T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

TÜRKÇE METİN VERİLERİNDE DOĞAL DİL İŞLEME VE DERİN ÖĞRENME YÖNTEMLERİ İLE KREDİ RİSKİ TAHMİNİ

Emre Rıdvan MURATLAR

DOKTORA TEZİ

İstatistik Anabilim Dalı

İstatistik Programı

Danışman

Dr. Öğr. Üyesi Doğan YILDIZ

Mart, 2026

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

TÜRKÇE METİN VERİLERİNDE DOĞAL DİL İŞLEME VE DERİN ÖĞRENME YÖNTEMLERİ İLE KREDİ RİSKİ TAHMİNİ

Emre Rıdvan MURATLAR tarafından hazırlanan tez çalışması 23/03/2026 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı, İstatistik Programı **DOKTORA TEZİ** olarak kabul edilmiştir.

Dr. Öğr. Üyesi Doğan YILDIZ
Yıldız Teknik Üniversitesi
Danışman

Jüri Üyeleri

Dr. Öğr. Üyesi Doğan YILDIZ, Danışman

Yıldız Teknik Üniversitesi _____

Prof. Dr. Ali Hakan BÜYÜKLÜ, Üye
Yıldız Teknik Üniversitesi _____

Dr. Öğr. Üyesi Erhan USTAOĞLU, Üye
Marmara Üniversitesi _____

Doç. Dr. Atıf Ahmet EVREN, Üye
Yıldız Teknik Üniversitesi _____

Prof. Dr. Ali Karun NEMLİÖĞLU, Üye
İstanbul Üniversitesi _____

Danışmanım Dr. Öğr. Üyesi Doğan YILDIZ sorumluluğunda tarafımca hazırlanan "Türkçe Metin Verilerinde Doğal Dil İşleme ve Derin Öğrenme Yöntemleri ile Kredi Riski Tahmini" başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim.

Bu tez çalışması sürecinde, tez sahibinin kontrolü altında dil ve ifade bütünlüğünü güçlendirmek, teknik süreçleri desteklemek, literatür taraması, kodlama ve veri görselleştirme amacıyla Üretken Yapay Zekâ (ÜYZ) araçlarından ChatGPT 5.2 kullanılmıştır.

ÜYZ araçlarının yanlış veya gerçek dışı bilgi (halüsinasyon) üretebilme riski taşıdığının farkında olduğumu, bu araçlardan elde edilen tüm çıktıları doğruluk, bütünlük, güvenilirlik ve etik standartlar açısından dikkatle kontrol ettiğimi beyan ederim.

ÜYZ kullanımı hiçbir aşamada yazarın entelektüel katkısının, bilimsel muhakemesinin ve özgün fikir üretiminin yerine geçmemiştir. Tez çalışmasında yer alan tüm bilgi, veri, analiz, iddia ve referansların doğruluğu, güncelliği ve bilimsel geçerliliğine ilişkin sorumluluk tamamen tarafıma aittir.

Beyanımın aksinin tespiti hâlinde doğabilecek her türlü idari ve hukuki sorumluluğu kabul ederim.

Emre Rıdvan MURATLAR

İmza

Aileme

ve

eşime

TEŞEKKÜR

Tez çalışmamda planlanmasında, araştırılmasında, yürütülmesinde ve oluşumunda ilgi ve desteğini esirgemeyen, engin bilgi ve tecrübelerinden yararlandığım, yönlendirme ve bilgilendirmeleriyle çalışmamı bilimsel temeller ışığında şekillendiren sayın hocam Dr. Öğr. Üyesi Doğan YILDIZ 'a teşekkürü bir borç bilirim ve sonsuz teşekkürlerimi sunarım.

Emre Rıdvan Muratlar

İÇİNDEKİLER

SİMGE LİSTESİ vi
KISALTMA LİSTESİ vii
ŞEKİL LİSTESİ viii
TABLO LİSTESİ ix
ÖZET x
ABSTRACT xii
1 GİRİŞ 1
1.1 Problem Tanımı 1
1.2 Kredi Riskinde Alternatif Veri İhtiyacı 3
1.3 Çağrı Merkezi Konuşmaları ve Metin Verileri 4
2 LİTERATÜR TARAMASI 6
2.1 Kredi Riski Modellemesi ve Temerrüt Tahmini 6
2.2 Alternatif Verilerin Kredi Riskinde Kullanımı 9
2.3 Metin Madenciliği ve Doğal Dil İşleme Yaklaşımları 10
2.4 Konuşmadan Metne Dönüştürme Teknolojileri 12
2.5 Metin Temsil Yöntemleri 14
2.6 Literatürdeki Boşluklar ve Çalışmanın Konumlandırılması 16
3 VERİ VE PROBLEM TANIMI 19
3.1 Problem Tanımı 19
3.2 Hedef Değişken Tanımı 24
3.3 Veri Ön İşleme ve Birleştirme 25
4 METİN TEMSİL YÖNTEMLERİ VE MODELLEME ÇERÇEVESİ 31
4.1 Genel Modelleme Yaklaşımı 31
4.2 Metin Temsil Yaklaşımı 32
4.3 Modelleme Çerçevesi 34
4.4 Performans Değerlendirme Kriterleri 36
5 SONUÇ 38
KAYNAKÇA 44
TEZDEN ÜRETİLMİŞ YAYINLAR 49

SİMGE LİSTESİ

C_{t-1}	Bir Önceki Hücre
O_t	Çıkış Kapısı (Output Gate)
i_t	Giriş Kapısı (Input Gate)
H_t	Gizli Durum (Hidden State)
C_t	Güncel Hücre
σ	Sigmoid Aktivasyon Fonksiyonu
f_t	Unutma Kapısı (Forget Gate)

KISALTMA LİSTESİ

ASR	Automatic Speech Recognition
BoW	Bag of Words
BERT	Bidirectional Encoder Representations from Transformers
EBA	European Banking Authority
FN	False Negative
FP	False Positive
LSTM	Long Short-Term Memory
NLP	Natural Language Processing
P2P	Peer to Peer
PR	Precision
PD	Probability of Default
RE	Recall
RNN	Recurrent Neural Network
STT	Speech to Text
TF–IDF	Term Frequency–Inverse Document Frequency
TP	True Positive
WER	Word Error Rate

ŞEKİL LİSTESİ

Şekil 3.1 Veri toplama aşamasından model performans ölçümü aşamasına kadarki sürecin görseli…..……... **40**

Şekil 3.2 Manuel veri düzeltme kod örneği……..…….. **42**

Şekil 4.1 Anlamsal ilişkileri gösteren kelime gömme örnekleri…….……... **47**

Şekil 4.2 LSTM algoritması görseli…….…….. **50**

Şekil 5.1 Eğitim veri kümesi üzerinde Self Embedding yöntemiyle eğitilmiş LSTM modelinin ROC eğrisi…… **54**

Şekil 5.2 Test veri kümesi üzerinde Self Embedding yöntemiyle eğitilmiş LSTM modelinin ROC eğrisi…… **54**

Şekil 5.3 Eğitim veri kümesi üzerinde FastText yöntemiyle eğitilmiş LSTM modelinin ROC eğrisi…… **55**

Şekil 5.4 Test veri kümesi üzerinde FastText yöntemiyle eğitilmiş LSTM modelinin ROC eğrisi…….…… **55**

TABLO LİSTESİ

Tablo 3.1 Anomimleştirilmiş örnek bir görüşmenin ham veri temsili……... **36**

Tablo 3.2 Verinin betimleyici istatistikleri…… **37**

Tablo 3.3 Bigram örnekleri inceleme tablosu…… **42**

Tablo 3.4 Veri temizleme ve gruplandırma işlemlerinden sonra modelleme için hazırlanan örnek veri…… **43**

Tablo 5.1 LSTM modelinin yapılandırması ve eğitim ayarları…….. **52**

Tablo 5.2 Test verilerinin performansını gösteren karışıklık matrisi……. **53**

Tablo 5.3 Eğitim ve test verilerini karşılaştırmak için performans tablosu…….... **53**

ÖZET

Türkçe Metin Verilerinde Doğal Dil İşleme ve Derin Öğrenme Yöntemleri ile Kredi Riski Tahmini

Emre Rıdvan MURATLAR

İstatistik Anabilim Dalı
Doktora Tezi

Danışman: Dr. Öğr. Üyesi Doğan YILDIZ

Bu tez çalışması, bankacılık sektöründe kredi riskinin yönetimine katkı sağlamak amacıyla, Türkçe çağrı merkezi telefon görüşmelerinden elde edilen metin verilerinin analiz edilmesini hedeflemektedir. Çalışmada, müşterilerin verdikleri ödeme taahhütlerini yerine getirip getirmeyeceklerinin tahmin edilmesi amaçlanmış, bu doğrultuda çağrı merkezi konuşmaları, yapılandırılmamış bir veri kaynağı olarak ele alınmıştır.

Analiz sürecinde, konuşmadan metne dönüştürme (speech-to-text) teknolojileri kullanılarak metne dönüştürülen konuşmalar üzerinde kapsamlı bir veri temizliği ve ön işleme süreci uygulanmıştır. Metinlerin sayısal temsile dönüştürülmesinde doğal dil işleme (NLP) tekniklerinden yararlanılmış, modelleme aşamasında 3 farklı metin gömme (embedding) yöntemi ile iki katmanlı Long Short-Term Memory (LSTM) mimarisi kullanılmıştır.

Elde edilen bulgular, ödeme taahhütlerini yerine getirmeyen müşterilerin konuşmalarında finansal zorluklara, sağlık veya ailevi nedenlere ilişkin ifadelerin daha sık yer aldığını, ek olarak güven tesis etmeye yönelik belirli söylem

kalıplarının öne çıktığını göstermektedir. Bu sonuçlar, çağrı merkezi konuşmalarından elde edilen metin verilerinin, müşterilerin ödeme davranışlarına ilişkin sinyaller sunabildiğine işaret etmektedir.

Çalışma, Türkçe çağrı merkezi verileri üzerinde farklı metin temsil yaklaşımlarının uygulanabilirliğini ortaya koymakta ve yapılandırılmamış metin verilerinin kredi riski tahmin modellerine entegre edilmesinin potansiyel faydalarını göstermektedir. Ayrıca geliştirilen yaklaşımın, kredi riski tahmin doğruluğunu artırmanın yanı sıra, çağrı merkezi süreçlerinin daha etkin yönetilmesine katkı sağlayabileceği değerlendirilmektedir. Gelecek çalışmalarda, veri temizleme süreçlerinin otomatikleştirilmesi ve konuşmadan metne dönüştürme teknolojilerinin geliştirilmesiyle model performansının iyileştirilmesi hedeflenebilir.

Anahtar Kelimeler: kredi riski, LSTM, veri temizliği, doğal dil işleme, Türkçe metin analizi, çağrı merkezi, müşteri davranışları

ABSTRACT

Credit Risk Estimation with Turkish Text Data Using Natural Language Processing and Deep Learning Methods

Emre Rıdvan MURATLAR

Department of Statistics
Doctor of Philosophy Thesis

Supervisor: Assist. Prof. Dr. Doğan YILDIZ

This thesis aims to analyze text data obtained from Turkish call center telephone conversations in order to contribute to the management of credit risk in the banking sector. The study aims to predict whether customers will fulfill their payment commitments, and call center conversations are treated as an unstructured data source. During the analysis process, speech-to-text technologies were used to convert the conversations into text, and a comprehensive data cleaning and preprocessing process was applied. Natural language processing (NLP) techniques were used to convert the texts into numerical representations, and a two-layered Long Short-Term Memory (LSTM) architecture with 3 different embedding methods was used in the modeling phase.

The findings show that customers who do not fulfill their payment commitments more frequently mention financial difficulties, health, or family reasons in their conversations, and that certain discourse patterns aimed at building trust are prominent. These results indicate that text data obtained from call center conversations can provide signals regarding customers' payment behavior. This study demonstrates the applicability of different text representation approaches to

Turkish call center data and shows the potential benefits of integrating unstructured text data into credit risk prediction models. Furthermore, it is assessed that the developed approach can contribute to more efficient management of call center processes, as well as increasing the accuracy of credit risk prediction. Future studies could aim to improve model performance by automating data cleaning processes and developing speech-to-text conversion technologies.

Keywords: credit risk, LSTM, data cleaning, natural language processing, Turkish text analysis, call center, customer behavior

1
GİRİŞ

1.1 Problem Tanımı

Bankacılık sektöründe, müşterilere verilen kredilerin zamanında geri ödenmesi, finansal risklerin yönetimi açısından kritik öneme sahiptir. Kredilerin geri ödeme olasılığı, bankaların likiditesini ve genel finansal sağlığını doğrudan etkilemektedir. Özellikle bireysel müşteri sayısının artması, her müşterinin kredi riskini geleneksel yöntemlerle değerlendirmeyi imkansız hale getirmiştir. Bu nedenle, bankalar kredi riskini değerlendirmek ve yönetmek için giderek daha fazla makine öğrenmesi ve yapay zeka tekniklerine yönelmektedir [1].

Etkin tahsilat süreçleri, takipteki kredi oranlarını düşürerek operasyonel ihtiyaçlarını karşılamalarını sağlar. Bu sayede, finansal sistemdeki riskler azaltılır ve genel ekonomik istikrar desteklenir. Bu da bankaların karlılığını artırır ve sermaye yeterlilik oranlarını olumlu yönde etkiler. Ayrıca, düzenleyici gerekliliklere uyum, yasal riskleri en aza indirir ve müşteri güvenini güçlendirir [2].

Son yıllarda, ekonomik durgunluklar ve çeşitli makroekonomik faktörler nedeniyle Avrupa'da kredi temerrüt oranları artmıştır. Avrupa'daki kredi temerrüt davranışı üzerine yapılan çalışmalar, borçlu özelliklerinin, krediye özgü değişkenlerin ve yerel ekonomik koşulların temerrüt üzerinde önemli etkilerinin olduğunu göstermiştir. Özellikle COVID-19'un ardından, küresel enflasyon ve faiz oranlarındaki artış nedeniyle müşteri ödeme davranışında belirgin bir değişim yaşanmıştır. Aynı nedenle, banka karlılığındaki düşüş göz önüne alındığında, veriye dayalı analizler ve tahmin süreçleri daha da önem kazanmıştır [3].

Finansal kurumlar için kredi riski, bireysel ve kurumsal müşterilerin yükümlülüklerini zamanında ve eksiksiz yerine getirememe olasılığı olarak tanımlanmaktadır. Kredi riski bankacılık sisteminin istikrarı açısından kritik bir

öneme sahiptir. Kredi riskinin doğru şekilde ölçülmesi ve yönetilmesi, yalnızca bireysel kredi kararlarını değil, aynı zamanda sermaye yeterliliği, portföy yönetimi ve fiyatlama stratejileri gibi makro düzeydeki karar süreçlerini de doğrudan etkilemektedir. Bu nedenle, kredi riski tahminine yönelik geliştirilen modeller, hem düzenleyici çerçeveler hem de rekabetçi piyasa koşulları doğrultusunda sürekli olarak geliştirilmektedir [4].

Geleneksel kredi riski modelleri ağırlıklı olarak finansal tablolar, demografik bilgiler ve geçmiş ödeme davranışları gibi yapılandırılmış veriler üzerine kuruludur [5]. Bu yaklaşımlar uzun yıllar boyunca finansal kurumlar tarafından kabul görmüş ve uygulama alanı bulmuştur. Ancak özellikle müşteri davranışlarının hızla değiştiği ve dijital etkileşimlerin yoğunlaştığı günümüzde, bu klasik veri kaynaklarının bazı sınırlamaları olduğu belirlenmiştir. Bu bağlamda, yapılandırılmış veriler müşterilerin temel finansal durumunu tanımlamak için uygun bir temel sunsa da müşteriye özgü niyet, beklenti, stres veya ödeme güçlüğü gibi davranışsal göstergeleri doğrudan yakalamada yetersiz kalabilmektedir [6].

Son yıllarda, bu sınırlılıkları aşmak amacıyla kredi riski modelleme süreçlerinde alternatif veri kaynaklarının kullanımı artmıştır [7]. Sosyal medya verileri, işlem açıklamaları, müşteri geri bildirimleri ve çağrı merkezi etkileşimleri gibi yapılandırılmamış veri kaynakları, müşterilerin finansal davranışlarına ilişkin daha erken ve daha güçlü sinyaller verme potansiyeline sahiptir. Bu kapsamda, özellikle çağrı merkezi konuşmaları, müşterilerin ödeme güçlüğü, borç yapılandırma talepleri veya finansal stres durumlarına ilişkin ipuçlarını doğrudan içermesi nedeniyle dikkat çekici bir veri kaynağı olarak öne çıkmaktadır [8] [9].

Ancak çağrı merkezi konuşmaları, doğası gereği ses verisi formatında olup, doğrudan kredi riski modellerinde kullanılabilir nitelikte değildir. Bu durum, söz konusu verilerin öncelikle metin formatına dönüştürülmesini ve sonrasında analiz edilebilir hale getirilmesini gerektirmektedir. Konuşmadan metne dönüştürme teknolojilerindeki gelişmeler, çağrı merkezi konuşmalarının yüksek doğrulukla metne dönüştürülmesini mümkün kılarken, bu metinlerin kredi riski modelleme süreçlerine nasıl entegre edileceği halen bir araştırma alanı olarak önemini korumaktadır [8].

