T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

SİGORTACILIK SEKTÖRÜNDE MAKİNE ÖĞRENMESİ İLE MÜŞTERİ KAYBI ANALİZİ

YÜKSEK LİSANS TEZİ

Hande Esin AKYİĞİT

Enstitü Anabilim Dalı	**:**	**BİLİŞİM SİSTEMLERİ MÜHENDİSLİĞİ**
Tez Danışmanı	**:**	**Dr. Öğr. Üyesi Tuğrul TAŞCI**

Temmuz 2021

T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

SİGORTACILIK SEKTÖRÜNDE MAKİNE ÖĞRENMESİ İLE MÜŞTERİ KAYBI ANALİZİ

YÜKSEK LİSANS TEZİ

Hande Esin AKYİĞİT

Enstitü Anabilim Dalı : BİLİŞİM SİSTEMLERİ MÜHENDİSLİĞİ

Bu tez 14.07.2021 tarihinde aşağıdaki jüri tarafından oybirliği / oyçokluğu ile kabul edilmiştir.

Jüri Başkanı **Üye** **Üye**

BEYAN

Tez içindeki tüm verilerin akademik kurallar çerçevesinde tarafımdan elde edildiğini, görsel ve yazılı tüm bilgi ve sonuçların akademik ve etik kurallara uygun şekilde sunulduğunu, kullanılan verilerde herhangi bir tahrifat yapılmadığını, başkalarının eserlerinden yararlanılması durumunda bilimsel normlara uygun olarak atıfta bulunulduğunu, tezde yer alan verilerin bu üniversite veya başka bir üniversitede herhangi bir tez çalışmasında kullanılmadığını beyan ederim

Hande Esin AKYİĞİT
25.05.2021

TEŞEKKÜR

Tüm eğitim öğretim hayatım süresi boyunca olduğu gibi yüksek lisans eğitimimde de beni destekleyip yanımla olan anne ve babama, her zorlukta yanımda olduğunu bildiğim eşime çok teşekkür ederim.

Tez dönemim boyunca kendisi ile beraber çalışmama fırsat veren, tüm aşamalarında sabırla, güleryüzüyle ve samimiyetiyle desteğini esirgemeyen değerli tez danışmanım Dr. Öğr. Üyesi Tuğul TAŞCI'ya teşekkürlerimi sunarım.

İÇİNDEKİLER

SİMGELER VE KISALTMALAR LİSTESİ

DVM	: Destek Vektör Makinesi
KNN	: K-NN Yakın Komşu
MİY	: Müşteri İlişkileri Yönetimi
NB	: Naive Bayes
RO	: Rastgele Orman

ŞEKİLLER LİSTESİ

TABLOLAR LİSTESİ

ÖZET

Anahtar Kelimeler: Müşteri Kaybı Analizi, Makine Öğrenmesi, Rastgele Orman Algoritması, Karar Ağacı Algoritması, K-En Yakın Komşu Algoritması

Hızlı büyüyen ve rekabet gücünün arttığı günümüzde, yeni müşteri edinme çabası ve maliyeti var olan müşteriyi kaybetmeme çabası ve maliyetinden fazla olması, firmaları var olan müşterilerin kaybedilmemesi gerektiği düşüncesine itmiştir. Güçlü rakiplerin olduğu sektörde tüketicilerin belli bir hizmet veya ürün için bir şirketi tercih ederek ihtiyaçlarını devamlı olarak bu şirket üzerinden karşılayan sadık müşteri portföyünü arttırmak amacıyla müşterinin tercih ettiği bir ürün veya hizmeti bırakması ihtimali üzerine analitik çalışmalar yapılmıştır. Mevcut müşterilerin profilleri ve davranışları incelenerek şirketi bırakma ihtimali olan müşterileri bulma, bu müşterilerin memnuniyetlerini arttırmayı hedefleyen müşteri kayıp analizi, stratejik karar verme ve planlama sürecinin en önemli aşamalarından biri olmaktadır. Bu çalışmada telekomünikasyon, bankacılık, online ticaret gibi müşteri sayısı ile gelir miktarının doğru orantılı olduğu sigortacılık sektöründe var olan şirketin verileri kullanılarak bir müşteriye ait içerisinde yaş, cinsiyet, doğum yeri gibi sosyodemografik bilgilerin yanı sıra kullanılan araç marka, model bilgilerinin de bulunduğu öznitelikler belirlenmiştir. Belirlenen öznitelikler makine öğrenmesi algoritmalarından Karar Ağacı (Desicion Tree) Algoritması, Rastgele Orman (Random Forest) Algoritması ve K-En Yakın Komşu (K Nearest Neighborhood) Algoritmaları ile terk eden müşterilerin profilleri analiz edilip, terk etme ihtimali olan müşteriler tahmin edilmiştir. Çalışmada en başarılı sonucu veren Rastgele Orman (Random Forest) Algoritması ile bu çalışma bir sınıfa dahil edilip son kullanıcı tarafından sürekli yapılmasına olanak sağlanmıştır.

CUSTOMER CHURN ANALYSIS WITH MACHINE LEARNING IN INSURANCE SECTOR

SUMMARY

Keywords: Customer Churn Analysis, Machine Learning, Random Forest Algorithm, Desicion Tree Algorithm, K-Nearest Neighbor Algorithm

In today's fast-growing and competitive world, the effort to acquire new customers and the effort of not losing the existing customer and its cost is more than the cost, there are thoughts to think companies to make an existing master. In the company where there are strong competitors, analytical research has been carried out on increasing the loyal customer portfolio that meets the consumers by choosing a company for a service or product that prefer a company for a certain service or product, and providing a product service preferred by the customer. Finding customers with the aim of researching campaigns and behaviors, customer loss analysis aiming to increase the satisfaction of these customers can be one of the most important stages of strategic decision making and planning. This phone is about the socio-demographic response of a customer such as the number of customers such as telecommunication, banking, online trade and the amount of income in the insurance industry, as well as the characteristics of the vehicle brand and model used, as well as the socio-demographic answer such as age, gender, place of birth. Desicion Tree Algorithm, Random Forest Algorithm and K-Nearest Neighbor (K Nearest Neighbor) Algorithms, attributes determined from machine learning algorithms are predicted by machine learning algorithms. With the Random Forest Algorithm, which gave the most successful results in the study, this study was included in a class and allowed to be continuously performed by the end user.

BÖLÜM 1. GİRİŞ

Son zamanlarda telekom, sigortacılık, hizmet gibi çeşitli sektörlerde faaliyet gösteren birçok şirket küreselleşme, artan rekabet koşullarında varlığını korumak ve devam ettirebilmek amacıyla çeşitli stratejiler geliştirmek zorunda kalmışlardır. Artan rekabet koşullarında müşteri davranışları, profilleri ve beklentileri oldukça değişmiştir. Firmalar bu rekabet ortamında üstünlük kazanmak için çeşitli yöntem arayışı içine girmişlerdir. Aranılan yöntemlerin ihtiyaca cevap verebilmesi için firmaların var olan müşterilerinin profillerini analiz edebilmesi, satın alma alışkanlıklarını izleyebilmesi, işletmenin amaç ve hedefleri doğrultusunda müşteri ilişkileri yönetimini faydalı bir biçimde kullanabilmesi ile mümkün olacaktır. Müşteriyi elde tutmak, müşteri merkezci yöntemler ile müşteriyi değerli kılmak, bağlılık arttırmak firmaların en büyük hedefleri arasında yer almaktadır.

Günümüzde sigortacılık, telekomünikasyon, finans, e-ticaret, hizmet gibi sektörlerde faaliyet göstermekte olan firmaların değerleri, sahip oldukları müşteri sayısı ve profili ile ölçülmektedir. Daha az maliyet ile daha fazla müşteri profiline ulaşıp, var olan müşterileri elde tutmayı hedefleyen firmalar sahip oldukları müşterileri kaybetmemek için öncelikli olarak aldığı hizmeti veya ürünü kullanmayı bırakma ihtimali olan müşterileri tahmin edebilmek ve bu müşterilere yönelik promosyon ve pazarlama çalışmaları yapmak. Yani ürün veya hizmeti bırakma ihtimali olan müşteri kitlesi tespit edilip, memnun edilebilirse müşteri kaybı minimuma indirilmiş olacaktır.

Günümüzün ekonomi dünyasını "müşterinin kral olduğu bir müşteri ekonomisi" şeklinde tanımlamak doğru olacaktır. (Keller ve Kotler, 2015). Böyle tanımlanan bir ekonomi dünyasında müşteriler şirketlerin var olma nedenleri arasındadır ve uzun vadede müşteri memnuniyetinin sağlanması, ancak etkili bir Müşteri İlişkileri

Yönetimi (MİY) süreci oluşturmakla mümkün olmaktadır. (Chorianopoulos ve Tsiptsis, 2009). Doymuş pazarlar nedeniyle ve mevcut koşullar altında yoğun rekabet ortamında güçlü durabilmek için işletmeler müşteri ilişkileri yönetimine odaklanarak, kârı yüksek olan müşteriler ile daha kârı az olan müşteriler arasında bir seçim olanağı kazanmıştır. Müşteriler ile ilgili verilerin toplanması, verilerden anlamlı bilgiler elde edilerek analiz edilmesi, müşterilerin hizmeti veya ürünü alma alışkanlıkları ve davranışlarına bakılarak firmaya kattığı değerlerin keşfedilmesi, müşterileri değerlerine göre sınıflandırmak, müşteriyle çok yönlü iletişimin sağlanması, müşteri bağlılığını arttıracak stratejilerin gerçekleştirilmesi ve çeşitli ürünlerin birlikte satışı ve üst satışlarının gerçekleştirilmesi gibi temel hedefler ile Müşteri İlişkileri Yönetimi (MİY) özetlenmiştir. (Bagheri ve Tarokh, 2015).

