmaksızın yapılan tahmin arasındaki farkı göstermektedir. Böylelikle i özniteliğinin marjinal katkısı belirlenmiş olmaktadır. S, i özniteliği olmaksızın oluşturulan bir alt öznitelik kümesidir. $S \cup i$ öznitelik altkümesine i'nin eklendiği durumu göstermektedir. M özniteliklerin tamamını göstermektedir. $S \subseteq M \setminus i$ ise i özniteliği hariç olası alt kümeleri göstermektedir.

4.3. DPModel-1 ve DPModel-2 için Yorumlanabilirlik Uygulaması

ML algoritmalarının temerrüt tahmininde kullanılması en çok yorumlanabilirliğinin olmamasından dolayı eleştirilmektedir. Bu eleştirinin makul gerekçeleri vardır. Örneğin bir bankanın bir şirketi temerrüt etme ihtimali olmadığı halde temerrüt edecek olarak değerlendirip kredi vermemesi ekonomik ve hukuki problemler doğurabilmektedir. Böyle bir tahminin gerekçelerinin açıklanmaya ihtiyacı vardır. IML algoritmaları karmaşık ML algoritmalarının bu eksikliğini gidermektedir.

Bu bölümde tez kapsamında önerilen DPModel-1 ve DPModel-2 için eksik olan sonuçların açıklanabilirliği IML algoritmalarıyla giderilmeye çalışılmıştır. Çünkü her iki tahmin modeli

en yüksek başarıyı karmaşık ağaç tabanlı algoritmalardan olan RF algoritmasıyla elde etmiştir.

IML algoritmaları olarak LIME ve SHAP kullanılmıştır. LIME sadece lokal yorumlanabilirlik sağladığı için sadece seçilen şirketler için lokal sonuçlar bulunmuştur. SHAP'le hem lokal hem de global sonuçlar elde edilmiştir. Özellikle SHAP'in global sonuçları temerrüt tahmininde hangi değişkenlerin en çok katkı sağladığını göstermesi açısından çok kıymetlidir. Çalışmanın yapıldığı sırada henüz BVP üzerinden dağıtık olarak çalışan bir IML algoritması olmadığı için çalışma lokal makine üzerinde yapılmıştır. Lokal makinede yapılmasından dolayı tüm veri setiyle çalışılamamıştır. Tüm veri setiyle çalışılamamasının bir diğer önemli sebebi SHAP öznitelik sayısı arttıkça hesaplama karmaşıklığının üstel olarak artması ve global yorumlanabilirlik için birçok örneklemin Shapley değerini ayrı hesaplamak zorunda kalmasıdır. Bu bölümdeki deneyler sadece 2017 yılı verisi kullanılarak gerçekleştirilmiştir.

Çalışmada DPModel-1'in kredi ve bilanço veri setlerinden oluşturulan alt modelleri ve DPModel-2 modeli IML algoritmalarıyla açıklanmıştır. Her iki model için hem global hem de lokal yorumlanabilirlikler incelenmiştir. Böylelikle lokal ve global olarak çözüme en çok katkı sağlayan öznitelikler belirlenmiştir.

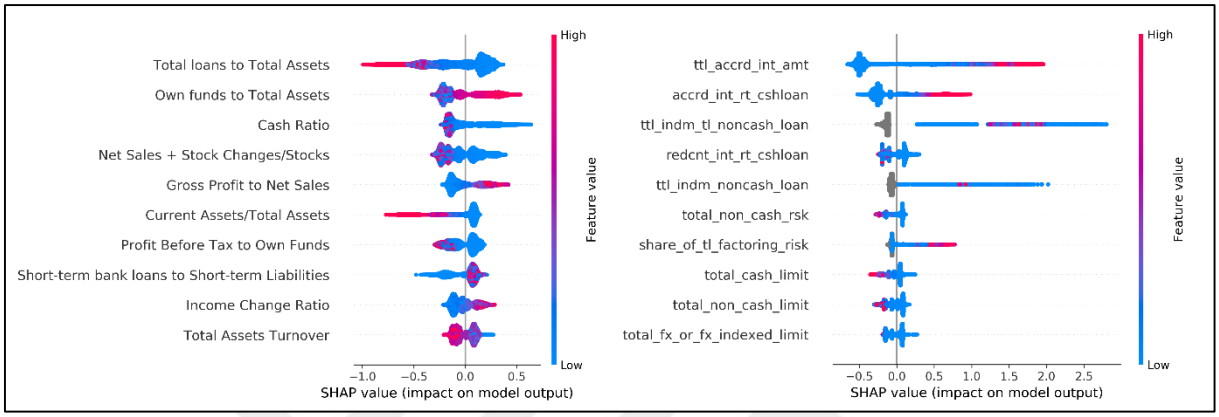
4.4. Sonuçlar ve Değerlendirmeler

SHAP ve LIME algoritmalarıyla DPModel-1 ve DPModel-2 modellerinin sonuçları için açıklanabilirlik aranmıştır. Sonuçlar global ve lokal yorumlanabilirlik olarak ayrılmıştır.

4.4.1. Global yorumlanabilirlik

Global yorumlanabilirlik, yorumlanabilirliğin veri setinin tamamı için arandığı durumdur. LIME algoritması sadece lokal yorumlanabilirlik sağladığı için bu bölümde SHAP algoritmasıyla sonuçlar elde edilmiştir. DPModel-1'de oluşturulan alt modeller ve DPModel-2 için en yüksek başarıyı elde ettiğimiz ML algoritması olan RF algoritmasının sonuçlarının SHAP değerleri hesaplanmıştır.

