Makine Öğrenmesi Yaklaşımı ile Acente Kredi Riski Hesaplama

Araştırma Makalesi/Research Article

Serkan KIRCA[1], Güldeniz CANATAN[2,3], Vedat GÜNEŞ[4,5]

[1]Anadolu Anonim Türk Sigorta Şirketi, İstanbul, Türkiye
[2]İstatistik, Çukurova Üniversitesi, Adana, Türkiye
[3]Anadolu Anonim Türk Sigorta Şirketi, İstanbul, Türkiye
[4] Anadolu Anonim Türk Sigorta Şirketi, İstanbul, Türkiye
[5] Elektronik ve Bilgisayar Mühendisliği, Altınbaş Üniversitesi, İstanbul, Türkiye
skirca@anadolusigorta.com.tr, gcanatan@anadolusigorta.com.tr, vgunes@anadolusigorta.com.tr
(Geliş/Received:24.12.2021; Kabul/Accepted:18.02.2022)
DOI: 10.17671/gazibtd.1039963

Özet— Ekonomik trendlerin hızla dijital süreçlere uyum sağladığı günümüzde finans sektöründe risk analizi ilgili kurumların sahip oldukları değerler, birlikte çalıştıkları paydaşlar ve faaliyet gösterdikleri sektörler için büyük önem arz etmektedir. Risk analizi herhangi bir anda müşterinin ya da çalışılan firmanın risk potansiyelini ölçümleyebilmek adına son derece önemli bir çalışma olup, sadece finans sektöründe değil, diğer sektörlerde faaliyet gösteren şirketlerin de değerlendirmesi gereken bir çalışma haline gelmiştir. Riskini önden bilme gayreti şirket finansal durumunu doğru yönetmek adına emek yoğun bir çaba haline dönüşmüştür. Bununla birlikte veri merkezli çalışmaların arttığı günümüzde risk değerlendirme çalışmaları da veri analitiği teknolojileri ile zenginleşmekte ve eskiye nazaran daha iyi sonuçlar üretmeye başlamışlardır. Geniş bir ekosistem ile müşterilerine hizmet vermeye gayret eden sigorta şirketleri risk konusunda iki temel alanda; müşterinin ve kendisine bağlı faaliyet gösteren acentelerin, risk değerlerini hesaplamak için veri merkezli analitik süreçler geliştirmektedir. Sigorta şirketleri için risk analizleri son derece önemli olup, finansal riski belirlenen müşteri ya da acenteye göre iş akışlarında kurallar tanımlamakta veya yaptırımlara varan aksiyonlar alabilmektedirler. Acente ağı sigortacılık sektörü için gelirlerin %70'i anlamına gelmektedir. Yapılan hesaplamalar sonucunda her bir acentenin sigorta şirketi için ifade ettiği finansal risk değeri tahminlenmiştir. Çok riskli bulunan acenteler bütün portföyün yaklaşık %0,2'si kadardır. Bu çalışmada acente risk değerlendirme süreci makine öğrenmesi yöntemleri kullanılarak hesaplanmıştır ve uygulanan faaliyetler paylaşılacaktır.

Anahtar Kelimeler— acenteler, acente riski, anadolu sigorta, finansal risk, risk tahminleme

Agency Credit Risk Calculation with Machine Learning Method

Abstract—Financial risk analysis is evolving according to the fast moving economical trends and it is getting more important for companies, ecosystem, and stakeholders. Risk prediction algorithms help companies to estimate their potential and define mitigation activities. Insurance is a risk centered topic in finance and risk managment is the basic activity to make profit. There are two main stakeholders in insurance; cusotmers and distributors. Distributors communicate with customers on behalf of insurance company for underwriting and claim processes. Distributors generate more than 70% revenue in total for insurance companies. Mainly, distributors collect money from customers and keep it for a while and then transfer to the company account. Anadolu Insurance have almost 5.000 distributors. This means more than 7 billion revenue generation in 2021. In this study, we developed a data analytics flow to predict the distributors which have high financial risk and define mitigation activities in order to handle the risk, identified. Also, it is a critical information for Anadolu Insurance to prevent the customers who are managed by the distributor, has high risk potential. Distributors can damage customer relations while they are trying to manage financials. In the flow we developed, we will share details about how Anadolu Insurance predicted financial risk value for each distributor by using machine learning algorithms. We identified 0.2% of the distributors have higher financial risk and Anadolu Insurance will try to handle this by not harm its customers.

Keywords— agency, agency risk, anadolu insurance, financial risk, risk prediction

1. GİRİŞ (INTRODUCTION)

Finans şirketleri, müşterilerine ürün satmanın yanında, müşterilerinin oluşturacak olduğu riskini de hesaplayarak, olası bir geri ödememe durumunu önceden öğrenerek, müşteriye göre aksiyon alınmasını sağlamaktadırlar [1].

Günümüzde sigorta şirketleri müşterilerine, poliçe yapma ya da sadece ürün (kasko, sağlık, trafik poliçeleri vb.) satmanın yanında, artık poliçelendirdikleri müşterilerinin veya poliçeyi yapan anlaşmalı acentelerinin, hemen hemen bütün finans şirketlerinde yapıldığı gibi risk potensiyellerini de bilmek isterler. Sigorta şirketleri için müşteri riskinin yanında acente riski de değerlendirilen ve hatta daha çok yönetme ihtiyacı duyduğu bir konu başlığıdır.

Sigorta şirketleri, müşterilerine dijital kanallar yoluyla direkt olarak ulaşabilmenin yanında, acente kanalı ile de müşterilerine ulaşmaktadırlar. Acenteler ile sigorta şirketleri arasında yapılan anlaşmalara göre acenteler müşterilere kestikleri poliçelerden tahsil ettikleri ücretleri hemen sigorta şirketinin hesaplarına aktarmamaktadır.