Bu tez çalışmasının temel amacı, çağrı merkezi konuşmalarından elde edilen yapılandırılmamış metin verilerinin, ödeme sözü veren gecikmeli müşterilerin ödeme performansı tahmininde ne ölçüde katma değer sağladığını incelemektir. Özellikle farklı metin temsil yaklaşımlarının karşılaştırılması ve bu temsillerin model performansına olan etkisinin analiz edilmesi, çalışmanın çıkış noktasını oluşturmaktadır.

1.2 Kredi Riskinde Alternatif Veri İhtiyacı

Kredi riski tahminine yönelik geliştirilen geleneksel modeller, ağırlıklı olarak yapılandırılmış veri kaynaklarına dayanmaktadır. Demografik bilgiler, gelir düzeyi, kredi geçmişi, ödeme performansı ve finansal bilgiler gibi değişkenler, uzun yıllardır kredi karar süreçlerinin temel bileşenleri olarak kullanılmaktadır. Bu tür veriler, müşterilerin geçmiş davranışlarını yansıtmakta ve modelleme açısından önemli avantajlar sunmaktadır. Ancak bu yaklaşım, müşterilerin finansal durumlarında meydana gelen ani değişimleri veya henüz finansal kayıtlara yansımamış davranışsal göstergeleri yakalamakta sınırlı kalabilmektedir [10] [11].

Özellikle ekonomik belirsizlik dönemlerinde veya bireysel finansal stresin arttığı durumlarda, müşterilerin ödeme güçlüğüne ilişkin ilk belirtiler çoğu zaman yapılandırılmış veri setlerinde gecikmeli olarak ortaya çıkmaktadır [12] [13]. Bu durum, kredi riskinin erken aşamada tespit edilmesini zorlaştırmakta ve alınan kararların geç kalmasına neden olabilmektedir [14]. Bu bağlamda, kredi riski modelleme süreçlerinin daha proaktif bir yapıya kavuşturulabilmesi için alternatif veri kaynaklarının kullanımı giderek daha fazla önem kazanmaktadır [15].

Alternatif veri, geleneksel finansal bilgilerin dışında kalan, müşterilerin davranışlarını, niyetlerini ve etkileşim biçimlerini yansıtan veri kaynaklarını kapsamaktadır [15]. Sosyal medya etkileşimleri, dijital kanallar üzerinden yapılan başvurular, işlem açıklamaları ve müşteri hizmetleri ile gerçekleştirilen iletişimler bu kapsama örnek olarak verilebilir [16]. Bu tür veriler, müşterilerin finansal durumlarına ilişkin dolaylı ancak zamanında göstergeler sunabilme potansiyeline sahiptir. Ancak alternatif verilerin yüksek boyutlu, gürültülü ve yapılandırılmamış olması, bu verilerin kredi riski modellerine entegrasyonunu metodolojik açıdan zorlaştırmaktadır [14].

Kredi riskinde alternatif veri kullanımına yönelik literatür, bu tür verilerin model performansını artırabildiğini göstermekte ancak elde edilen kazanımların veri türüne, temsil yöntemine ve modelleme stratejisine bağlı olarak değişkenlik gösterdiğini ortaya koymaktadır [17] [18]. Bu durum, alternatif verilerin tek başına değil, geleneksel değişkenlerle birlikte ve kontrollü bir biçimde değerlendirilmesini gerekli kılmaktadır. Aksi halde, model performansında elde edilen iyileşmelerin kaynağını doğru biçimde yorumlamak güçleşmektedir.

Bu tez çalışması çağrı merkezi etkileşimlerini ele almaktadır. Çağrı merkezi konuşmaları, müşterilerin finansal sorunlarını, beklentilerini ve ödeme davranışlarına ilişkin ipuçlarını doğrudan ifade ettikleri bir kanal olması nedeniyle, kredi riski açısından anlamlı göstergeler barındırmaktadır. Bununla birlikte, bu göstergelerin modele katkısının ölçülebilmesi, kullanılan metin temsil yöntemlerine ve bu temsillerin finansal değişkenlerle nasıl birleştirildiğine bağlıdır. Bu nedenle, alternatif veri kullanımının kredi riski tahminine olan katkısının sistematik ve karşılaştırmalı bir çerçevede incelenmesi gerekmektedir.

1.3 Çağrı Merkezi Konuşmaları ve Metin Verileri

Bankacılık sektöründe çağrı merkezleri, müşteriler ile finansal kurumlar arasındaki en yoğun etkileşim kanallarından biridir [19]. Kredi başvuruları, ödeme gecikmeleri, borç yapılandırma talepleri ve finansal zorluklara ilişkin bildirimler çoğu zaman çağrı merkezi üzerinden iletilmekte, bu etkileşimler müşterilerin mevcut ve gelecekteki finansal davranışlarına dair önemli ipuçları barındırmaktadır. Bu yönüyle çağrı merkezi konuşmaları, kredi riski tahmini açısından zengin ancak yeterince değerlendirilmemiş bir veri kaynağı olarak değerlendirilebilir.

Çağrı merkezi etkileşimleri, müşterilerin sadece gerçekleştirdikleri finansal işlemleri değil, aynı zamanda bu işlemlerin nedenlerini, müşterilerin beklentileri ve endişeleri de içermektedir [20]. Müşterilerin ödeme güçlüğüne ilişkin ifadeleri çoğu zaman yapılandırılmış veri setlerine yansımadan önce çağrı merkezi konuşmaları aracılığıyla ortaya çıkmaktadır. Bu durum, söz konusu konuşmaların kredi riskinin erken sinyallerini barındırma potansiyeline sahip olduğunu göstermektedir [21] [22].

Bununla birlikte, çağrı merkezi konuşmaları ses verisi formatında olup, doğrudan geleneksel kredi riski modelleme süreçlerinde kullanılabilir nitelikte değildir [23]. Ses verilerinin analiz edilebilmesi için öncelikle metin formatına dönüştürülmesi gerekmektedir. Konuşmadan metne dönüştürme teknolojileri, konuşmaları otomatik olarak yazılı metne dönüştürmeyi amaçlayan yöntemler sunmakta, son yıllarda derin öğrenme tabanlı yaklaşımlar sayesinde bu alanda önemli ilerlemeler kaydedilmektedir. Özellikle büyük veri setleri üzerinde eğitilmiş modeller, çağrı merkezi gibi gürültülü ortamlarda dahi kabul edilebilir doğruluk seviyelerine ulaşabilmektedir [24] [25].

Konuşmadan metne dönüştürme süreci, çağrı merkezi konuşmalarının analizinde kritik bir ara adım olup, elde edilen metinlerin kalitesi doğrudan modelleme sonuçlarını etkilemektedir. Konuşma sırasında ortaya çıkan duraksamalar, tekrarlar, argo ifadeler ve bağlama özgü terimler, metin verisinin gürültülü bir yapıya sahip olmasına neden olabilmektedir [26]. Bu durum, elde edilen metin çıktılarının doğrudan kullanılmasından önce, uygun ön işleme çalışmalarının yapılmasını ve verinin analiz edilmesini gerektirmektedir.

Bu tez çalışmasında, çağrı merkezi konuşmalarından elde edilen metin verileri, kredi riski tahmininde kullanılabilecek bir alternatif veri kaynağı olarak ele alınmaktadır. Amaç, söz konusu metinlerin hangi temsil yöntemleri aracılığıyla daha anlamlı hale getirilebileceğini ve bu temsillerin model performansına ne ölçüde katkı sağladığını incelemektir.

2
LİTERATÜR TARAMASI

2.1 Kredi Riski Modellemesi ve Temerrüt Tahmini

Kredi riski modellemesi, bir borçlunun belirli bir zaman periyodunda yükümlülüklerini yerine getirememe olasılığının tahmin edilmesini amaçlamaktadır. Bankacılıkta bu problem genellikle ikili sınıflandırma çerçevesinde ele alınır. Özellikle olasılık temelli çıktılar karar süreçlerinde temel bir bileşen olarak kullanılır. Düzenleyici çerçeveler kredi riskinin bu bileşenler üzerinden ölçülmesini bekler, Temerrüte düşme olasılığı (probabiliry of default, PD) ölçümü ve sermaye hesaplamasında standart bir referans noktasıdır [27].

Kredi riski modellemesinin "temerrüt" kavramı etrafında tanımlanması, hem akademik literatürde hem de uygulamada tanımın standardizasyonunu gerektirir. Basel çerçevesi IRB yaklaşımı altında temerrüdün ne olduğuna dair asgari koşulları tarif ederek, veri etiketleme ve performans karşılaştırmalarında tutarlılık sağlar. Bu standardizasyon, model çıktılarını farklı portföyler ve dönemler arasında kıyaslamayı mümkün kıldığı için model yönetişiminin de temelini oluşturur [28].

Kredi riski modellemesi literatürü, iflas ve bireysel kredi temerrüdü problemlerinin uzun yıllardır istatistiksel yöntemlerle ele alındığını göstermektedir. Erken dönem çalışmalar, finansal oranlara dayalı diskriminant analizi yaklaşımıyla temerrüt riskini tahmin etmeye odaklanmıştır. Bu tür yöntemler, seçili oran kombinasyonlarının firmaları "temerrüt/temerrüt değil" biçiminde ayırabildiğini göstermiştir [29].

Daha sonra temerrüt gibi ikili olayların olasılıksal çerçevede modellenmesi yaygınlaşmış ve lojistik regresyon gibi modeller kredi skorlama uygulamalarında güçlü bir "benchmark" haline gelmiştir. Ohlson'un çalışması, temerrüt tahminini koşullu olasılık yaklaşımıyla ele alarak, olasılık temelli tahminin yorumlanabilirliğini ve karşılaştırılabilirliğini güçlendiren referanslardan biridir [30].

Kredi skorlama alanındaki kapsamlı çalışmalar, lojistik regresyon ve karar ağaçları gibi farklı sınıflandırma yöntemlerinin kredi skorlama uygulamarında örneklem seçimi, model kararlılığı, tekrar edilebilirlik kısıtları gibi konuları ayrıntılı biçimde tartışır. Hand ve Henley 1997 yılında yayınladıkları çalışma, kredi skorlama için istatistiksel sınıflandırma yöntemlerini ele alarak, alanın hem metodolojik hem de operasyonel zorluklarını sistematik biçimde özetleyen temel kaynaklar arasında yer alır [10].

Güncel kredi riski modelleme yaklaşımları, bir yandan klasik istatistiksel modellerin yorumlanabilirliğini ve yönetişim avantajlarını korurken, diğer yandan makine öğrenmesi yöntemlerinin artan öngörü gücünden faydalanmayı hedefler. Kredi verileri genellikl temerrüt oranları düşük olduğu için dengesiz sınıf yapısına sahip olduğu için, sınıflandırma algoritmalarının bu koşullarda nasıl performans verdiği kritik hale gelir. Dengesiz sınıflı kredi skorlama veri setlerinde farklı sınıflandırma algoritmalarını deneysel olarak kıyaslayan çalışmalar, model seçimi ve performans metriklerinin doğru kullanılmasının önemini vurgular [31].

Makine öğrenmesi, büyük veri setlerinden anlamlı bilgiler çıkararak tahmin modelleri oluşturmak için kullanılan güçlü bir araçtır. Özellikle derin öğrenme yöntemleri ve LSTM (Long Short-Term Memory) gibi tekrarlayan sinir ağları, zaman serisi verilerini ve doğal dil işleme problemlerini etkili bir şekilde analiz edebilme kapasitesine sahiptir [32][33].

Tahsilat süreçlerinde makine öğrenmesinin önemi, mortgage ve tarım kredileri gibi farklı tipteki krediler özelinde de araştırılmış ve başarılı sonuçlar elde edilmiştir. Sirignano ve arkadaşlarının çalışması [34], mortgage riskinin analizinde derin öğrenme modellerinin kullanımını vurgulayarak borçlu davranışlarının ve kredi performansının tahmininde yüksek doğruluk oranları elde etmiştir. Bu çalışma, büyük veri setleri kullanarak kredi temerrüt riskinin dinamiklerini ortaya koymuştur.

Tarım sektöründe yaşanan ekonomik durgunluk ve çiftçilerin geri ödeme kapasiteleri konusunda oluşan endişeler üzerine, tarımsal kredi gecikme tahminlerinde makine öğrenme yöntemlerinin kullanılması finansal riski daha doğru tahmin etmede önemli katkılar sağlamıştır. Bu bağlamda çeşitli çalışmalar, tarım

sektöründeki kredi riski tahmininde makine öğrenme algoritmalarının etkili olduğunu göstermiştir [35].

Makine öğrenmesinin kredi riski tahminindeki rolü, sadece tek bir algoritmanın "en iyi" olmasından ziyade, (i) veri kalitesi ve temsil gücü, (ii) modelleme stratejisi, (iii) validasyon tasarımı ve (iv) operasyonel kullanılabilirlik gibi boyutlarla birlikte değerlendirilmelidir. Avrupa örneklemi üzerinde temerrüt davranışını makine öğrenmesiyle ele alan çalışmalar, borçlu özellikleri, kredi özellikleri ve yerel ekonomik koşulların temerrüt üzerinde anlamlı etkileri olabildiğini göstererek, modelleme sürecinin veri kapsamı ve değişken seçimi açısından çok boyutlu ele alınması gerektiğini ortaya koyar [31].

Bu çerçevede kredi riski modellemesinde hedef değişken gözlem süresi ve hedef değişken tanımı da belirleyicidir. Uygulamada, temerrür olasılığı genellikle bir yıllık zaman periyodunda derecelendirme notuna bağlanmış bir olasılık olarak ele alınır. Bu tanım, model çıktılarının sermaye hesaplaması ve risk raporlaması gibi süreçlere entegrasyonunda standardizasyon sağlar. Bu nedenle, akademik çalışmalarda kurulan hedef değişkenin düzenleyici tanımlarla uyumu, hem bulguların bankacılık uygulamasına aktarılabilirliğini hem de yöntemsel sağlamlığı artırır [27].

LSTM algoritması, zaman serisi verilerini etkili bir şekilde analiz ederek kredi riski tahmininde üstün performans sergilemektedir [36]. Gao ve arkadaşları, XGBoost ve LSTM modellerinin birleşimi olan XGBoost-LSTM modelini kullanarak kredi kartı kullanıcılarının temerrüt tahmininde yüksek doğruluk oranları elde etmişlerdir. Bu model, özellik mühendisliği(data engineering) gerektirmeden yüksek sınıflandırma doğruluğu sağlamıştır [37].

Loughran ve McDonald'ın 2011 yılında yaptığı çalışmada, finansal raporlarda kullanılan dilin, şirketlerin mali durumları hakkında daha iyi tahminler yapılmasına yardımcı olabileceğini öne sürmüşlerdir [38].

Fu ve arkadaşları [39] çevrimiçi P2P platformlarının varsayılan risklerini tahmin etmek için derin öğrenme yöntemlerinin kullanılmasının etkili olduğunu göstermiştir. Özellikle, BiLSTM tabanlı modellerin, yatırımcı yorumlarından anahtar kelimeleri çıkartarak ve bu kelimeleri kullanarak platformların varsayılan risklerini başarılı bir şekilde tahmin edebildiğini bulmuşlardır.

2019 yılında Netzer ve arkadaşlarının yayınladığı “When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications” makalesi, kredi başvurusu esnasında başvuranın amacını belirten başvuru mektuplarının, kredi temerrüdünü tahmin etmekte nasıl kullanılabileceğini araştırır. Yazarlar, metin madenciliği ve makine öğrenimi araçlarını kullanarak 120 binden fazla kredi başvurusunu analiz etmiş ve metinsel bilgilerin, kredi temerrüdünü tahmin etmede önemli olduğunu tespit etmişlerdir. Makalenin sonucunda, makine öğrenimi ile metin verilerini birleştirmenin kredi skorlama modellerinin doğruluğunu artırmada değerli olduğunu vurgulanmıştır [36].

2.2 Alternatif Verilerin Kredi Riskinde Kullanımı

Temerrüt tahmini, çoğu zaman borçlunun finansal durumuna ilişkin değişkenlerin yanı sıra, davranışsal ve çevresel değişkenlerin de birleşimiyle şekillenen bir risk sürecini temsil eder. Bu nedenle, kredi riski modelleme literatüründe tek bir veri tipine dayalı yaklaşımlar yerine, farklı veri kaynaklarının birlikte değerlendirilmesi giderek daha fazla önem kazanmıştır. Özellikle kredi skorlamada dengesiz veri yapısı gibi nedenler, model seçiminde salt doğruluk oranı yerine ayrıştırma gücü ve kalibrasyon gibi ölçütlerin dikkate alınmasını gerektirir [10].

Bu tez bağlamında kredi riski modellemesi, temerrüt etme davranışına neden olan süreçleri daha iyi yakalayabilmek için, yapılandırılmış verilerin yanında alternatif veri kaynaklarının da modellere nasıl katkı sağlayabileceği sorusuna cevap aranmaktadır.

