www.turkishstudies.net/appliedsciences

Turkish Studies - Information Technologies and Applied Sciences

eISSN: 2667-5633

Research Article / Araştırma Makalesi

INTERNATIONAL
BALKAN
UNIVERSITY
Sponsored by IBU

Makine Öğrenmesi ile Sigorta Hasarlarında Sahtecilik Tespiti

Fraud Detection in Insurance Claims with Machine Learning

Yaşar Geren*

Abstract: Fraud has been rising recent years all over the world. Especially, increase of the internet usage and use of credit cards more increase to fraud cases as well. Much work has been done and measures have been taken in the banking sector to detect and prevent credit and credit card frauds. Similarly, counterfeit claims are also common in the insurance sector. It is seen that the insurance sector is a little more difficult to predict fake claims than the bank sector and the studies in this area are less. As computers' transaction volumes and speeds progress, human beings have started to use newer technologies in every area of their business. Nowadays, we are talking about the writing, learning and implementation of the algorithm by the program itself. Machine learning is an application of artificial intelligence and it achieves results by developing what it has learned. Machine learning has also started to be used in finance and insurance sector like it is used in many industry. Insured customer informations and claim informations are collected in the big data pool and reach serious dimensions over time. If this data is used and analyzed with correct machine learning algorithms, counterfeit claims can be estimated to a large extent. In this study, using the machine learning algorithms on the claim data set obtained from an insurance company, the predictive scores of the fraud cases of the claims will be compared. By running 7 different machine learning algorithms on the same data set with the same test and training rates, the accuracy rates and performances will be compared and the results will be shown.

Structured Abstract: Insurance is designed according to the principle of mutual solidarity clause against unexpected losses. Fake insurance claims, cause to consumption of funds generated by the insured that is gathered to cover real losses and it cause great damage to the system. Machine learning applications are the newest solutions against these measures. Machine learning tries to reveal the patterns in the datasets, and its importance has increased with the growth of the data. Machine learning models are generally divided into two as supervised and unsupervised. Classifications and regression algorithms can be given as example for supervised models. Unsupervised models consist mostly of clustering algorithms. It is seen that the supervised learning models are mostly used for insurance losses, because the predictions are realized from data sets formed from very determinant areas. It seems in Turkey some of commercial products are being used to detect fraud in the insurance sector. Also, big data teams in the information technology units of the insurance companies have also started to be established and they are trying to predict false damage determinations with the applications they have developed.

Machine learning is a branch of computer science developed in 1959 from numerical learning and model recognition studies in artificial intelligence. Machine learning is a system that investigates the operation

* Yüksek Lisans Öğrencisi, Aydın Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği ABD
Master Student, Aydın University, Institute of Science, Computer Engineering
ORCID 0000-0001-7632-5715
yasargeren@gmail.com

Cite as/ Atıf: Geren, Y. (2020). Makine öğrenmesi ile sigorta hasarlarında sahtecilik tespiti. *Turkish Studies - Applied Sciences*, 15(2), 195-209. <https://dx.doi.org/10.29228/TurkishStudies.42855>

Received/Geliş: 11 April/Nisan 2020

Accepted/Kabul: 20 June/Haziran 2020

Copyright © INTAC LTD, Turkey

Checked by plagiarism software

Published/Yayın: 25 June/Haziran 2020

CC BY-NC 4.0

and development of algorithms that can learn as structural functions and make predictions based on data. Supervised learning is a learning model built to make prediction, given an unforeseen input instance. A supervised learning algorithm takes a known set of input dataset and its known responses to the data to learn the regression/classification model. Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses. The most common unsupervised learning method is cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping in data. Semi-supervised learning is exactly between the two concepts above and it is a method in which tagged big data and untagged small amounts of data are used together. The method by which a system is evaluated as true or false for the result produced by the system is reinforcement learning. Briefly, reinforcement learning is an interaction based learning in cause and effect environment.

The purpose of this study is to predict the counterfeit damages statements with machine learning in vehicle insurances notifications and compare success of the machine learning algorithms. Algorithms will be compared in terms of score and calculation times also the algorithm that gives the best values will be determined. For the study, a data set was prepared from the auto damage records of a private insurance company. The data set consisting of 3000 damage records contains rule sets prepared by the claims department. Rule_1, Rule_3, Rule_6, Rule_12 and Rule_13 rule sets contain critical data related to damage. Some of these rule sets may have values as 0 and 1, while others may take higher values. For example, Rule_1 is related to the presence of alcohol in the blood and its value can be 0 or 1. Another rule is Rule_12 and this rule states the number of previous claims in the policy. In the light of this data, machine learning algorithms will be compared and their values will be given graphically.

The developed application in this study is written with python language. Machine learning algorithms are run with python and the results are compared. Real claims data from a private insurance company were used. It was used as Visual Studio Code and JupyterLab editor during application development. The algorithms compared are random forest, logistic regression, linear regression, k-nearest neighbors, naive bayes, decision tree and support vector machines algorithms. Testing and learning process was carried out by k-fold cross-validation method. 5 is taken as the K parameter value. With this method, the data set is divided 5 different sets to training and testing processes. Accuracy rates and test calculation times of all algorithms are shown in Table-1.