Acenteler sigorta şirketlerine göndermediği poliçe primlerini kendi hesaplarında tutup sigorta şirketinin hesabına aktarmayarak şirket için bir risk meydana getirmektedirler. Sigorta şirketleri alacaklarını yaptıkları anlaşmalara göre acentelerden poliçe ücretlerini tahsil edebileceklerinin planlamalarını yapmaktadırlar. Acentelerin kötü gidişatı, müşteriden aldıkları poliçe ücretini farklı kanallarda kullanmaları (inşaat, borsa, bitcoin vb.) acentelerin sigorta şirketine yapacakları ödemelerde gecikme yaşanmasına hatta bazı zamanlarda ödenememesine sebep olmaktadır. Sigorta şirketleri de bu gibi durumla karşılaşmamak adına acentelerin kendine ödeyecek olduğu tutar üzerinden risk analizleri yapmaktadırlar.

Son yıllarda finans sektörleri yapay zekâ algoritmalarının farkına varmakla kalmayıp [2], yapay zekâ algoritmalarını sistemlerine hızlı bir şekilde entegre ederek yapay zekâ ile elde edilen başarılı sonuçlardan yüksek oranda faydalanmaya başladılar [3]. Risk analizleri de artık yapay zekâ temelli algoritmalar [4] ile desteklenerek işlem hacimlerinde artış, maliyet kısmında azalış trendleri elde edildi. Bu çalışmada da yapay zekâ algoritmalarının gücü kullanılarak finansal risk tahminleme çalışması yapılmıştır [5].

Bu tarz çalışmaların yapılabilmesi için farklı yazılım dillerinin en az birine ihtiyaç vardır (Python [6], KNIME Analitik Platform, R, Julia vb.). Python yazılım dili özellikle açık kaynaklı bir dil olması nedeni ile veri bilimi ve yapay zekâ konularında sıklıkla kullanılan bir yazılım dilidir [7]. KNIME Analitik Platformu ise gerek kullanım kolaylığı gerekse sürükle bırak (node) mantığında olmasından kaynaklı kullanıcıya oldukça kolay kullanım olanağı sağlamaktadır [8]. Bu çalışmada hem python (keşifsel veri analizleri ve görselleştirmeler için) hem de KNIME Analitik Platform kullanılmıştır.

Yapılan çalışmada, 2600 adet acente analiz edilmiş olup, 2017-2021 dönemleri (Her yıl içerisinde 4 dönem olacak şekilde) analiz edilmiştir. Çalışmada 5 farklı grup oluşturularak, her grup acente kodu seviyesinde tekildir. Çalışma çıktısında her bir acentenin (risk seviyesi olarak) tahmin edilen dönemde, sigorta şirketine ne kadar borcu olacağı tespit edilmektedir. Borç tahmini üzerinden bir risk hesaplaması yapılmaktadır (Regresyon) [9]. Tahmin edilen borç tutarları üzerinden de kümeleme (clustering) [10] işlemi yapılarak, acentelerin risklerine göre gruplandırılma çalışması yapılmıştır.

2. MATERYAL ve METOD (MATERIAL and METHOD)

2.1. Verilerin Hazırlanması (Data Preparation)

Makine öğrenmesi algoritmalarının uygulanabilmesi için düzenlenmiş ve amaç için uygun hale getirilmiş veriye ihtiyaç duyulmaktadır [11]. Dolayısıyla verilerin hazırlanması, temizlenmesi ve makine öğrenmesi için uygun hale getirilmesi süreçleri uzun süreli ve uğraş gerektiren adımlardan geçerek yapılır.

Çalışmada veriler yapısal veri kaynaklarında tutulan değişkenlerden oluşmaktadır. Çalışmayla ilgili elde edilen değişkenler farklı tablolardan Structured Query Language (SQL) [12] sorgu dili ile elde edilmiştir.

Veriler her bir acente için her bir dönem bazında tutulmuştur. Örneğin; 201903 ile belirtilen dönemdeki acente verisi, 2019 yılı 3. ay acente verisini ifade etmektedir.

Tablo 1. Dönem ve acente kodu örnek tablo
(Period and Agency Code Sample Table)

Dönem	Acente Kodu
202012	11111
202009	11112
201906	21111
201906	21113

Tablo 1 incelendiğinde; her bir dönem için acente kodu tekil olmaktadır. 3. ve 4. satırlar incelendiğinde, aynı acente farklı dönemlerde gözükmektedir. Çalışmada dönem ve acente kodunun birleşimi (örneğin; 202012-11111) tekil değer olarak (ID) düşünülebilir. ID içeren kolonlar makine öğrenmesi algoritmalarında hataya sebep olduğu için veri setinden çıkartılacaktır [13]. (Acente Kodu örnek olması açısından rastgele oluşturulmuştur.)

2.2. Uygulanan Makine Öğrenmesi Türü (Type of Applied Machine Learning Method)

Makina öğrenmesi algoritma türleri;

- ✓ Denetimli (Supervised) Öğrenme,
  - o Regresyon,
  - o Sınıflandırma,
- ✓ Denetimsiz (Unsupervised) Öğrenme,
  - o Kümeleme,
  - o Boyut Azaltma,

- ✓ Takviyeli – Pekiştirmeli (Reinforcement) Öğrenme,

olmak üzere üçe ayrılmaktadır [14]. Yapılan çalışmada denetimli öğrenme türlerinden olan regresyon ve denetimsiz öğrenme türlerinden kümeleme algoritmaları kullanılmıştır.

Denetimli öğrenme yönteminin bir diğer adı da etiketli öğrenme yöntemidir. Regresyon problemlerinde girdi değişkenlerini (X), çıktı değişkenine (Y) eşleme işlevini öğrenmek için etiketli eğitim verileri kullanılır [15]. Bu makalenin devamında, girdi değişkenleri bağımsız değişken, çıktı değişkeni ise hedef değişken olarak isimlendirilecektir. Ayrıca veri seti Anadolu Sigorta şirketi veri tabanlarında bulunan tablolardan elde edilmiştir.