Müşteri İlişkileri Yönetim (MİY) kavramının teknoloji ve yönetim olarak ikiye ayrılan yapısı, oluşan kavramı operasyonel, stratejik ve analitik yönleri ile ele alınmasını gerektirmektedir. (Maklan ve Buttle, 2015). Bu süreçleri ele alacak bir Müşteri İlişkileri Yönetimi (MİY) yapısının aynı zamanda bazı kaynaklara odaklanması gerekmektedir. Bu anlamda faydalı bir Müşteri İlişkileri Yönetimi (MİY) süreci, müşterilerin kazanımı, elde tutulması, kaybı ve geri kazanımı olmak üzere 4 ana başlık çerçevesi ile incelenebilmektedir. (Petersan ve Kumar, 2012).

Bu dört ana başlıklar arasında var olan müşteri kaybı, en basit ifadesiyle "rekabet nedeniyle müşterilerin firmayı tercih etmekten vazgeçmeleri" anlamına gelmektedir (Nettleton, 2014).

Müşteri kaybı tanımı rekabet koşullarında firmaların müşterilerin tutumlarında meydana geldiği gözlemlenen değişimler sonucunda bulunduğu firmayı bırakmalarının artmasıyla daha fazla gündeme girmiştir. Müşteri kayıplarının yüksek olmasının sebeplerinde, rakip firmalarda tatminsizlik yaşayan müşterileri farklı özellikleri vurgulayarak farklı kampanyalar oluşturarak müşteriyi kendi firmasına çekme oldukça etkili olmaktadır. "Müşterilerin son teknolojilere erişme imkânları, Müşteriyi düşünüp, faydalarını gözeten çalışan kadrosu, daha düşük ücretlendirme, değişim maliyeti, yapılan tanıtımların etkisi, konum ve çeşitli hizmet teklifleri

sebebiyle rakip firmalara geçme yatkınlığı göstermiştir. (Farquad, Ravi ve Raju, 2014). Geçiş eğilimi gösteren müşterileri önceden tespit edebilmek ve o müşteriler üzerinden farklı kampanyalar uygulayarak müşterinin tatminsizliğini azaltıp yeniden firmaya bağlamak günümüz işletmeleri için çok önemli bir yer edinmiştir.

Paralelinde bilişim sektöründe yürütülen gelişmeler sayesinde işletmeler, verileri saklayabilir, kolay erişebilir, işleyebilir ve işlenen veriler ile anlamlı bilgiler elde edilme imkanına sahip olabilmektedirler. Çeşitli sektörlerde bulunan işletmeler, müşterilerinin satın alma davranışlarının yanı sıra müşteriye ait farklı özellikleri de depolamaktadırlar. Depolanan veri yığınlarından makine öğrenmesi yöntemleri kullanılarak anlamlı bilgiler, isabetli tahminler elde edilebilmektedir. Elde edilen sonuçlar, müşteri odaklı uygulamaların geliştirilebilmesine ve müşteri kaybını engellemek için önemlemler alınmasına fayda sağlamaktadır.

Makine öğrenmesi, büyük veriler üzerinden anlamlı çıkarımlar yapabilen, yapısal olarak öğrenebilen algoritmalara verilen isimdir. Günümüzde verilerin insan gücü ile işlenemeyecek boyutlara gelmiş olması sebebiyle makine öğrenmesi konusu oldukça önemli bir hale gelmektedir. Firmalar tuttukları müşteri bilgilerini makine öğrenmesi yöntemleri sayesinde anlaşılabilir, sınıflandırılmış ve ilişki kurulmuş anlamlı bilgiler elde etmektedirler.

Literatür incelemeleri yapıldığında kayıp olacak müşterileri tahmin etmeye yönelik çalışmaları yürüten ve yakından takip eden sektörlerin başında telekomünikasyon ve iletişim sektörü gelip devamında ise bankacılık sektörü, perakendecilik sektörü, enerji sektörü ile oyun ve eğlence sektörlerine ait önde gelen firmalar makine öğrenmesi ile müşteri kayıp analizi yapılmaktadır.

Günümüz sigortacılığında da diğer sektörler gibi rekabet ortamının artması ve rakip firmaların müşterileri elde etme çabası sonucu firmalar müşterilerini kaybetmemek için çalışmalar yapmaya başlamıştır. Yapılan literatür incelemeleri sonucunda sigortacılık sektöründe müşteri kaybı analizinin diğer sektörlere oranla daha az yapıldığının fakat bu çalışmaya müsait veri kapasitelerinin olduğu tespit edilmiştir.

Bu sebeple bu çalışmada artan rekabet koşullarında var olmak ve tutulan veri kapasitelerinin müşteri kayıp analizine uygun olması sebebiyle bir sigortacılık sektöründe özel bir şirkete ait müşteri verileri üzerinde makine öğrenmesi yöntemleri ile müşteri kayıplarına yönelik modeller geliştirilerek tahminler yapılmış ve bu modellerin çıktılarını ilgili olan herkesin istediği zaman görebilmek istemesi sebebiyle bir arayüz oluşturulmuştur. Bu çıktılar ile birimlerin kendi müşterilerini kaybetmemesi için gerekli aksiyonların alınmasına imkan sağlamaktadır.

Büyüyen rekabet ortamında var olabilme, kayıp olabilecek müşterileri tahmin edip kaybetmemek için gerekli aksiyonların alınması hedeflenen bu çalışmada ilk olarak müşteri kayıp analizini yapan çalışmalar incelenmiş, bu çalışmalarda hangi yöntemlerin kullanıldığı, hangilerinde başarı oranının yüksek olduğu bilgilerine erişilmiştir. İncelenen kaynaklar doğrultusunda müşteri kayıp analizinde yüksek başarı oranı veren makine öğrenmesi algoritmalarından Rastgele Orman (Random Forest) Algoritması, Karar Ağacı (Decision Tree) Algoritması ve K-En Yakın Komşu (K-Nearest Neihbors) Algoritmalarından bahsedilmiştir. Bir sonraki aşamada Sigortacılık Sektöründe Makine Öğrenmesi Yöntemleri İle Müşteri Kaybı Analizi çalışmasında kullanılan veri setinden ve veri ön işleme aşamalarından bahsedilip sonrasında uygulamanın nasıl yapıldığı anlatılmıştır. Son bölümde sonuçların nasıl değerlendirilmesi gerektiğine değinilip, bu çalışmada kullanılan algoritmalar değerlendirilip aralarındaki başarı oranları karşılaştırılmış ve en yüksek başarı oranı veren algoritma ile yapılan uygulamadan bahsedilip çalışma kaynaklar da eklenerek sonlandırılmıştır.

BÖLÜM 2. LİTERATÜR TARAMASI VE İLGİLİ ÇALIŞMALAR

Son dönemlerde makine öğrenmesi ile müşterinin kaybedilmesi, sadakatinin ölçülmesi, geri kazanımının yolları, geleceğe yönelik tahminlerin yapılması pek çok makalenin konusu olmaktadır. Bu çalışmadaki literatür taraması çalışması sırasında müşteri kaybı analizi yapan, kayıp müşteri profillerini çıkaran ve bu müşterileri tahmin etmeye çalışan çalışmalar tespit edilip kullandıkları yöntemler detaylı incelenmiştir.

Literatür taramasında incelenen ilk çalışmalardan Türk iletişim sektöründe, son 6 ay içinde yükleme yapılmayan kontörlü hatları iptal eden bir mobil iletişim şirketi verileri kullanılarak müşteri kaybı analizi yapılmıştır. Karar ağaçları ve regresyon modellemesi kullanılarak her müşterisine bir puan atayıp bu puanların analizi sonucu 6 ay sonra kaybedilecek müşterileri tahmin etmeye çalışıp, bu müşteriler içinden yüksek kontör yükleyenlerin şirkete sağladığı katkılar değerlendirilerek müşteri kaybı analizi çalışmalarının artması ve yayılması sağlanmıştır. (Özmen, 2006). Yine aynı yıl içinde Yapı Kredi Bankası müşterilerinden kredi kartı kullanan müşterilerin çeşitli özelliklerdeki bilgileri incelenerek, kaybedilmiş bir müşteri profilinin veri madenciliği yöntemleri ile ortaya çıkarılmıştır. C Programlama dili ile Karar Ağacı (Decision Tree) Algoritma'sı kullanılarak farklı kurallar doğrultusunda eşik değerleri belirlenmiş ve kurallar elde edilmiştir. Bu kurallar inceliterek, müşteriyi kaybetme sebepleri ve ne zaman kaybedildiği bilgisine erişilmeye çalışılmıştır. (Tosun, 2006).

2009 yılında yapılan bir çalışmada sınıflandırma yöntemi olarak Rastgele Orman (Random Forest) ve Lojistik Regresyonu kullanılarak müşteri kaybı analizi konusunda çalışma yapılmıştır. Çalışmada veri seti 6 farklı kategoride ele alınarak nasıl işleneceği ve nasıl makine öğrenmesi yöntemlerinde kullanılacağı üzerine çalışılmıştır. (Burez ve Van den Poel, 2009).

Müşteri kaybı analizi yapılan sigortacılık, telekom sektörlerinin yanı sıra kozmetik sektörüne ait bir firmanın müşterileri üzerinde müşteri kaybına yönelik bir çalışma yapılmıştır. Sınıflandırma ve kümeleme algoritmaları kullanılarak yapılan çalışmada müşterilere ait veriler analiz edilerek ayrılma ihtimali yüksek olan müşteriler belirlenip bu müşteri grubu için strateji ve kampanyalar geliştirilmiştir. Çalışmada kullanılan algoritmalar arasından en yüksek başarıyı veren algoritmanın J48 olduğu tespit edilmiştir. (Aydoğan, Gencer ve Akbulut, 2009).