Table 1: Algorithms Test Results

Model	Successful Data	Accuracy	Process Times (sn)
Random Forest	2578	86%	4.41
Logistic Regression	2113	71%	2.41
Linear Regression	2099	70%	0.08
K Nearest Neighbors	2264	75%	0.51
Naive Bayes	2089	70%	0.04
Decision Trees	2267	76%	0.17
Support Vector Machines	2120	71%	1.99

When the results are analyzed, it is seen that the highest accuracy rate is obtained by random forest algorithm. Decision trees follow this algorithm with a rate of 76%. When looking at the test performances of algorithms, there are very small differences. When the results are examined, it is seen that the highest accuracy rate is obtained by random forest algorithm. Decision trees follow this algorithm with 76%. The K Nearest Neighbors algorithm also performed at the same levels with the Decision Trees algorithm with a 75% accuracy rate. When looking at the process times performances of algorithms, it seems different values. Despite Random Forest algorithm has the highest accuracy rate, it has the lowest performance in terms of processing time. When we compare the process times, the fastest algorithm is Naive Bayes. Linear Regression and Decision Trees algorithms follow it. Even these tests are done with a limited number of data, it can be an indicator for larger environments.

When the outputs of the algorithms were compared, it was seen that the best results were obtained with random forests and decision trees. Random Forests and Decision Trees algorithms are supervised machine learning algorithms and they are using classification techniques. It is showing us that our data set is more suitable for supervised models. When the tests performed by changing the default parameter values of Random Forest Algorithm, it didn't show a significant increase. Other algorithms are also run with their default values. Due to small number of such studies related to auto damages in the insurance sector, I think that this study will serve as a model for experts or academic staff who want to work in this field.

Keywords: Machine Learning; Machine Learning Algorithms; Fraud Detection; Fake Claim; Big Data;

Öz: Dolandırıcılık son zamanlarda tüm dünyada artış göstermektedir. Özellikle internet kullanımının artması ve kredi kartlarının daha yoğun kullanılması ile sahtecilik olayları da artmaktadır. Bankacılık sektöründe kredi ve kredi kartı sahteciliklerinin tespiti ve önlenmesi ile ilgili bir çok çalışma yapılmış ve önlemler alınmıştır. Benzer şekilde sahte hasarlar sigorta sektöründe de yaygın bir şekilde görülmektedir. Sigorta sektöründe sahte hasarların tahmin edilmesi banka sektörüne göre biraz daha zor olduğu ve bu alanda yapılan çalışmaların daha az olduğu görülmüyor. Bilgisayarların işlem hacimleri ve hızları ilerledikçe insanoğlu daha yeni teknolojileri iş hayatının her alanında kullanmaya başlamıştır. Günümüzde artık algoritmanın programın kendisi tarafından yazılmasından, öğrenmesinden ve uygulamasından bahsediyoruz. Makine öğrenmesi yapay zekanın bir uygulamasıdır ve öğrendiklerini deneyerek, geliştirerek sonuç elde eder. Makine öğrenmesi bir çok alanda kullanıldığı gibi finans ve sigorta alanlarında da kullanılmaya başlanmıştır. Sigortalı bilgileri ve hasar bilgileri büyük veri havuzunda toplanmakta ve zamanla ciddi boyutlara ulaşmaktadır. Bu verilerin doğru makine öğrenmesi algoritmaları ile iyi analiz edilerek kullanılması durumunda sahte hasarlar büyük oranda tahmin edilebilir. Bu çalışmada bir sigorta şirketinden alınan hasar veri seti üzerinden makine öğrenmesi algoritmaları kullanılarak, hasarların sahte olma durumlarının tahmin skorları karşılaştırılacaktır. 7 değişik makine öğrenmesi algoritmasını aynı veri seti üzerinde, aynı test ve eğitim oranlarında çalıştırıp, doğruluk oranları ve performansları karşılaştırılacak ve nasıl sonuçlar verdiği gösterilecektir.

Anahtar Kelimeler: Makine Öğrenmesi; Makine Öğrenmesi Algoritmaları; Sahte Hasar; Büyük Veri; Sahtecilik Tespiti;