Veri setinde; yaklaşık 120,000 adet veri, 29 adet bağımsız değişken ve 1 adet hedef değişken bulunmaktadır. Veri seti incelendiğinde kayıp değer (missing value) bulunmamaktadır. Bağımlı değişken sayısal değişken türü olup, bağımsız değişkenlerde, sayısal ve kategorik değişken türleri bulunmaktadır.

Hedef değişken Target olarak isimlendirilmiştir. "Acente model kodu", "Sadece AS", "Plaza acentesi mi", "KSM ile çalışıyor mu" ve "Sirket tipi" kategorik değişkenlerdir. Bunların dışında kalan değişken türleri ise sayısal değişken türleridir.

Sayısal değişken türlerinden olan "Toplam Ödeme Süresi" (TÖS) değişkeni denklem (1) ile elde edilmiştir.

TÖS = Ödeme Süresi + Ek Ödeme Süresi (1)

Türetilen diğer bir sayısal değişken türü de denklem (2) deki gibi elde edilmiştir.

Net Risk = Toplam Borç – (Teminat – Çek Tutarı) (2)

2.3. Keşifsel Veri Analizi (Exploratory Data Analiysis)

Keşifsel veri analizi makine öğrenmesi algoritmalarının en önemli kısmıdır [16]. Bu sebeple keşifsel veri analizi oldukça uğraş gerektiren ve zaman isteyen bir iştir. Bu çalışmada da keşifsel veri analizleri yapılmış olup bu analizler;

- ✓ Yaşlandırma,
- ✓ Yaş hesaplama,
- ✓ Değişken Türetme,
- ✓ Kategorik ve Sayısal değişkenler için özellik mühendisliği (feature engineering) analizleri,
- ✓ Veriyi normalize etme,

yapılmıştır. Yaşlandırma işlemi, geçmiş dönemlerdeki borç tutarlarının günümüzdeki karşılığını yani değerini hesaplayabilmek için yapılmıştır [17]. Yaşlandırma işleminde, dolar kuru, asgari ücret veya TÜFE oranları kullanılmaktadır. Bu çalışmada TÜFE üzerinden yaşlandırma işlemi yapılarak geçmiş dönemdeki borçlar, günümüzdeki değerine getirilmiştir [18].

Yaş hesaplama işleminde ise farklı tarihlerde kurulan acentelerin kuruluş tarihinden günümüze kadar geçen süre (yıl) baz alınarak acente yaşı hesaplanmıştır.

Acente Yaşı = Sistem Tarihi – Acente Kuruluş Tarihi (3)

Değişken türetme işleminde, ödeme süresi ve acente yaşlarına göre kategorik değişken şekil 1'deki gibi türetilmiştir.

```
$ODEME_SURESI$ = 0.0 AND $Acente_Yasi$ <= 10.0 => "zeroyung"
$ODEME_SURESI$ = 0.0 AND ( 10.0 < $Acente_Yasi$ AND $Acente_Yasi$ < 17.0) => "zeromiddle"
$ODEME_SURESI$ = 0.0 AND ( 17.0 <= $Acente_Yasi$) => "zeroold"
(NOT $ODEME_SURESI$ = 0.0) AND $Acente_Yasi$ <= 10.0 => "notzeroyung"
(NOT $ODEME_SURESI$ = 0.0) AND ( 10.0 < $Acente_Yasi$ AND $Acente_Yasi$ < 17.0) => "notzeromiddle"
(NOT $ODEME_SURESI$ = 0.0) AND ( 17.0 <= $Acente_Yasi$) => "notzeroold"
```

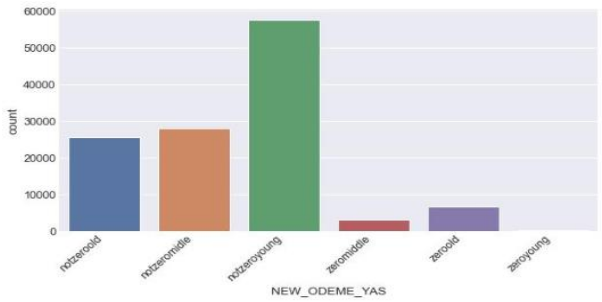
Şekil 1. Ödeme süresi ve acente yaşı değişkenlerini kullanarak değişken türetme
(Variable derivation using payment term and age of agent variables)

Şekil 1 incelendiğinde, ödeme süresi 0 olan ve acente yaşı 10 ve 10 'dan küçük olan verilere "zeroyong" isminde bir veri türetilmiştir. Şekil 1'de görüldüğü üzere yeni değişken 6 farklı kategoriden oluşmaktadır. Oluşturulan bu yeni değişkenin ismi "New Odeme Yas" değişkeni ve türü kategoriktir. "New Odeme Yas" değişkeninin veri seti içerisindeki dağılımı aşağıdaki gibidir.

Tablo 2. Türetilen değişkenin dağılımı
(Distribution of the derived variable)

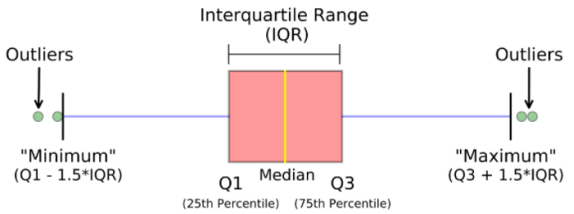
Türetilen Değişkenler	Değişkenlerin Dağılımı
notzeroyoung	57604
notzeromiddle	28065
notzeroold	25472
zeroold	6670
zeromiddle	3052
zeroyoung	149

Tablo 2 incelendiğinde, "zeroyoung" ve "zeromiddle" kategorik verileri, veri setinde çok az bulunduğu görülmektedir. Bu fark şekil 2'de daha açık görülmektedir.



Şekil 2. Türetilen new odeme yaş değişkeninin gösterimi
(Representation of the derived new peyment age variable)

Sigorta şirketleri çok fazla sayıda acenteler ile çalışmaktadır. Çalıştıkları acentelerin içerisinde borcunu ödemeyen acenteler fazla sayıda bulunmadığından dolayı, borcunu ödemeyen acentelerin değerleri üzerinden veri setinde bulunan aykırı değerler (outliers) [19] şekil 3'te gösterildiği gibi tespit edilmiştir.