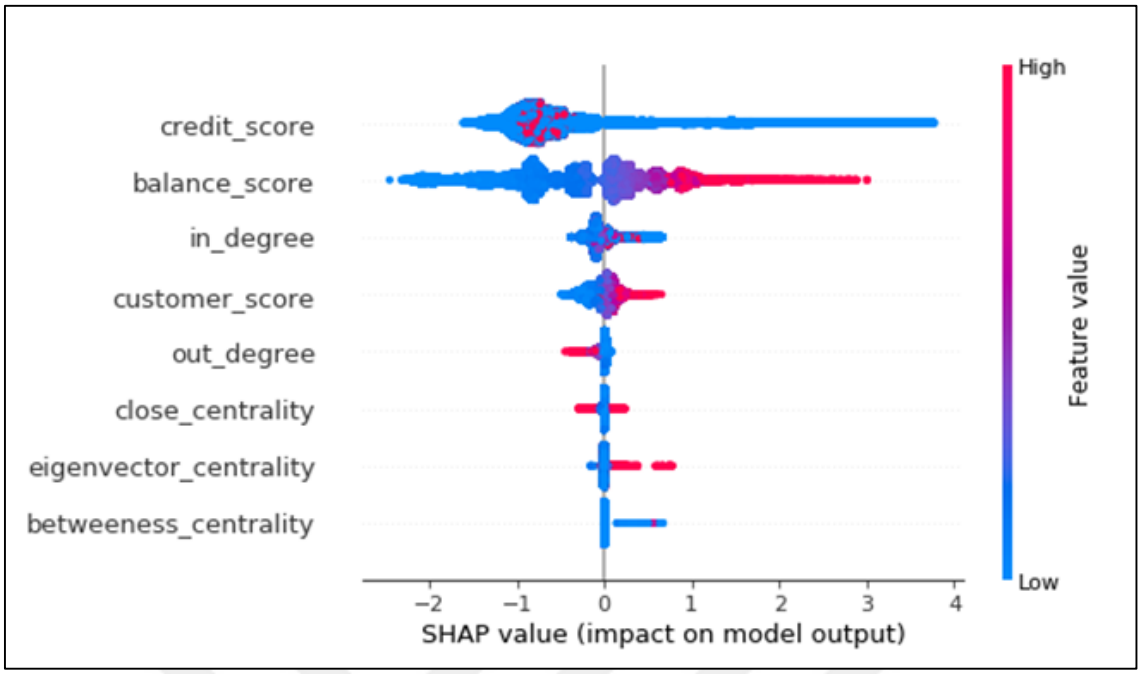
DPModel-1'in ilk alt modeli bilanço veri setiyle gerçekleştirilmiştir. Bu veri seti 61 adet bilanço değişkeni içermektedir. İkinci alt model ise 57 adet kredi değişkeni için oluşturulmuştur. Her iki alt model RF algoritması kullanılarak tekrar çalıştırılmış ve sonuçlar SHAP algoritmasıyla açıklanmıştır. Şekil 4.3'de alt modellerde tahmine en çok katkı sağlayan 10 değişkene ait sonuçlar görülmektedir.



(a) (b)

Şekil 4.3. DPModel-1 bilanço (a) ve kredi (b) alt model SHAP değerleri

Şekil 4.3 (a)'da bilanço veri seti sonuçları, Şekil 4.3 (b)'de ise kredi veri seti sonuçları verilmiştir. En üst satır en yüksek katkıyı sağlayan değişkeni göstermektedir. Başka bir ifadeyle temerrüt tahminine en çok katkı sağlayan değişkendir. Grafikteki her nokta bir öznitelik için bir örneklemi temsil etmektedir. Noktaların rengi özniteliğin küçükten büyüğe değerini göstermektedir. DPModel-1'in alt modelleri için bilanço değişkenleri arasında en yüksek katkıyı "total loans to total assets" değişkeni, kredi'de ise "ttl accrd int amt" değişkeni sağlamaktadır.



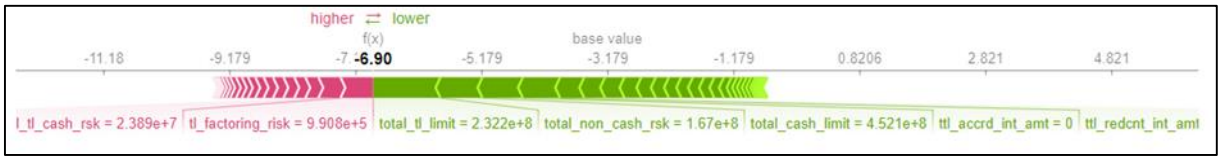
Şekil 4.4. DPModel-2 SHAP değerleri

DPModel-2 için SHAP grafiği Şekil 4.4'de görüldüğü gibidir. Tahmin için en yüksek katkıyı DPModel-1'den gelen kredi ve bilanço skorları sağlamaktadır. Bunun devamında ise müşterilerinin sayısı ve çizge teorisinden hesaplanan ağırlıklandırılmış müşteri skoru sağlamıştır.

4.4.2. Lokal yorumlanabilirlik

Lokal yorumlanabilirlik veri setindeki her bir örneklem için değişkenlerin tahmine olan katkısını göstermektedir. Lokal yorumlanabilirlik her bir model ve her bir örneklem için ayrı olduğundan bu bölümde sadece bir örnek şirket seçilerek SHAP ve LIME algoritmalarının sonuçları grafiklerde verilmiştir. Ancak her iki algoritmada bir kez çalıştırıldığında veri setindeki tüm örneklemler için lokal yorumlanabilirlik sunmaktadır.

Şekil 4.5'de seçilen örnek şirket için SHAP algoritmasının DPModel-1'in kredi veri setiyle oluşturulan alt modeli için sonuçları görülmektedir.