Giriş

Sigorta bilindiği üzere beklenmeyen kayıplara karşı karşılıklı dayanışma ilkesine göre tasarlanmıştır. Kişilerin kendilerini muhtemel zarar ve beklenmeyen masrafların mali etkilerine karşı koruma ihtiyacından doğmuştur (Demirci, 2019). Sahte sigorta hasarları ise gerçek kayıpları karşılamak üzere sigortalılar tarafından ödenen primlerle oluşan fonların tüketilmesi sonucunu doğurmakta ve sisteme büyük zarar vermektedir. Sigorta suistimalleri, birçok dürüst sigortalının mağdur olmasına neden olan ve dikkatle ele alınması gereken bir suçtur (Yıldırım, 2013). Dürüst sigortalıların büyük çoğunluğu ödedikleri yüksek primlerle yapılan suistimalin bedelini ödemektedir. Bu nedenle sigorta sektörü bu sorunun çözülmesi için her türlü önlemi almaya çalışmaktadır. Makine öğrenmesi ile geliştirilen uygulamalar da bu önlemler içindedir. Makine öğrenmesi veri setleri içindeki örüntüleri ortaya çıkarmaya çalışarak verilerde belirli bir kalıp arar ve bu çıkarımlardan bilinmeyen sonuçlara dair tahminlerde bulunur. Örüntü tanıma, bilgisayar algoritmalarını kullanarak verilerdeki düzenliliklerin otomatik olarak bulunması ve bu düzenliliğin verilerin farklı kategorilere göre sınıflandırılması için kullanılmasıdır (Bishop, 2006). Makine öğrenmesi modelleri genelde denetimli ve denetimsiz olarak ikiye ayrılmaktadır. Denetimli olan modeller altında sınıflandırma ve regresyon algoritmalarını sayabiliriz. Denetimsiz olan modeller ise daha çok kümeleme algoritmalarından oluşmaktadır (Kızılkaya ve Oğuzlar, 2018). Sigorta hasarlarından oluşan veri setlerinde genelde çok belirleyici alanlardan oluşturulan kural setleri ile tahminler gerçekleştiği için denetimli olan öğrenme modellerinin daha çok kullanıldığı görülmektedir (Günbatır, 2019). Türkiye'de ticari olarak bazı ürünlerin de sigorta alanında sahtecilik tespiti için kullanılmaya başlandığı görüldüğü gibi sigorta şirketlerinin bilgi teknolojileri birimlerinde bulunan büyük veri

ekipleri de kurulmaya başlanmış ve kendi geliştirdikleri uygulamalar ile sahte hasar tespitlerini öngörmeye çalışmaktalar (www.sas.com, 2018), (www.turkishtimedergi.com, 2016).

Literatür Taraması

Bu alanda yapılan çalışmalar incelendiğinde daha çok bankacılık sektöründe kredi kartı sahte işlemleri tespiti ve sigorta sektöründe de sağlık alanında sahte işlemlerin tespiti ile ilgili örnekler çokça rastlanmaktadır. Elementer branş dediğimiz oto hasarlarında sahtecilik tespiti ile ilgili çok örnek bulunmamaktadır. Sigorta dergileri ve şirketlerin yayınları üzerinden yapılan araştırmalarda insuretech adı altında yeni teknolojilerin inovasyonlarından ve maliyet avantajlarından yararlanılmaya çalışıldığı görülmekte. Yabancı yazılımlar incelendiğinde sigorta sektörü için SAS Software, Friss fraud detection, Scorto InsuSafe, NS8, Ravelin ve Araxxe isimli üreticilerin ürünleri göze çarpmaktadır. Altta akademik alanda yapılan bazı örnek tez ve çalışmalar verilmiştir. Hazım (2018) 507937 numaralı tezinde kredi kartı dolandırıcılık tespiti için dört sınıflandırma yöntemini test etmiştir (Naive Bayesian, destek vektör makinesi, K-en yakın komşu ve rastgele orman). Çalışmasında yüksek dengesizlik sınıfı nedeniyle alt örneklemeyi kullanmış ve modellerinin doğruluğunu ölçmek için çapraz validasyon uygulayarak bu dört algoritmayı test etmiş. Sonuçlarında rasgele ormanın diğer üç algoritmaya göre daha iyi performans gösterdiğini belirtmiştir. Aşuk (2010) 258531 nolu tezinde sağlık sigorta şirketleri için veri madenciliği tabanlı susitimal tespit sistemi incelemesi yapmış. Çalışmasında linear kernel fonksiyon tabanlı ve gaussian kernel fonksiyon tabanlı modellerin incelenmesine yer vermiştir. Sonuç olarak üçe ayırdığı suistimal tiplerine göre Linear kernel tabanlı fonksiyonları ve Gaussian kernel fonksiyonlarının hangisinin hangi tip suistimal için başarılı olduğunu incelemiştir. Jagdish Pathak (Accounting Systems and IT Auditing Area, Odette School of Business Administration, University of Windsor, Windsor, Ontario, Canada) 2005 yılında fuzzy logic math kombinasyonlarını kullanarak sigorta hasar bildirimlerinde sahte bildirimlerin tespiti ile ilgili bir çalışma yapmıştır. Kanada gizlilik mevzuatları nedeniyle gerçek veriler kullanılamamış olsada test ettikleri fuzzy logic kombinasyonlarının sigorta sahte hasar tespitlerinde kullanımının iyi sonuçlar verdiğini belirtmiştir. Günbatar (2019) 587949 numaralı tezinde otomobil sigortası sahtekarlıklarının makine öğrenmesi ile tespit edilmesine dair çalışma yapmış. Tezinde lojistik regresyon, karar ağaçları, yapay sinir ağları, destek vektör makineleri ve bayes ağlarını test etmiş sahte hasarların tespitine yönelik bütün algoritmaların yüksek performans gösterdiği ve hiçbirinin diğerine çok üstün gelmediği sonucuna varmıştır. Karamanlı (2019) 579835 nolu tez çalışmasında makine öğrenmesi algoritmalarını kullanarak metin madenciliği ve duygu analizi ile müşteri deneyiminin geliştirilmesi konusunu ele almış. Çalışmasının sonucunda pozitif ve negatif yorum tahminleme birlikte dikkate alındığında en iyi sonucu Destek Vektör Makineleri algoritması ile elde ettiğini belirtmiştir. Aleskeov, Freisleben ve Rao (1997) yaptıkları çalışmada kredi kartı sahtekarlığı için sinir ağı öğrenme modeline dayanan veri tabanı madenciliğini kullanmışlar ve sinir ağı öğrenme modeli ile elde edilen sonuçların çok başarılı tespit oranları verdiğini belirtmişlerdir.