Şekil 3. Aykırı değer hesaplama [20]
(Outlier calculation)

Tespit edilen aykırı değerlere farklı yöntemler uygulanabilir. Bu çalışmada aykırı değerler veri setinin dağılımına uygun olarak en küçük ve en büyük değere indirgeme işlemi yapılmıştır.

Veri setindeki "KSM ile çalışıyor mu" ve "Plaza acentesi mi" kategorik değişkenleri kullanarak "New Plaza KSM" değişkeni elde edilmiştir. Elde edilen değişkenin verideki dağılımları tablo 3'te gösterilmektedir.

Tablo 3. New plaza KSM değişeninde bulunan verilerin dağılımı
(Distribution of data found in the new plaza KSM variable)

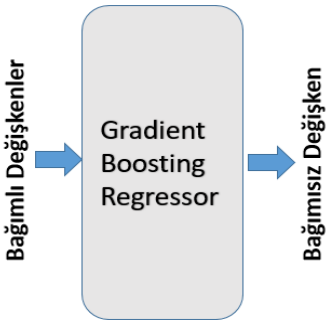
New Plaza KSM	Adet	Frekans
1	105844	0.875
2	10384	0.086
3	4784	0.04

Elde edilen New Plaza KSM değişkeni, tablo 3'teki frekanslara bakıldığında 3 kategori elde edilmiştir. Bu kategorilerde frekanslarına göre kendi içlerinde birleştirilerek New Plaza KSM değişkeni iki kategoriden oluşan bir değişken olacak şekilde (1 ve (2-3) = 0) elde edilmiştir.

Sayısal değişkenlerde ölçek farklılığı olduğu durumlarda, veri setinde bulunan sayısal değişkenler normalize edilerek değişkenlerin birbiri üzerindeki farklılıklar ortadan kaldırılır. Elde edilen veri kümesinde, sayısal değişkenler arasında ölçek farklılığı bulunduğundan dolayı sayısal değişkenler normalize edilmiştir. Normalizasyon işleminde sayısal veriler 0 ile 1 arasında normalize edilmiştir. Diğer taraftan veri kümesini normalleştirme işlemi çoğu makine öğrenmesi algoritmasının başarımını artırmakta olup [20], normalize edilen veriler üzerinde daha iyi sonuç vermektedirler.

2.4. Model İçin Oluşturulan Bağımlı ve Bağımsız Değişkenler (Dependent and Independent Variables Created for the Model)

Veri setindeki bağımlı ve bağımsız değişkenler üzerinde yapılan keşifsel veri analizi ve özellik çıkarımı çalışmaları sonucunda makine öğrenmesi algoritmasına hazır olan veri kümesindeki bağımlı ve bağımsız değişkenler şekil 6 'da gösterilmiştir. Gradient boosting regressor algoritması seçildiğinden (Bölüm 3'e bakınız) şekil 4'teki görselde model yerine, gradient boosting regressor algoritması yazılmıştır.



Şekil 4. Modeli besleyen bağımlı değişkenler ve model çıktısı olan bağımsız değişken gösterimi
(Dependent variables feeding the model and independent variable representation as model output)

3. MAKİNE ÖĞRENMESİ MODELİ KURULMASI (DETERMINE THE MACHINE LEARNING MODEL)

Son yıllarda bilgisayarların işlem gücü ve kapasitesinin artmasıyla birlikte makine öğrenmesi algoritmalarının popülerliği de artmıştır [21]. Bu çalışmada makine öğrenmesi algoritmalarından regresyon algoritmalarından, Lineer Regresyon, Random Forest Regresyon, Gradient Boosting Regresyon ve XGBoost Regresyon algoritmaları denenmiş ve sonuçlar tablo 4'te gösterilmiştir. Model başarımını ölçmek adına eldeki veri 20% doğrulama, 80% eğitim verisi olacak şekilde ayrılmıştır. Eğitilen model üzerinden doğrulama verisi kullanarak, model doğrulama çalışması yapılmaktadır.

Regresyon algoritmalarından oluşturulan model başarımı farklı parametreler üzerinden ölçümlenebilmektedir. R^2 değeri, ortalama mutlak hata (mae) ve ortalama karesel hatanın karekökü (rmse) parametreleri üzerinden oluşturulan model başarımı değerlendirilir. R^2 değeri yerine genellikle adjusment R^2 değeri üzerinden model seçimi yapılır. Bunun nedeni, veri setine değişken eklendikçe, eklenen değişkenlerin model başarımını değiştirmemesi, hatta model başarısını olumsuz yönde etkilese bile, R^2 değerinin artması, ama adjusment R^2 değeri eklenen değişkeni cezalandırarak, model başarımına göre artıp azalmaktadır.

$$R^2 = 1 - \frac{Hata\ Karelerinin\ Toplamı}{Ortalama\ Uzaklığın\ Kareleri\ Toplamı} \qquad (4)$$

Denklem (4) incelendiğinde pay kısmında bulunan hata karelerinin toplamı ifadesi sıfıra yaklaştığında (yani hata değeri çok az olduğunda), R^2 değeri 1 'e yaklaşmaktadır. Bu ifadeden anlaşılacağı üzere R^2 değeri 1 'e ne kadar yakın olursa model başarısı o kadar iyi denilebilir. Dolayısıyla denklem (4) 'te R^2 değerinin 0 ile 1 arasında değişmekte olduğu (0 <= R^2 <= 1) görülmektedir.

Tablo 4. Veri setinde denenen makine öğrenmesi algoritmalarının sonuçları
(Resoults of machine learning algorithms tried on the dataset)

Algoritma	R^2	MAE	RMSE
Lineer Regresyon	0.745	8,200	70,997
Random Forest	0.757	4,724	69,853
Gradient Boosted	0.837	4,071	56,258
XGBoost	0.746	7,973	70,316

Tablo 4 incelendiğinde algoritma kolonunda uygulanmış makine öğrenmesi algoritmalarını, R^2 kolonunda algoritmaların R^2 sonuçlarını, MAE (Mean Absolute Error) ortalama mutlak hatayı ve RMSE (Root Mean Squared Error) kolonunda ise, ortalama karesel hatanın karekökü sonuçları gösterilmektedir.