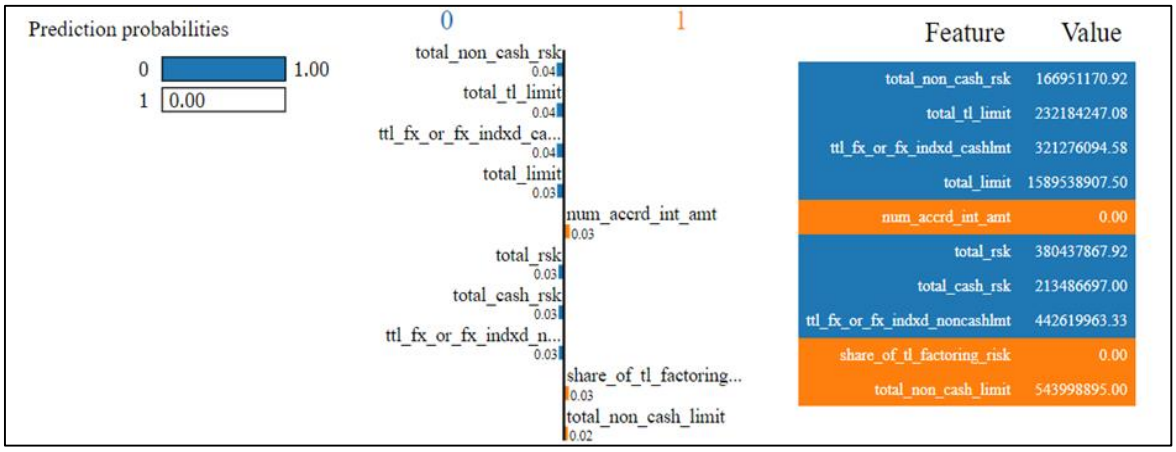


Şekil 4.5. Örnek şirket için SHAP lokal yorumlanabilirlik sonuçları

Seçilen örnek şirketin SHAP değeri -6,9 olarak belirlenmiştir. Tüm şirketlerin ortalama değeri (base value) -3,179 olarak görülmektedir. Örnek şirketin değeri bundan küçük olduğu için şirket temerrüt etmeme ihtimali daha yüksek olduğu anlaşılmaktadır. Şekil 4.5'de şirketin temerrüt etme ihtimalini artıran öznitelikler kırmızıyla, temerrüt etme ihtimalini düşüren öznitelikler yeşille gösterilmiştir. Her iki ihtimali artıran ve düşüren özniteliklerin önem derecesi ise -6,9 olan SHAP değerini oluşturmuştur.

Şekil 4.5'e bakıldığında modelin bu şirket için yaptığı tahminde hangi özniteliklerin daha önemli olduğu ve önem derecesi kolaylıkla görülmektedir. Örneğin bu şirkette "total tl limit" temerrüt etme ihtimalini en çok düşüren ö.znitelikken, "tl factoring risk" bu ihtimali en çok artıran özniteliktir. Bu sonuçlar Şekil 4.3 (b) bölümünde verilen kredi veri setinin global yorumlanabilirliğiyle karşılaştırıldığında farklılık görülmektedir. Global olarak tahmine en çok katkı sağlayan öznitelik "ttl accrd int amt" iken seçilen şirket için lokalde en çok katkı sağlayan öznitelik "total tl limit" olmaktadır. Bu durum SHAP algoritmasının her bir örneklemi lokal olarak değerlendirmesi ve global sonuçları lokallerin birleşiminden elde etmesinden kaynaklanmaktadır. SHAP'in lokal değerlendirme yapabilmesi LR, DT gibi yorumlanabilir algoritmalar içinde kullanılabilir hale getirmektedir. Böylelikle bu algoritmaların sonuçları lokalde yorumlanabilir olmaktadır.

Şekil 4.6 aynı şirket için LIME algoritmasından elde edilen sonuçları göstermektedir. LIME grafiği, SHAP grafiğine benzer şekilde en çok katkı sağlayan öznitelikleri ve bu özniteliklerin tahmine olan katkısının büyüklüğünü göstermektedir.



Şekil 4.6. Örnek şirket için LIME lokal yorumlanabilirlik sonuçları

Aynı şirketin kredi veri setiyle yapılan tahminin SHAP ve LIME sonuçları karşılaştırıldığında benzer özniteliklerin seçildiği görülmektedir. Aynı şekilde öznitelikleri temerrüt etme ihtimalini artırma veya azaltma yönü de benzerlik göstermektedir. Bu durum IML algoritmalarının lokal yorumlanabilirlikte tutarlı davrandığını göstermektedir. Çalışmada ayrıca DPModel-1'in bilanço alt modeli ve DPModel-2 içinde lokal yorumlanabilirliklere bakılmıştır. Bu modeller içinde kredi veri setindeki örnek şirkete benzer sonuçlar alındığı görülmüştür. Her bir şirket ve her bir model için ayrı sonuçlar olduğundan tez kapsamında tüm sonuçlar verilememiştir.

Global ve lokal yorumlanabilirlik sonuçlarına bakıldığında IML algoritmalarının amacı olan RF gibi yorumlanamaz bir algoritmayla geliştirilen DPModel-1 ve DPModel-2'ye, kolay anlaşılabilir ve tutarlı yorumlanabilirlik sağladığı görülmüştür. Böylelikle literatürde ML algoritmalarıyla temerrüt tahmininin en çok eleştirilen eksikliği olan yorumlanabilirliğinin olmaması giderilmiştir. Bu çalışma temerrüt tahmini gibi sadece yüksek tahmin doğruluğunun değil aynı zamanda açıklanabilirliğin önemli olduğu problemlerde oluşan doğruluk ve yorumlanabilirlik arasındaki seçim yapma zorunluluğunu ortadan kaldırılmıştır. Yüksek tahmin başarısı sağlayan karmaşık algoritmaların yorumlanabilirlik eksikliği IML algoritmaları ile giderilebileceği gösterilmiştir.

4.5. IML ile Öznitelik Seçimi ve Yorumlanabilir Algoritmaların Karşılaştırılması

ML tahminlerinin açıklanabilir olması için çok uzun yıllardır öznitelik seçimi ve LR, DT gibi sonuçları yorumlanabilen ML algoritmaları kullanılmaktadır. Son yıllarda ise IML

algoritmaları kara kutu olarak adlandırılan RF, XGBOOST gibi doğruluğu yüksek ancak yorumlanabilir olmayan karmaşık ML algoritmaları için yorumlanabilirlik aramaktadır. IML algoritmaları bunu yaparken öznitelik seçimi veya yorumlanabilir algoritmalar gibi özniteliklerin tahmine olan katkısını ölçmeye çalışmaktadır. Bu bölümde IML algoritmaları ile öznitelik seçimi ve yorumlanabilir algoritmaların seçtiği özniteliklerin benzerliği karşılaştırılmıştır. Böylelikle IML algoritmalarının seçtiği özniteliklerle, öznitelik seçimi ve yorumlanabilir modellerin seçtiği özniteliklerin benzerliklerine bakılarak IML algoritmalarının açıklamadaki başarısı ölçülmüştür.