Amaç

Bu çalışmanın amacı, araç sigortalarında yapılan hasar bildirimlerinden makine öğrenmesi ile sahte olanları tahmin etmek ve algoritmaları karşılaştırmaktır. Algoritmalar doğruluk oranları ve hesap süreleri açısından karşılaştırılacak ve en iyi değerleri veren algoritma belirlenmeye çalışılacak. Uygulamaya veri setinde bulunan sahte hasarlar ile öğretim yapılacak ve test verisi üzerinden de gerçekten sahte hasar olanları ne kadar doğru tespit ettiği ölçülecek. Hesap süresi olarak algoritmaların işlem süreleri ölçülerek ne kadar hızlı çalıştıkları karşılaştırılacaktır. Çalışma için özel bir sigorta şirketinin oto hasar kayıtlarından bir veri seti hazırlanmıştır. Veri seti gerçek hasar kayıtlarından oluşmaktadır. Hasar birimi tarafından sahte hasar tespit çalışmaları için hazırlanmış veri setidir. Rasgele seçilmiş 3000 adet hasar kaydından oluşan veri setinde hasar birimi tarafından hazırlanmış kural setleri bulunmaktadır.

Veri Seti

Veri setimiz hasar departmanı tarafından sahte hasar tespiti için yapılacak bir çalışmaya özel hazırlanmıştır. Hasar bildirimleri esnasında sigortalının kendisi ve hasarın detayı ile ilgili çok sayıda veri alınmaktadır. Hazırlanan bu veri seti normalize edilmiş ve uygulama tarafından kullanılacak hale getirilmiştir. Veri setimizde hasarların sahte mi yoksa gerçek bir hasar mı olduğuna dair bilgi bulunmaktadır. Veri setinde bulunan alanlar; dosya_no, fraud, hasar_neden_aciklama, teminat_bedel, yeni_is_yenileme, ozel_tuzel, kullanim_tarzi, model_yasi, sigortalı_il, sig_cinsiyet, Rule_1, Rule_3, Rule_6, Rule_12, Rule_13 ve sig_yas. Rule ile başlayan alanların açıklamaları Tablo 1 de verilmiştir.

Tablo 1: Veri Setinde Bulunan Kuralların Açıklamaları

Kural	Açıklama
Rule_1	Kandaki alkol oranı. 0 alkol olmadığı anlamına gelir.
Rule_3	Hasar geçmişi. 1 ise araştırılmış başka hasarı var.
Rule_6	Sigortalının sahip olduğu sigortalı araç sayısı.
Rule_12	Poliçedeki hasar sayısı.
Rule_13	Police başlangıç tarihi ile hasar tarihi arasındaki fark.

Kapsam

Bu çalışmanın kapsamı, oto branşı için sigorta hasarlarında sahte hasar tespitinin makine öğrenmesi ile yapılması ve hangi algoritmaların daha iyi sonuç verdiğinin tespit edilmesidir.

Yöntem

Çalışmada geliştirilen uygulama python dilinde yazılmıştır. Makine öğrenmesi algoritmaları python ile çalıştırılarak sonuçlar karşılaştırılmıştır. Özel bir sigorta şirketinden alınan gerçek hasar verileri kullanılmıştır. Uygulamanın geliştirilmesi esnasında Visual Studio Code kullanılmıştır. Debug işlemleri için uygulamayı satır satır çalıştırıp çıktılarını görebildiğim için JupyterLab editörden de faydalandım. Karşılaştırılan algoritmalar; random forest, logistic regression, linear regression, k-nearest neighbors, naive bayes, decision tree ve support vector machines algoritmalarıdır. 3000 adet hasar verisinden oluşan veri kümesinin test ve eğitim sınıflandırması için k parçalı çapraz doğrulama yöntemi kullanılmıştır (K-Fold Cross Validation). Böylece veri seti belirlenen sayıda kümeler bölünerek bütün veri seti bu kümeler üzerinden teste tabi tutulmuştur. Bu yöntemde veri seti k adet parçaya bölünür ve her seferde farklı veri kümesi test için ayrılır ve kalan k-1 adet veri kümesi uygulama tarafından eğitim amacıyla kullanılır. Bu işlem k defa tekrar edilir ve her tekrarda test kümesi değiştirilerek sınıflandırıcı eğitilir (Narin, 2014). K parçalı çapraz doğrulama dizin sonuçları da “k-CV Değerleri” olarak görünecek şekilde her algoritma için verilmiştir. Uygulamada algoritmaların işlem süreleri ve doğruluk oranları hesaplanacaktır. İşlem sürelerini ölçmek için her algoritmanın işlem başlangıcında ve bitişinde alınan iki zaman değerinin arasındaki fark hesaplanmıştır. Uygulamada doğruluk oranları hasar verilerinin fraud olanlarının tespiti üzerinden ölçülecektir. Her algoritmaya fraud olan kayıtlar için bir tahmin yaptırılacak ve sonuçlar karşılaştırılacaktır.