Klasik kredi riski modelleri öncelikli olarak kredi geçmişi ve finansal değişkenler gibi yapılandırılmış veri kaynaklarına dayanırken, bu veriler özellikle belirli müşteri segmentlerinde yetersiz kalabilmektedir. Bu durum, hem kredi değerlendirme süreçlerinde bilgi asimetrisi yaratmakta hem de geleneksel veri kaynaklarının kapsamadığı müşteri gruplarında risk tahminini zorlaştırmaktadır. Bu kapsamda alternatif veri kaynaklarının kullanımı son yıllarda hem teorik hem de uygulamalı akademik çalışmaların önemli bir konusu hâline gelmiştir [9].

Alternatif veriler, finansal verilerin dışında kalan, bireylerin davranışsal, dijital ve sosyal etkileşimlerine dair bilgi sağlayan veri türlerini içerir. Bu kapsamda telekomünikasyon kayıtları, sosyal medya etkinlikleri, e-posta kullanımı ve dijital

ayak izleri gibi yeni veri türleri, bireylerin kredi riskini tahmin etmede geleneksel veri setlerine üstünlük sağlayabilecek bilgiler sunabilmektedir.

Akademik literatürde alternatif verinin kredi riski tahminine etkisine ilişkin güçlü kanıtlar bulunmaktadır. Örneğin, demografik verilerle birlikte psikometrik göstergeleri içeren modellerin, yalnızca demografik verilerle oluşturulan modellere kıyasla daha yüksek tahmin doğruluğu sunduğu gösterilmiştir, bu durum, alternatif verinin tahmin gücünü artırabileceğini ortaya koymaktadır [40]. Benzer şekilde, çağrı detay kayıtları gibi telekomünikasyon verileri kullanılarak oluşturulan modellerin, kredi kayıt bürosu verileri temelli modellere göre daha iyi performans verdiği bulguları literatürde raporlanmıştır [41].

Alternatif veri kullanımı aynı zamanda krediye erişim eşitliğini artırma potansiyeline sahiptir. Geleneksel kredi geçmişi olmayan veya sınırlı kredi kayıtlarına sahip tüketiciler alternatif verilerle desteklenen modeller sayesinde daha doğru kredi değerlendirmesine tabi tutulabilir, bu da krediye erişimi genişletici bir etki yaratabilir. Bu bağlamda, alternatif verilerin entegrasyonunun finansal kapsayıcılığı artırdığı sonucuna varılmıştır [42].

Bununla birlikte alternatif veri kullanımının metodolojik ve etik zorlukları da bulunmaktadır. Alternatif verilerin doğası gereği yüksek boyutlu, gürültülü ve heterojen olması, uygun ön işleme, temsil ve modelleme tekniklerinin gerekliliğini doğurmaktadır. Ayrıca veri gizliliği, kullanıcı onayı ve algoritmik adalet gibi etik konular, alternatif verilerin kredi riskine entegrasyonunda dikkate alınması gereken başlıca hususlardır [9].

Özetle, alternatif veriler geleneksel kredi veri setlerini tamamlayıcı nitelikte olup, kredi risk modellerinin tahmin performansını artırma potansiyeline sahiptir. Bununla birlikte bu potansiyelin gerçekleştirilebilmesi için uygun metodolojik çerçevelerin benimsenmesi, model validasyonu ve etik gibi konuların titizlikle ele alınması gerekmektedir.

2.3 Metin Madenciliği ve Doğal Dil İşleme Yaklaşımları

Metin madenciliği ve doğal dil işleme (Natural Language Processing, NLP), yapılandırılmamış metin verilerinden anlamlı bilgi ve örüntülerin otomatik olarak çıkarılmasını amaçlayan yöntemler bütünüdür. Doğal dilin bağlama duyarlı, çok

anlamlı ve düzensiz yapısı, metin verilerinin doğrudan geleneksel istatistiksel yöntemlerle analiz edilmesini güçleştirmektedir. Bu nedenle NLP yaklaşımları, metinlerin analitik modellere girdi olabilecek sayısal temsillere dönüştürülmesini hedeflemektedir [23].

NLP tabanlı analizler genellikle üç temel aşamadan oluşmaktadır. Bunlar metin ön işleme, metin temsili ve modellemedir. Metin ön işleme aşamasında yazım farklılıklarının giderilmesi, durak kelimelerin (stopwords) temizlenmesi ve kelime köklerinin belirlenmesi gibi işlemlerle gürültünün azaltılması amaçlanır. Temsil aşamasında ise metinler, belirli bir vektör uzayında sayısal biçimde ifade edilir. Elde edilen bu temsiller, sınıflandırma problemlerinde model girdisi olarak kullanılmaktadır.

Metin madenciliğinin erken dönemlerinde en yaygın kullanılan yöntemler, kelime sıklıklarına dayalı temsil yaklaşımları olmuştur. Bag-of-Words (BoW) yöntemi, bir metni içerdiği kelimelerin frekansları üzerinden temsil ederken, kelimeler arasındaki sıralama ve bağlam bilgisini dikkate almamaktadır. Bu yaklaşımın geliştirilmiş bir uzantısı olan TF–IDF (Term Frequency–Inverse Document Frequency) yöntemi ise, kelimelerin belge içindeki ve belge koleksiyonundaki ayırt ediciliğini hesaba katarak ağırlıklandırma yapmaktadır [43].

Bu tür frekans temelli yöntemler, hesaplama açısından basit ve yorumlanabilir olmaları nedeniyle uzun süre yaygın biçimde kullanılmıştır. Ancak kelimeler arasındaki anlamsal ilişkileri ve bağlam bilgisini yakalayamamaları, özellikle uzun ve karmaşık metinlerde bu yaklaşımların sınırlı kalmasına yol açmıştır.

Dağıtımsal anlambilim yaklaşımı, bir kelimenin anlamının, birlikte kullanıldığı kelimeler üzerinden öğrenilebileceği varsayımına dayanmaktadır. Bu çerçevede geliştirilen kelime gömme (word embedding) yöntemleri, kelimeleri düşük boyutlu sürekli vektörler olarak temsil ederek anlamsal benzerliklerin vektör uzayında yakalanmasını mümkün kılmıştır. Word2Vec modeli, büyük metin koleksiyonları üzerinde eğitilerek kelimeler arasındaki anlamsal ilişkileri öğrenen öncü çalışmalardan biridir [44] .

Benzer şekilde GloVe (Global Vectors) yöntemi, kelime birlikte oluşum istatistiklerini kullanarak anlamsal yapıyı vektör uzayına yansıtmaktadır [45]. Bu yaklaşımlar, klasik BoW ve TF–IDF yöntemlerine kıyasla daha güçlü anlamsal

temsiller sunmakla birlikte, çoğunlukla bağlamdan bağımsız olmaları nedeniyle bir kelimenin farklı bağlamlardaki anlamlarını ayırt etmekte sınırlı kalmaktadır.

Derin öğrenme yöntemlerinin NLP alanında yaygınlaşması, metin temsillerinin bağlamsal bilgiyi dikkate alacak şekilde öğrenilmesini mümkün hale getirmiştir. Tekrarlayan sinir ağları (RNN) ve özellikle Long Short-Term Memory (LSTM) mimarileri, kelime dizilerindeki sıralı bağımlılıkları modelleyerek cümle ve doküman düzeyinde daha anlamlı temsiller üretmiştir [33].

Daha yakın dönemde geliştirilen Transformer mimarisi, dikkat (attention) mekanizmaları aracılığıyla metnin farklı bölümleri arasındaki ilişkileri eş zamanlı olarak öğrenebilmektedir [46]. Transformer tabanlı önceden eğitilmiş dil modelleri arasında yer alan BERT, çift yönlü bağlam bilgisini kullanarak kelime temsilleri üretmekte ve NLP alanında önemli bir yaklaşım değişimini temsil etmektedir [47].

Finansal uygulamalarda metin madenciliği ve NLP yöntemleri, çoğunlukla yapılandırılmış finansal değişkenleri tamamlayıcı nitelikte bir bilgi kaynağı olarak ele alınmaktadır. Metin tabanlı veriler, borçluların davranışsal eğilimlerine ve niyetlerine ilişkin ek bilgiler sunabilmektedir. Literatürde elde edilen bulgular, metin verisinin kredi riski modellerine katkısının kullanılan temsil yöntemine, veri kalitesine ve entegrasyon stratejisine bağlı olarak değiştiğini göstermektedir [48].

Bu nedenle kredi riski bağlamında temel soru, metin verisinin kullanılıp kullanılmamasından ziyade, hangi metin temsil yaklaşımının ve hangi modelleme çerçevesinin tercih edilmesi gerektiği şeklinde ortaya çıkmaktadır. Bu tez çalışması, NLP literatüründe yer alan farklı temsil yaklaşımlarını kredi riski problemi özelinde ele alarak, meitn verilerinin hangi yöntemlerle daha anlamlı temsillere dönüştürülebileceğini incelemektedir.

2.4 Konuşmadan Metne Dönüştürme Teknolojileri

Konuşmadan metne dönüştürme (speech-to-text, STT) ya da otomatik konuşma tanıma (Automatic Speech Recognition, ASR) teknolojileri, konuşmaların yazılı metne dönüştürülmesini amaçlayan yöntemler bütünüdür. ASR sistemleri, akustik sinyalin özelliklerini dilsel birimler ile eşleştirerek kelime dizilerini tahmin etmeye çalışır. Son yıllarda derin öğrenme tabanlı yaklaşımların yaygınlaşmasıyla birlikte, ASR sistemlerinin doğruluğu ve ölçeklenebilirliği önemli ölçüde artmıştır [24].

Güncel ASR mimarileri genellikle akustik model, dil modeli ve çözümleme (decoding) mekanizması olmak üzere üç temel bileşenden oluşmaktadır. Akustik model, seslerden elde edilen özellikleri fonem veya karakter düzeyindeki birimlere eşlerken, dil modeli, olası kelime dizilerinin dilsel olasılıklarını hesaba katar. Çözümleme aşaması ise bu bileşenleri birleştirerek en olası metin çıktısını üretir [23].

Derin öğrenme yöntemleri, ASR alanında özellikle DNN-HMM hibrit mimarilerinin yaygınlaşmasıyla belirgin bir performans artışı sağlamıştır. Bu yaklaşımda derin sinir ağları, HMM durumlarının olasılıklarını tahmin ederek klasik GMM tabanlı akustik modellemenin yerini almıştır [24].

Daha yakın dönemde uçtan uca ASR yaklaşımları, konuşma metinleriyle doğrudan kelime dizisi üretmeyi hedefleyerek akustik ve dil bileşenlerini tek bir öğrenme çerçevesinde birleştirmiştir. Bu kapsamda önemli çalışmalardan biri, yinelemeli sinir ağlarıyla (RNN) konuşmadan metne doğrudan çözümleme yaklaşımını göstermektedir [49].

Transformer mimarisinin dikkat (attention) mekanizmalarına dayalı yapısı, uzun bağımlılıkların daha etkin modellenmesine ve paralel hesaplamaya uygunluğuna katkı sağladığı için ASR alanında da yaygın biçimde benimsenmiştir. Transformer tabanlı ASR örneklerinden biri olan Speech-Transformer, yinelemeli yapı kullanmadan tamamen dikkat mekanizmalarıyla konuşma tanıma yapan bir model olarak sunulmuştur [50].

ASR sistemlerinin çıktısı, duraksamalar, tekrarlar, araya giren konuşmalar, arka plan gürültüsü ve alana özel terimler nedeniyle hatalar içerebilir. ASR performansının yaygın bir değerlendirme ölçütü olan Word Error Rate (WER), referans metin ile üretilen metin arasındaki ekleme, silme ve değiştirme hatalarını dikkate alan edit-distance tabanlı bir metriktir [23].

Bu nedenle STT çıktıları çoğu zaman ham metin olarak nihai analize doğrudan dahil edilmek yerine, uygun ön işleme ve temsil adımları ile modele aktarılması gereken bir ara ürün olarak ele alınır. Özellikle çağrı merkezi konuşmalarında görülen konuşma dili, kesintiler ve bağlama özgü finansal terimler, hem WER'ı hem de metin temsil kalitesini etkileyebilmektedir.

Finansal çağrı merkezlerinden elde edilen STT verisi, müşterinin ödeme niyeti ve finansal riskle ilişkili davranışsal ipuçlarını içerebilir, konuşma verilerinin finansal risk tahmininde kullanılabileceğini gösteren çalışmalar bu potansiyeli desteklemektedir [51]. Ancak bu verilerin analitik modellerde faydalı hâle gelebilmesi, ASR çıktısındaki hataların downstream NLP görevlerini olumsuz etkileyebilmesi nedeniyle STT/ASR kalitesi kadar, metin temsilinin ve uçtan uca modelleme çerçevesinin dikkatle kurgulanmasına bağlıdır [52] [53]. Alternatif ve yapılandırılmamış veri kaynaklarının kredi değerlendirme ve risk yönetiminde artan rolü de bu yaklaşımı desteklemektedir [54]. Bu tez bağlamında STT, çağrı merkezi konuşmalarını metne dönüştürerek kredi riski tahminine girdi sağlayan kritik bir ön adım olarak konumlandırılmaktadır.

2.5 Metin Temsil Yöntemleri

Metin verisinin analitik modellerde kullanılabilmesi için metinlerin sayısal temsillere dönüştürülmesi gerekmektedir. “Metin temsili” kavramı, bir dokümanın bir vektör uzayında ifade edilmesi anlamına gelir. Bu temsiller, metindeki anlamı ve ayırt edici bilgiyi farklı düzeylerde yakalayabilir. Literatürde metin temsilleri genel olarak frekans temelli temsiller, yoğun (dense) statik gömmeler, bağlam duyarlı (contextual) gömmeler ve cümle-doküman düzeyi gömmeler olarak sınıflandırılabilir [55].

Bu tez çalışmasının temel motivasyonu, çağrı merkezi konuşmalarından üretilen metinlerinin farklı temsil yöntemleriyle vektörleştirilmesi ve bu temsillerin kredi riski tahminine katkısının karşılaştırmalı biçimde değerlendirilmesidir. Bu nedenle metin temsili yöntemlerinin kavramsal çerçevesi, kullanılan modelleme stratejilerinin anlaşılabilmesi açısından kritik önemdedir.

Metinlerin en temel temsilleri, kelime sıklıklarına dayalı yaklaşımlardır. Bu tür temsiller, kolay uygulanabilir, yorumlanabilir ve birçok problemde güçlü bir başlangıç sağlar. Ancak bağlam bilgisini sınırlı ölçüde yansıtmaları nedeniyle özellikle dilin ince anlam farklılıklarının önemli olduğu senaryolarda yetersiz kalabilir. Bu sınırlılıklar, öğrenilmiş (learned) yoğun temsillere geçişin temel gerekçelerinden biridir [56].

Yoğun (dense) temsillerde amaç, kelimeleri düşük boyutlu sürekli vektörler olarak ifade ederek anlamsal benzerlikleri vektör uzayında yakalayabilmektir. Bu yaklaşımlarda metin gömme çoğunlukla bağlamdan bağımsızdır, yani bir kelime, farklı bağlamlarda aynı vektör ile temsil edilir. Bu durum pratikte bir basitleştirme sağlarken çok anlamlılık gibi dil olgularında sınırlılık yaratabilir [47].

Bu çizgide, FastText yaklaşımı özellikle Türkçe gibi morfolojik açıdan zengin dillerde önemli bir avantaj sunar. Yöntem, kelimeleri yalnızca "tek birim" olarak değil, karakter n-gram alt-birimlerinin toplamı olarak temsil ederek nadir kelimeler ve yazım varyasyonları karşısında daha dayanıklı gömmeler üretir. Bu yaklaşım "subword information" kullanımıyla literatürde standart bir referans hâline gelmiştir [57].

Ayrıca FastText, metin sınıflandırma tarafında da verimli bir başlangıç noktası olarak önerilmiş, hızlı eğitim ve güçlü performans özellikleri vurgulanmıştır [55].

Bu tez kapsamında FastText ve benzeri statik gömmeler, metinlerinde sık görülen yazım hataları, çekim eki çeşitliliği ve nadir kelimeler gibi olgular karşısında bir avantaj sağlayabilir.

Kelime düzeyi gömmelerin doküman ölçeğine taşınması, bir metnin tamamını sabit boyutlu bir vektörle temsil etme ihtiyacını doğurur. Bu amaçla önerilen yaklaşımlardan biri Paragraph Vector (Doc2Vec) olup, değişken uzunluktaki metin parçalarından sabit boyutlu temsiller öğrenmeyi hedefler [56].

Doc2Vec gibi yöntemler, doküman temsili üretse de modern baglama duyarlı modellerin yaygınlaşmasıyla birlikte çoğu senaryoda kıyas rolünde kullanılmaktadır.