Müşteri kaybı analizinde programlama dillerinin yanı sıra açık kaynaklı yazılımlar da oldukça yaygın kullanılmaktadır. 2011 yılında yapılan çalışmada veri madenciliği teknikleri kullanılarak analizi ve bilgi çıkarımı zor olan veri seti incelenerek içerisinden gizli kalmış faydalı ve anlamlı veriler açığa çıkartılmıştır. WEKA, Tanagra ve Scikit-Learn gibi açık kaynaklı yazılımlar kullanılarak çeşitli algoritmalar denenmiş ve bu algoritmaların karşılaştırılmaları yapılmıştır. (Coskun ve Bayrak, 2011).

Sigortacılık sektöründe yapılmış olan bir çalışmada bireysel emeklilik yaptıran müşterilerin çeşitli niteliklerde bilgileri incelenerek kaybedilen müşteri profili Karar Ağacı (Decision Tree) yöntemleriyle ortaya çıkarılmıştır. Çıkarılan kaybedilmiş müşteri profiline göre kural tablosu oluşturularak bu kurallar sonucu müşteri kayıplarının sebepleri ve ne zaman gerçekleştiği bilgilerine ulaşılmıştır. (Koçtürk, 2010).

Bankacılık sektöründe 2015 yılında müşteri kaybı analizinin yanı sıra kaybedilmemiş fakat kar olarak azalmış müşteriler tespit etmeye çalışılmıştır. Karar ağaçları ve lojistik regresyon yöntemleri kullanılan çalışmada veri seti ön işleme aşamalarından geçirilmiş, oluşan veri seti ile model oluşturulmuştur. Model oluştuktan sonra karışıklık matrisine göre doğruluk oranının %89 olduğu belirtilmiştir. (Karaağaç, 2015).

Karar ağaçları, rastgele orman algoritması, lojistik regresyon kullanılan müşteri kaybı analizi çalışmalarına ek olarak 2015 yılında telekom sektöründe yapışan bir çalışmada yapay sinir ağlaru, destek vektör makineleri, karar ağaçları, naive bayes ve regresyon analizi uygulanarak müşteri kayıp analizi yapılmıştır. Uygulanan algoritmalar arasındaki sonuçların karşılaştırılmasıyla birlikte benzer çalışmalardan farklı olarak uygulanan her yöntemden sonra performans yükseltici (boosting algorithm) uygulanmıştır. Performans yükseltici (boosting algorithm) kullanılmadan önce en yüksek başarı oranını veren algoritmanın yapay sinir ağları olduğu gözlemlenirken, performans yükseltici kullanıldıktan sonra en yüksek başarı oranını destekçi vektör makinelerinin verdiği gözlemlenmiştir. (Vafeiadis ve ark., 2015).

Bizim çalışmamıza benzer olarak yapılan, 3 milyondan fazla müşteriye sahip olan Hollandalı sağlık sigorta şirketi Centraal Ziekenfonds verileri kullanılarak müşteri kaybı tahminlemesi yapmıştır. Rekabetçi ve dinamik bir ortamda var olan müşterileri elde tutabilmek için çalışmada öncelikli olarak kaybedilecek müşterilerin önceden tahmin edilmesi gerektiğine inanılmaktadır. Bu sebeple veri madenciliği tekniklerinden Lojistik Regresyon, Karar Ağacı (Decision Tree), Sinir Ağları ve Destek Vektörü Makine kullanıalrak Knime üzerinde kayıp müşteri tahminlemesi yapılmıştır. (Huigevoort, 2015).

2017 yılında yapılan çalışmada Destek Vektör Makineleri (DVM), Yapay Sinir Ağları (YSA) ve Naive Bayes (NB) gibi çeşitli sınıflama yöntemleri kullanılarak telekomünikasyon sektöründe müşteri kaybını tahmin etmeye yönelik bir analiz gerçekleştirilmiştir. Bu analiz sonucu sadık ya da terk eden müşterileri en iyi tahmin eden yöntemin yapay sinir ağları olduğu tespit edilmiştir. (Kaynar, Tuna, Gömer ve Deveci, 2017).

Müşteri kaybı analizi üzerine yapılan çalışmalar incelendiğinde bu çalışmaların Karar Ağaçları, Yapay Sinir Ağları, Rastgele Orman, Destek Vektör Makinesi, Naive Bayes, K-NN Yakın Komşu, Regresyon gibi farklı yöntemler kullanılarak yapıldığı bilgisine erişilmiştir. Çalışmalarda tek bir yöntem üzerinde durulduğu gibi hibrit modellere de yer verilmiştir. Bunlar arasında daha sık tercih edilenlerin ise destek

vektör makineleri, karar ağaçları ve derin öğrenme olarak da geçen yapay sinir ağları olduğu görülmektedir.

Yapılan çalışmalar ve başarı oranları kaynak alınarak oluşturulan bu çalışmada müşteri kaybı analizinin en yaygın olarak kullanıldığı telekom sektöründen sonra gelen sigortacılık sektöründe bir firma müşterileri ile yapılmıştır. Çalışma sırasında makine öğrenmesi algoritmalarından iyi başarı sonucu veren rastgele orman algoritması, Karar Ağacı (Desicion Tree) Algoritması ve K-En Yakın Komşu Algoritma'sı kullanılmıştır. Bu çalışmadaki amaç önceki çalışmalar gibi algoritmaları uygulayıp başarı oranı karşılaştırmanın yanı sıra en iyi sonucu veren algoritmayı bir sınıfa dahil edip son kullanıcı tarafından tekrar tekrar kullanabilmesine olanak sağlamaktır.

BÖLÜM 3. MAKİNE ÖĞRENMESİ

Yapay zeka branşlarından olan makine öğrenmesi, yapısal olarak öğrenebilen, büyük ve karmaşık veriler üzerinde anlamlı çıkarımlar yapabilen bilgisayar algoritmalarına verilen isimdir. Dijital veri üretiminin kurumsal ve bireysel düzeyde hızla artış gösterdiği günümüz dünyasında, verilerinin insan gücü ile işlenemeyecek boyuta gelmiş olması, klasik algoritmalar ile verilerin sınıflandırılmasının mümkün olmaması ve ileriye dönük tahminlemenin yapılamaması sebebiyle makine öğrenmesi ve yöntemleri önemli bir hale gelmiştir. Arthur Samuel 1959 Yılında makine öğrenmesini "Açıkça programlanmadığı halde makinelere öğrenme yeteneği kazandıran disiplin." olarak tanımlamıştır. Makine öğrenmesi büyük veriler üzerinde anlamlı çıkarımlar yapabilen, yapısal olarak öğrenebilen algoritmalara verilen isim olmaktadır.

Makine öğrenmesi, Gözetimli (Supervised), Gözetimsiz (Unsupervised) ve Takviyeli (Reinforcement) olmak üzere üç temel gruba ayrılmaktadır.

Hedef değişkeni belli olan öğrenme yöntemi gözemli öğrenme(supervised learning) olarak adlandırılmaktadır. Tahmin edilmek istenen sınıflar bellidir, buradaki amaç girdi değerleri ile hedef değişken arasında bir bağlam öğrenerek yeni gelen değerlerde bu bağlamdan yola çıkarak tahminler yapmaktır. Veri setindeki hedef değişkeni kategorik ise sınıflandırma (classification), nümerik ise regresyon (regression) algoritmaları kullanılmaktadır.

Gözetimsiz öğrenme (unsupervised learning) yöntemi şeklinde hedef değişkeni bulunmamakta, sadece girdi bulunmaktadır. Amaç girdiler arasındaki yakınlıklar, düzenlilikleri bulmaktır. Gözetimsiz öğrenme ile kümele ve kestirim yapılabilmektedir.

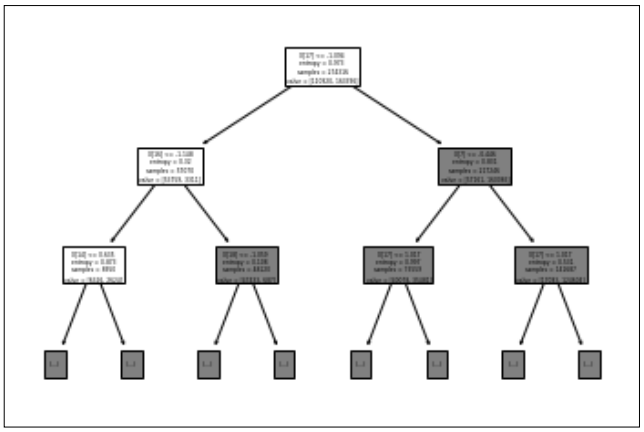
Gözetimli ve gözetimsiz öğrenmeden daha farklı olan takviyeli öğrenme (reinforcement learning) yönteminde ödül-ceza olarak adlandırılan sistem bulunmaktadır. Burada makinenin ana hedefi elde etmek istenilen eyleme ulaşırken kullanılan en doğru yolu bulmaktır. Doğru yolu bulmaya çalışırken yaptığı hatalardan çıkarımlar yaparak belli bir ödül-ceza sistem temeli üzerinde çalışmakta ve çıkarımlardan en optimize yol ile doğru eylem bulunmaya çalışılmaktadır.

Bu çalışmada, kullanılan veri setinde elde edilen hedef değişkeninin kategorik olması sebebiyle Gözetimli (Supervised) öğrenme yöntemlerinden K-En Yakın Komşu (K-Nearest Neihbors) algoritması, Karar Ağacı (Decision Tree) Algoritması ve bu karar ağaçlarının birleşmesi ile oluşturulan Rastgele Orman (Random Forest) algoritması kullanılmıştır.