Makine Öğrenmesi Nedir

Makine öğrenimi, örnek verileri veya geçmiş deneyimleri kullanarak bir performans ölçütünü optimize etmek için bilgisayarı programlamaktır. Öğrenme, eğitim verilerini veya geçmiş deneyimleri kullanarak sahip olunan modelin parametrelerini optimize etmek için bir bilgisayar programının yürütülmesidir (Alpaydın, 2020)

Denetimli Öğrenme

Denetimli öğrenme birbiri ile etkileşimli belirli bir veri seti üzerinden belirli bir düzende organize edilerek öğrenmenin gerçekleştirildiği yöntemdir. Diğer bir deyişle makine öğrenmesini girdiler verip bir dizi çıktı aldığımız karakutu olarak düşünelim. Verdiğimiz girdilere istinaden aldığımız çıktı kümesi için geçmiş veriler varsa elimizde bu verilere dayalı öğrenmeye denetimli öğrenme denir. En klasik denetimli öğrenme modeli sınıflandırmadır (Joshi, 2020:10).

Denetimsiz Öğrenme

Veriler arasında belirli bir etkileşim olmayan ve sınıf bilgisi barındırmayan verilerin içindeki grupların irdelendiği yöntemdir. Denetimsiz öğrenmede etiketli veriler mevcut değildir. En klasik örnek kümelemedir (Joshi, 2020:11).

Yarı denetimli Öğrenme

Bu kavram tam olarak yukarıdaki iki kavramın arasında yer alır ve etiketlenmiş büyük veri ile etiklenmemiş küçük miktarlarda verinin beraber kullanıldığı yöntemdir. Yarı denetimli öğrenme, insanlar gibi bilgisayarların ve doğal sistemlerin hem etiketlenmiş hem de etiketlenmemiş verilerin varlığında nasıl öğrendikleri ile ilgili bir öğrenme paradigmasıdır. Geleneksel olarak öğrenme ya tüm verilerin etiketsiz olduğu denetimsiz paradigmada (örneğin kümeleme, aykırı tespit) ya da tüm verilerin etiketlendiği denetimli paradigmada (örn. Sınıflandırma, regresyon) incelenmiştir. Yarı denetimli öğrenmenin amacı, etiketli ve etiketsiz verilerin birleştirilmesinin öğrenme davranışını nasıl değiştirebileceğini anlamak ve böyle bir kombinasyondan yararlanan algoritmalar tasarlamaktır. (Zhu ve Goldberg, 2009).

Takviyeli Öğrenme

Sistemin ürettiği sonuç için doğru ya da yanlış olarak bir değerlendirmesidir. Takviyeli ya da pekiştirmeli öğrenme, neden sonuç ortamında etkileşime dayalı bir öğrenmedir ve denetimli-denetimsiz yöntemlerden ayrı olarak ele alınması gerekir. Takviye öğrenimi çevreden gelen geribildirimleri içerir ve tam olarak denetlenemez. Yani eğitim için etiketli bir veri dizisi yoktur (Joshi, 2020:11).

Rasgele Orman (Random Forest)

Rasgele Orman algoritması torba karar ağaçlarının biraz daha gelişmiş versiyonudur. CART (Classification and Regression Trees) karar ağacı modellerinin problemi aç gözlü algoritmalar olmasıdır. Aç gözlü algoritmaların zayıf yönü çok fazla korelasyona sahip olmalarıdır. Alt modellerden gelen tahminler zayıf korelasyonlu olabilir. Rasgele orman tüm ağaçlardan elde edilen tahminlerin daha az korelasyona sahip olması için alt ağaçların öğrenilme şekline ilişkin algoritmayı değiştirir (Oshiro, 2012). Rasgele orman ile yaptığımız çalışmada varsayılan parametreler ile elde edilen değerler altta listelenmiştir.

Basarili data sayisi : 2578

RandomForest dogruluk orani : 86%

RF Test Suresi (sn) : 4.41

RF k-CV Değerleri: [0.87166667, 0.85166667, 0.85666667, 0.845, 0.855]

Şekil 1: Rasgele Orman Hata Matrisi (Confusion Matrix)

Doğrusal Regresyon (Linear Regression)

Doğrusal regresyon analizi, bir değişkenin değerini başka bir değişkenin değerine göre tahmin etmek için kullanılır. Tahmin etmek istediğiniz değişken, bağımlı değişken olarak adlandırılır. Diğer değişkenin değerini tahmin etmek için kullandığınız değişken ise bağımsız değişken olarak adlandırılır. Bu analiz biçimi, bağımlı değişkenin değerini en iyi öngören bir ya da daha fazla bağımsız değişkeni kullanarak lineer denklemin katsayılarını tahmin eder. Lineer regresyon, öngörülen ve gerçek çıkış değerleri arasındaki uyumsuzlukları en aza indiren düz bir çizgi ya da yüzeye yerleştirir. Bir çift eşleştirilmiş veri kümesi için en uygun satırı keşfetmek üzere "en küçük kareler" yöntemini kullanan basit lineer regresyon hesaplayıcıları vardır. Daha sonra, y' den (bağımsız değişken) X' in (bağımlı değişken) değerini tahmin edersiniz (www.ibm.com, 2020). Dolayısıyla regresyon tekniği giriş x ve tahmin edilen çıkış y arasında aşağıdaki ilişkiyi doğrusal denklem şeklinde tanımlar (Joshi, 2020:34). Doğrusal regresyon algoritmasının doğruluk oranı hata matrisi ile elde edilen başarılı veri sayısı üzerinden hesaplanmıştır.

$$y = 0_1 + 0_2.X$$

Doğrusal regrasyon analizi ile yaptığımız çalışmada elde edilen değerler altta verilmiştir.