Çalışma sırasında IML algoritması olarak SHAP ve ELI5, öznitelik algoritmaları olarak Karşılıklı Bilgi (Mutual Information / MI) ve Ardışık İleri Yönde Seçim (Sequential Forward Selection / SFS) ve yorumlanabilir algoritmalar olarak DT ve LASSO kullanılmıştır.

4.5.1. Kullanılan veri seti

Deneysel çalışmalar DPModel-1'de kullanılan bilanço veri seti üzerinden seçilen örneklem veriler üzerinde yapılmıştır. Veri seti 2015-2017 yılları arasında faaliyet gösteren şirketlerden rastgele seçilmiştir. 61 adet değişkene ve 43 318 adet örnekleme sahiptir. Veri setindeki temerrüt oranı %7'dir.

4.5.2. Değerlendirme metriği

Farklı amaçlar için kullanılan birçok benzerlik ölçüm yöntemi vardır. Bizim karşılaştırmalarımız aynı uzunlukta 0 ve 1 içeren listeler olduğu için benzerlik ölçüm metriği olarak Jacard indeks kullanılmıştır. Jacard indeks sonlu kümelerde kümelerin kesişiminin birleşimine bölümü ile hesaplanmaktadır.

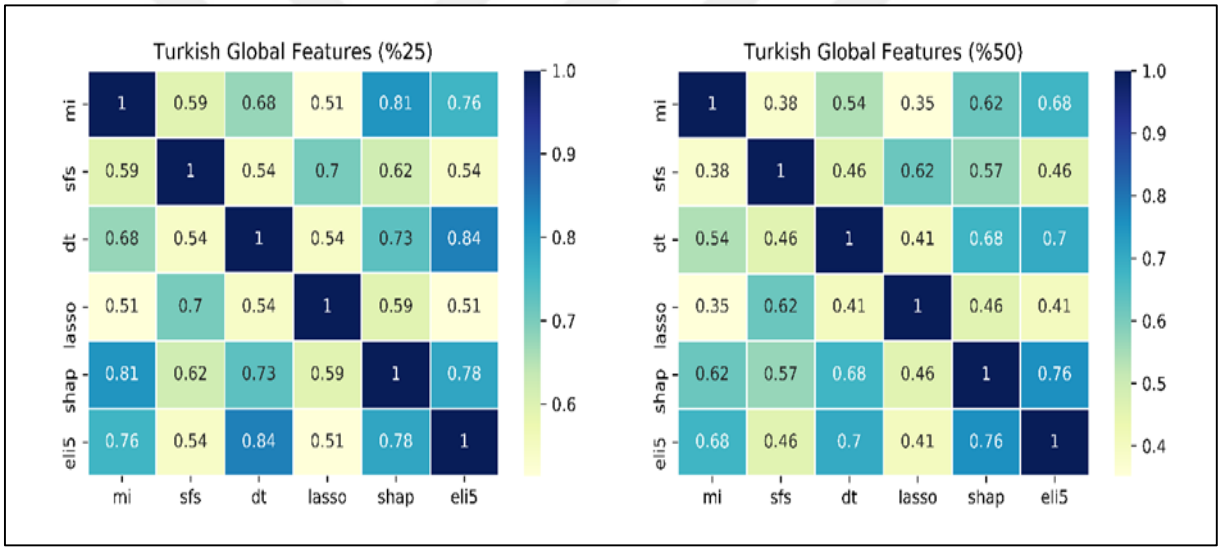
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{4.4}$$

4.5.3. Özniteliklerin benzerliğinin karşılaştırılması

IML algoritmaları ML algoritmalarının sonuçları üzerinde çalışmaktadır. Bu nedenle öncelikle bir ML algoritmasının tahmini gerekmektedir. Bunun için veri setimiz RF

algoritmasıyla modellenmiş ve sonuçları üzerinden IML algoritmaları özniteliklerin önem (feature importance) değerini belirlemiştir. Öznitelik seçimi ve yorumlanabilir modeller doğrudan veri setiyle çalıştırılmış ve bunlar içinde özniteliklerin önem değeri belirlenmiştir. 3 farklı yöntemin her birinden seçilen 2'şer adet algoritmadan elde edilen öznitelik önem değeri bir eşik değer belirlenerek öznitelik seçimi yapılmıştır. Çalışmada 2 farklı eşik değerle öznitelik seçim sonuçları elde edilmiştir. Bunlar öznitelik sayısının %25'i ve %50'si olarak belirlenmiştir.

Şekil 4.7'de sunulan matrislerde 6 algoritmanın seçtiği özniteliklerin %25 ve %50'lik eşik değerlerde benzerlikleri karşılaştırılmıştır. Matrisler, algoritmaların seçtiği özniteliklerin Jacard indeks ölçütüne göre hesaplanmış benzerlik oranlarını göstermektedir.