3.1. Karar Ağaçları

Tümevarım metodunu kullanarak verilerin sınıflandırılmasını veya sonuç tahmini yapılmasını sağlayan karar ağaçları, özellik ve hedefe göre karar düğümleri (decision nodes) ve yaprak düğümlerinden (leaf nodes) oluşan ağaç yapısı formunda bir model oluşturan makine öğrenmesi sınıflandırma yöntemlerindendir. Karar ağaçları kök düğüm, düğüm ve dalladan meydana gelmektedir. Kurallar doğrultusunda sorular sorulup bu kuralların cevapları bir araya getirilerek, yeni kurallar oluşturulmaktadır.
Peşi sıra sorulan soruların cevabına göre belirlenen değişken ağaç yapısının ilk aşaması olan kök düğüm oluşturulmaktadır. Bağımlı değişkeni temsil eden kök ile başlanan ağacın yapısında aşağı doğru gidildikçe veri kümeleri daha küçük parçalara ayrılarak dalları ve yeni düğümleri oluşturmaktadır. Her bir soruyu bir düğüm temsil etmekte ve verilen cevap sayısı kadar dar oluşmaktadır. Ağaçta bulunan dallar, 'eğer-ise' kurallarını oluşturmakta ve seçilen adıma göre bir sonraki düğüme aktarılmaktadır.

Farklı bir soru meydana gelene kadar ağaç yapısı bu şekilde meydana gelmektedir. Son olarak bir sınıfı temsil eden Şekil 3.1.'de örnek olarak gösterilen son düğüme ulaşılmaktadır.



Şekil 3.1. Örnek Karar Ağacı (Desicion Tree) Modeli

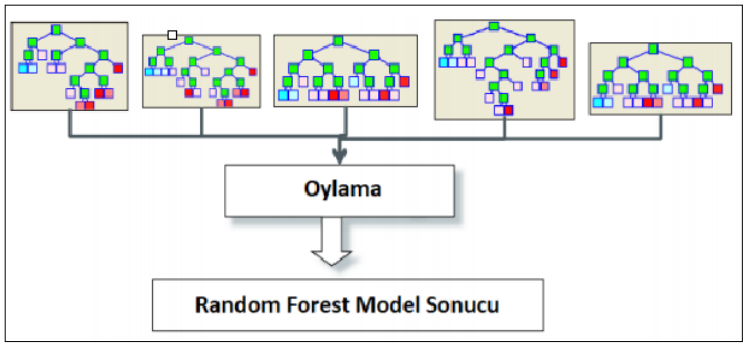
Karar ağaçları, diğer istatistiksel çözümlere göre oluşturulması daha kolay, ağaç şeklinde sınıflandırıcılardır. Modelin ağaca benzeyen yapısı, bağımlı ve bağımsız değişkenler arası ilişkiyi göstermektedir. Ağaçta meydana gelen yaprak düğümleri hedef niteliğinin değeri olmaktadır. Karar düğümü ise nitelikte uygulanan test değerini göstermektedir. Karar ağacı verilen örneği kökten yaprağa kadar inceleyip sınıflandırmaktadır. Karar ağaçlarının öğrenme algoritmaları, bir hipotezi göstermek için karar ağacı kullanmaktadırlar. Öğrenme kümesinde, ham veri incelenerek mümkün olan en iyi şekilde sınıflandırılmaktadır. (Tosun, 2006). Algoritma bu işlemi recursive olarak tekrar edip, son ortaya çıkardığı karar ağacı en son hipotezi oluşturmaktadır. İdeal olan karar ağacı, öğrenme kümesi dışındaki verilerde de aynı kuralları oluşturur ya da az hata payıyla aynı hipotez sonuçlarını ortaya çıkarmaktadır. (Moore, 2001).

Karar ağaçları hem kategotik hem de sayısal verileri işleyip anlamlandırabilmektedir aynı zamanda bağımsız ve bağımlı değişkenler sebebiyle eksik veya kayıp değerlerden etkilenmemektedirler.

Bazı durumlarda ağaç modeli oluşturma ve ağaç budama karmaşıklığı çok olabilmekte veya az girdi ile oluşturulan karar ağacı modeli anlamlı bilgi yansıtamayabilmektedir.

3.2. Rastgele Orman (Random Forest) Algoritması

Karar ağacı modellerinin en büyük problemleri arasında veriyi ezberleme – aşırı öğrenme bulunmaktadır. Rastgele orman modeli bu problemi önlemek amacıyla veri setinden rassal olarak 100'lerce farklı alt ağaçlar seçip bunları eğitmektedir. Bu yöntem sebebiyle 100'lerce karar ağacı modeli oluşturulup, oluşturulan karar ağaçları bireysel olarak tahminde kullanılmaktadır. Gün sonunda ise problem regresyonsa karar ağaçlarının tahminlerinin ortalaması, problem sınıflandırmaysa tahminler arasında en çok oy alan seçilmektedir. Bu yöntem en büyük problem olan aşırı öğrenme ve ezberlemenin de önüne geçip, farklı veri setlerinden dolayı aykırı veri (outlier) problemini de minimum seviyeye inmektedir.



Şekil 3.2. Örnek Rastgele Orman (Random Forest) Modeli

Birden çok karar ağacı modeli ile oylama yapılarak oluşturulan rastgele orman algoritması, karar ağaçlarında görülmekte olan aşırı öğrenmenin önüne geçmiş olmaktadır. Her bir karar ağacı için kullanılan farklı veri setleri sebebiyle aykırı veri (outlier) problemini de mininmum seviyeye indirmektedir. Rastgele Orman (Random Forest) yöntemi bir bagging yöntemi olmaktadır. Bagging yönteminde birden fazla yöntem paralel olarak farklı veri seti kümeleri ile eğitilmekte ve tüm modellerin oluşturduğu sonuç oylamaya tabii tutularak nihai sonuç ortaya

çıkarılmaktadır. Basit rastgele orman algoritması modeli Şekil 3.2.'de gösterilmektedir.

3.3. KNN (K-Nearest Neihbors) Algoritması

KNN algoritması içerisinde tahmin edilecek değerin bağımsız değişkenlerinin oluşturduğu vektörün en yakın komşularının hangi sınıfta yoğun olduğu bilgisi üzerinden sınıfı tahmin etmeye dayanmaktadır. KNN (K-Nearest Neihbors) algoritması denetimli öğrenme yöntemlerinden biri olarak hem sınıflama hem de regresyon tarafında oldukça kullanılmaktadır. Özet olarak tanımlamak gerekirse sınıfı henüz belirlenmemiş veri eğitim setinde bulunan diğer veriler ile karşılaştırılarak uzaklık ölçümü yapılıp bu ölçüme göre en optimal sınıf bulunmaktadır.

KNN (K-Nearest Neihbors) Algoritması iki temel değer üzerinden tahmin yapmaktadır ve uzaklık hesaplamaları sırasında kullanılması amacıyla Oklid (Euclidean), Manhattan ve Minkowski Uzaklık Fonksiyonları gibi farklı fonksiyonlar belirlenmektedir. Bu fonksiyonlar arasında yaygın olarak kullanılan Öklid Uzaklık Fonksiyonu olmaktadır. Fonksiyon ile p boyutlu bir uzayda i ve j noktaları arasında bulunan uzaklık Denklem 3.1.'de verilen formül ile elde edilmektedir:

$$d(i,j) = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2} \tag{3.1}$$

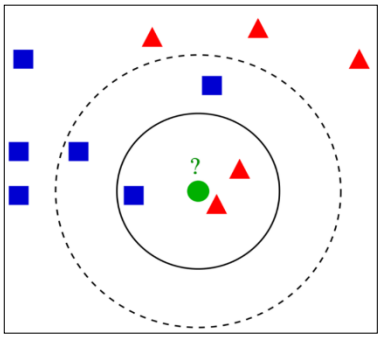
Veri setinde bulunan değişkenlerin sayısının ikiden fazla olduğu durumlarda Standardize Edilmiş Öklid Uzaklık Fonksiyonu kullanılmaktadır. Her bir değişken kendi içerisinde z dönüşümü uygulanarak standardize edilip, eşitlik formüle yerleştirilmektedir. Böylelikle değişkenler arasındaki ölçüm farklılıkları ortadan kalkmış olmaktadır.

KNN (K-Nearest Neihbors) Algoritması uygulanırken k seçimi oldukça önemli bir kısımdır. Komşuluk değerini ifade eden değerin (k) küçük bir sayı olduğu

durumlarda gözlemlerin şiddetli vurgulanmasına bağlı olarak algoritmanın performansında bir bozulma görülebilmektedir.

KNN (K-Nearest Neihbors) algoritması kullanılarak elde edilen KNN Kayıp Değer Atama yöntemi ilk olarak Dixon tarafından 1979 yılında sınıflandırma problemlerinde kayıp veri sorunu için kullanılmıştır. Daha sonrasında mikrodizilim veri setlerinde ve yapay sinir ağları gibi dallarda da kullanılmaya başlanmıştır.

Literatürde Acuna ve Rodriguez tarafından geliştirilen atama algoritmasında uzaklık fonksiyonunun hesaplandığı gözlem satırlarında kayıp veri içermesine izin verilmemektedir. Geliştirilen kayıp değer algoritmasından bahsedilen Şekil 3.3.'de grafikte, K=3 (düz çizginin olduğu yer) seçilirse sınıflandırma algoritması "?" işareti ile gösterilen noktayı, kırmızı üçgen sınıfı olarak tanımlaması beklenmektedir. Fakat K=5 (kesikli çizginin olduğu alan) seçilirse sınıflandırma algoritması mavi kare sınıfı olarak tanımlayacaktır.