Transformer mimarisi, dikkat mekanizmalarıyla dizilerdeki ilişkileri modelleyerek bağlam bilgisini güçlü biçimde yakalayabilmektedir [46]. Bu mimari üzerine inşa edilen BERT, çift yönlü bağlamı kullanarak token temsilleri üretir ve birçok NLP görevinde güçlü bir temel model olarak kabul edilir [47]. BERT türü modellerde önemli bir pratik nokta, doküman/cümle temsiline nasıl geçileceğidir. BERT çıktısı token düzeyinde olduğundan, aşağı akış görevlerinde sabit boyutlu temsil üretmek için genellikle özel bir token temsili ya da token temsillerinin havuzlanması gibi stratejiler kullanılır [47].

BERT'in token temelli ve görev bağımlı yapısı, özellikle büyük ölçekli benzerlik araması ve kümeleme gibi senaryolarda doğrudan kullanımını zorlaştırabilir. Bu nedenle Sentence-BERT (SBERT), BERT tabanlı bir modeli Siamese ve triplet ağ yapıları ile eğiterek doğrudan karşılaştırılabilir cümle gömmeleri üretmeyi amaçlar [58]. SBERT, cümle temsili üretme problemine model mimarisi ve eğitim hedefi üzerinden yaklaşarak, klasik BERT'e göre daha verimli benzerlik hesaplaması yapılabilir.

Konuşmadan elde edilen metinler, yazılı metinlere göre farklı hata kaynakları içerir. Kelime düzeyinde yanlış tanıma, kesintili cümle yapısı, tekrarlar ve alana özel terimler gibi zorluklar bulunur. Bu nedenle temsil seçimi yapılırken alt birim duyarlılığı, [57] bağlamı yakalama kapasitesi [47], doküman ve cümle ölçeğinde sabit temsil üretme stratejisi [58] gibi unsurlar belirleyici olabilir.

Bu tez çalışması, yukarıdaki temsil yaklaşımlarını bir arada ele alarak, çağrı merkezi konuşmalarından türetilen metin temsillerinin kredi riski tahminine katkısını değerlendirmeyi hedeflemektedir.

2.6 Literatürdeki Boşluklar ve Çalışmanın Konumlandırılması

Kredi riski tahmini literatürü, uzun yıllardır ağırlıklı olarak yapılandırılmış finansal ve davranışsal değişkenler üzerinden gelişmiştir. Sistematik literatür taramaları, kredi skorlama çalışmalarında lojistik regresyonun güçlü bir temel yaklaşım olmaya devam ettiğini, bununla birlikte ensemble ve makine öğrenmesi yöntemlerinin performans artışı sağlayabildiğini, ancak yorumlanabilirlik, sınıf dengesizliği ve model yönetişimi gibi konuların devam eden metodolojik zorluklar olarak öne çıktığını göstermektedir [59].

Son yıllarda alternatif veri kaynakları ile kredi riskinin zenginleştirilmesi önemli bir araştırma alanı hâline gelmiştir. Özellikle kredi geçmişi sınırlı müşteri gruplarında alternatif verinin potansiyel faydaları üzerine araştırmalar devam etmektedir [9].

Metin tabanlı kredi riski çalışmalarının önemli bir kısmı, P2P platformlarındaki kullanıcı açıklamaları gibi metinlerden risk göstergesi çıkarımına odaklanmaktadır. Bu literatürde metnin modele eklenmesinin tahmin performansını artırabildiği gösterilmiştir [60].

Benzer şekilde Transformer tabanlı metin temsillerinden risk göstergesi türetme yaklaşımı da P2P kredi verilerinde tartışılmaktadır [61].

Buna karşın, bankacılık çağrı merkezi konuşmalarının metin çıktılarının kredi riski tahmini bağlamında sistematik biçimde ele alındığı çalışma sayısı görece sınırlıdır.

STT kaynaklı "gürültü" metin temsil kalitesini doğrudan etkiler, bu etki literatürde çoğu zaman ikincil kalmaktadır. STT çıktıları, yanlış tanıma, kesintili cümle yapısı, tekrarlar ve alana özel terimler gibi nedenlerle gürültülü olabilir. Bu nedenle STT metinlerinde "hangi temsil yöntemi daha dayanıklıdır?" sorusu kritik hâle gelir. P2P metinlerinde görülen yazım çeşitliliği ile STT hataları benzer görünse de, STT'nin hata yapısı farklı olduğu için doğrudan genelleme yapmak her zaman mümkün değildir. Bu bağlamda temsil yöntemi karşılaştırması tez düzeyinde açık bir çerçeve gerektirir.

Metin temsillerinin katkısı çoğu çalışmada "tek yöntem" üzerinden raporlanır ve temsil ailesi bazında karşılaştırma daha sınırlıdır.

Kredi riski çalışmalarında metin kullanımı çoğu zaman belirli bir etrafında sunulmakta, farklı temsillerin aynı problem tanımı ve aynı değerlendirme protokolü ile sistematik kıyaslandığı çalışmalar daha azdır. Bu durum, metin verisinin katkısının hangi koşullarda ortaya çıktığını anlamayı zorlaştırabilir. Sistematik kredi skorlama taramaları da yöntem karşılaştrma ihtiyacını dolaylı biçimde işaret etmektedir [59].

Bankacılıkta hedef değişken tanımı ve düzenleyici uyum, akademik çalışmalarda açıkça standardize edilmeyebilir.

Özellikle banka içi çalışmalarda, temerrüt tanımı ve etiketleme yaklaşımının düzenleyici beklentilerle uyumu model geliştirme ve sonuçların yorumlanması açısından önemlidir. Avrupa Bankacılık Otoritesi (EBA) kılavuzları, temerrüt tahmini ve temerrüt tanımı uygulamasına ilişkin beklenen analiz ve dokümantasyon pratiklerini tanımlar [62].

Bu nedenle, tez kapsamında hazırlanan hedef değişkenin ve değerlendirme tasarımının açık biçimde tanımlanması, çalışmanın bankacılık uygulamasına aktarılabilirliğini artıracaktır.

Bu tez, yukarıdaki boşlukları dikkate alarak çağrı merkezinden elde edilen metinlerinin kredi riski tahminine katkısını, temsil yöntemleri odağında karşılaştırmalı biçimde incelemektedir. Tezin çıkış noktası, yayımlanmış çalışmada elde edilen bulguların tez formatında daha geniş bir metodolojik çerçeveye oturtulmasıdır [31].

Bu kapsamda tez, literatüre aşağıdaki açılardan katkı sunmayı hedeflemektedir:

1. Veri kaynağı açısından, çağrı merkezi konuşmalarının STT çıktıları üzerinden metin temsili üretme problemini ele alması [31].
2. Yöntemsel açıdan, metin temsillerini "tek bir model" yerine temsil ailesi bazında karşılaştırmalı bir değerlendirme çerçevesinde analiz etmesi [59].
3. Uygulama ve yönetişim açısından, bankacılık bağlamında hedef değişken tanımı, değerlendirme protokolü ve dokümantasyonun düzenleyici beklentilerle tutarlı şekilde kurgulanmasına önem vermesi [62].

Sonuç olarak bu tez, metin verisinin kredi riski tahminine katkısını yalnızca model performansı üzerinden değil, veri kaynağının doğası, temsil stratejisi ve bankacılık bağlamındaki uygulanabilirlik eksenlerinde konumlandırarak literatürdeki önemli bir boşluğu hedeflemektedir.

3
VERİ VE PROBLEM TANIMI

Bu bölümde, tez çalışmasında kullanılan veri seti, problem tanımı ve analiz kapsamında izlenen genel çerçeve sunulmaktadır. Çalışmada kullanılan veri, bankacılık çağrı merkezi konuşmalarından elde edilen metin verilerine dayanmaktadır. Amaç, çağrı merkezi konuşmalarından elde edilen metin temsillerinin kredi riski tahminine olan katkısını karşılaştırmalı bir deneysel kurgu içerisinde incelemektir.

3.1 Problem Tanımı

Bankacılık sektöründe kredi riski yönetimi, finansal kurumların sürdürülebilirliği ve sistemik istikrarın sağlanması açısından kritik bir rol oynamaktadır. Geleneksel kredi riski tahmin yaklaşımları ağırlıklı olarak finansal tablolar, demografik bilgiler ve geçmiş ödeme davranışları gibi yapılandırılmış veri kaynaklarına dayanmaktadır. Bu tür veriler, müşterilerin finansal durumuna ilişkin önemli bilgiler sunmakla birlikte, müşteriye özgü niyet, beklenti ve kısa vadeli ödeme davranışlarını yansıtan sinyalleri sınırlı ölçüde yakalayabilmektedir.

Son yıllarda, dijitalleşmenin artmasıyla birlikte, çağrı merkezi görüşmeleri gibi yapılandırılmamış veri kaynakları, kredi riski modelleme süreçlerinde tamamlayıcı bir bilgi kaynağı olabilme potansiyeli taşımaktadır. Ancak bu tür veriler, doğası gereği ses formatında olup, analitik modellerde doğrudan kullanılabilir nitelikte değildir.

Bu tez çalışmasının temel amacı, Türkçe çağrı merkezi telefon görüşmelerinden elde edilen konuşma verilerinin, modellemede kullanılabilecek anlamlı metin temsillerine dönüştürülmesi ve bu temsillerin müşterilerin ödeme taahhütlerini yerine getirip getirmeyeceklerini tahmin etmede ne ölçüde etkili olduğunun incelenmesidir. Bu kapsamda, konuşma verilerinin konuşmadan metne dönüştürme teknolojileri aracılığıyla metne dönüştürülmesi, metinlerin uygun ön işleme adımlarından geçirilmesi ve doğal dil işleme teknikleri kullanılarak sayısal temsillere dönüştürülmesi gerekmektedir.

Dolayısıyla bu çalışmada ele alınan temel problem, çağrı merkezi konuşmalarından türetilen yapılandırılmamış metin verilerinin, kredi riski bağlamında ayırt edici bilgi içerip içermediğinin ve bu tür verilerin mevcut kredi riski tahmin yaklaşımlarına nasıl entegre edilebileceğinin araştırılmasıdır. Çalışma, özellikle Türkçe çağrı merkezi verileri üzerinde, metin tabanlı modelleme yaklaşımlarının uygulanabilirliğini ve potansiyel katkılarını değerlendirmeyi amaçlamaktadır.

3.1.1 Veri Kaynağı ve Genel Özellikler

Bu tez çalışmasında kullanılan veri seti, bankacılık operasyonları kapsamında temerrüt durumundaki müşteriler ile gerçekleştirilen Türkçe çağrı merkezi telefon görüşmelerinden elde edilmiştir. Bankanın mevcut stratejik ve operasyonel kararları doğrultusunda, belirli bir gecikme gününü aşan müşteriler çağrı merkezi aracılığıyla aranmakta ve bu görüşmeler sırasında müşterilerden ödeme taahhütleri alınmaktadır. Bu kapsamda gerçekleştirilen telefon görüşmeleri, çalışmanın temel veri kaynağını oluşturmaktadır.

Çağrı merkezi görüşmelerine ait ses kayıtları, konuşmadan metne dönüştürme teknolojileri kullanılarak metin formatına çevrilmiş ve bankanın operasyonel veri altyapısı içerisinde yer alan veritabanında saklanmıştır. Veritabanında, görüşmeler sırasında yer alan her bir konuşmacının ifadeleri, konuşmanın doğal akışını koruyacak biçimde sıralı olarak kayıt altına alınmıştır. Bu yapı, diyalogların zamansal bütünlüğünün korunmasına olanak tanımaktadır.

Veri toplama süreci, bankanın çağrı merkezi sistemleri aracılığıyla, gerçek operasyonel koşullar altında gerçekleştirilmiştir. Görüşmeler, banka temsilcileri ile müşteriler arasında doğal iletişim ortamında yapılmış olup, kontrollü veya deneysel bir senaryoya dayanmamaktadır. Dönüştürülen metin verileri, sistematik bir şekilde saklanmış, her bir konuşmacının ifadeleri, diyalog yapısını ve konuşma sırasını yansıtacak biçimde düzenlenmiştir. Buna ek olarak, arama tarihi, görüşme süresi ve müşteri hesap bilgileri gibi çeşitli meta veriler de veri setinde yer almaktadır.

Çağrı merkezi görüşmeleri, genel olarak üç ana etkileşim türünü içermektedir. Bunlar müşterilerden alınan ödeme taahhütlerine ilişkin ifadeler, gecikmede bulunan kredi veya ürün bilgileri ve görüşme sırasında gerçekleşen iletişim detaylarıdır. Bu bileşenler, müşterilerin ödeme davranışlarını, finansal durumlarına

ilişkin söylemlerini ve iletişim kalıplarını analiz edebilmek açısından önemli bilgiler sunmaktadır.

Bu tür çağrı merkezi verileri, yalnızca ödeme tahmin modellerinin geliştirilmesi açısından değil, aynı zamanda müşteri iletişim stratejilerinin iyileştirilmesi açısından da değerli içgörüler sağlayabilir. Veri setinin kontrollü laboratuvar ortamları yerine gerçek dünya koşullarında toplanmış olması, elde edilen bulguların uygulamaya aktarılabilirliğini artırmaktadır. Ayrıca, görüşmelerin sıralı ve ayrıntılı biçimde saklanması, diyalog dinamiklerinin ve zamansal örüntülerin analiz edilmesine olanak tanımaktadır.

Ham veri setinde her bir konuşma sırayla müşteri iletişim merkezi çalışanı ve müşteri olacak şekilde Tablo 3.1'de görüleceği şekilde tutulmaktadır.

Veri gizliliği ve kurumsal hassasiyetler göz önünde bulundurularak, çalışmada kullanılan veri seti hedef değişken dağılımı dengeli olacak şekilde oluşturulmuştur. Bu kapsamda, ödeme taahhüdünü yerine getiren ve getirmeyen müşterilere ait gözlemler yaklaşık %50–%50 oranında temsil edilecek biçimde bir alt örneklem seçilmiştir. Kullanılan metin verileri, bankanın gerçek çağrı merkezi görüşmelerinden elde edilmiş olmakla birlikte, uygulanan bu örnekleme stratejisi nedeniyle oluşturulan veri seti bankanın gerçek müşteri dağılımını birebir yansıtmamaktadır.

Bu tercih, çalışmanın temel amacının operasyonel bir kredi riski tahmin modeli geliştirmekten ziyade, farklı metin saysallaştırma yöntemlerinin metin temsili üretme ve ayırt edici bilgi yakalama kapasitelerinin karşılaştırmalı olarak değerlendirilmesi olmasından kaynaklanmaktadır. Hedef değişken dağılımının dengelenmesi, sınıf dengesizliğinden kaynaklanabilecek yanlılıkları azaltarak, kullanılan metin gömme yöntemlerinin göreli performanslarının da daha sağlıklı biçimde analiz edilmesine olanak tanımaktadır. Bu doğrultuda elde edilen bulgular, mutlak tahmin performansından ziyade, yöntemler arasındaki göreli farkların incelenmesine odaklanmaktadır.

Tablo 3.1 Anomimleştirilmiş örnek bir görüşmenin ham veri temsili

Müşteri Numarası	Konuşmacı	Başlangıç Zamanı	Görüşme Numarası	Orjinal İfade
12345	Agent	1117	6460	TEŞEKKÜR EDERİM GÖRÜŞMELER KAYIR ALTINDA… KULLANDIĞINI KREDİ KATRI ON … GÜN KREDİLER ON BİR ON BİR VE 4 GÜN GECİKMİŞ … TANE BAKIYORUM HEMEN … KREDİ KARTI ASGARİSİ ON BİN ALTMIŞ
12345	Customer	5450	6460	HIHI
12345	Customer	4355	6460	TAMAM
12345	Customer	3727	6460	BUGÜN MAAŞ YATARSA YARIN HEPSİNİ HALLEDER
12345	Agent	48	6460	İYİ GÜNLER
12345	Customer	383	6460	BUYRUN BENİM
12345	Agent	4643	6460	TAMAM O TAMAM BİRİ ZATEN HEPSİNİ SALI DİYORUM ON BİRİ O ZAMAN YARIN … ON ÜÇ BİN ALTI YÜZ ON ÜÇ … BİR KREDİ KARTI … YARIN ÖDEMEYİ ONAYLIYOR MUSUNUZ
12345	Customer	4250	6460	HA … YARIN DİYELİM GÜNCELLEYELİM ORADAKİ TALİMAT YANLIŞ OLMASIN
12345	Agent	5750	6460	GECİKME FAİZİ GÜNLÜK … ÖZELLİKLER KREDİLER EKSTRENİZ TAKAN … GECİKME … FİNANSAL İŞLEMİNİZE KARŞINIZA ÇIKAR … YARIN DÖRT ÖDEMEYİ … OLACAĞIZ
12345	Customer	5680	6460	EVET
12345	Agent	4021	6460	TAMAM YARIN MI DİYELİM
12345	Customer	7142	6460	İYİ GÜNLER
12345	Agent	452	6460	ÖNCE BABA İSMİNİZİ ÖĞRENEBİLİRMİYİM

Ham veri ile ilgili betimleyici istatistikler Tablo 3.2'de verilmiştir.