Seçilecek olan algoritma bu üç kolon üzerinden herhangi bir değer baz alınarak örneğin, R^2 değeri yüksek veya MSE değeri düşük olan değerler üzerinden seçim yapılabilir. Bu çalışmada doğrulama verisine göre ortalama karesel hatanın karekökü (rmse) değeri en küçük olan değere sahip algoritma (Gradient Boosted) seçilmiştir. Tablo 4 dikkatli incelendiğinde aslında bir kritere göre yapılan algoritma seçimi diğer kriterleri de sağlamaktadır. Daha sonra aynı algoritma test veri setinde de uygulamış ve hata durumu incelenmiştir.

$$RMSE = \sum_{k=0}^{n} \frac{(tahmin(y_k) - \text{gercek}(y_k))^2}{n} \quad (5)$$

Denklem 5 incelendiğinde tahmin (y_k) değeri tahmin edilen k'ıncı regresyon sonucunu, gerçek (y_k) değeri ise, k'ıncı gerçek değeri ifade etmekte olup, n değeri ise test verisindeki örnek sayısını ifade etmektedir. Önemli olan sonuç test verisinin sonucudur. Eğitim verisinde model veriye aşırı öğrenmiş (overfit), ya da eksik öğrenmiş (underfit) olabilir. Bunu öğrenmenin yolu eğitim verisi sonucu ile test verisi sonucunu karşılaştırmaktır.

3.1. Kurulacak Modelin Test Sonucu (Test Resoult of the Model Determine)

Seçilen gradient boosting regresyon algoritması üzerinden test veri kümesinde bulunan veriler değerlendirilmiş olup, sonuç tablo 5'te gösterilmiştir.

Tablo 5. Seçilen modelin test verisi üzerindeki sonucu
(Resoult of the selected model on the test data)

Algoritma	**RMSE**
Gradient Boosting	58,069

Ortalama karesel hatanın karekök (rmse) değeri incelendiğinde test veri kümesindeki veriler üzerindeki değerin beklendiği gibi biraz arttığı gözlemlenmiştir. Eğitilmiş model üzerindeki RMSE değeri genellikle test verisi üzerindeki RMSE değerinden düşük olması makine öğrenmesi algoritmalarında istenilen hatta olması gereken bir durumdur. Bu farkın yüksek olması istenmemektedir. Fark ne kadar yüksek olursa model, eğitim veri setindeki verilere aşırı eğilmiş bir diğer deyişle ezberlemiştir. Böylece görmediği veri modele verildiğinde ise bu veri üzerinde model düşük performans göstermektedir. Bu çalışmada, eğitim ve test verisi karşılaştırıldığında, model ezberlemesi gibi bir durum söz konusu olmadığı görülmüştür.

3.2. Model Optimizasyonu (Model Tuning)

Model optimizayonu (hyperparameter tuning), veri kümesi üzerinden kurulan makina öğrenmesi algoritmalarının daha iyi performans gösterebilmesi için (daha düşük hata veya daha iyi sınıflandırma) yapılan bir çalışmadır [22]. Model optimizasyonu genellikle veriye uygun algoritma seçildikten sonra, makine öğrenmesi algoritmasının eğitim sonucunda elde edilen performansları iyileştirmek için yapılmaktadır. Algoritma parametre ayarları yapıldıktan sonra (parametre ayarı eğitim (train) veri kümesindeki veriler üzerinden yapılır.) elde edilen model parametreleri modele verilerek, parametre ayarı yapılmış model test veri kümesinde denenerek bir de test verisi sonucundaki başarımına göre değerlendirme yapılır. Bir diğer önemli bir konu model başarımı her zaman test verisi üzerinden değerlendirilmelidir.

Parametre ayarı zor gibi gözükse de python programla dilinde sklearn kütüphanesi altında bulunan GridSearchCV fonksiyonu ile kolayca yapılabilmekte ve en iyi algoritma sonucunu veren parametre sonuçları elde edilmektedir.

Yapılan çalışmada en iyi performans gösteren gradient boosting algoritması için parametre ayarlama işlemi yapılmıştır. Gradient boosting algoritması için [23],

- ✓ Maksimum derinlik (Max depth)
- ✓ Öğrenme oranı (learning rate)

parametreleri kullanılarak, en iyi başarımı veren parametre sonuçları tablo 6'da paylaşılmıştır.

Tablo 6. Parametre sonuçları tablosu
(Resoults of the parameter's table)

Parametre	**Sonuç**
Maksimum derinlik	6
Öğrenme oranı	0.2

Parametre ayarları yapılan model test veri setindeki veriler ile denenmiş ve tablo 7'deki ortalama karesel hata sonucu elde edilmiştir.

Tablo 7. Parametre ayarlaması yapılan modelin test verisi üzerindeki sonucu
(The result of the parameter adjusted model on the test data)

Algoritma	**RMSE**
Gradient Boosting	48,381.015

Tablo 7'de gösterilen sonuca göre parametre ayarı yapılmış model üzerinden elde edilen hata miktarı, parametre ayarı yapılmamış model üzerinden elde edilen hata miktarından beklenildiği gibi düşük çıkmış olup, elde edilen bu sonuç beklenildiği gibi model başarısını arttırmıştır. Model parametre ayarı yapılarak test verisindeki hata yaklaşık olarak %20 azalmıştır.

Gradient boosting algoritmasında ayarlanacak parametreler, sadece maksimum derinlik (max_depth) ve öğrenme oranı (learning rate) değildir. Bunlardan başka ayarlanacak parametreler de bulunmaktadır [24]. Ancak gerek KNIME analitik platformunun yavaşlığından gerekse parametre ayarlama işlemlerinin KNIME analitik platformunun yeni versiyonlarında daha da geliştirilecek olmasından kaynaklı olarak, sadece iki parametre üzerinden model ayarlaması yapılmıştır. Bu durum ek olarak öneriler kısmında belirtilmiştir.