Şekil 3.3. K değerinin önemi

KNN (K-Nearest Neighbors) Algoritması ile üretilmiş bir modelin başarısını ölçmek için genel olarak kullanılan 3 adet indikatör bulunmaktadır. Bunlardan ilki Jaccard Index adı verilen, doğru tahmin kümesi ile gerçek değer kümesinin kesişim kümesinin bunların birleşim kümesine oranı olmaktadır. 1 ile 0 arası değer alıp, 1 en iyi başarı anlamına gelmektedir. Confusion Matriks üzerinden hesaplanan Precission ve Reccal değerlerinden hesaplanan F1-Score, 1 ile 0 arası değer alıp, 1 en iyi başarı anlamına gelmektedir. Logistic Regresyon sonunda tahminlerin olasılıkları üzerinden

ise Logloss değeri hesaplanmaktadır. 1 ile 0 arası değer alıp, Jaccard Index ve F1-Score değerlerinden farklı olup 0 en iyi başarı anlamına gelmektedir.

Yapılan çalışmada en yaygın kullanılan yöntemleri ile makine öğrenmesi kullanılarak müşteri kaybı analizi yapılıp, bir sonraki bölümde adımlarından bahsedilmektedir.

BÖLÜM 4. SİGORTACILIK SEKTÖRÜNDE MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE MÜŞTERİ KAYBI ANALİZİ

Sınıflandırma ve tahminleme algoritmalarında kullanılan verinin kalitesi ve boyutu yapılan tahminleme sonucunda başarı oranı ile doğru orantılıdır. Bu sebeple modeli eğitmek için kullanılacak veri setinin hazırlanması en önemli aşamalardandır. (Kunt 2019)

Bu çalışmada modeli oluşturma, eğitme ve kullanma aşamalarından önce modeli oluştururken ve tahminleme yaparken kullanacağımız veri seti özelllikleri çıkarılmış ve uygun formata getirilmiştir.

Bu çalışma bir sigorta şirketine ait anonim veriler ile yapılmıştır. İlgili sigorta şirketinin arabalarına kasko poliçesi yaptıran müşterileri arasında sadece veri işleme izni olan müşterilerine ait veriler anonimleştirilip, KVKK kapsamında kullanıma uygun hale getirilmiştir.

Müşteri kaybı analizi için poliçesini iptal eden ve poliçesini yenileyen olmak üzere iki farklı tipte müşteri profiline ihtiyaç duyulup bu formatta veri seti hazırlanmıştır. Müşterilerin poliçesini yenileyip yenilemediğini belirleyen bu öznitelik yanında, kayıp olan müşteri profillerini belirlemek amacı ile sosyodemogrofik bilgiler, kullanılan araç, marka, model bilgileri, poliçesini daha önce kaç kez yenilediği bilgileri de yer almaktadır.

4.1. Veri Setinin Oluşturulması

Tahminleme için kullanacağımız model içerisindeki veri setinde, kasko poliçesini yenileyen ve yenilemeyen müşterilerin sosyodemografik bilgileri, kullandığı aracın öznitelikleri, poliçe özellikleri gruplanarak tutulmaktadır.

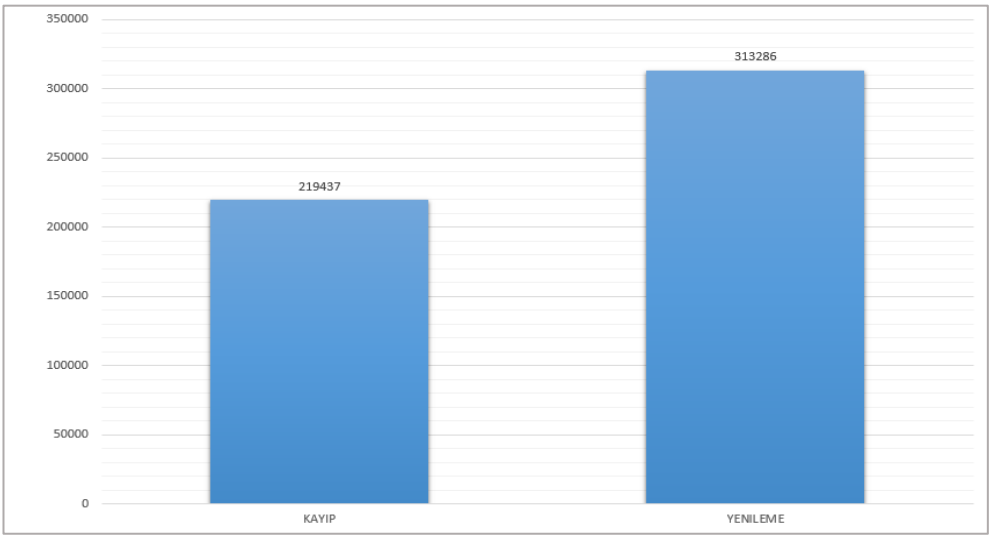
Müşterileri ait Tablo 4.1.'deki maddelerde yer alan öznitelikler çıkarılmış ve veri seti oluşturulmuştur.

Tablo 4.1. Veri setinde kullanılaran öznitelikler ve açıklamaları

Öznitelik	Açıklama
Yaş	Müşterinin yaş bilgisi
Cinsiyet	Müşterinin cinsiyeti
Medeni Hal	Müşterinin medeni hali
Çalışma Durumu	Müşterinin çalışma durumu
Meslek	Müşterinin mesleği
Yaşadığı İl	Müşterinin yaşadığı il
Eğitim Seviyesi	Müşterinin eğitim seviyesi
Müşterinin Ödediği Toplam Prim	Müşterinin tüm poliçeleri için ödediği toplam prim
Plaka Sayısı	Müşterinin şimdiki zamana yaptığı araç/araçları plakalarının sayısı
Marka	Müşterinin güncel kasko poliçesi yaptırdığı aracın marka bilgisi
Marka Sayısı	Müşterinin şimdiki zamana kadar tüm poliçeleri dahilinde yaptığı araç/araçları markalarının sayısı
Model	Müşterinin güncel kasko poliçesi yaptırdığı aracın model bilgisi
Model Sayısı	Müşterinin şimdiki zamana kadar tüm poliçeleri dahilinde yaptığı araç/araçları modellerinin sayısı
Kullanım Tarzı	Aracın kullanım tarzı
Model Yılı	Müşterinin güncel kasko poliçesi yaptırdığı aracın model yılı bilgisi
Plaka İl Kodu	Müşterinin güncel kasko poliçesi yaptırdığı aracın plaka il kodu bilgisi
Ortalama Hasarsızlık Kademesi	Müşterinin şimdiki zamana kadar tüm poliçeleri dahilinde belirlenen hasarsızlık kademesi
Hasarsızlık Kademe	Müşterinin güncel kasko poliçesi için belirlenen hasarsızlık kademesi
Araç Yaşı	Müşterinin güncel kasko poliçesi yaptırdığı aracın yaş bilgisi
Satış Kanalı	Müşterinin güncel kasko poliçesini nereden yaptırdığı bilgisi
Unsur Tip	Müşterinin güncel kasko poliçesindeki aracın unsur tipi
Yenileme	Müşterinin poliçesinin yenilenip yenilenmediğinin bilgisi

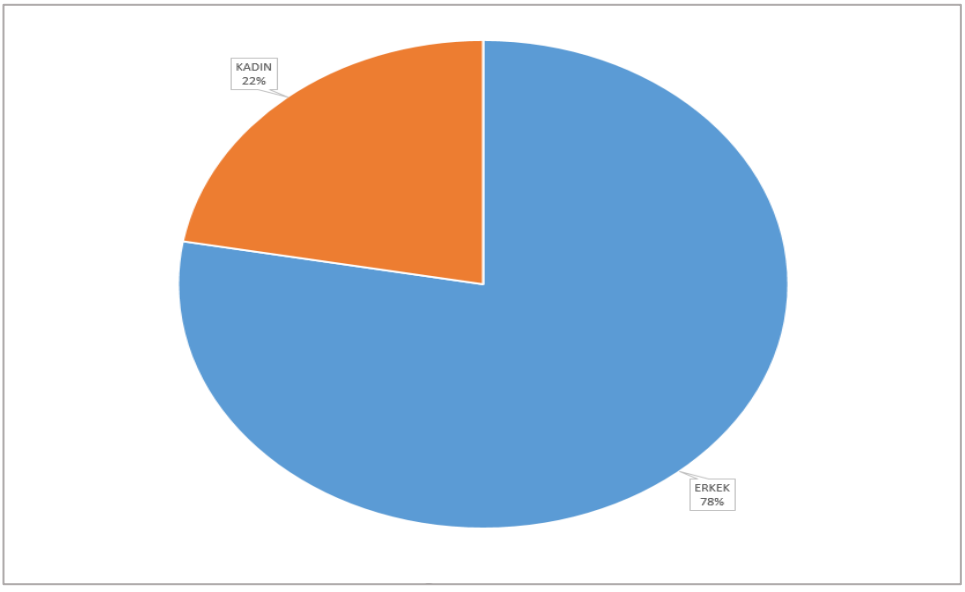
Veri setinde incelenmek üzere alınan 532723 müşterinin ve her bir müşteri için 22 öznitelik bulunmaktadır. Var olan müşteriler arasında 219437 müşteri poliçesini yenilememiş, 313286 müşterinin ise poliçesini yenilemiş olduğu gözlemlenmektedir. Şekil 4.1.'de çalışmada kullanılacak veri setine ait sınıfların müşteri kayıp adetleri

gösterilmiştir. Sınıflar arası farka bakıldığında veri setinin dengeli olduğu gözlemlenmektedir.