Tablo 3.2 Verinin betimleyici istatistikleri

Özellik	Değer
Veri toplama dönemi	01.12.2022 – 16.12.2023
Toplam telefon görüşmesi sayısı	3.884.198
Görüşmeye katılan müşteri temsilcisi sayısı	407
Görüşülen toplam müşteri sayısı	1.153.903
Ortalama görüşme süresi	2 dakika 52 saniye
Maksimum görüşme süresi	14 dakika[1]
Ortalama müşteri gecikme günü	33 gün
Görüşmelerin zaman dağılımı	%97'si hafta içi 09:00–17:00 saatleri arasındadır.
Günlere göre görüşme dağılımı	Uniform
Müşteri temsilcisi başına ortalama günlük görüşme adedi	109
Ödeme sözü alınan görüşmelerin oranı	55%
Modelleme kitlesi	Ödeme sözü alınan %55'lik görüşmelerdir.
Gerçekleşen ödeme oranı	%50 (alınan sözlerin yarısı gerçekleşmiştir)

[1] 99. yüzdelik dilimin üzerindeki değerler aykırı (outlier) olarak değerlendirilmiştir.

Sunulan betimleyici istatistikler, çağrı merkezi görüşmelerinin genel yapısına ve müşterilerin ödeme taahhütlerine ilişkin davranışsal örüntülere dair önemli bilgiler sunmaktadır. Tablolarda yer alan görüşme sayıları, görüşme süreleri, kelime sayıları ve ödeme taahhütlerine ilişkin göstergeler, veri setinin kapsamı ve içeriği hakkında genel bir çerçeve oluşturmaktadır. Bu göstergeler, müşteri görüşmelerinin

heterojen yapısını ve farklı müşteri profillerine ait etkileşimlerin çeşitliliğini ortaya koymaktadır.

Elde edilen istatistikler, veri setinin yalnızca sayısal büyüklük açısından değil, aynı zamanda içerik çeşitliliği açısından da zengin bir yapıya sahip olduğunu göstermektedir. Görüşme sıklıkları, konuşma süreleri ve metin uzunluklarındaki farklılıklar, müşterilerin ödeme süreçlerine yaklaşımlarında ve iletişim biçimlerinde gözlemlenen çeşitliliği yansıtmaktadır. Bu durum, çağrı merkezi konuşmalarının kredi riski bağlamında potansiyel olarak ayırt edici bilgiler barındırabileceğine işaret etmektedir.

Bu betimleyici istatistikler, gecikmiş ödemelere ilişkin müşteri iletişim süreçlerinin genel özelliklerini anlamak açısından önemli bir başlangıç noktasıdır. Ancak bu aşamada sunulan analizler, yalnızca veri setinin yapısal özelliklerini tanımlamaya yönelik olup, herhangi bir nedensel çıkarım veya modelleme sonucunu içermemektedir. Betimleyici bulgular, sonraki bölümlerde gerçekleştirilecek veri ön işleme ve modelleme adımlarının gerekçelendirilmesine zemin hazırlamaktadır.

İlerleyen bölümlerde, ham veri seti analiz ve modelleme süreçlerinde kullanılmak üzere yeniden yapılandırılacaktır. Bu kapsamda, çağrı merkezi görüşmeleri müşteri bazında birleştirilecek, her bir diyalog tek bir gözlem satırı olacak şekilde düzenlenecek ve veri temizliği ile ön işleme adımları uygulanacaktır.

3.2 Hedef Değişken Tanımı

Bu tez çalışmasında hedef değişken, çağrı merkezi görüşmeleri sırasında müşterilerden alınan ödeme taahhütlerinin, belirlenen süre içerisinde yerine getirilip getirilmediğini gösteren ikili bir değişken olarak tanımlanmıştır. Amaç, çağrı merkezi konuşmalarından elde edilen metin verileri kullanılarak, müşterinin verdiği ödeme sözünü tutup tutmayacağının önceden tahmin edilmesidir.

Bankanın operasyonel süreçleri doğrultusunda, belirli bir gecikme gününü aşan müşteriler çağrı merkezi aracılığıyla aranmakta ve bu görüşmeler sırasında müşterilerden ödeme taahhütleri alınmaktadır. Bu çalışmada, söz konusu çağrı merkezi görüşmeleri referans alınmış, her bir müşteri için alınan ödeme taahhüdü, ilgili taahhüdün yerine getirilip getirilmediğine göre etiketlenmiştir. Hedef değişken, müşterinin taahhüt edilen ödeme tarihine kadar ilgili yükümlülüğünü

yerine getirmesi durumunda pozitif sınıf, taahhüdün yerine getirilmemesi durumunda ise negatif sınıf olarak tanımlanmıştır.

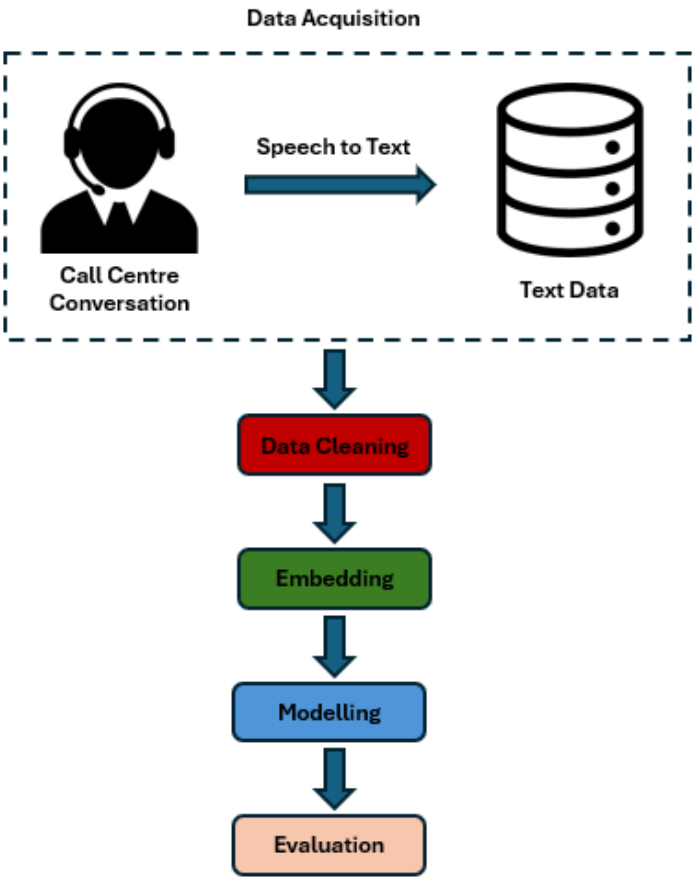
Hedef değişkenin oluşturulmasında, çağrı merkezi görüşmesi sonrasında gerçekleşen ödeme davranışı esas alınmış, çağrı sırasında veya öncesinde mevcut olan bilgiler hedef etiketin belirlenmesinde kullanılmamıştır. Bu yaklaşım, modelleme sürecinde ileriye dönük bilgi kullanımının (data leakage) önlenmesini ve tahmin probleminin gerçekçi bir biçimde ele alınmasını amaçlamaktadır.

Bu hedef değişken tanımı, çağrı merkezi konuşmalarının müşterilerin ödeme davranışlarına ilişkin ayırt edici bilgiler içerip içermediğinin değerlendirilmesine olanak tanımaktadır. Aynı zamanda, kullanılan hedef değişkenin bankacılık operasyonlarındaki ödeme taahhüdü süreçleriyle uyumlu olması nedeniyle, elde edilen bulgular uygulamaya aktarılabilir.

3.3 Veri Ön İşleme ve Birleştirme

Bu tez çalışmasında izlenen analiz süreci, Şekil 3.1'de sunulan genel çerçeve doğrultusunda veri toplama, veri temizliği, metin gömme, modelleme ve model değerlendirme aşamalarından oluşmaktadır. Veri toplama süreci önceki bölümlerde ayrıntılı olarak açıklanmıştır. Bu bölümde ise, çağrı merkezi görüşmelerinden elde edilen ham metin verilerinin analiz ve modelleme süreçlerinde kullanılabilir hale getirilmesi amacıyla uygulanan veri ön işleme ve temizleme adımları ele alınmaktadır.

Veri temizliği, metin madenciliği ve doğal dil işleme çalışmalarında kritik bir aşamadır [63]. Ham metin verileri, doğası gereği yazım hataları, gereksiz karakterler ve tutarsız biçimlendirmeler içerebilmektedir. Bu nedenle, modelleme aşamasına geçilmeden önce metinlerin belirli bir standartta ve tutarlı bir yapıda dönüştürülmesi gerekmektedir.



Şekil 3.1 Veri toplama aşamasından model performans ölçümü aşamasına kadarki sürecin görseli

Çalışma kapsamında ilk olarak, ham verileri Şekil 3.2'de gösterildiği gibi analiz edilebilir bir biçime dönüştürmek amacıyla temel temizlik işlemleri uygulanmıştır. Bu aşamada, metinlerde yer alan gereksiz boşluklar, noktalama işaretleri ve özel karakterler (örneğin nokta, virgül, parantez gibi) temizlenmiştir. Söz konusu işlemler, metin içerisindeki gürültünün (noise) azaltılmasını ve kelime temelli analizlerin daha sağlıklı biçimde gerçekleştirilmesini amaçlamaktadır.

İkinci aşamada, veri setindeki tüm kelimeler küçük harfe dönüştürülmüştür. Büyük ve küçük harf farklılıkları, metin madenciliği süreçlerinde aynı kelimenin farklı biçimlerde algılanmasına ve kelime frekanslarının hatalı hesaplanmasına yol açabilmektedir. Bu nedenle, harf dönüşümü işlemi ile metinlerin biçimsel tutarlılığı artırılmış ve analiz sürecinde tutarlı bir metin temsili sağlanmıştır.

Veri temizleme sürecinin en uzun ve detaylı aşamasında ise, metinlerde yer alan yazım hataları ve dil bilgisi sorunlarının giderilmesine yönelik kapsamlı işlemler

gerçekleştirilmiştir. Bu kapsamda, monogram ve bigram kontrolleri kullanılarak yaygın yazım hataları tespit edilmiş ve manuel düzeltmeler yapılmıştır. Ayrıca, otomatik konuşmadan metne dönüştürme sürecinden kaynaklanan ve anlam bütünlüğü taşımayan ifadeler incelenmiş, anlaşılamayan kelime sayısının yüksek olduğu ve metin içeriğinin analize uygun olmadığı tespit edilen gözlemler veri setinden çıkarılmıştır.

Uygulanan bu ön işleme ve temizleme adımları sonucunda çağrı merkezi görüşmelerinden elde edilen ham metin verilerin, metin gömme ve modelleme aşamalarında kullanılabilecek tutarlı ve güvenilir bir yapıda olması sağlanmıştır. Bir sonraki başlıkta, temizlenen metin verilerinin farklı metin temsil yöntemleri kullanılarak sayısallaştırılması ve modelleme sürecine dahil edilmesi ele alınacaktır.

```
df['Phrase'] = df['Phrase'].str.replace(' noise ','')
df['Phrase'] = df['Phrase'].str.replace(' ya ','')
df['Phrase'] = df['Phrase'].str.replace(' vocalized noise ','')
df['Phrase'] = df['Phrase'].str.replace(' ha ','')
df['Phrase'] = df['Phrase'].str.replace(' he ','')
df['Phrase'] = df['Phrase'].str.replace(' dakka ',' dakika ')
df['Phrase'] = df['Phrase'].str.replace(' yo ',' yok ')
df['Phrase'] = df['Phrase'].str.replace(' hihievet ',' evet ')
df['Phrase'] = df['Phrase'].str.replace(' geliyo ',' geliyor ')
df['Phrase'] = df['Phrase'].str.replace(' eveteveteveteve evet evet')
df['Phrase'] = df['Phrase'].str.replace(' imdi simdi ')
df['Phrase'] = df['Phrase'].str.replace(' iyigunler ',' iyi günler ')
df['Phrase'] = df['Phrase'].str.replace(' ateme ',' ATM ')
df['Phrase'] = df['Phrase'].str.replace(' tamamsagolun ',' tamam sagolun ')
df['Phrase'] = df['Phrase'].str.replace(' nasi ',' nasil ')
df['Phrase'] = df['Phrase'].str.replace(' tamamtesekkur ',' tamam tesekkurler ')
df['Phrase'] = df['Phrase'].str.replace(' vocalized noisetamam ',' tamam ')
```

Şekil 3.2 Manuel veri düzeltme kod örneği

Monogram ve bigram kontrolleri Tablo 3.3'te örnek olarak gösterildiği gibi, ödeme sözünü tutan ve tutmayan müşteriler için ayrı ayrı incelenmiş. Bu sayede hem veri seti hakkında bilgi sahibi olunmuş hem de manuel veri temizliği süreci daha detaylı çalışılmıştır.

Tablo 3.3 Bigram örnekleri inceleme tablosu

Bigram	Adet
iyi gunler	62.724
tamam tamam	61.115
evet evet	49.181
tesekkur ederim	29.072
ne kadar	21.324
evet benim	11.676
su anda	11.594
buyrun benim	10.577
değil mi	9.267
kredi karti	8.232
tamam onayliyorum	7.934
ayin on	7.549
cuma gunu	6.188
uc yuz	6.128
yuz lira	6.010

Betimleyici istatistiklerin incelenmesi sonucunda, veri setinde yer alan en kısa ve en uzun konuşmalar ayrıntılı olarak analiz edilmiştir. Yapılan incelemeler, bu uç değerlerin büyük ölçüde teknik sorunlar nedeniyle gerçekleştiğini veya kısa sürede hat kopması gibi nedenlerle tamamlanamayan çağrılardan oluştuğunu göstermiştir. Bu tür görüşmelerin, müşterilerin gerçek iletişim örüntülerini yeterince yansıtmadığı gözlemlenmiştir.

Veri setinin, tamamlanmış ve analize uygun görüşmelerden öğrenmesini sağlamak amacıyla, konuşma uzunlukları açısından aykırı gözlemler belirlenmiş ve veri setinden çıkarılmıştır. Bu kapsamda, konuşma uzunluğu dağılımına göre 0.1 persentilden daha kısa ve 0.9 persentilden daha uzun olan görüşmeler aykırı gözlem olarak tanımlanmış ve analiz dışında bırakılmıştır. Bu yaklaşım, metinlerdeki tutarlılığı artırmayı ve modelleme sürecinde daha homojen bir veri yapısı elde etmeyi amaçlamaktadır.

Uygulanan bu filtreleme işlemleri sonucunda, veri setindeki metinlerin yapısal tutarlılığı artırılmış ve analiz için daha elverişli bir ortam oluşturulmuştur. Temizleme sürecinin ardından, görüşmeler konuşma sırasına göre yeniden düzenlenmiştir. Çağrı merkezi temsilcilerinin büyük ölçüde standartlaştırılmış ifadeler kullandığı göz önünde bulundurularak, veri setinin boyutunu ve karmaşıklığını azaltmak amacıyla yalnızca müşterilere ait konuşma bölümleri

korunmuştur. Bu varsayım doğrultusunda, temsilci konuşmalarının çıkarılmasının bilgi kaybına yol açmadığı kabul edilmiştir.

Tablo 3.4 Veri temizleme ve gruplandırma işlemlerinden sonra modelleme için hazırlanan örnek veri

Müşteri Numarası	Hedef Değişken	Görüşme Numarası	Düzenlenmiş İfade
12345	1	6460	buyrun benim bugun maas yatarsa yarin hepsini hallederim yarın diyelim guncelleyelim ordaki talimati yanlis olmasin evet evet tamam iyi gunler

Temizlenen veri setinde, her bir müşteri görüşmesi tek bir gözlem satırı olacak şekilde gruplanmıştır. Gruplandırılmış örnek ifade Tablo 3.4'te gösterilmiştir. Bu işlemler sonucunda elde edilen veri seti toplam 1.701.782 satırdan oluşmaktadır. Veri setinde ortalama kelime sayısı 47.65, minimum kelime sayısı 13, maksimum kelime sayısı 127 ve medyan kelime sayısı 40 olarak hesaplanmıştır. Hedef değişkenin veri seti içerisindeki dağılımı incelendiğinde, pozitif sınıf oranının yaklaşık %57 olduğu görülmüştür.

Veri setinin boyutu ve metinlerin karmaşıklığı dikkate alınarak, modelleme sürecinde hesaplama maliyetlerini yönetilebilir seviyede tutmak amacıyla örnekleme yaklaşımı benimsenmiştir. Bu doğrultuda, veri setinin yaklaşık %15'ine karşılık gelen ve rastgele seçilmiş 250.000 satırlık bir alt örneklem oluşturulmuştur. Seçilen örneklem, modelleme sürecinde temsil gücünü koruyacak biçimde rastgele örnekleme yöntemi ile belirlenmiştir.

Oluşturulan veri seti, model eğitimi ve performans değerlendirmesi amacıyla eğitim ve test gruplarına ayrılmıştır. Bu kapsamda, verinin %80'i modelin öğrenme sürecinde kullanılmış, kalan %20'lik bölüm ise model performansının değerlendirilmesi için ayrılmıştır. Model, eğitim verisi üzerinde öğrenilmiş ve test verisi üzerinde performans ölçümleri gerçekleştirilmiştir.