(a) (b)

Şekil 4.7. %25 (a) ve %50 (b) eşik değerlerde Jacard indeks benzerlikleri

Matrislere bakıldığında IML algoritmalarının seçtiği özniteliklerin, öznitelik seçimi ve yorumlanabilir algoritmaların seçtiği özniteliklerle genel olarak %50'den daha fazla benzerliğinin olduğu görülmektedir. %25 ile %50 eşik değerleri karşılaştırıldığında %25'lik öznitelik seçimi daha yüksek bir benzerlik göstermektedir. Bunun en önemli sebebi seçilen (%25) ve seçilmeyen (%75) özniteliklerin seçilmeyenler yönünde dengesizlik oluşturmasından kaynaklanmaktadır. %50 oranında öznitelik seçimi bu dengesizliği ortadan kaldırmaktadır. %50'lik öznitelik seçim benzerlik sonuçları hala IML algoritmalarının yeterli düzeyde diğer algoritmalara benzerliğini ortaya koymaktadır. Bir diğer önemli sonuç ise SHAP ve ELI5 algoritmalarının benzerliğinin tüm testlerde yüksek olmasıdır. Bu durum

bize IML algoritmalarının her ikisinin de tutarlı sonuçlar ürettiğini kanıtlamaktadır. Sonuçlar, IML algoritmalarının diğerleriyle benzerlik oranlarına bakıldığında, bu algoritmaların karmaşık modellerdeki yorumlanabilirliğin giderilmesinde güvenli şekilde kullanılabileceğini göstermiştir.

5. SONUÇ VE ÖNERİLER

Temerrüt tahmini çok uzun yıllardır çalışılan ve gelecekte çalışılmaya devam edecek önemli bir gerçek dünya problemdir. Birçok tahmin probleminde olduğu gibi temerrüt tahmininde de ML algoritmaları yüksek başarı elde etmektedir. Bu tez çalışmasında temerrüt tahmini için BVP ve ML algoritmaları kullanılarak, ülkemizde faaliyet gösteren 1 milyondan fazla şirketin bir sonraki yılda temerrüde düşme tahmini yapılmıştır.

Çizelge 5.1. Önerilen modellerdeki en yüksek başarılar

	DPModel-1		DPModel-2	
	AUC	Doğruluk	AUC	Doğruluk
Ortalama Başarı	0,81	0,84	0,82	0,92
En Yüksek Başarı (2017 yılı)	0,87	0,92	0,89	0,94

Tezin ilk aşamasında şirketlere ait kredi ve bilanço veri setleri üzerinde LR, DT, RF ve GB algoritmaları kullanılarak tahminler yapılmıştır. Her iki veri setiyle ayrı alt modeller oluşturulmuş, nihai tahmin için bu alt modeller birleştirilmiştir. Şirketlerin yaklaşık %6'sı temerrüt ettiği için veri seti dengesizdir. Bu nedenle performans değerlendirmelerinde AUC metriği kullanılmıştır.

Çizelge 5.1'de önerilen modellerden elde edilen ortalama başarı oranı ve yıllar içindeki en yüksek başarı gösterilmiştir. DPModel-1'den elde edilen ortalama model başarıları arasındaki en yüksek başarıyı 0,81 AUC ve 0,84 doğruluk değeriyle RF algoritması göstermektedir. Yıllar içindeki en yüksek başarıyı ise 2017 yılında 0,87 AUC ve 0,92 doğruluk ile yine RF algoritması gerçekleştirmiştir. Kullanılan algoritmalar arasındaki en düşük başarıyı ise LR algoritması göstermektedir. Yıllar itibariyle bakıldığında giderek artan bir başarı görülmektedir. Bu durum veri kalitesiyle ilişkilidir. Şirketler son yıllarda muhasebe ve bilançoları için daha çok otomasyon kullanmaya başladıklarından veri kalitesi artmıştır. Ayrıca e-fatura, e-beyanname gibi yeni uygulamalar veri kalitesini artıran bir diğer önemli nedendir. Algoritmaların başarıları genel olarak değerlendirildiğinde benzeri birçok çalışmaya göre yüksektir. Bunun birkaç sebebi vardır. (i) BVP platformunun sağladığı yüksek performans sayesinde her bir algoritma için parametre optimizasyonu yapılmıştır. (ii) BVP ile modellerde k-fold CV kullanarak sonuçların ortalaması alınmıştır. (iii)

Metriklerin hesaplanması sırasında varsayılan eşik değer yerine f1 skoru maksimum yapan eşik değer kullanılmıştır.

Örneklem veri seti üzerinde yaygın olarak kullanılan DL algoritmalarından olan LSTM, RNN ve GRU algoritmalarıyla yapılan deneysel çalışmalarda başarının yüksek olduğu görülmektedir. Başarı oranı birbirine yakın olmakla birlikte LSTM algoritması en yüksek başarıyı göstermiştir. Bu ek çalışma DL algoritmalarının temerrüt tahminine önemli katkılar sağlayacağını göstermiştir.

Tezin ikinci aşamasında fatura veri seti üzerinde oluşturulan çizge yapısı analiz edilerek temerrüt tahmini için yeni değişkenler üretilmiştir. Düğümlerin şirketleri, kenarların ise şirketlerin arasındaki alım satım ilişkisini temsil ettiği fatura miktarıyla ağırlıklandırılmış yönlü çizge oluşturulmuştur. Bu çizge üzerinden ağırlıklandırılmış müşteri skoru ve merkezilik değişkenleri üretilmiş ve temerrüt tahmini için kullanılmıştır. Tahminde DPModel-1'de kullanılan algoritmaların aynısı kullanılmıştır. Sonuçlara bakıldığında tüm algoritmaların başarısı yeni değişkenlerle arttığı gözlemlenmiştir. DPModel-2'den elde edilen ortalama model başarıları arasındaki en yüksek başarıyı 0,82 AUC ve 0,92 doğruluk değeriyle RF algoritması göstermektedir. Yıllar içindeki en yüksek başarıyı ise 2017 yılında 0,89 AUC ve 0,94 doğruluk ile yine RF algoritması gerçekleştirmiştir. Bu başarılar bir şirketin temerrüde düşmede iş ilişkisi içinde oldukları diğer şirketlerinde önemli olduğunu göstermiştir.