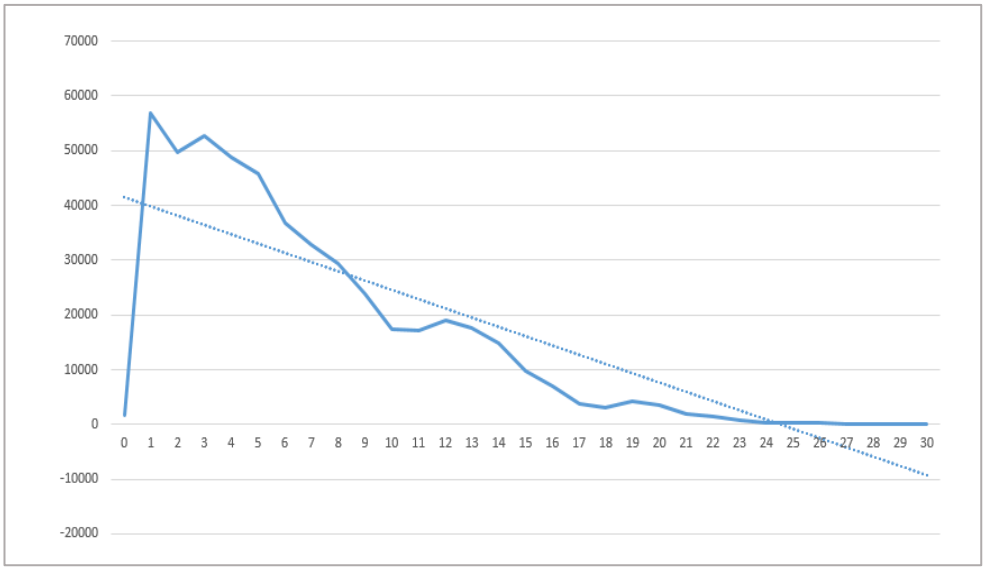
Şekil 4.1. Veri setindeki müşteri kayıp dağılımı

Şekil 4.2.'de veri setinde bulunan ve poliçesini yenileyen müşterilerin cinsiyet dağılımı verilmektedir. Grafiğe göre poliçesini yenileyen müşterilerin %78'inin erkek olduğu gözlemlenmektedir.



Şekil 4.2. Veri setindeki kayıp müşterilerin cinsiyet dağılımı

Şekil 4.3.'de veri setinde kasko poliçesi yaptırılan araçların trend dağılımı verilmektedir. Grafiğe göre aracın yaşı arttıkça kasko sigortası yaptırma oranının da doğrusal bir şekilde düştüğü gözlemlenmektedir.



Şekil 4.3. Veri setindeki poliçe yaptırılan araçların yaşlarının trendi

Veri seti istatiksel olarak incelenip kullanılacak veri iyice tanındıktan sonra birçok ön hazırlık aşamasından geçip model oluşuna uygun hale getirilmesi gerekmektedir. Eksik verilerin incelenip gerektiğinde uygun koşullarda doldurulması, veri seti içindeki verilerin uyumu için aykırı verilerin(outliers) tespit edilip düzenlenmesi, veri setindeki alanların diğer alanlarla arasındaki pozitif veya negatif yönlü ilişkinin (korelasyon) tespiti gibi ön hazırlık aşamaları veri setini modele hazırlayıp modelden maksimum başarı alınmasını sağladığı için veri madenciliği ve makine öğrenmesi yöntemleri süreçlerinde oldukça büyük öneme sahip olmaktadır. Çalışmamızda yapılan veri ön hazırlık çalışmaları bir sonraki başlık olan yöntem altında detaylandırılıp, uygulama başlığı altında ise nasıl uygulandığı anlatılmıştır.

4.2. Veri Ön İşleme

Hazırlanan veri setine modelin uygulanabilmesi, uygulandığında yüksek başarı elde edebilmek için veri seti veri ön işleme aşamalarından geçilmiştir.

4.2.1. Eksik verilerin temizlenmesi

Veri setinde eksik değerlerin olup olmadığı kontrol edilmiştir. Makine öğrenmesi algoritmalarında kullanılan veri setlerinde bulunmakta olan eksik (null) kayıtlar, modeli oluşturma, eğitme ve tahminleme çalışmalarında yer alabilecek tutarsız sonuçların meydana gelmesine sebebiyet vermektedir. Eksik veri olan veri setlerinde genel olarak aşağıdaki adımlar uygulanmaktadır.

- Veri setinde eksik verilerin bulunduğu sütunun veya satırın silinmesi. Eğer sütun yani özniteliğin veya satırın belirli bir yüzde oranı üzerinde eksik değerler (missing value) içeriyorsa veri setinden komple çıkarılması

- Veri setinde bulunan eksik verilerin sabit bir değer ile doldurulması

- Eksik verilerin bulunduğu sütunun ortalaması alınarak eksik verilerin bu ortalama ile doldurulması

- Süreklilik gerektiren sayısal değerler bulunan özntelikleri kategorikleştirilmesi. Yani çok geniş aralıklarda sayısal verilerin belli aralıklara göre kategorik hale getirilmesidir. “0’dan küçük”, “0-50”, “50-100” şeklinde kategorikleştirilebilinmektedir

- Veri setinde eksik olan veriyi yüksek ilişkisi (high correlation) olduğu başka bir özniteliğin değerine göre tahmin edilmesi

Bu çalışmada veri setindeki boş veri içeren sütunların satır bazında boş adetlerinin genel adede oranı alınmıştır. Bu oran boş verilerin analizini yapmak için kullanılmıştır. Sütun bazında %1 ‘in altında boş olan veriler silinmiş, %1’in üstünde boş olan sütunlarda ise string değerler var ise sütunun modu, numerik değerler var ise ortalama değerler ile doldurulmuştur.
Veri setinde bulunan değişkenler arasındaki ilişki, bu ilişkinin yönü ve şiddeti ile bilgileri sağlayan istatistiksel yöntem olan korelasyon analizi uygulanmaktadır. İki

ya da daha fazla değişkenler arasındaki ilişkinin matematiksel bağıntısı "Regresyon Analizi", ilişkinin yönü ve derecesi ise "Korelasyon Analizi" ile incelenmektedir.

4.2.2. Korelasyon analizi

Korelasyon katsayısı, bağımlı ve bağımsız değişkenler arasındaki ilişkinin gücünü göstermektedir. Örneğin; bir insanın kalp krizi geçirme riskiyle sigara kullanımı arasındaki ilişki veya insanlardaki eğitim seviyesi (X) ile içinde bulunduğu coğrafi konum (Y) arasındaki ilişki korelasyon katsayısı ile incelenmektedir. Korelasyon katsayısı değişkenler arasındaki ilişkinin nasıl olduğu hakkında bilgi vermektedir

Korelasyon katsayısı, açıklanan varyans ve açıklanmayan varyans oranı olarak tanımlanmaktadır.

Korelasyon katsayısı iki değişken arasındaki doğrulsal ilişkinin ölçüsü olup incelenen değişkenler birbirinden bağımsız ve $-1 \leq r \leq 1$ arasında bulunmaktadır. Korelasyon katsayısının 0 ile 1 arasında bir değer alması değişkenler arasındaki ilişkinin pozitif yönde bir ilişki olduğunu, -1 ile 0 arasında yer alması ise negatif yönde bir ilişki olduğunu göstermektedir. Tablo 4.2.'de korelasyon aralığı ve ilişki düzeyleri verilmiştir.

Tablo 4.2. Korelasyon Aralığı

Korelasyon analizi	İlişki Düzeyi
(-0,25) – 0 ve 0 – (0,25)	Çok Zayıf
(-0,49) – (-0,26) ve (0,26) – (0,49)	Zayıf
(-0,69) – (-0,50) ve (0,50) – (0,69)	Orta
(-0,89) – (-0,70) ve (0,70) – (0,89)	Yüksek
(-1) – (-0,90) ve (0,90) – 1	Çok Yüksek

4.2.3. Özellik ölçeklendirme

Bir veri kümesindeki özelliklerin değerlerini, mesafe hesaplamasına orantılı olarak katkıda bulunacak şekilde ölçeklendirme işlemidir. En yaygın olarak kullanılan özellik ölçekleme tekniği Standardizasyon (veya Z-Skoru Normalleştirme) ve Min-Max ölçeklendirmedir.

Standardizasyon: Z-Puanı Normalleştirme / Standardizasyon olarak da bilinmektedir. χ özelliklerinin yeniden ölçeklendirilmesi işlemidir, böylece $\mu = 0$ ve $\sigma = 1$ olmaktadır. Teknik olarak, standardizasyon ortalamayı çıkararak ve standart sapmaya bölerek verileri merkezler ve normalleşmektedir. Elde edilen değerlere standart puan (veya z-puanı) denir ve Denklem 4.1.'deki gibi hesaplanabilmektedir.

$$Z = \frac{X - \mu}{\sigma} \tag{4.1}$$

Formüldeki μ ortalama ve σ ise ortalamadan standart sapmadır. Özelliklerin standart sapması 1 ile 0 civarında ortalanacak şekilde standartlaştırılması, farklı birimleri olan ölçümleri karşılaştırdığımızda değil, aynı zamanda birçok makine öğrenme algoritması için genel bir gereklilik olarak da önemli olmaktadır.

4.2.4. Eğitim ve test kümesi oluşturma

Uygulamada veri setinin %33'lük dilimi 135112 olan satır test veri seti olarak ayrılmış, kalan 274316 satır veri ise eğitim veri seti olarak ayrılmıştır. Veri setimizde bulunan yenilenen müşteri ve kayıp müşteri dağılımı dengesiz olmamaktadır, dengesiz olma durumunda çözümlemek için az olan sınıfa ait veri örneği arttırılabilir, fazla olan sınıfa ait veri örneği azaltılabilir veya farklı performans metriklerine göre başarı ölçümlenebilmektedir.

BÖLÜM 5. BULGULAR VE DEĞERLEDİRME

Bu bölümde makine öğrenmesi yöntemlerinin değerlendirme ölçütlerinden, Sigortacılık Sektöründe Makine Öğrenmesi İle Müşteri Kaybı Analizi uygulamasından bahsedilmektedir.