Veri temizleme sürecinde kritik öneme sahip unsurlardan biri, konuşmadan metne dönüştürme araçlarının doğru ve tutarlı biçimde çalışmasıdır. Bu araçların doğruluk düzeyi, elde edilen metin verilerinin kalitesini doğrudan etkilemekte, dolayısıyla metin tabanlı modelleme süreçlerinde elde edilen sonuçlar üzerinde belirleyici bir rol oynamaktadır. Konuşma tanıma sürecinde oluşabilecek hatalar, metinlerde anlamsal bozulmalara yol açarak model performansını olumsuz yönde etkileyebilmektedir.

Buna ek olarak, çağrı merkezi görüşmelerinin içeriği ve yapısı, yalnızca kullanılan teknolojilerden değil, aynı zamanda banka çağrı merkezi çalışanlarının müşterilerle kurduğu iletişim biçiminden de etkilenmektedir. Görüşmeler sırasında müşterilerin doğru yönlendirilmesi ve tutarlı bir iletişim yöntemi benimsenmesi, elde edilen metin verilerinin analitik açıdan daha anlamlı olmasına olanak sağlamaktadır.

Bu çalışma kapsamında kullanılan çağrı merkezi konuşmalarının analizi, müşteri ifadelerinde öne çıkan dilsel örüntülerin incelenmesine ve dolaylı olarak müşteri iletişim süreçlerine ilişkin değerlendirmeler yapılabilmesini mümkün kılmaktadır. Bu yönüyle elde edilen bulgular, çağrı merkezi süreçlerinin iyileştirilmesine ve müşteri etkileşimlerinin daha etkin yönetilmesine yönelik potansiyel katkılar sunmaktadır.

4
METİN TEMSİL YÖNTEMLERİ VE MODELLEME ÇERÇEVESİ

Bu bölümde, çağrı merkezi konuşmalarından elde edilen metin verilerinin sayısal temsile dönüştürülmesi amacıyla kullanılan yöntemler ile bu temsillerin kredi riski tahmini probleminde nasıl modellendiği ayrıntılı olarak ele alınmaktadır. Doğal dil işleme uygulamalarında, metin verilerinin doğrudan makine öğrenmesi veya derin öğrenme modellerine girdi olarak kullanılması mümkün olmadığından, metinlerin anlamsal ve bağlamsal özelliklerini koruyacak şekilde sayısallaştırılması kritik bir öneme sahiptir. Bu amaçla kullanılan metin gömme yöntemleri, kelime ve cümle düzeyinde anlamsal ilişkilerin vektör uzayında temsil edilmesini sağlamaktadır.

Bu çalışma kapsamında, metin verilerinin temsili için Self Embedding, FastText ve BERT (Bidirectional Encoder Representations from Transformers) olmak üzere üç farklı metin gömme yaklaşımı değerlendirilmiştir. Her bir yöntemin teorik temelleri, avantajları ve sınırlılıkları incelenmiş, elde edilen metin gömme temsilleri sabit bir modelleme mimarisi altında karşılaştırmalı olarak analiz edilmiştir.

4.1 Genel Modelleme Yaklaşımı

Bu tez çalışmasında problem, gecikmede olan müşterilerin çağrı merkezi görüşmesi sonrasında ödeme taahhüdünü yerine getirip getirmeyeceğinin tahmin edilmesi olarak ele alınmıştır. Problem, ikili sınıflandırma çerçevesinde tanımlanmış olup, her bir müşteri için tek bir hedef etiket bulunmaktadır.

Modelleme yaklaşımı, uçtan uca bir yapı üzerine kurulmuştur. Bu yapı içerisinde, metin temsilleri modelin bir parçası olarak öğrenilmekte ve hedef değişkeni doğrudan optimize edecek şekilde eğitilmektedir. Böylece metinlerin sayısal temsilleri, önceden sabitlenmiş gömmeler yerine, probleme özel bir temsil uzayında öğrenilmektedir.

Model girdisi olarak, müşteri bazında birleştirilmiş çağrı merkezi konuşmalarından elde edilen metinler kullanılmış, model çıktısı ise müşterinin ödeme davranışını temsil eden ikili sınıf etiketi olmuştur.

4.2 Metin Temsil Yaklaşımı

Bu tez çalışmasında kullanılan metin temsil yaklaşımları, çağrı merkezi konuşmalarından elde edilen metin verilerinin doğası ve kredi riski problemine özgü gereksinimler dikkate alınarak belirlenmiştir. Çağrı merkezi konuşmaları yazılı metinlere kıyasla daha gürültülü, düzensiz ve bağlama duyarlı bir yapı sergilemektedir. Bu nedenle tek bir metin temsil yaklaşımına bağlı kalmak yerine, farklı temsil stratejilerinin karşılaştırmalı olarak değerlendirilmesi tercih edilmiştir.

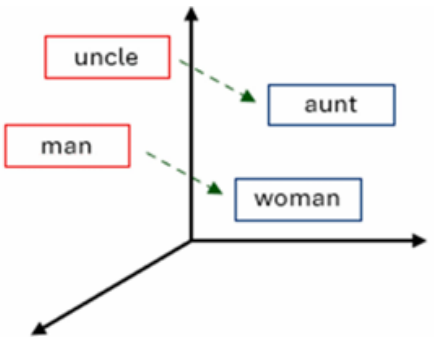
Bu kapsamda çalışmada üç farklı metin temsil yaklaşımı ele alınmıştır:

(i) göreve özel olarak öğrenilen Self Embedding temsilleri,

(ii) önceden eğitilmiş FastText kelime gömmeleri ve

(iii) bağlam-duyarlı BERT tabanlı temsiller.

Bu yöntemlerin birlikte değerlendirilmesi, metin temsilinin kredi riski tahminine katkısının farklı düzeylerde analiz edilmesini mümkün kılmaktadır. Şekil 4.1, kelimeler arasındaki anlamsal ilişkileri vurgulayan kelime gömülerine bir örnek göstermektedir.



Şekil 4.1 Anlamsal ilişkileri gösteren kelime gömme örnekleri

4.2.1 Self Embedding Yaklaşımı

Self Embedding yöntemi, kelime ve cümle temsillerinin doğrudan modelin kendi öğrenme süreci içerisinde oluşturulduğu, veri odaklı bir metin gömme yaklaşımıdır.

Bu yöntemde, önceden eğitilmiş sabit kelime vektörleri kullanılmak yerine, metin gömme katmanı model mimarisinin bir parçası olarak öğrenilmekte ve metin temsilleri görev-özel bir biçimde şekillenmektedir. Bu sayede model, kullanılan veri setine özgü dilsel kalıpları ve anlamsal ilişkileri daha etkin bir şekilde yakalayabilmektedir [64].

Self Embedding yaklaşımı, bağlam duyarlılığı sağlaması nedeniyle aynı kelimenin farklı bağlamlarda farklı temsiller almasına olanak tanır. Özellikle uzun ve ardışık metin yapılarının bulunduğu çağrı merkezi konuşmalarında, bağlamın korunması kredi riski tahmini açısından kritik bir avantaj sunmaktadır. Ayrıca bu yöntem, büyük ölçekli önceden eğitilmiş veri setlerine olan bağımlılığı azaltarak, alan ve veri setine özgü öğrenmeyi ön plana çıkarmaktadır.

Bu çalışmada Self Embedding yöntemi, Türkçe gibi morfolojik açıdan zengin ve görece az kaynaklı bir dilde, veri setine özgü ince dilsel ayrıntıların yakalanabilmesi açısından avantajlı bir yaklaşım olarak değerlendirilmiştir. Bununla birlikte, bu yöntemin etkin bir şekilde öğrenilebilmesi yüksek hesaplama maliyeti gerektirmekte ve sınırlı veri setlerinde genelleme kabiliyeti görece sınırlı kalabilmektedir.

4.2.2 FastText

FastText, kelime temsillerini oluştururken kelimeleri yalnızca bütün hâlleriyle değil, aynı zamanda karakter n-gram'ları aracılığıyla alt birimlerine ayırarak öğrenen bir metin gömme yöntemidir. Bu yaklaşım, özellikle yazım hataları, eklemeli dil yapısı ve düşük frekanslı kelimelerin bulunduğu veri setlerinde daha dayanıklı temsiller üretmektedir [57].

Türkçe'nin eklemeli yapısı göz önüne alındığında, FastText'in alt birim temelli yaklaşımı önemli bir avantaj sunmaktadır. Çağrı merkezi konuşmalarında sıkça karşılaşılan yazım hataları, kısaltmalar ve konuşma diline özgü ifadeler, FastText tarafından daha etkin bir şekilde temsil edilebilmektedir. Bununla birlikte, FastText bağlamdan bağımsız kelime temsilleri ürettiği için, aynı kelimenin farklı bağlamlardaki anlam ayrımlarını yakalama konusunda sınırlı kalabilmektedir.

Bu çalışmada FastText, hesaplama verimliliği, Türkçe dil yapısına uyumu ve yazım hatalarına karşı dayanıklılığı nedeniyle güçlü bir karşılaştırma yöntemi olarak değerlendirilmiştir.

4.2.3 BERT

BERT, Transformer mimarisi üzerine inşa edilmiş, bağlama duyarlı kelime ve cümle temsilleri üreten güçlü bir dil modelidir. BERT'in temel ayırt edici özelliği, kelimelerin yalnızca kendisinden önce gelen kelimelerle değil, aynı zamanda kendisinden sonra gelen kelimelerle olan ilişkilerini de dikkate alarak iki yönlü bağlamsal temsiller üretmesidir. Bu özellik, özellikle doğal konuşma metinlerinde anlamın bağlama bağlı olarak değiştiği durumlarda önemli bir avantaj sağlamaktadır [47].

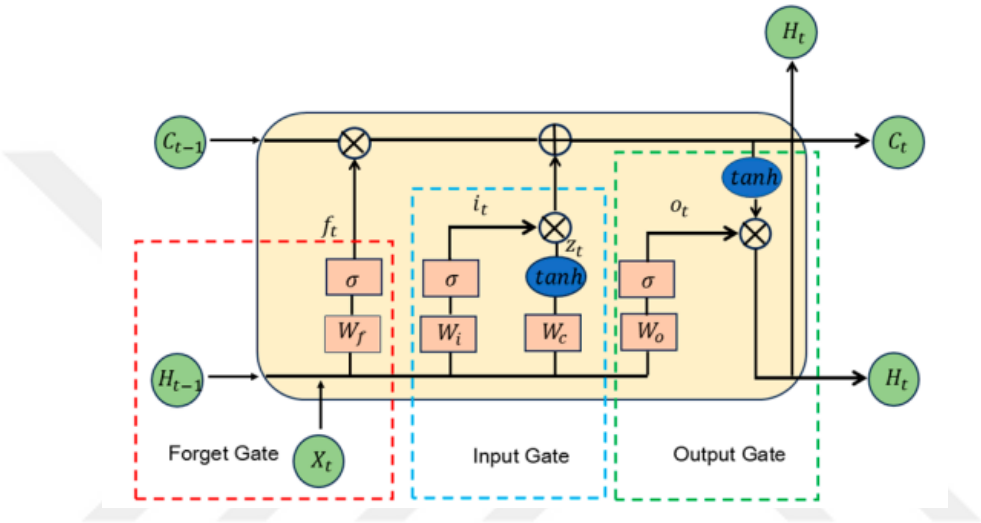
Çağrı merkezi konuşmaları, kısa ifadeler, eksik cümleler ve bağlama dayalı anlam içeren söylemlerden oluştuğu için, BERT'in bağlama duyarlı yapısı teorik olarak güçlü bir temsil kapasitesi sunmaktadır. Ancak BERT modellerinin yüksek parametre sayısına sahip olması, ciddi hesaplama maliyeti doğurmakta ve özellikle Türkçe için yeterli büyüklükte ve kalitede önceden eğitilmiş modellerin sınırlı olması uygulama sürecinde bazı kısıtlamalara yol açmaktadır. Bu çalışmada BERT, güçlü bağlamsal temsil yeteneğine rağmen, hesaplama maliyeti ve optimizasyon zorlukları nedeniyle sınırlı ölçüde değerlendirilmiştir.

4.3 Modelleme Çerçevesi

Metin verilerinin temsil edilmesinin ardından, kredi riski tahmini problemi için uygun bir modelleme yaklaşımı belirlenmiştir. Çağrı merkezi konuşmaları zamansal ve ardışık bir yapı sergilediğinden, metin içerisindeki kelime sıralarının ve uzun dönemli bağımlılıkların modellenebilmesi kritik önem taşımaktadır. Bu nedenle çalışmada, sıralı veri yapılarında etkin performans gösteren Long Short-Term Memory (LSTM) tabanlı bir derin öğrenme mimarisi tercih edilmiştir.

LSTM, sıralı veri ve zaman serisi analizlerinde yaygın olarak kullanılan bir tür tekrarlayan sinir ağıdır (RNN). Geleneksel RNN'lerin uzun dönemli bağımlılıkları öğrenme konusundaki zorluklarını aşmak amacıyla geliştirilmiştir. LSTM, geçmiş bilgiyi daha uzun süre koruyabilmek için üç ana kapı mekanizması kullanır. Bunlar unutma kapısı, giriş kapısı ve çıkış kapısıdır. Bu kapılar sayesinde, LSTM hem kısa hem de uzun vadeli bağımlılıkları etkili bir şekilde öğrenir, bu da onu dil modelleme ve tahmin görevlerinde güçlü kılar [33].

Çalışmada 3 farklı metin gömme yönteminin performanslarını ölçmek amacıyla 3 farklı model kurulmuş, tüm modellerde 2 katmanlı LSTM mimarisi kullanılmıştır. Model mimarisinin sabit tutulması, farklı metin gömme yöntemlerinin performanslarını doğrudan karşılaştırmayı mümkün kılmış ve elde edilen sonuçların yalnızca kullanılan metin gömme tekniklerinden kaynaklanan farklılıkları yansıtmasını sağlamıştır. Böylece, metin temsillerinin kredi risk tahmini üzerindeki etkisi daha sağlıklı ve güvenilir bir şekilde analiz edilmiştir.



Şekil 4.2 LSTM algoritması görseli

Şekil 4.2'deki Temel Bileşenler:

1. Unutma Kapısı (Forget Gate, f_t): Geçmiş hücre durumundan (C_{t-1}) ne kadarının unutulacağını belirler.
2. Giriş Kapısı (Input Gate, i_t): Yeni bilgi hücre durumuna (C_t) ne kadar ekleneceğini belirler.
3. Güncelleme Kapısı (Update Gate, C_t): Unutma kapısı ve giriş kapısından gelen bilgilere göre bellek ünitesini günceller.
4. Çıkış Kapısı (Output Gate, O_t): Yeni hücre durumundan (C_t) ne kadarının çıkışa (H_t) aktarılacağını belirler.
5. Aktivasyon Fonksiyonları (Activation Functions): Sigmoid (σ) ve tanh fonksiyonları hücredeki bilgiyi işleyerek çıkışı belirler [65].

Bu çalışma kapsamında, üç farklı metin gömme yöntemi için aynı mimari yapı korunarak ayrı ayrı modeller kurulmuştur. Tüm modellerde iki katmanlı LSTM

mimarisi kullanılmış, böylece metin gömme yöntemleri arasındaki performans farklarının model yapısından değil, yalnızca metin temsillerinden kaynaklanması sağlanmıştır. İki katmanlı LSTM yapısı, ilk katmanda temel dilsel örüntülerin, ikinci katmanda ise daha soyut ve üst düzey temsillerin öğrenilmesine olanak tanımıştır. Bu yaklaşım, kredi riski tahmini probleminde hem doğruluk hem de genelleme performansını artırmıştır.

LSTM tabanlı metin temsili, nihai sınıflandırma aşamasında kullanılan tam bağlantılı (dense) katmanlara girdi sağlamaktadır. Bu katmanlar, metin temsili üzerinden müşterinin ödeme davranışına ilişkin tahmin üretmektedir.

Modelin çıktı katmanı, ikili sınıflandırma problemine uygun şekilde yapılandırılmıştır. Model eğitimi sırasında, tahmin edilen sınıf olasılıkları ile gerçek etiketler arasındaki farkı minimize etmeyi amaçlayan bir kayıp fonksiyonu kullanılmıştır. Böylece model, çağrı merkezi konuşmalarından elde edilen metin temsillerini, ödeme davranışını ayırt edecek biçimde öğrenmektedir.

Modelin tüm bileşenleri birlikte eğitilmekte olup, metin temsili üretimi ile sınıflandırma süreci arasında açık bir ayrım bulunmamaktadır. Bu uçtan uca yapı, metin verisinin hedef değişken ile doğrudan ilişkilendirilmesini sağlamaktadır.

Bu bölümde açıklanan metin temsil ve modelleme çerçevesi, bir sonraki bölümde sunulacak deneysel değerlendirmelerin temelini oluşturmaktadır. Model performansının nasıl ölçüldüğü, hangi değerlendirme metriklerinin kullanıldığı ve elde edilen sonuçların nasıl yorumlandığı, ilgili bölümde ayrıntılı olarak ele alınacaktır.