DPModel-2 sonuçlarına bakıldığında DPModel-1'e göre AUC skorda önemli bir artış gözlenmektedir. Bu artış çizge teorisinden elde edilen ağırlıklandırılmış müşteri skoru ve merkezilik değişkenlerinden kaynaklanmaktadır. Çizge teorisinden elde edilen ağırlıklandırılmış müşteri skoru, şirketin temerrüde düşmesinde müşterisinin temerrüt olasılığı ve ticari ilişkisinin büyüklüğünün etkili olduğunu ortaya koymaktadır. Başka bir ifadeyle temerrüde düşen veya düşme ihtimali yüksek olan müşterileri fazla olan şirketin temerrüde düşme ihtimali artmakta, tam tersi durumda ise temerrüde düşme ihtimali azalmaktadır. Benzer şekilde merkezilik değişkenleri de şirketin alım satım ilişkisi içinde olduğu tedarikçi ve müşterilerinin sayısı, bu şirketlerin çizge içindeki önemi gibi değişkenler temerrüt tahminine olumlu katkı sağlamaktadır. Önerilen değişkenlerin temerrüt tahminine olan yıllara göre katkısı AUC skorda %2-6 artış olarak ölçülmüştür. Bu artış çalışılan şirket sayısı düşünüldüğünde çok büyük miktarda şirketin yanlış sınıflandırmasını önlemektedir.

Tezin son kısmında ise ML ve çizge teorisiyle önerilen temerrüt tahmini modellerinin IML algoritmalarıyla global ve lokal yorumlanabilirliği araştırılmıştır. Böylelikle literatürde ML, DL gibi karmaşık algoritmaların temerrüt tahminine getirilen en önemli eleştiri olan sonuçların yorumlanabilir olmaması için çözüm sunulmuştur. Çalışmada SHAP ve LIME olmak üzere iki farklı IML algoritması kullanılmıştır. Kullanılan algoritmaların henüz BVP üzerinde dağıtık olarak çalışan bir versiyonu olmadığı için çalışmalar lokal makine üzerinde yapılmıştır. Lokal makinenin kapasitesi tüm veri setiyle çalışmak için yetersiz olduğundan deneysel çalışma örneklem üzerinde yapılmıştır. Her iki model için global yorumlanabilirlik SHAP algoritmasıyla ölçülmüş ve global olarak tahmine en çok katkı sağlayan değişkenler bulunmuştur. Ayrıca lokal yorumlanabilirlik ile şirket bazlı olarak önemli değişkenler belirlenmiştir. Lokal yorumlanabilirlik için SHAP ve LIME algoritmaları kullanılmıştır. Örnek şirketler için sonuçlar paylaşılmış ve her iki algoritmanın sonuçlarının birbirine benzediği görülmüştür.

Son bölümde yapılan bir diğer çalışma ise IML algoritmalarının güvenilirliği üzerinedir. Çalışmada IML algoritmalarının seçtiği değişkenlerle, öznitelik seçimi ve yorumlanabilir algoritmaların seçtiği değişkenlerin benzerliği karşılaştırılmıştır. IML algoritması olarak SHAP, LIME ve ELI5, öznitelik seçim algoritması olarak MI ve SFS, yorumlanabilir algoritma olarak ise LASSO ve DT kullanılmıştır. DPModel-1'de kullanılan bilanço veri setinden seçilen örneklemle yapılan çalışmada 61 değişken kullanılmıştır. Tüm algoritmalarla değişken sayısının %25 ve %50 oranında değişken seçimi yapılmıştır. Her bir algoritmanın benzerlikleri Jacard indeks benzerlik metriğiyle karşılaştırılmıştır. Sonuçlara bakıldığında IML algoritmalarının, öznitelik seçimi ve yorumlanabilir algoritmalarla benzer ve tutarlı şekilde çalıştığı görülmektedir. Bu durum IML algoritmalarının sonuçlarına olan güvenilirliği artırmaktadır.

Tez kapsamında BVP üzerinde yapılan temerrüt tahmini çalışması genel olarak küçük değişikliklerle iflas tahmini ya da kredi derecelendirme gibi birçok çalışmaya uyarlanabilir. Temerrüt tahmini için gelecekte yapılabilecek çalışmalar şu şekilde sıralanabilir:

- Bu tez kapsamında temerrüt tahmini yapılırken sektör veya ölçek ayrımı yapılmamıştır. Ancak özellikle bilançolar sektörel olarak farklılıklar göstermektedir. Çalışma, sektör veya ölçek ayrımında yapılarak geliştirilmeye açıktır.

- Çalışmada fatura veri setinde oluşturulan çizge yapısından elde edilen bilgi miktarını artırabilmek için farklı değişkenler elde edilebilir. Örneğin toplam müşteri sayısı, toplam tedarikçi sayısı, ithalat veya ihracat yapma durumu gibi çizge üzerinde elde edilecek farklı değişkenlerin tahmine katkısı ölçülebilir.
- Çizge yapısı kullanılarak temerrüt eden şirketlerin bağlı oldukları şirketlere verdikleri hasar ve temerrüdün yayılma hızı ölçülebilir.
- Geleneksel veri setlerinin dışında haber siteleri ya da sosyal medya gibi kaynaklardan temerrüt tahmininin başarısını artıracak yeni veri kaynaklarıyla çalışılabilir.
- Şirketlerin bağlı ortaklık yapılarının bilinerek grup şirketlerinin durumunun özel olarak değerlendirilmesi ile temerrüt tahmininin geliştirilebileceği öngörülmektedir.
- Bu çalışmada örneklem veri seti ile yapılan DL çalışmasının da gösterdiği gibi temerrüt tahmininde DL'nin başarıyı artıracağı tahmin edilmektedir.
- IML algoritmalarıyla yapılan çalışmanın BVP'de çalışabilen kütüphaneleri geliştirildiğinde tüm veriyle yapılması öngörülmektedir.

KAYNAKLAR

1. Beaver, W.H. (1966). Financial Ratios As Predictors of Failure. *Journal of Accounting Research*, 4, 71-111.

2. Altman, E.I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, 23(4), 589.

3. Ohlson, J.A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, 18(1), 109.

4. Fu, X., Ouyang, T., Chen, J.and Luo, X. (2020). Listening to the investors: A novel framework for online lending default prediction using deep learning neural networks. *Information Processing & Management*, 57(4), 102236.