5.1. Sonuçların Değerlendirilmesi

Makine öğrenmesinde kullanılan sınıflandırma modellerinin performansını değerlendirmek için hedef niteliğine ait tahminlerin ve bu tahminlerin gerçek değerleri analiz etmek için yaygın olarak kullanılan bir takım ölçütler yer almaktadır. Ölçütlerin hesaplamasında karışıklık matrisi (confusion matrix) kullanılmaktadır. Tablo 5.1.'de gösterilen, bu matriste satırlarda yer alan değerler veri setimizdeki gerçek değerleri, sütunlar ise modelimizin çalışması sonrası oluşan sınıflandırma / tahmin değerlerini içermektedir.

Tablo 5.1. Karışıklık Matrisi (Confusion Matrix)

		Tahmin Edilen Sınıf (Predicted Class)	
		Sınıf = 1	Sınıf = 2
Gerçek Sınıf Değeri (Actual Class)	Sınıf = 1	True Positive (TP)	False Negative (FN)
	Sınıf = 2	False Positive (FP)	True Negative (TN)

Doğruluk (Accuracy): Çalışma sonucunda doğru olarak tahmin edilen / sınıflandırılan verilerin, veri setindeki tüm örnek verilerin sayısına oranıdır. Formülü Denklem 5.1.'deki gibidir.

$$Doğruluk = \frac{TP+TN}{TP+TN+FP+FN} \quad (5.1)$$

Hata Oranı (Error Rate): Çalışma sonucunda yanlış olarak tahmin edilen / sınıflandırılan verilerin, veri setindeki tüm örnek verilerin sayısına oranıdır. Formülü Denklem 5.2'deki gibidir.

$$Hata\ Oranı = \frac{FP+FN}{TP+TN+FP+FN} \quad (5.2)$$

Duyarlılık (Sensitivity): Çalışma sonucunda doğru olarak tahmin edilen / sınıflandırılan pozitif örnek sayılarının, pozitif tüm örnek sayısına oranıdır. Formülü Denklem 5.3'deki gibidir.

$$Duyarlılık = \frac{TP}{TP+FN} \quad (5.3)$$

Kesinlik (Precision): Çalışma sonucunda doğru olarak tahmin edilen / sınıflandırılan pozitif örnek sayılarının, pozitif sınıflandırılmış tüm örnek sayısına oranıdır. Formülü Denklem 5.4'daki gibidir.

$$Kesinlik = \frac{TP}{TP+FP} \quad (5.4)$$

F-Ölçütü (F-Measure): Çalışma sonucunda değerlendirilmesi için duyarlılık ve kesinlik değerlerinin tek başına anlam ifade etmemesi sebebiyle bu iki değerin harmoni ortalamaları alınarak f-ölçütü bulunmaktadır. Formülü Denklem 5.5.'de verilmektedir.

$$F - \text{Ölçüt} = \frac{2\ x\ Kesinlik\ x\ Duyarlılık}{Kesinlik+Duyarlılık} \quad (5.5)$$

5.2. Problemin Çözümü İçin Uygulanan Model

Çalışmamızda problemin çözümü için nesne yönelimli, yorumlamalı, birimsel ve etkileşimi yüksek seviyeli bir programlama dili olan Pyhton ve veri analizi yaparkan temel olarak kullanılan başlıca kütüphaneler; pandas, numpy ve matplotlib kullanılmaktadır.

Pandas kütüphanesi, genelde veri işleme ve temizleme çalışmalarında oldukça efektif şekilde kullanılan bunun yanı sıra makine öğrenmesi gibi alanlarda da oldukça fazla kullanılan bir python kütüphanesidir. Numpy kütüphanesi, Python programlama dilinde, Bilimsel hesaplamalar yapılırken kullanılan temel bir kütüphanedir. Çok boyutlu diziler (array), maskelenmiş diziler ve matrisler, dizilerdeki hızlı işlemler için matematiksel, mantıksal, şekil işleme, sıralama, seçme gibi işlemleri içeren bir kütüphanedir. Çalışmada kullanılan diğer bir kütüphane ise Matplotlib kütüphanesi, verileri görselleştirebilen bir python kütüphanesidir, neredeyse bütün görselleştirme metodlarına entegre olarak çalışmaktadır. Aynı zamanda Matplotlib ile verileri etkileşimli olarak görselleştirebilir,herhangi bir yerde paylaşmaya uygun yüksek kalitede çıktılar hazırlanabilmektedir. Aynı zamanda hem iki boyutlu hemde üç boyutlu çıktılar hazırlanabilmektedir.

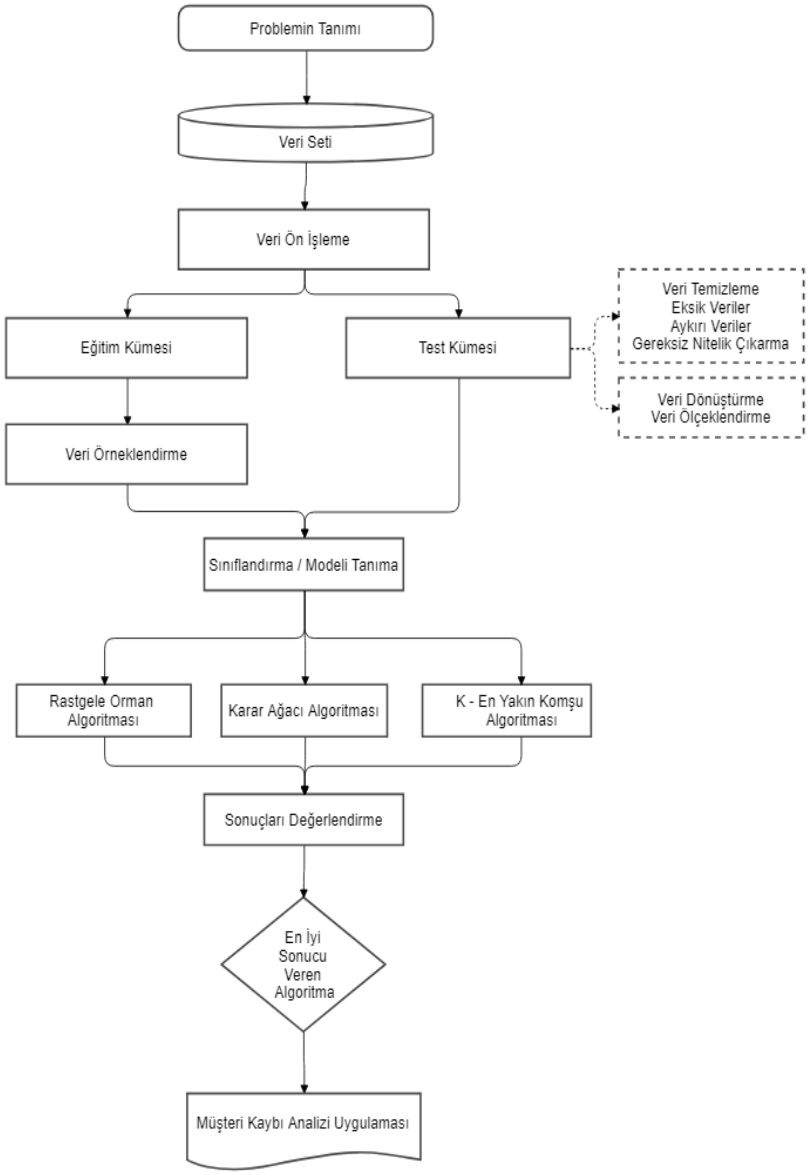
Bir makine öğrenmesi modelinde öncelikle eldeki veri ile neler yapılabileceğine karar verip, nasıl temizleneceğinin planları yapılmaktadır. Ardından eldeki verileri makine öğrenmesine hazır hale getirmek için belli ön işleme metodları uygulanmaktadır. Boş veriler doldurulup, aykırı veya yanlış değerler düzeltilip veri seti modele hazır hale gelince veri seti test ve eğitim olarak ikiye ayrılmaktadır. İki ayrılan veri setinden eğitim veri setine model öğretilip ardından test için ayrılan veri ile bu modelin doğruluk oranı ölçülmektedir.

5.3. Uygulama

Müşteri kaybı; müşterilerin artan rekabet ortamında bulundukları firmayı tercih etmekten vazgeçmeleri anlamına gelmektedir. Müşteri İlişkileri Yönetimi (MİY) ölçeklerinden olan müşteri kaybı; müşterinin son alışverişini yaptıktan sonra odağını diğer firmalara kaydırmasından bahsetmektedir. Nitekim firmalar artan rekabet ortamında yerini ve müşterilerini korumak amacıyla çeşitli stratejik kararlar alırken öncelikli olarak sadık olarak adlandırılan müşterilerinden gelecek istikrarlı gelire odaklanmaları sebebiyle var olan müşterilerin elde tutulmasına daha çok önem vermişlerdir. Eldeki müşterileri firmada tutmak, yeni bir müşteri kazanmak ve onu sadık müşteri yapmaktan daha maliyetli olmaktadır. Nitekim müşteri kaybı yönetimi

adı verilen süreçte istenen, firmayı tercih etmeyi bırakma ihtimali olan müşterileri tespit etmektir. Bu sebeple müşteri kaybının doğru yönetilmesi, ancak firmadan vazgeçme ihtimali bulunan müşterinin doğru şekilde belirlenmesi ile mümkün olmaktadır. Bu noktada temel amaç, müşterileri kayıp olacak ve kayıp olmayacak müşteri şeklinde sınıflandırmaktır.

Uygulamaya ait akış şeması Şekil 5.1. ‘de gösterilmektedir.