Elde edilen model sonuçları değerlendirildiğinde, acentenin sigorta şirketine ödeyecek olduğu borç tutarı RMSE değeri üzerinden hesaplanmış olup, ortalama 48,381 TL hata ile gelecek dönemki borç tahmini yapılabilmektedir. Bu sonuç iş birimiyle görüşülmüş olup acente borçları tutarları göz önünde bulundurularak, makul ve kabul edilebilir olduğu iş birimi tarafından belirtilmiştir.

Parametre ayarı yapılan model, 2021 yılına ait olan dönemlerdeki borç tutarlarının tahmin edilmesinde de denenmiş olup her dönem için başarılı sonuçlar vermiştir.

3.3. Tahmin Sonuçlarına Göre Kümeleme İşlemi (Clustering Process According to Estimation Results)

Elde edilen tahmin sonuçlarına göre artık acentelerin gelecek dönemlerdeki sigorta şirketine ödeyecekleri borçlar başarılı bir şekilde tahmin edildi. Bu tahminler üzerinden, borçları birbirine yakın olan acenteler tablo 8'deki gibi kümelenerek bu küme sonuçlarına göre, acentelerin risk grupları belirlenmiştir.

Kümeleme yönteminde makine öğrenmesi türlerinden olan denetimsiz (unsupervise) öğrenme yöntemi [25] kullanılarak, varsayılan olarak ayarlanan küme sayısı, Tablo 8 incelendiğinde 5 adet risk grup olduğundan dolayı, küme sayısı da 5 olarak seçilmiş ve tablo 8'de gösterilmiştir.

Tablo 8. Riskli olan acentelerin ait oldukları kümelerin belirlenmesi
(Determining the clusters to which risky agencies belong)

Küme No	**Risk Durumu**
1 'nolu küme	Risksiz
2 'nolu küme	Düşük riskli
3 'nolu kume	Orta riskli
4 'nolu küme	Riskli
5 'nolu küme	Yüksek riskli

Tablo 8'deki değerler incelendiğinde, örneğin acentenin tahmin edilen borcu sıfıra eşit veya 0'dan küçük ise bu acentenin sigorta şirketine borcunun olmadığı anlamına gelir ve acente risksiz olan kümeye dahil olur. Acentenin hangi kümeye ait olacağı k-means kümeleme algoritması sonucuna göre [26] belli olacaktır.

Tahmin edilen borç tutarlarına göre acenteler, belirlenen kümelere ekleneceklerdir. Acentenin tahmin edilen borç tutarı ne kadar yüksek olursa, elemanı olacağı küme numarası da yükselecek ve böylelikle acentenin risk durumu artmış olacaktır.

Örneğin; acentenin tahmin edilen borcu 20,000 TL olduğunda 3 numaralı kümeye yani orta riskli kümeye ait olup, ilgili acentenin durumu orta riskli olarak belirlenecek, başka bir acentenin tahmin edilen borcu 500,000 TL olduğunda 5 numaralı kümeye yani yüksek riskli kümenin elemanı olup, ilgili acentenin risk durumu yüksek riskli olarak belirlenecek ve acentelerin elamanı oldukları kümelere göre iş birimi gerekli aksiyonları alacaktır.

4. SONUÇLAR VE ÖNERİLER (RESULTS AND SUGGESTIONS)

Yapılan çalışmada elde edilen sonuçlar iş birimi ile paylaşılmış olup model değerlendirme başarısı iş birimine de anlatılmış ve uzman görüşlerine göre (burada uzman görüşü çalışmayı kullanacak olan iş biriminin görüşüdür) çalışma başarılı olarak değerlendirilmiştir. Bu kısımda bahsedilen uzman kişiler Anadolu Sigorta A.Ş. bünyesinde bulunan Risk Kontrol Müdürlüğü'ndeki uzman arkadaşlardır. Çalışma çıktıları irdelendiğinde, çok yüksek risk grubu bazında 10 adet acente bulunmaktadır. Zaten çalışmada acentenin çok küçük bir kısmının çok riskli olması beklenmektedir. Diğer risk grubu dağılımına bakıldığında sonuçlar Risk Kontrol Müdürküğü'nde çalışan uzmanlar tarafından kabul görmüştür.

Yapılan çalışmanın başarısından çok elde edilen hataların analizlerinin yapılıp paylaşılması hem iş birimi hem de bu çalışmayı referans alıp yapılacak olan diğer çalışmalara yol göstermesi açısından son derece önem arz etmektedir. Genellikle bu tarz çalışmalarda yapılması gereken en önemli kısım model hatasını maksimize eden örneğin ya da örneklerin veri kümesinde tespit edilip bu örneklerin üzerine eğilmektir.

Örneğin elde edilen makine öğrenmesi algoritması herhangi bir acentenin gerçekte olan borcu 20,000 TL iken bunu 25,000 TL ya da 16,000 TL tahmin ediyor olsun. Bu durumda zaten acentelerin üretimlerine ve sigorta şirketine ödeyecek olduğu borçlarına bakıldığında küçük bir değer olarak kalacak ve kabul edilebilir olacaktır. Ama başka bir acentenin gerçekte 2,000,000 TL olan borcunu, 25,000 TL olarak tahmin ediyorsa işte burada algoritma kabul edilemez bir hata yapmış olacaktır. Gerçekte 2,000,000 TL olan acentenin borcunun 25,000 TL olarak tahmin edilmesi sonucu acente düşük riskli kümenin elemanı olarak belirleneceği için, iş biriminin bu acenteyi inceleyememesine ve bu acenteye özgü riski belirleyememesine neden olacaktır. Bu sebepten dolayı, hataların analizleri yapılarak bu yüksek hataya neyin sebep olduğu, hangi durumlarda bunun gibi hatalarla karşılaşılabilir olacağının açıklanması gerekmektedir. Elde model test verisi ile değerlendirildikten sonra, ek olarak yeni modeli 2021 yılı verileri kullanılarak çalıştırılmış ve tablo 9'daki sonuçlar elde edilmiştir.