Basarili data sayisi : 2099

LN dogruluk orani : 70%

LN hesap suresi (sn): 0.08

LN k-CV Değerleri: [0.14172893, 0.1927147, 0.19500589, 0.23342316, 0.24026156]

Şekil 2: Doğrusal Regresyon Hata Matrisi (Confusion Matrix)

Lojistik Regresyon (Logistic Regression)

Lojistik Regresyon, doğru yada yanlış, 1 yada 0, evet veya hayır gibi ikili değere sahip bir değişkenin sonuçlarını açıklamak için lojistik fonksiyonu kullanan bir makine öğrenmesi yöntemidir. Lojistik regresyon y değişkeninin olma olasılığı ile veri setindeki x öznitelikleri arasındaki ilişkiyi anlamamıza yardımcı olur. Normalde lineer regresyon $-\infty$, $+\infty$ arasında sonuçlar verir. Oysa problemde olasılık değeri elde etmeye çalıştığımız için değerleri 0 ve 1 aralığına çevirmek için sigmoid fonksiyonu kullanılmaktadır (Joshi, 2020:37).

$$P(y = 1|x)$$

$$\sigma(x) = 1/(1 + e^{(-x)})$$

Lojistik regresyon analizi varsayılan parametreleri ile çalıştırıldığında yaklaşık 71% oranında bir doğruluk skoru elde edilmiştir.

Basarili data sayisi : 2113

LR dogruluk orani : 71%

LR hesap suresi (sn): 2.41

LR k-CV Değerleri : [0.72, 0.71333333, 0.70166667, 0.695, 0.69833333]

Şekil 3: Lojistik Regresyon Hata Matrisi (Confusion Matrix)

K-En Yakın Komşular (K Nearest Neighbors)

En yakın komşular, mevcut tüm vakaları saklayan ve yeni vakaları benzerlik ölçüsüne göre sınıflandıran basit bir algoritmadır. Bir vaka, komşularının çoğunluk oyu ile sınıflandırılır ve bir mesafe fonksiyonu ile ölçülen en yakın K komşuları arasında en yaygın sınıfa atanır. $K = 1$ ise, vaka en yakın komşusunun sınıfına atanır. Bu algoritma sınıflamasını Manhattan, Öklid ve Minkowski mesafe fonksiyonları ile benzerlik ölçüsüne dayalı olarak gerçekleştirir (Hazım, 2018). En iyi k değerini bulmak için uygulamada veri seti büyüklüğünde tekrar eden bir döngü ile skorlar bir dizine yazdırılmış ve en yüksek değeri veren skorun index değeri alınmıştır. En iyi komşu değeri 1 olarak elde edilmiştir. Ayrıca k için 1-12 arası değerler elle verilerek testler yapılmış ve k değeri arttıkça algoritmanın performansının ve doğruluk oranının düştüğü görülmüştür. Bu nedenle her iki yöntemde de en iyi performans ve doğruluk oranını veren 1 değeri alınmıştır.

K en yakın komşu algortiması çalıştırıldığında alttaki değerler elde edilmiştir.

En iyi n komsu degeri (k) : 1

Basarili data sayisi : 2264

KNN dogruluk orani : 75%

KNN hesap suresi (sn) : 0.51

KNN k-CV Değerleri : [0.72, 0.765, 0.77, 0.74, 0.74166667]

Şekil 4: K-En Yakın Komşu Hata Matrisi (Confusion Matrix)

Naive Bayes

Naive Bayes algoritması İngiliz matematikçi Thomas Bayes'ten ismini almıştır. Naive Bayes algoritması örüntü problemlerine oldukça kısıtlayıcı gibi görülen bir önerme ile kullanılan olasılıkcı bir yaklaşımdır. Buradaki önerme, örüntü tanımada kullanılan her bir tanımlayıcı öznelilik ya da parametrenin istatistik açıdan bağımsız olması gerekliliğidir. Bu önerme Naive Bayes sınıflandırmanın kullanımını kısıtlasa da, istatistiksel bağımsızlık koşulu esnetilerek kullanıldığında karmaşık yapay sinir ağları gibi metotlarla karşılaştırılabilir sonuçlar vermektedir (Aggarwal, 2014:67). Bayesian olasılığı altta veirilmiştir. Burada $P(Y|X)$ hipotezin olasılığını yani fraud olup olmadığını belirtir. $P(Y)$ hipotezin önceki deneyimlere sahip olma olasılığını gösterir. Son olarak formüldeki $P(X|Y)$ ise olasılık oranıdır (Hazım, 2018).

$$P(Y|X) = P(X|Y) \cdot P(Y) / P(X)$$

Naive Bayes algoritması varsayılan parametreleri ile çalıştırıldığında 70% doğruluk oranı elde edilmiştir.