5. Stevenson, M., Mues, C.and Bravo, C. (2021). The value of text for small business default prediction: A Deep Learning approach. *European Journal of Operational Research*, 295(2), 758-771.

6. He, H.and Fan, Y. (2021). A novel hybrid ensemble model based on tree-based method and deep learning method for default prediction. *Expert Systems with Applications*, 176, 114899.

7. Sigrist, F.and Hirnschall, C. (2019). Grabit: Gradient tree-boosted Tobit models for default prediction. *Journal of Banking & Finance*, 102, 177-192.

8. Zhu, L., Qiu, D., Ergu, D., Ying, C.and Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. *Procedia Computer Science*, 162, 503-513.

9. De Castro Vieira, J.R., Barboza, F., Sobreiro, V.A.and Kimura, H. (2019). Machine learning models for credit analysis improvements: Predicting low-income families' default. *Applied Soft Computing*, 83, 105640.

10. Wang, Y., Zhang, Y., Lu, Y.and Yu, X. (2020). A Comparative Assessment of Credit Risk Model Based on Machine Learning: A case study of bank loan data. *Procedia Computer Science*, 174, 141-149.

11. Li, J.-P., Mirza, N., Rahat, B.and Xiong, D. (2020). Machine learning and credit ratings prediction in the age of fourth industrial revolution. *Technological Forecasting and Social Change*, 161, 120309.

12. Antulov-Fantulin, N., Lagravinese, R.and Resce, G. (2021). Predicting bankruptcy of local government: A machine learning approach. *Journal of Economic Behavior & Organization*, 183, 681-699.

13. Chaudhuri, A.and De, K. (2011). Fuzzy Support Vector Machine for bankruptcy prediction. *Applied Soft Computing*, 11(2), 2472-2486.

14. Kim, H., Cho, H.and Ryu, D. (2020). Corporate Default Predictions Using Machine Learning: Literature Review. *Sustainability*, 12(16), 6325.

15. Kim, H.S.and Sohn, S.Y. (2010). Support vector machines for default prediction of SMEs based on technology credit. *European Journal of Operational Research*, 201(3), 838-846.

16. Zhou, L., Lai, K.K.and Yen, J. (2014). Bankruptcy prediction using SVM models with a new approach to combine features selection and parameter optimisation. International *Journal of Systems Science*, 45(3), 241-253.

17. Wang, G., Ma, J.and Yang, S. (2014). An improved boosting based on feature selection for corporate bankruptcy prediction. *Expert Systems with Applications*, 41(5), 2353-2361.

18. Tsai, C.-F., Hsu, Y.-F.and Yen, D.C. (2014). A comparative study of classifier ensembles for bankruptcy prediction. *Applied Soft Computing*, 24, 977-984.

19. Barboza, F., Kimura, H.and Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 83, 405-417.

20. Nehrebecka, N. (2018). Predicting the default risk of companies, comparison of credit scoring models: Logit vs support vector machines. *Econometrics*, 22(2), 54-73.

21. Hosaka, T. (2019). Bankruptcy prediction using imaged financial ratios and convolutional neural networks. *Expert Systems with Applications*, 117, 287-299.

22. Jing, J., Yan, W. and Deng, X. (2021). A hybrid model to estimate corporate default probabilities in China based on zero-price probability model and long short-term memory. *Applied Economics Letters*, 28(5), 413-420.

23. Yeh, S. J., Chen, H. C., Lu, C. W., Wang, J. K., Huang, L. M., Huang, S. C., Wu, M. H. (2015). National database study of survival of pediatric congenital heart disease patients in Taiwan. *Journal of the Formosan Medical Association*, 114(2), 159-163.

24. Luo, C., Wu, D.and Wu, D. (2017). A deep learning approach for credit scoring using credit default swaps. *Engineering Applications of Artificial Intelligence*, 65, 465-470.

25. Addo, P., Guegan, D.and Hassani, B. (2018). Credit Risk Analysis Using Machine and Deep Learning Models. *Risks*, 6(2), 38.

26. Narayanan, V. (2014). Using big-data analytics to manage data deluge and unlock real-time business insights. *The Journal of Equipment Lease Financing (Online)*, 32(2), 1.

27. Kolanovic, M., Krishnamachari, R.T. (2017). Big Data and AI Strategies. *JP Morgan Report*, Newyork. 1-278.

28. Dulger, U. (2015). *Stratejik Büyük Veri Yönetiminin Yatırımlar Üzerindeki Etkileri*. Yüksek Lisans Tezi, İstanbul Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, 1-153.

29. Khemais, Z., Nesrine, D.and Mohamed, M. (2016). Credit Scoring and Default Risk Prediction: A Comparative Study between Discriminant Analysis & Logistic Regression. *International Journal of Economics and Finance*, 8(4), 39.

30. Midi, H., Sarkar, S.K.and Rana, S. (2010). Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics*, 13(3), 253-267.

31. Raileanu, L.E. and Stoffel, K. (2004). Theoretical Comparison between the Gini Index and Information Gain Criteria. *Annals of Mathematics and Artificial Intelligence*, 41(1), 77-93.

32. Bhardwaj, R. and Vatta, S. (2013). Implementation of ID3 algorithm. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(6), 845-851.

33. Svetnik, V., Liaw, A., Tong, C., Christopher Culberson, J., Sheridan, R.P.and Feuston, B.P. (2003). Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences*, 43(6), 1947-1958.

34. Ali, J., Khan, R., Ahmad, N.and Maqsood, I. (2012). Random forests and decision trees. *International Journal of Computer Science Issues*, 9(5), 272-278.

35. Friedman, J.H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4), 367-378.

36. Rodano, G., Serrano-Velarde, N.and Tarantino, E. (2016). Bankruptcy law and bank financing. *Journal of Financial Economics*, 120(2), 363-382.