Tablo 9. 2021 yılına ait dönemlere göre elde edilen sonuçlar
(Results obtained according to periods in 2021)

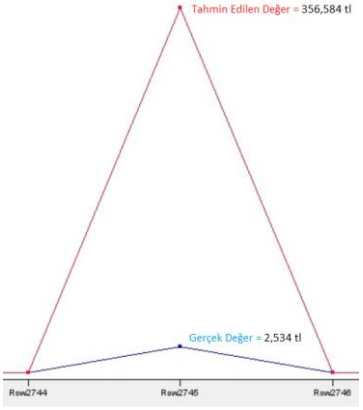
Donem (Yıl-Ay)	**Sonuç (RMSE)**
2021-03	53,693.809
2021-06	43,201.365
2021-09	33,777.917

Elde edilen sonuçlar değerlendirildiğinde, model performansı oldukça başarılıdır.

4.1. Hata Analizi (Error Analysis)

Yapılan bu çalışmada, test veri setinde model değerlendirilirken tahmin edilen borç tutarı ile (tahmin

(y_k)), gerçekte olan borç tutarları (gerçek (y_k)) karşılaştırılmıştır. Yüksek hata oranına sahip olan veriler belirlenmiştir. Belirlenen bu veriler üzerinden hangi durumlarda bunun gibi sonuçlarla karşılaştırılabilecek olduğu durumlar incelenerek, şekil 5'de örnek bir tanesi gösterilmektedir.



Şekil 5. Modelin örnek bir veri üzerindeki gerçek ve tahmin edilen değeri
(Actual and predicted value of the model on a sample data)

Şekil 5 incelendiğinde, mavi nokta acentenin ilgili dönemdeki gerçek borcunu göstermekte olup (2,534 tl), kırmızı noktada tahmin edilen borç değerini (356,584 tl), Row2745 ilgili acentenin ilgili dönemini (örneğin 2745 numaralı acente gibi düşünülebilir) göstermektedir.

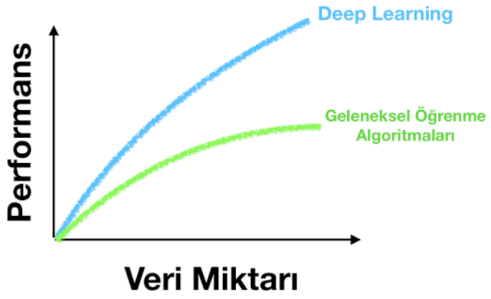
Bu dönem incelendiğinde ilgili acentenin ilgili dönemine kadarki davranışlarında, acentenin sigorta şirketine hiç borcu bulunmamaktadır. Sadece tahmin edilen dönemde acente sigorta şirketine aniden borçlanmıştır. Acentenin tahmin edilen dönemden önceki dönemdeki borçları da veri setinde değişken olarak oluşturulduğundan ve ilgili değişkenler hedef değişken ile yüksek korelasyonu bulunduğundan dolayı, nadiren de olsa şekil 8'de gösterilen durumlar otaya çıkabilmektedir.

Yapılan değerlendirme sonucunda çok yüksek hataya sahip olunan veriler analiz edilmiş ve analiz sonucunda, daha önce hiç borcu olmayıp sadece tahmin edilen dönemde acentenin sigorta şirketine, birden yüksek borcu meydana gelmiştir.

Ya da tersi bir durum yani acentenin geçmişte sigorta şirketine borçları var ve bu borçlar katlanarak çoğalmış ya da sabit kalmış ya da çok küçük miktarda azalmış (acente sigorta şirketine küçük bir miktar ödeme yapmış) ama tahmin edilen dönemde, acente sigorta şirketine olan borcunun tamamını kapatmış olduğundan dolayı hesaplanan hata değeri yüksek olmaktadır.

4.2. Öneriler (Suggestions)

Veri seti incelendiğinde yaklaşık 120,000 kayıt olduğu bölüm 2'de bahsedilmişti. Bu veriye bakıldığında derin öğrenme yöntemi kullanılarak bir model oluşturulduğunda, derin öğrenme ile oluşturulan model başarımının (r^2, rmse başarı parametreleri göz önünde bulundurulduğunda) daha da artacağı şekil 9'de gözlemlenmektedir.



Şekil 6. Derin öğrenme ve makine öğrenmesi artan veri performans grafiği [27]
(Deep learning and machine learning incremental data performance graph)

Şekil 6 incelendiğinde, veri miktarının artması belirli bir seviyeden sonra artık klasik makine öğrenmesi algoritmalarında başarımı çok fazla arttırmamaktadır. Ancak veri sayısının artması derin öğrenme yöntemlerinde başarıyı arttırmaya devam etmektedir [28]. Böylece eldeki veri miktarı büyük olduğundan derin öğrenme yöntemi ile de model oluşturulup, elde edilen model sonuçları değerlendirilebilir.

KAYNAKLAR (REFERENCES)

[1] Z. T. Aloğlu, **Bankacılık sektörünün karşılaştığı riskler ve bankacılık krizler üzerindeki etkileri**, Uzmanlık Yeterlilik Tezi, Türkiye Cumhuriyet Merkez Bankası, 2005.

[2] E. Gümüş, B. Medetoğlu, S. Tutar, "Finans ve Bankacılık Sisteminde Yapay Zeka Kullanımı: Kullanıcılar Üzerine bir Uygulama", *Bucak İşletme Fakültesi Dergisi*, 3(1), 28-53, 2020.

[3] A. M. Suresha, "Machine Learning for Mining Weather Patterns and Weather Forecasting", *ResearchGate*, 95(6), 42-51, 2020.