4.4 Performans Değerlendirme Kriterleri

Bu çalışma, müşterilerin çağrı merkezi telefon görüşmelerinde kullandıkları cümleler üzerinden verdikleri ödeme sözlerini yerine getirip getirmeyeceklerinin tahmin edilebileceğini ortaya koymaktadır. Modelin sınıflandırma performansı, hedef değişkenin dengeli dağılmadığı durumlar da dikkate alınarak doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ve F1 skoru gibi standart performans ölçütleri kullanılarak değerlendirilmiştir. Bu metrikler, modelin tahmin başarımının farklı yönlerden kapsamlı bir şekilde analiz edilmesini sağlamaktadır.

Doğruluk, model tarafından yapılan tüm tahminler içerisinde doğru sınıflandırılan gözlemlerin oranını ifade ederken, kesinlik, pozitif olarak sınıflandırılan gözlemler içerisindeki gerçek pozitiflerin oranını göstermektedir. Duyarlılık, modelin gerçek pozitif sınıfı doğru biçimde tespit edebilme yeteneğini ölçerken, F1 skoru, kesinlik ve duyarlılık ölçütlerinin harmonik ortalaması olarak, özellikle sınıf dağılımının dengesiz olduğu problemlerde daha dengeli bir performans değerlendirmesi sunmaktadır [66][67]. İlgili metrikler Denklem (1)–(4)'te tanımlanan ifadeler kullanılarak hesaplanmıştır.

$$\text{Doğruluk Oranı} = \frac{\text{Doğru tahmin sayısı}}{\text{Toplam tahmin sayısı}} \tag{4.1}$$

$$\text{Kesinlik (PR)} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i+FP_i} \tag{4.2}$$

N sınıf sayısı, TP_i i. sınıf için doğru pozitif sayısı ve FP_i ise i. sınıf için yanlış pozitif sayısıdır.

$$\text{Duyarlılık (RE)} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i+FN_i} \tag{4.3}$$

FN_i, i sınıfı için yanlış negatiflerin sayısıdır.

$$\text{F1-Score} = \frac{1}{N}\sum_{i=1}^{N}\frac{2.PR_i.RE_i}{PR_i+RE_i} \tag{4.4}$$

PR_i ve RE_i, sırasıyla i sınıfı için kesinlik ve duyarlılık oranlarını ifade etmektedir.

Buna ek olarak, sınıflandırıcının farklı hata türlerini ne ölçüde ayırt edebildiğini daha ayrıntılı biçimde incelemek amacıyla karışıklık matrisleri oluşturulmuştur. Karışıklık matrisi, doğru ve yanlış sınıflandırmaların dağılımını görsel olarak sunarak, modelin güçlü yönlerinin ve iyileştirilmesi gereken alanlarının daha net bir şekilde anlaşılmasına olanak tanımaktadır [68].

5
SONUÇ

Bu çalışmada, kredi riski tahmininde iki katmanlı bir LSTM mimarisi kullanılmıştır. Bir LSTM'de, her katman, modelin girdi verilerini işleyerek belirli korelasyonları öğrenmesini sağlayan hesaplama birimlerinden oluşur. Her katman, zaman serisi verilerindeki ardışık bağımlılıkları yakalar ve bu bilgiyi bir sonraki katmana iletir.

İki katmanlı bir LSTM, modelin daha derin ve karmaşık kalıpları öğrenmesini sağlar. Birinci katman, girdi verilerinden temel kalıpları çıkarırken, ikinci katman bu temel bilgiyi daha üst düzey temsiller haline dönüştürür. Bu yapı, özellikle dilsel verilerdeki ardışık bağımlılıkları yakalamada önemli bir avantaj sağlar. Bu nedenle, model kredi riski tahmininde daha yüksek doğruluk oranlarına ulaşmış ve genelleme yeteneğini artırmıştır.

LSTM modelinin tam hiperparametre konfigürasyonu Tablo 5.1'de sunulmuştur. Tüm deneylerde, mimari sabit tutulmuş ve yalnızca karşılaştırma amacıyla gömme yöntemleri değiştirilmiştir.

Tablo 5.1 LSTM modelinin yapılandırması ve eğitim ayarları

Bileşen	Değer / Ayar
Tokenizer	Kelime seviyesinde, eğitim verisi üzerinde eğitildi (fit_on_texts)
Sözlük Boyutu	Eğitim verisine dayalı (açıkça sınırlandırılmamıştır)
Dizi Uzunluğu	Maksimum uzunluk = 127 token (padding = post, truncation = post)
Embedding Katmanı	Embedding(input_dim = tokenizer.word_index + 1, output_dim = 300, input_length = 127, trainable = True)
Tekrarlayan Katmanlar	1. Katman: Bidirectional LSTM, 64 birim, return_sequences = True
	2. Katman: LSTM, 64 birim
Dropout	0.5 (ilk LSTM katmanından sonra uygulanmıştır)
Yoğun (Dense) Katman	1 birim, aktivasyon = sigmoid
Kayıp Fonksiyonu	Binary cross-entropy
Optimizasyon Algoritması	Adam

Tablo 5.1 LSTM modelinin yapılandırması ve eğitim ayarları (devamı)

Değerlendirme Metrikleri	Accuracy, Precision, Recall, F1 Skoru, ROC Eğrisi, AUC Skoru
Batch Size	32
Epoch Sayısı (Maks.)	25 (erken durdurma ile)
Erken Durdurma	monitor = val_loss, patience = 5, restore_best_weights = True
Doğrulama Yapısı	validation_data = (test_x, test_y) eğitim sırasında kullanılmıştır
Donanım	40 CPU çekirdeği, GPU kullanılmamıştır
Yazılım	Python 3.11

Model performanslarının değerlendirilmesi sonucunda, test veri seti üzerinde %79,54 doğruluk oranı elde edilmiştir. Modelin sınıflandırma başarımına ilişkin detaylı performans ölçütleri Tablo 5.2 ve Tablo 5.3'te gösterilmektedir. Tablolar incelendiğinde, Self Embedding tabanlı iki katmanlı LSTM modelinin, FastText ve BERT tabanlı modellere kıyasla hem ödeme taahhüdünü yerine getiren müşteri **(**target = 1**)** hem de ödeme taahhüdünü yerine getirmeyen müşteri **(**target = 0**)** sınıfları için daha dengeli ve tutarlı sonuçlar ürettiği gözlemlenmiştir.

Tablo 5.2 Test verilerinin performansını gösteren karışıklık matrisi

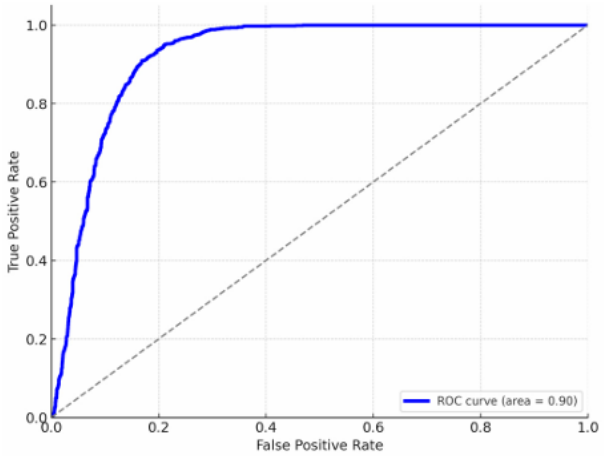
		Self Embedding		**FastText**		**BERT**	
		Tahmin Edilen		**Tahmin Edilen**		**Tahmin Edilen**	
		Negatif	**Pozitif**	**Negatif**	**Pozitif**	**Negatif**	**Pozitif**
Gerçekleşen	**Negatif**	42,69%	7,37%	43,95%	5,95%	0,00%	49,90%
	Pozitif	13,09%	36,85%	19,88%	30,23%	0,00%	50,10%

Tablo 5.3 Eğitim ve test verilerini karşılaştırmak için performans tablosu

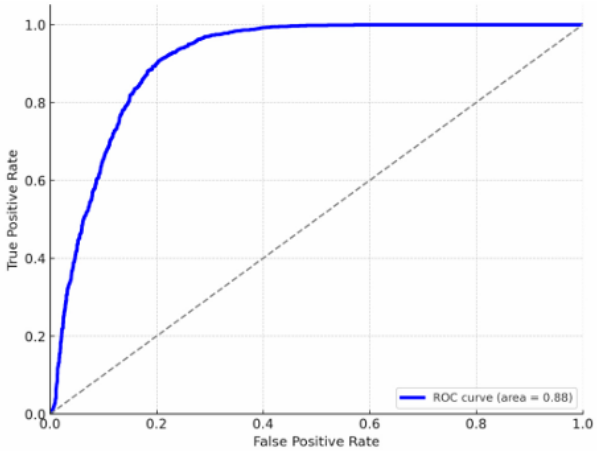
	Self Embedding		**FastText**		**BERT**	
	Eğitim	**Test**	**Eğitim**	**Test**	**Eğitim**	**Test**
Doğruluk	87,62%	79,54%	73,42%	74,17%	49,93%	50,10%
Kesinlik	89,24%	83,33%	86,11%	83,55%	49,93%	50,10%
Duyarlılık	85,57%	73,79%	55,76%	60,33%	100,00%	100,00%
F1 Skoru	87,37%	78,27%	67,69%	70,07%	66,60%	66,76%

Modellerin ayırt edici gücünü daha kapsamlı bir şekilde değerlendirmek için, en iyi performans gösteren iki mimari (Self Embedding LSTM ve FastText tabanlı LSTM) için ROC eğrileri ve karşılık gelen AUC puanları sunulmuştur. Bu sonuçlar, yukarıda sunulan temel sınıflandırma ölçütlerinin ötesinde, modellerin performansına ilişkin daha bütünsel bir değerlendirme sağlar.

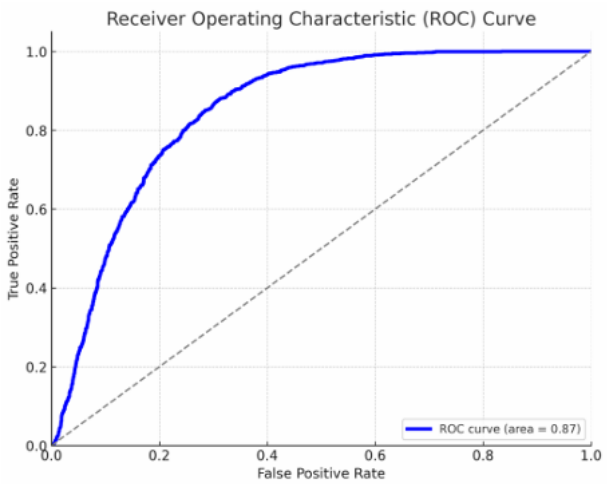
Tablo 5.2 ve 5.3 ile Şekil 5.1-5.4 incelendiğinde, Self Embedding tabanlı 2 katmanlı LSTM modelinin, FastText ve BERT tabanlı modellerden daha iyi sonuçlar verdiği ve hem hedef değişeken 1 hem de hedef değişken 0 için daha dengeli sonuçlar elde ettiği gözlemlenmiştir.



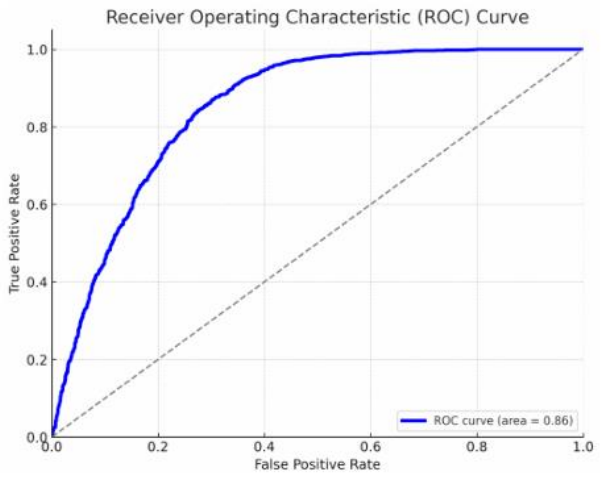
Şekil 5.1 Eğitim veri kümesi üzerinde Self Embedding yöntemiyle eğitilmiş LSTM modelinin ROC eğrisi



Şekil 5.2 Test veri kümesi üzerinde Self Embedding yöntemiyle eğitilmiş LSTM modelinin ROC eğrisi



Şekil 5.3 Train veri kümesi üzerinde FastText yöntemiyle eğitilmiş LSTM modelinin ROC eğrisi



Şekil 5.4 Test veri kümesi üzerinde FastText yöntemiyle eğitilmiş LSTM modelinin ROC eğrisi

FastText tabanlı model, özellikle yazım hataları ve morfolojik çeşitlilik içeren kelimelerde dayanıklılık göstermiş olsa da, bağlamdan bağımsız yapısı nedeniyle sınıflar arasındaki ayrımı sınırlı ölçüde yakalayabilmiştir. Buna karşın, Self Embedding yaklaşımı, veri setine özgü dilsel kalıpları öğrenebilmesi sayesinde daha dengeli bir sınıflandırma performansı sergilemiştir.

BERT tabanlı modelde ise, tahmin dağılımında belirgin bir dengesizlik oluştuğu ve modelin çoğunlukla target sınıfı "1" yönünde tahmin üretme eğilimi gösterdiği tespit edilmiştir. Bu durum, Tablo 5.2 ve Tablo 5.3'te yer alan performans

metriklerinde açıkça gözlemlenmektedir. Söz konusu dengesizliğin, BERT'in önceden eğitilmiş büyük ölçekli bir dil modeli olması ve sınırlı, alana özel veri üzerinde yapılan modelleme sürecinde sınıf dağılımına yeterince adapte olamaması ile ilişkili olabileceği değerlendirilmektedir. Özellikle eğitim sürecinde baskın olan sınıfın özelliklerinin daha güçlü öğrenilmesi, modelin diğer sınıfa karşı önyargı geliştirmesine yol açmış olabilir.

Bu çalışmada elde edilen bulgular, çağrı merkezi telefon görüşmelerinden elde edilen metin verilerinin kredi riski tahmini probleminde anlamlı ve ayırt edici bilgiler sunduğunu göstermektedir. Telefon görüşmelerine dayalı metinler, yazılı metinlere kıyasla daha düzensiz, eksik ve gündelik dile yakın bir yapıya sahiptir. Bu özellikler, metin temsili ve modelleme sürecinde kullanılan yöntemlerin performansını doğrudan etkilemektedir.

Elde edilen sonuçlar, konuşma temelli metin verilerinde veri setine özgü olarak öğrenilen kelime temsillerinin daha avantajlı olduğunu göstermektedir. Özellikle Self Embedding tabanlı iki katmanlı LSTM modelinin, FastText ve BERT tabanlı modellere kıyasla daha dengeli ve tutarlı sınıflandırma sonuçları üretmesi, veri doğasına uyumlu temsil öğreniminin önemini vurgulamaktadır. Çağrı merkezi konuşmalarında yer alan eksik cümleler, tekrarlar, konuşma diline özgü ifadeler ve bağlamsal anlamlar, genel amaçlı ve önceden eğitilmiş dil modelleri için ek zorluklar yaratabilmektedir.

FastText yöntemi, kelimeleri alt birimlerine ayırarak temsil etmesi sayesinde Türkçe'nin morfolojik yapısına uyum sağlamış ve yazım hatalarına karşı belirli bir dayanıklılık göstermiştir. Bununla birlikte, bağlamdan bağımsız yapısı nedeniyle kelimelerin cümle içerisindeki anlam farklılıklarını yeterince ayırt edememiştir. BERT tabanlı model ise bağlama duyarlı güçlü bir temsil kapasitesine sahip olmasına rağmen, sınırlı ve alana özel veri üzerinde yapılan modelleme sürecinde tahmin dağılımında dengesizlik göstermiştir. Bu durum, büyük ve önceden eğitilmiş modellerin, veri miktarı ve sınıf dağılımı gibi koşullara duyarlılığını bir kez daha ortaya koymaktadır.

Modelleme sürecinde yalnızca sınıflandırma doğruluğu değil, model çıktılarının yorumlanabilirliği de önemlidir. Bu kapsamda, ödeme taahhüdünü yerine getiren ve getirmeyen müşterilerin kullandıkları dilsel kalıplar incelenmiş, müşterilerin

ödeme davranışları ile kullandıkları dil arasında anlamlı ilişkiler olduğu tespit edilmiştir. Ödeme güçlüğü yaşayan müşterilerin ifadelerinde kişisel ve ailevi sorunlar, finansal zorluklar, hukuki problemler ve güvenilirlik vurgusu içeren söylemler öne çıkmaktadır. Buna karşın, ödeme yapma olasılığı yüksek müşterilerin daha net, kesin ve olumlu ifadeler kullandıkları gözlemlenmiştir.

Bu bulgular, çağrı merkezi konuşmalarının yalnızca operasyonel süreçler için değil, aynı zamanda risk yönetimi ve müşteri iletişim stratejilerinin geliştirilmesi açısından da önemli bir bilgi kaynağı olduğunu göstermektedir. Dil analizi yoluyla elde edilen bu içgörüler, çağrı merkezi süreçlerinin iyileştirilmesi, müşteri segmentasyonunun zenginleştirilmesi ve daha kişiselleştirilmiş iletişim stratejilerinin geliştirilmesi için kullanılabilir.