37. Câmara, A., Popova, I.and Simkins, B. (2012). A comparative study of the probability of default for global financial firms. *Journal of Banking & Finance*, 36(3), 717-732.

38. Fuller, K.P., Yildiz, S.and Uymaz, Y. (2018). Credit default swaps and firms' financing policies. *Journal of Corporate Finance*, 48, 34-48.

39. Li, C., Lou, C., Luo, D.and Xing, K. (2021). Chinese corporate distress prediction using LASSO: The role of earnings management. *International Review of Financial Analysis*, 76, 101776.

40. Dumitrescu, E., Hué, S., Hurlin, C.and Tokpavi, S. (2021). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, 296(3), 16218.

41. Dastile, X., Celik, T.and Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, 91, 106263.

42. Ahn, J.J., Byun, H.W., Oh, K.J.and Kim, T.Y. (2012). Using ridge regression with genetic algorithm to enhance real estate appraisal forecasting. *Expert Systems with Applications*, 39(9), 8369-8379.

43. Sohn, S.Y.and Kim, H.S. (2007). Random effects logistic regression model for default prediction of technology credit guarantee fund. *European Journal of Operational Research*, 183(1), 472-478.

44. Olson, L.M., Qi, M., Zhang, X. and Zhao, X. (2021). Machine learning loss given default for corporate debt. *Journal of Empirical Finance*, 64, 144-159.

45. Moscatelli, M., Parlapiano, F., Narizzano, S.and Viggiano, G. (2020). Corporate default forecasting with machine learning. *Expert Systems with Applications*, 161, 113567.

46. Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q.and Niu, X. (2018). Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research and Applications*, 31, 24-39.

47. Liu, W., Fan, H.and Xia, M. (2021). Step-wise multi-grained augmented gradient boosting decision trees for credit scoring. *Engineering Applications of Artificial Intelligence*, 97, 104036.

48. Zhou, L., Fujita, H., Ding, H.and Ma, R. (2021). Credit risk modeling on data with two timestamps in peer-to-peer lending by gradient boosting. *Applied Soft Computing*, 110, 107672.

49. Lee, J.W., Lee, W.K.and Sohn, S.Y. (2021). Graph convolutional network-based credit default prediction utilizing three types of virtual distances among borrowers. *Expert Systems with Applications*, 168, 114411.

50. Duan, J. (2019). Financial system modeling using deep neural networks (DNNs) for effective risk assessment and prediction. *Journal of the Franklin Institute*, 356(8), 4716-4731.

51. Xu, L., Cui, L., Weise, T., Li, X., Wu, Z., Nie, F., . . . Tang, Y. (2021). Semi-supervised multi-Layer convolution kernel learning in credit evaluation. *Pattern Recognition*, 120, 108125.

52. Fischer, T.and Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2), 654-669.

53. Park, S.and Yang, J.-S. (2021). Relationships between capital flow and economic growth: A network analysis. *Journal of International Financial Markets, Institutions and Money*, 72, 101345.

54. Téllez-León, I.-E., Martínez-Jaramillo, S., O. L. Escobar-Farfán, L.and Hochreiter, R. (2021). How are network centrality metrics related to interest rates in the Mexican secured and unsecured interbank markets? *Journal of Financial Stability*, 55, 100893.

55. Temizsoy, A., Iori, G.and Montes-Rojas, G. (2017). Network centrality and funding rates in the e-MID interbank market. *Journal of Financial Stability*, 33, 346-365.

56. Bhattacharya, M., Inekwe, J.N.and Valenzuela, M.R. (2020). Credit risk and financial integration: An application of network analysis. *International Review of Financial Analysis*, 72, 101588.

57. Gofman, M.and Wu, Y. (2021). Trade credit and profitability in production networks. *Journal of Financial Economics*, 142(1), 1-26.

58. Kuzubaş, T.U., Ömercikoğlu, I.and Saltoğlu, B. (2014). Network centrality measures and systemic risk: An application to the Turkish financial crisis. *Physica A: Statistical Mechanics and its Applications*, 405, 203-215.

59. Tsai, C.-F.and Wu, J.-W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, 34(4), 2639-2649.

60. El Shawi, R., Sherif, Y., Al-Mallah, M.and Sakr, S. (2019). Interpretability in HealthCare A Comparative Study of Local Machine Learning Interpretability Techniques. *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, 275-280

61. İnternet: Grath, R.M., Costabello, L., Van, C.L., Sweeney, P., Kamiab, F., Shen, Z.and Lécué, F. (2018). Interpretable Credit Application Predictions With Counterfactual Explanations. URL: https://arxiv.org/abs/1811.05245, Son Erişim Tarihi:01.09.2021.

62. İnternet: Zablocki, É., Ben-younes, H., Pérez, P. and Cord, M. (2021). Explainability of vision-based autonomous driving systems: Review and challenges. URL: https://arxiv.org/abs/2101.05307, Son Erişim Tarihi:01.09.2021.

63. İnternet: Demajo, L.M., Vella, V.and Dingli, A. (2020). Explainable AI for Interpretable Credit Scoring. URL: https://arxiv.org/abs/2012.03749, Son Erişim Tarihi:01.09.2021.

64. Bussmann, N., Giudici, P., Marinelli, D. and Papenbrock, J. (2020). Explainable AI in Fintech Risk Management. *Frontiers in Artificial Intelligence*, 3-26.

65. Futagami, K., Fukazawa, Y., Kapoor, N. and Kito, T. (2021). Pairwise acquisition prediction with SHAP value interpretation. *The Journal of Finance and Data Science*, 7, 22-44.

66. İnternet: Fan, A., Jernite, Y., Perez, E., Grangier, D., Weston, J. and Auli, M. (2019). ELI5: Long Form Question Answering. URL: https://arxiv.org/abs/1907.09190, Son Erişim Tarihi:01.09.2021.

67. Ribeiro, M.T., Singh, S. and Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144.

68. Lundberg, S.M.and Lee, S.I. (2017). A Unified Approach to Interpreting Model Predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768-4777.

GAZİ GELECEKTİR...