[4] B. E. Katı, E. U. Küçüksille, "Oracle ve MS SQL Server Veri Tabanları İçin Veri Tabanı Yönetim Sistemleri Güvenlik Kontrol İlkelerinin Takip Edilmesi ve Uygulanması", *Uluslararası Teknolojik Bilimler Dergisi*, 10(2), 22-46, 2018.

[5] A. Geron, **Scikit-Learn, Keras ve Tensorflow ile Uygulamalı Makine Öğrenmesi**, Mustafa Murat Arat & Vedat Çelik, Ankara, Türkiye, 2021.

[6] Internet: Türkiye Cumhuriyet Merkez Bankası Tüketici Fiyatları, https://www.tcmb.gov.tr/wps/wcm/connect/TR/TCMB+TR/Main+Menu/Istatistikler/Enflasyon+Verileri/Tuketici+Fiyatlari, 12.12.2021.

[7] D. Karasoy, N. Tuncer, "Outliers in Survival Analysis", *Alphanumeric Jurnal*, 3(2), 139-152, 2015.

[8] Internet: M. Bektaş, Aykırı Değerlerin Tespiti ve Bu Değerlerle Mücadele Yöntemleri, https://medium.com/@mbektas/ayk%C4%B1r C4%B1-de%C4%9Ferlerin-tespiti-ve-bu-de%C4%9Ferler-ile-m%C3%BCcadele-y%C3%B6ntemleri-4f0cf76737d1, 01.12.2021.

[9] Y. Poyraz, S. Sevgen, "GPU Programlama Tekniği ile Yüksek Performanslı Araç Takibi", *Bilişim Teknolojileri Dergisi*, 10(3), 255-261, 2017.

[10] S. Kırca, K. Halatçı, V. Güneş, "Acente Performans Ölçümleme Çalışması", *Veri Bilimi Dergisi*, 4(2), 49-56, 2021.

[11] Internet: Sklearn Ensemble GredientBoostingRegressor, https://scikit-learn.org/stable/modules/generated/sklearn.ensemble GradientBoostingRegressor.html, 24.11.2021.

[12] Hwang H., Jung T., Suh E. "An LTV Model and Customer Segmentation Based on Customer Value: A Case Study on the Wireless Telecommunication Industry", *Jurnal of Interactive Marketing*, 13(3), 2-12, 2004.

[13] Jain A. K., Murty M. N., Flynn P. J. "Data Clustering: A Review", *ACM Computing Surveys*, 31(3), 264-323, 1999.

[14] A. Anbar, Kredi riski yönetim aracı olarak kredi türevleri ve türk bankacılık sektöründe uygulanabilirliği, Doktora Tezi, Uludağ Üniversitesi, 2005.

[15] C. Lombardi, G. F. Tassi, G. Pizzocolo, F. Donato, "Clinical Significance of a Multiple Biomarker Assay in Patients with Lung Cancer: A Study with Logistic Regression Analysis", *Chest*, 97(3), 639-644, 1990.

[16] Doğan B., Buldu A., Demir Ö., Ceren E. B. "Sigortacılık Sektöründe Müşteri İlişki Yönetimi İçin Kümeleme Analizi", *Karaelmas Fen ve Mühendislik Dergisi*, 8(1), 11-18, 2018.

[17] Saruman G. "Veri Madenciliğinde Kümeleme Teknikleri Üzerine Bir Çalışma: K-Means ve KMedoids Kümeleme Algoritmalarının Karşılaştırılması", *Fen Bilimleri Enstitüsü Dergisi*, 15(3), 192-202, 2011.

[18] Çakmak D, Baştürk F. H. "Türk Sigortacılık Sektörünün 2007-2018 Yıllarına Ait Performansının Oran Analizi Yöntemi ile Ölçülmesi ve Sektörün Ekonomik Büyüme Üzerindeki Etkisi". *Çanakkale Onsekiz Mart Üniversitesi Uluslararası Sosyal Bilimler Dergisi*, 4(2), 235-264, 2019.

[19] J. Robert, S. M. Turnbull, "The Intersection of Market and Credit Risk", *Journal of Banking & Finance*, 24(1), 271-299, 2000.

[20] T. İldaş, "Kredi Riski Ölçüm Modellerinin Değerlendirilmesi", *Finansal Araştırmalar ve Çalışmalar Dergisi*, 13(25), 516-547, 2021.

[21] H. Budak, S. Erpolat, "Kredi Riski Tahmininde Yapay Sinir Ağları ve Logistic Regresyon Karşılaştırılması", *AJIT-e: Online Academic Jurnal of Information Technology*, 3(9), 23-30, 2012.

[22] B. Donel, **Yapay sinir ağı yöntemi ile kredi skorlama**, Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi, 2012.

[23] Internet: B. Ay, BVYZLab, http://buyukveri.firat.edu.tr/veri-setleri/, 01.02.2021.

[24] M. Talo, B. Ay, S. Makinist, G. Aydin, "Bigailab-4race-50K: Race Classification with a New Benchmark Dataset", **International Conference on Artificial Intelligence and Data Processing (IDAP)**, 1–4, 2018.

[25] P. G. Shambharkar, A. Anand and A. Kumar, "A Survey Paper on Movie Trailer Genre Detection", **2020 International Conference on Computing and Data Science (CDS)**, 2020, 238-244, doi: 10.1109/CDS49703.2020.00055.

[26] S. Oktay, H. Temel, "Basel II Kriterleri Ekseninde Ticari Bankalarda Kredi Riski Yönetiminin Karşılaştırılmasına Yönelik Bir asha Çalışması", *Uluslararası Yönetim İktisat ve İşletme Dergisi*, 3(6), 163-186, 2007.

[27] Internet: Deep Learning (Derin Öğrenme) Nedir?, **https://www.kodsihirbazi.com/deep-learning-derin-ogrenme-nedir/,** 13.12.2021.

[28] A. Baykal, "Veri Madenciliği Uygulama Alanları", *Dicle Üniversitesi Ziya Gökalp Eğitim Fakültesi Dergisi*, 1(7), 95-107, 2006.