Basarili data sayısı : 2089

NB dogruluk orani : 70%

NB hesap suresi (sn) : 0.04

NB k-CV Değerleri : [0.71333333, 0.68666667, 0.695, 0.70333333, 0.67666667]

Şekil 5: Naive Bayes Hata Matrisi (Confusion Matrix)

Karar Ağaçları (Decision Trees)

Karar ağacı algoritması, verilerin belirli bir parametreye göre sürekli olarak bölündüğü bir denetimli makine öğrenimidir. Karar ağaçları nodes, edges/branch ve leaf nodes olmak üzere üç bölümden oluşur. Karar ağaçları, classification ve regression olmak üzere iki ana sınıfa ayrılmaktadır. Classification ağaçları yes/no şeklinde çıktılar verirken, regression ağaçları ise değişkenin sürekli değerler alabildiği bir sınıflandırmadır (Barros, 2015). Karar ağacı ile yapılan çalışmamızda alttaki değerler elde edilmiştir.

Basarili data sayisi : 2267

DT dogruluk orani : 76%

DT hesap suresi (sn): 0.17

DT k-CV Değerleri : [0.735, 0.76833333, 0.78666667, 0.75833333, 0.75833333]

Şekil 6: Karar Ağacı Hata Matrisi (Confusion Matrix)

Destek Vektör Makineleri (Support Vector Machines)

Destek Vektör Makineleri, istatistiksel öğrenmeye dayanan kontrollü bir sınıflandırma algoritmasıdır. DVM' in matematiksel algoritmaları, başlangıç için iki sınıflı doğrusal verileri sınıflandırma problemi için tasarlanmış fakat daha sonra çok sınıflı ve doğrusal olmayan verilerin sınıflandırılması için genelleştirilmiştir. DVM sınıflandırma ve örüntü tanıma problemlerinin çözümü Vapnik tarafında geliştirilmiş ve Vapnik Chervonenkis teorisine dayanır. Şekil 7' de gösterilen ağ yapısında $K(x, x_1)$ çekirdek fonksiyonunu ve α lagrange çarpanlarını ifade etmektedir. Çekirdek fonksiyonlar ile girdilerin iç çarpımları bulunur ve Lagrange çarpanları ise ağırlıkları gösterir. Buradan yola çıkarak DVM ile bir örnek için çıktı değeri, girdilerin iç çarpımları ve lagrange çarpanlarının kombinasyonlarının toplamına eşittir diyebiliriz (Ayhan ve Erdoğan, 2014: 175-198).

Şekil 7: Destek Vektör Makinesi Yapısı (Ayhan & Erdoğan, 2014).

Destek Vektör Makineleri ile yapılan çalışmada 71% doğruluk oranı elde edilmiştir.

Basarili data sayisi : 2120

SVM dogruluk orani : 71%

SVM hesap suresi (sn) : 1.99

SVM kCV Değerleri : [0.705, 0.71166667, 0.70166667, 0.705, 0.70666667]

Şekil 8: Destek Vektör Makineleri Hata Matrisi (Confusion Matrix)

Bulgular ve Değerlendirme

Sigorta sektörü oto hasarları sahte hasar tespitinde makine öğrenmesi algoritmalarının kullanımı çalışmamızda üç bin kayıtlık bir veri seti kullanılmıştır. K parçalı çapraz doğrulama yöntemiyle 7 algoritma karşılaştırılmıştır. Ayrıca algoritmaların işlem süreleri de karşılaştırılmış ve performans açısından da bir değerlendirme yapılmıştır. Bu veri kümemizden fraud ve no fraud şeklinde iki sonuç alınmaktadır. Tüm algoritmaların doğruluk oranları Şekil-9 ve Tablo-2 de gösterilmiştir. Ayrıca başarılı data sayısı Şekil-10 da ve işlem süreleri de Şekil-11 de gösterilmiştir.

Tablo 2: Tüm Algoritmaların Sonuçları

Model	Başarılı Data	Doğruluk	İşlem Süresi (sn)
Random Forest	2578	86%	4.41
Logistic Regression	2113	71%	2.41
Linear Regression	2099	70%	0.08
K Nearest Neighbors	2264	75%	0.51
Naive Bayes	2089	70%	0.04
Decision Trees	2267	76%	0.17
Support Vector Machines	2120	71%	1.99

Şekil 9: Doğruluk Oranları **Şekil 10:** Başarılı Data Oranları **Şekil 11:** İşlem Süreleri

Sonuçlar incelendiğinde en yüksek doğruluk oranının rasgele orman algoritması ile elde edildiği görülmektedir. Bu algoritmayı %76 oranla karar ağaçları izlemektedir. K En Yakın Komşu algoritması da %75 başarı oranı ile Karar Ağaçları algoritmasıyla aynı seviyelerde performans göstermiştir. Algoritmaların işlem sürelerinin performanslarına bakıldığında birbirinden farklı değerler görülmektedir. Rasgele Orman algoritması doğruluk oranı en yüksek algoritma olurken işlem süresi olarak performansı en düşük algoritma olmuştur. İşlem süreleri açısından karşılaştırma yaptığımızda en hızlı algoritma Naive Bayes görülmektedir. Onu Doğrusal Regresyon ve Karar

Ağaçları algoritmaları izlemektedir. Belirli sayıda veri ile yapılan bu testlerde süreler her ne kadar düşük olsa da daha büyük ortamlar için bir gösterge olabilir.