Çalışmanın önemli bir bileşeni olan veri temizliği ve konuşmadan metne dönüştürme süreci, elde edilen sonuçların kalitesini doğrudan etkilemiştir. Türkçe gibi morfolojik açıdan zengin bir dilde, konuşma temelli metinlerin doğru biçimde temizlenmesi ve yapılandırılması, model performansı açısından kritik bir rol oynamıştır. Düşük kaliteli transkriptler ve yetersiz ön işleme adımları, modelin öğrenme kapasitesini sınırlayabileceğinden, bu süreçlerin dikkatle tasarlanması oldukça önemlidir.

Son olarak, çalışmanın bazı sınırlılıkları bulunmaktadır. Kullanılan veri seti tek bir bankaya ait çağrı merkezi görüşmeleriyle sınırlıdır ve yalnızca Türkçe dilinde test edilmiştir. Bu nedenle elde edilen bulguların farklı bankalar, sektörler veya diller için doğrudan genellenmesi mümkün değildir. Bununla birlikte, çalışma, çağrı merkezi konuşmalarından elde edilen metin verilerinin kredi riski tahmininde kullanılabileceğini göstermesi açısından literatüre anlamlı bir katkı sunmaktadır.

KAYNAKÇA

[1] I. Brown ve C. Mues, “An experimental comparison of classification algorithms for imbalanced credit scoring data sets”, *Expert Syst. Appl.*, c. 39, sy 3, ss. 3446-3453, Şub. 2012, doi: 10.1016/j.eswa.2011.09.033.

[2] S. Poledna, A. Hinteregger, ve S. Thurner, “Identifying Systemically Important Companies by Using the Credit Network of an Entire Nation”, *Entropy*, c. 20, sy 10, s. 792, Eki. 2018, doi: 10.3390/e20100792.

[3] L. Barbaglia, S. Manzan, ve E. Tosetti, “Forecasting Loan Default in Europe with Machine Learning”, *J. Financ. Econom.*, c. 21, sy 2, ss. 569-596, Mar. 2023, doi: 10.1093/jjfinec/nbab010.

[4] O. Oyetola, M. Joseph, O. Adenike, A. Omolara, ve A. Godwin, “Credit Risk and its Management in the Banks: A Conceptual Review”.

[5] C. Tas, “A THESIS SUBMITTED TO THE GRADUATE SCHOOL OF INFORMATICS OF THE MIDDLE EAST TECHNICAL UNIVERSITY”.

[6] J. Alvi, I. Arif, ve K. Nizam, “Advancing financial resilience: A systematic review of default prediction models and future directions in credit risk management”, *Heliyon*, c. 10, sy 21, s. e39770, Kas. 2024, doi: 10.1016/j.heliyon.2024.e39770.

[7] G. Yan, “Research on the Application of Alternative Data in Credit Risk Management”, *Highlights Bus. Econ. Manag.*, c. 40, ss. 1156-1160, Eyl. 2024, doi: 10.54097/vn32pp64.

[8] Y. Sun, L. Liu, Y. Xu, X. Zeng, ve Y. Shi, “A Survey on Alternative Data in Finance and Business: Emerging Applications and Theory Analysis”, *SSRN Electron. J.*, 2023, doi: 10.2139/ssrn.4148628.

[9] The World Bank, The Use of Alternative Data in Credit Risk Assessment: Opportunities, Risks, and Challenges, Washington, DC, USA: World Bank, 2024.

[10] D. J. Hand and W. E. Henley, "Statistical classification methods in consumer credit scoring: A review," Journal of the Royal Statistical Society: Series A, vol. 160, no. 3, pp. 523–541, 1997.

[11] Basel Committee on Banking Supervision, International Convergence of Capital Measurement and Capital Standards: A Revised Framework (Comprehensive Version), Bank for International Settlements, Jun. 2006.

[12] Basel Committee on Banking Supervision, Guidance on Credit Risk and Accounting for Expected Credit Losses, Bank for International Settlements, Dec. 2015.

[13] A. Fuster, P. Goldsmith-Pinkham, T. Ramadorai, ve A. Walther, “Predictably Unequal? The Effects of Machine Learning on Credit Markets”, *J. Finance*, c. 77, sy 1, ss. 5-47, Şub. 2022, doi: 10.1111/jofi.13090.

[14] S. Mullainathan ve J. Spiess, “Machine Learning: An Applied Econometric Approach”, *J. Econ. Perspect.*, c. 31, sy 2, ss. 87-106, May. 2017, doi: 10.1257/jep.31.2.87.

[15] T. Berg, “On the Rise of FinTechs: Credit Scoring Using Digital Footprints”.

[16] S. Lessmann, B. Baesens, H.-V. Seow, ve L. C. Thomas, “Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research”, *Eur. J. Oper. Res.*, c. 247, sy 1, ss. 124-136, Kas. 2015, doi: 10.1016/j.ejor.2015.05.030.

[17] A. E. Khandani, A. J. Kim, ve A. W. Lo, “Consumer credit-risk models via machine-learning algorithms”, *J. Bank. Finance*, c. 34, sy 11, ss. 2767-2787, Kas. 2010, doi: 10.1016/j.jbankfin.2010.06.001.

[18] P. Addo, D. Guegan, ve B. Hassani, “Credit Risk Analysis Using Machine and Deep Learning Models”, *Risks*, c. 6, sy 2, s. 38, Nis. 2018, doi: 10.3390/risks6020038.

[19] Z. Aksin, M. Armony, ve V. Mehrotra, “The Modern Call Center: A Multi-Disciplinary Perspective on Operations Management Research”, *Prod. Oper. Manag.*, c. 16, sy 6, ss. 665-688, Kas. 2007, doi: 10.1111/j.1937-5956.2007.tb00288.x.

[20] C. Sánchez, S. Maldonado, ve C. Vairetti, “Improving debt collection via contact center information: A predictive analytics framework”, *Decis. Support Syst.*, c. 159, s. 113812, Ağu. 2022, doi: 10.1016/j.dss.2022.113812.

[21] Z. Huang, Y. Yang, C. Liao, ve P. Du, “How to say? Voice analytics of debt collection strategies”, *Manag. Decis. Econ.*, c. 43, sy 4, ss. 1091-1104, Haz. 2022, doi: 10.1002/mde.3441.

[22] J. Kim ve P. Kang, “Late payment prediction models for fair allocation of customer contact lists to call center agents”, *Decis. Support Syst.*, c. 85, ss. 84-101, May. 2016, doi: 10.1016/j.dss.2016.03.002.

[23] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed., Jan. 2026.

[24] G. Hinton *vd.*, “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups”, *IEEE Signal Process. Mag.*, c. 29, sy 6, ss. 82-97, Kas. 2012, doi: 10.1109/MSP.2012.2205597.

[25] A. Graves, A. Mohamed, ve G. Hinton, “Speech recognition with deep recurrent neural networks”, içinde *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada: IEEE, May. 2013, ss. 6645-6649. doi: 10.1109/ICASSP.2013.6638947.

[26] E. Shriberg, “To ‘errrr’ is human: ecology and acoustics of speech disfluencies”, *J. Int. Phon. Assoc.*, c. 31, sy 1, ss. 153-169, Haz. 2001, doi: 10.1017/S0025100301001128.

[27] Basel Committee on Banking Supervision, Calculation of RWA for credit risk: CRE32 – IRB approach: risk components, Bank for International Settlements, Jan. 2023.

[28] Basel Committee on Banking Supervision, Calculation of RWA for credit risk: CRE36 – IRB approach: minimum requirements to use IRB approach, Bank for International Settlements, Jan. 2023.

[29] E. I. Altman, "FINANCIAL RATIOS, DISCRIMINANT ANALYSIS AND THE PREDICTION OF CORPORATE BANKRUPTCY", *J. Finance*, c. 23, sy 4, ss. 589-609, Eyl. 1968, doi: 10.1111/j.1540-6261.1968.tb00843.x.

[30] J. A. Ohlson, "Financial Ratios and the Probabilistic Prediction of Bankruptcy", *J. Account. Res.*, c. 18, sy 1, s. 109, 1980, doi: 10.2307/2490395.

[31] E. R. Muratlar, D. Yildiz, ve E. Ustaoglu, "Turkish Telephone Conversations in Credit Risk Management: Natural Language Processing and LSTM Approach", *Appl. Sci.*, c. 16, sy 1, s. 108, Ara. 2025, doi: 10.3390/app16010108.

[32] A. Graves, "Generating Sequences With Recurrent Neural Networks", 05 Haziran 2014, *arXiv*: arXiv:1308.0850. doi: 10.48550/arXiv.1308.0850.

[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[34] J. Sirignano, A. Sadhwani, ve K. Giesecke, "Deep Learning for Mortgage Risk", 10 Mart 2018, *arXiv*: arXiv:1607.02470. doi: 10.48550/arXiv.1607.02470.

[35] J. Chen, A. L. Katchova, ve C. Zhou, "Agricultural loan delinquency prediction using machine learning methods", *Int. Food Agribus. Manag. Rev.*, c. 24, sy 5, ss. 797-812, Tem. 2021, doi: 10.22434/IFAMR2020.0019.

[36] O. Netzer, A. Lemaire, ve M. Herzenstein, "When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications", *J. Mark. Res.*, c. 56, sy 6, ss. 960-980, Ara. 2019, doi: 10.1177/0022243719852959.

[37] J. Gao, W. Sun, ve X. Sui, "Research on Default Prediction for Credit Card Users Based on XGBoost-LSTM Model", *Discrete Dyn. Nat. Soc.*, c. 2021, ss. 1-13, Ara. 2021, doi: 10.1155/2021/5080472.

[38] T. Loughran ve B. Mcdonald, "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks", *J. Finance*, c. 66, sy 1, ss. 35-65, Şub. 2011, doi: 10.1111/j.1540-6261.2010.01625.x.

[39] X. Fu, T. Ouyang, J. Chen, ve X. Luo, "Listening to the investors: A novel framework for online lending default prediction using deep learning neural networks", *Inf. Process. Manag.*, c. 57, sy 4, s. 102236, Tem. 2020, doi: 10.1016/j.ipm.2020.102236.

[40] V. B. Djeundje, J. Crook, R. Calabrese, ve M. Hamid, "Enhancing credit scoring with alternative data", *Expert Syst. Appl.*, c. 163, s. 113766, Oca. 2021, doi: 10.1016/j.eswa.2020.113766.

[41] R. Hlongwane, K. K. K. M. Ramaboa, ve W. Mongwe, "Enhancing credit scoring accuracy with a comprehensive evaluation of alternative data", *PLOS ONE*, c. 19, sy 5, s. e0303566, May. 2024, doi: 10.1371/journal.pone.0303566.

[42] T. Bradford, "'Give Me Some Credit!': Using Alternative Data to Expand Credit", 2023.

[43] G. Salton ve C. Buckley, "Term-weighting approaches in automatic text retrieval", *Inf. Process. Manag.*, c. 24, sy 5, ss. 513-523, Oca. 1988, doi: 10.1016/0306-4573(88)90021-0.

[44] T. Mikolov, K. Chen, G. Corrado, ve J. Dean, "Efficient Estimation of Word Representations in Vector Space", 07 Eylül 2013, *arXiv*: arXiv:1301.3781. doi: 10.48550/arXiv.1301.3781.

[45] J. Pennington, R. Socher, ve C. Manning, "Glove: Global Vectors for Word Representation", içinde *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, 2014, ss. 1532-1543. doi: 10.3115/v1/D14-1162.

[46] A. Vaswani *vd.*, "Attention Is All You Need", 02 Ağustos 2023, *arXiv*: arXiv:1706.03762. doi: 10.48550/arXiv.1706.03762.

[47] J. Devlin, M.-W. Chang, K. Lee, ve K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", 24 Mayıs 2019, *arXiv*: arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805.

[48] F. Haag, S. Zürcher, ve J. Lienert, "Enhancing the elicitation of diverse decision objectives for public planning", *Eur. J. Oper. Res.*, c. 279, sy 3, ss. 912-928, Ara. 2019, doi: 10.1016/j.ejor.2019.06.002.

[49] A. Graves ve N. Jaitly, "Towards End-to-End Speech Recognitionwith Recurrent Neural Networks".

[50] L. Dong, S. Xu, ve B. Xu, "Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition", içinde *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB: IEEE, Nis. 2018, ss. 5884-5888. doi: 10.1109/ICASSP.2018.8462506.

[51] Y. Yang, Y. Qin, Y. Fan, ve Z. Zhang, "Unlocking the Power of Voice for Financial Risk Prediction: A Theory-Driven Deep Learning Design Approach", *MIS Q.*, c. 47, sy 1, ss. 63-96, Mar. 2023, doi: 10.25300/MISQ/2022/17062.

[52] R. Southwell *vd.*, "Challenges and Feasibility of Automatic Speech Recognition for Modeling Student Collaborative Discourse in Classrooms".

[53] S. Alharbi *vd.*, "Automatic Speech Recognition: Systematic Literature Review", *IEEE Access*, c. 9, ss. 131858-131876, 2021, doi: 10.1109/ACCESS.2021.3112535.

[54] Y. Sun *vd.*, "Alternative data in finance and business: emerging applications and theory analysis (review)", *Financ. Innov.*, c. 10, sy 1, s. 127, Ara. 2024, doi: 10.1186/s40854-024-00652-0.

[55] A. Joulin, E. Grave, P. Bojanowski, ve T. Mikolov, "Bag of Tricks for Efficient Text Classification", 09 Ağustos 2016, *arXiv*: arXiv:1607.01759. doi: 10.48550/arXiv.1607.01759.

[56] Q. V. Le ve T. Mikolov, "Distributed Representations of Sentences and Documents", 22 Mayıs 2014, *arXiv*: arXiv:1405.4053. doi: 10.48550/arXiv.1405.4053.

[57] P. Bojanowski, E. Grave, A. Joulin, ve T. Mikolov, "Enriching Word Vectors with Subword Information", *Trans. Assoc. Comput. Linguist.*, c. 5, ss. 135-146, Ara. 2017, doi: 10.1162/tacl_a_00051.

[58] N. Reimers ve I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", 27 Ağustos 2019, *arXiv*: arXiv:1908.10084. doi: 10.48550/arXiv.1908.10084.

[59] X. Dastile, T. Celik, ve M. Potsane, "Statistical and machine learning models in credit scoring: A systematic literature survey", *Appl. Soft Comput.*, c. 91, s. 106263, Haz. 2020, doi: 10.1016/j.asoc.2020.106263.

[60] J. Kriebel ve L. Stitz, "Credit default prediction from user-generated text in peer-to-peer lending using deep learning", *Eur. J. Oper. Res.*, c. 302, sy 1, ss. 309-323, Eki. 2022, doi: 10.1016/j.ejor.2021.12.024.

[61] M. Sanz-Guerrero ve J. Arroyo, "Credit Risk Meets Large Language Models: Building a Risk Indicator from Loan Descriptions in P2P Lending", *Intel. Artif.*, c. 28, sy 75, ss. 220-247, Haz. 2025, doi: 10.4114/intartif.vol28iss75pp220-247.

[62] European Banking Authority, *Guidelines on PD Estimation, LGD Estimation and the Treatment of Defaulted Exposures (EBA/GL/2017/16)*, Nov. 2017.

[63] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed., Jan. 2026.

[64] M. E. Peters *vd.*, "Deep contextualized word representations", 22 Mart 2018, *arXiv*: arXiv:1802.05365. doi: 10.48550/arXiv.1802.05365.

[65] K. Li ve X. Chen, "Machine Learning-Based Lithium Battery State of Health Prediction Research", *Appl. Sci.*, c. 15, sy 2, s. 516, Oca. 2025, doi: 10.3390/app15020516.

[66] K. Górny, P. Kuwałek, ve W. Pietrowski, "Increasing Electric Vehicles Reliability by Non-Invasive Diagnosis of Motor Winding Faults", *Energies*, c. 14, sy 9, s. 2510, Nis. 2021, doi: 10.3390/en14092510.

[67] D. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation".

[68] T. Fawcett, "An introduction to ROC analysis", *Pattern Recognit. Lett.*, c. 27, sy 8, ss. 861-874, Haz. 2006, doi: 10.1016/j.patrec.2005.10.010.

TEZDEN ÜRETİLMİŞ YAYINLAR

Makale

1. E. R. Muratlar, D. Yildiz, ve E. Ustaoglu, “Turkish Telephone Conversations in Credit Risk Management: Natural Language Processing and LSTM Approach”, Appl. Sci., c. 16, sy 1, s. 108, Ara. 2025, doi: 10.3390/app16010108. (SCI Expanded)

Konferans Bildirisi

1. E. R. Muratlar, D. Yildiz. “E-Ticarette Alternatif Metin Kümeleme Yöntemleri Üzerine Bir Uygulama” 7th International Researchers, Statisticians, and Young Statisticians Congress (IRSYSC2023), İstanbul, Türkiye, 02 - 05 Kasım 2023, ss.59