Sonuç

Algoritmaların çıktıları karşılaştırıldığında en iyi sonuçların rasgele orman ve karar ağaçları ile alındığı görülmüştür. Rasgele Orman ve Karar Ağaçları denetimli makine öğrenmesi algoritmalarından ve sınıflandırma tekniklerini kullanan algoritmalar. Bu da bize veri setimizin bu modellere daha uygun olduğunu göstermektedir. En iyi skoru veren rasgele orman algoritmasının varsayılan parametre değerleri ile oynanarak daha iyi bir skor elde etmek için yapılan testlerde doğruluk oranı kayda değer bir yükseliş göstermemiştir. Bu nedenle bütün algoritmalar varsayılan değerleri ile çalıştırılmıştır. Sigorta sektöründe özellikle oto hasarları ile ilgili bu tarz çalışmaların sayısının az olması nedeniyle yaptığım çalışmanın bu alanda çalışmak isteyen uzmanlara veya akademik çalışanlara örnek teşkil edeceğini düşünmekteyim.

Kaynakça

- Aggarwal, C.C. (Eds.) (2015), Data Classification Algorithms and Applications, no. 67.
- Alpaydın, E. (2020), Introduction to Machine Learning, no. 3.
- Ayhan, S. ve Erdoğan, Ş. (2014), “Destek Vektör Makineleriyle Sınıflandırma Problemlerinin Çözümü İçin Çekirdek Fonksiyonu Seçimi”, Eskişehir Osmangazi Üniversitesi İİBF Dergisi, 9(1), 175- 198
- Barros R.C., de Carvalho A.C.P.L.F., Freitas A.A. (2015). Decision-Tree Induction. In: Automatic Design of Decision-Tree Induction Algorithms. No. 7-9. https://doi.org/10.1007/978-3-319-14231-9_2
- Bishop, M. (2006), Pattern Recognition and Machine Learning, no. 1.
- Demirci, S. (2019), Sigortacılıkta Yeni Bir Yaklaşım: Katılım Sigortacılığı, İnönü Üniversitesi Hukuk Fakültesi Dergisi – İnÜHFD 10(1): 25-39 (2019).
- Günbatar, E. (2019), Türkiye’ de Otomobil Sigortası Sahtekarlıklarının Makine Öğrenmesi Yöntemleri ile Tespit Edilmesi, yüksek lisans tezi, Hacettepe Üniversitesi Endüstri Mühendisliği Anabilim Dalı.
- Hazım, L.R. (2018), Four Classification Methods Naive Bayesian, Support Vector Machine, K-Nearest Neighbors and Random Forest Are Tested For Credit Card Fraud Detection, yüksek lisans tezi, Altınbaş Üniversitesi Fen Bilimleri Enstitüsü.
- Joshi, A. V. (2019), Machine Learning and Artificial Intelligence, no. 10-11, 34, 37, <https://dx.doi.org/10.1007/978-3-030-26622-6>
- Kızılkaya, Y.M. ve Oğuzlar A. (2018), Bazı Denetimli Öğrenme Algoritmalarının R Programlama Dili ile Kıyaslanması, 37(37), 90-98. <https://doi.org/10.17498/kdeniz.405746>
- Narin, A, İşler, Y, Özer, M. (2014). Konjestif Kalp Yetmezliği Teşhisinde Kullanılan Çapraz Doğrulama Yöntemlerinin Sınıflandırıcı Performanslarının Belirlenmesine Olan Etkilerinin Karşılaştırılması. Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi, 16 (48), 1-8. [Çevrimiçi]. <https://dergipark.org.tr/en/pub/deumffmd/issue/40797/492155>
- Oshiro T.M., Perez P.S., Baranauskas J.A. (2012) How Many Trees in a Random Forest?. Department of Computer Science and Mathematics Faculty of Philosophy, Sciences and Languages at Ribeirão Preto University. https://doi.org/10.1007/978-3-642-31537-4_13

-
- Yıldırım, İ. (2013), Türk Sigortacılık Sektörünün Yumuşak Karnı: Sigorta Suistimalleri Sorunu, *Sosyal ve Beşeri Bilimler Dergisi*, 5(1), 331-340.
- Zhu, X. ve Goldberg A.B. (2009), Introduction to Semi-Supervised Learning, <https://doi.org/10.2200/S00196ED1V01Y200906AIM006>
- www.ibm.com, [Çevrimiçi]. Available: <https://www.ibm.com/tr-tr/analytics/learn/linear-regression>
- www.turkishtimedergi.com, [Çevrimiçi]. Available: <http://www.turkishtimedergi.com/sigorta/sigorta-sektorunde-sahtecilik-buyuk-veriyle-sucustu>
- www.sas.com, [Çevrimiçi]. Available: https://www.sas.com/tr_tr/customers/sigorta-bilgi-merkezi.html