Şekil 5.1. Çalışmanın akış diyagramı

5.3.1. Veri seti hazırlaması

Artan rekabet dünyasında mevcut müşteri portföyünü korumak, rakip firmaların stratejisine yenik düşmeden müşterilerin gitme ihtimalini önceden tespit edip, bu durumun önüne geçmek için gitme ihtimali olan müşterilerin gitme sebeplerini çözümleyip, müşterileri memnun etmek gerekmektedir. Bu çalışmada belirlenen bu problem üzerine yapılmıştır.

Müşteri kaybı analizi yapılan bu çalışmada özel bir sigortacılık şirketinin verileri kullanılarak 2017 ve 2020 yılları arasında yer almış müşterilere ait içerisinde sosyodemografik bilgilerin yanı sıra kullanılan araç marka, model bilgilerinin de içinde yer aldığı 23 öznitelik belirlenmiştir.

Veri seti Oracle veri tabanında bulunan tablolardan PLSQL veri tabanı sorgusu kullanılarak hazırlanıp, csv formatında export alınmıştır. Pyhton programlama dilinde bulunan veri okuma, veri ön işleme ve veri temizleme aşamalarının yapılmasına olanak sağlayan pandas kütüphanesi ile hazırlanan veri seti proje içerisine alınmıştır.

5.3.2. Veri ön işleme

Öznitelikler belirlenirken Plaka, TCKN , VKN gibi müşteriye özel olan bilgilerin eğitim modeline faydası olmayacağı gerekçesiyle çalışmaya dahil edilmemiştir. Bunun yanı sıra müşteri özelinde poliçesinin iptal olup olmadığı bilgisi ve yenileme sayısı gibi öznitelikler de modelde ezbere sebebiyet verdiği için yer almamıştır.

Öznitelikler belirlendikten sonra kullanılan arayüze'e (Spyder) dahil edilen verisetinde veri ön işleme aşamalarına başlanmıştır. İlk olarak veri setindeki sütunların satır bazında boş adetlerinin genel adede oranı alınmıştır. Bu oran boş verilerin analizini yapmak için kullanılmıştır. Sütun bazında %1 'in altında boş olan veriler silinmiş, %1'in üstünde boş olan sütunlarda ise kategorik değerler var ise sütunun modu, numerik veriler var ise ortalama değerler ile doldurulmuştur.

Veri seti üzerinde ki boş olan verileri doldurma yöntemleri ile boş veriler temizlendikten sonra modelin başarı oranını arttırmak için, kategorik değişkenleri işlemek için kullanılan kodlama yöntemi olan Label Endocing ile string türündeki sütunlardaki her bir veriye alfabetik sıralamaya göre benzersiz bir tam sayı atanmıştır.

Tüm sütunlardaki verilerin sayısal olduğuna emin olduktan sonra veri seti üzerinde aykırı değerlerin tespiti için z score puanları hesaplanmıştır. Z score'ün -3 ile 3 arasında olması gerekmektedir. –3 ün altındaki ve 3'ün üstündeki değerler aykırı değerler olacağı için veri setinden çıkarılmıştır.

Veri işleme adımlarına devam ederken veri setinde bulunan sütunların birbirleri ile aralarındaki ilişkilerin yönü ve derecesini incelemek için Korelasyon Analizi uygulanmıştır. Bu analiz sonucu belirlenen korelasyon katsayısı sütunlar arasındaki ilişkinin bağımlı değişken mi yoksa bağımsız değişken mi olduğunu göstermektedir. Korelasyon katsayısı -1 ile 0 arasında çıkan sütunlar arasında bağımsız değişken, 0 ile 1 arasında çıkan sütunlar arasında ise bağımlı değişken ilişkisinin olduğu bilgisine erişilmiştir. Bu analiz sonucunda korelasyon katsayısı 0 olan verilerin ilişkileri kurulamadığı, modelin öğrenimine bir fayda sağlamayacağı gerekçesiyle temizlenmelidir.

Tüm veri setinin kesirli sayılara (float) çevrildiğinden emin olduktan sonra verilerin dağılımı ve özellikler arasındaki ölçek farklılıkları sebebiyle veri setine standart scaler uygulanıp veri seti veri temizleme aşamasını tamamlamıştır.

5.3.3. Veri setini eğitim ve test olarak ayırma

Veri temizleme aşaması tamamlanan veri setinden rastgele seçilen %33'lük dilimi 135112 olan satır test veri seti olarak ayrılmış, kalan 274316 satır veri ise eğitim veri seti olarak ayrılmıştır. Modellerin karşılaştırılması sırasında veri setinden doğacak üstünlüklerin engellenmesi sebebiyle test ve eğitim veri setlerinin bölme işlemi bir kez yapılmış ve her model için aynı eğitim ve test verileri kullanılmıştır.

5.3.4. Veri setine algoritmaların uygulanması

Veri setinden rastgele seçilen %67'lik kısım, 274316 satır eğitim verisi, çalışmalar dahilinde belirlenen algoritmalar ile modele dahil edilip, sonuç test için ayrılan veri seti ile karşılaştırılıp, değerlendirilmiştir.

5.3.4.1. Karar Ağacı Algoritması'nın uygulanması

Eğitim olarak ayrılan 274316 satır veriye Karar Ağacı Algoritması uygulanıp müşterilerin kayıp olup olmayacağı tahmin edilmiştir, test veri kümesiyle ise bu algoritmanın doğruluğu ölçülmüştür.

Tablo 5.2. Karar Ağacı Algoritması ile oluşturulan modelin karmaşıklık matrisi

		Öngörülen	
		S=0	S=1
Gerçek Sınıf	S=0	43629	11048
	S=1	11025	69410

Tablo 5.2.'de gösterilen, model sonucu oluştrulan karmaşıklık matrisi incelendiğinde tabloda DN=43629 DP=69410, YN= 11048 YP=11025 olarak görülmektedir. Doğru olarak sınıflanan ayrılmayan müşteri sayısı 43629 ayrılan müşteri sayısı 69410'dur. Toplamda 113039 müşteri doğru sınıflandırılmıştır. 11048 müşterinin gerçekte ayrılmış olup sınıflandırma sonucunda ayrılmamış olarak ve 11025 müşterinin gerçekte ayrılmamış olup sınıflandırma sonucunda ayrılmış olarak etiketlendiğini görmekteyiz. Bu modelin genel başarı oranı %84'dür. Sınıflara ait performans metrikleri Tablo 5.3.'de gösterilmiştir. Modelin kod blokları ise aşağıda verilmektedir.

Tablo 5.3. Karar Ağacı Algoritması ile oluşturulan modelin sonuçları

	Kesinlik	Duyarlılık	F-Ölçütü
S=0	0.80	0.80	0.80
S=1	0.86	0.86	0.86

5.3.4.2. Rastgele Orman Algoritması'nın uygulanması

Eğitim olarak ayrılan 274316 satır veriye Rastgele Orman (Random Forest) algoritması uygulanıp müşterilerin kayıp olup olmayacağı tahmin edilmiştir, test veri kümesiyle ise bu algoritmanın doğruluğu ölçülmüştür.

Tablo 5.4. RO Algoritması ile oluşturulan modelin karmaşıklık matrisi

		Öngörülen	
		S=0	S=1
Gerçek Sınıf	S=0	45839	8838
	S=1	8665	71770

Tablo 5.4'de gösterilen, model sonucu oluştrulan karmaşıklık matrisi incelendiğinde tabloda DN=45839 DP=71770, YN= 8838 YP=8665 olarak görülmektedir. Doğru olarak sınıflanan ayrılmayan müşteri sayısı 45839 ayrılan müşteri sayısı 71770'dur. Toplamda 117609 müşteri doğru sınıflandırılmıştır. 8838 müşterinin gerçekte ayrılmış olup sınıflandırma sonucunda ayrılmamış olarak ve 8665 müşterinin gerçekte ayrılmamış olup sınıflandırma sonucunda ayrılmış olarak etiketlendiğini görmekteyiz. Bu modelin genel başarı oranı %87'dir. Sınıflara ait performans metrikleri Tablo 5.5.'de gösterilmiştir. Modelin kod blokları ise aşağıda verilmektedir.

Tablo 5.5. RO Algoritması ile oluşturulan modelin sonuçları

	Kesinlik	Duyarlılık	F-Ölçütü
S=0	0.84	0.84	0.84
S=1	0.89	0.89	0.89

5.3.4.3. K-En Yakın Komşu Algoritması'nın uygulanması

Eğitim olarak ayrılan 274316 satır veriye K- En Yakın Komşu Algoritması uygulanıp müşterilerin kayıp olup olmayacağı tahmin edilmiştir, test veri kümesiyle ise bu algoritmanın doğruluğu ölçülmüştür.

Tablo 5.6. KNN Algoritması ile oluşturulan modelin karmaşıklık matrisi

		Öngörülen	
		S=0	S=1
Gerçek Sınıf	S=0	38325	16352
	S=1	16959	63476

Tablo 5.6’da gösterilen, model sonucu oluştrulan karmaşıklık matrisi incelendiğinde tabloda DN=38325 DP=63476, YN= 16352 YP=16959 olarak görülmektedir. Doğru olarak sınıflanan ayrılmayan müşteri sayısı 38325 ayrılan müşteri sayısı 63476’dır. Toplamda 101801 müşteri doğru sınıflandırılmıştır. 16352 müşterinin gerçekte ayrılmış olup sınıflandırma sonucunda ayrılmamış olarak ve 16959 müşterinin gerçekte ayrılmamış olup sınıflandırma sonucunda ayrılmış olarak etiketlendiğini görmekteyiz. Bu modelin genel başarı oranı %75’dir. Sınıflara ait performans metrikleri Tablo 5.7.’de gösterilmiştir. Modelin kod blokları ise aşağıda verilmektedir.

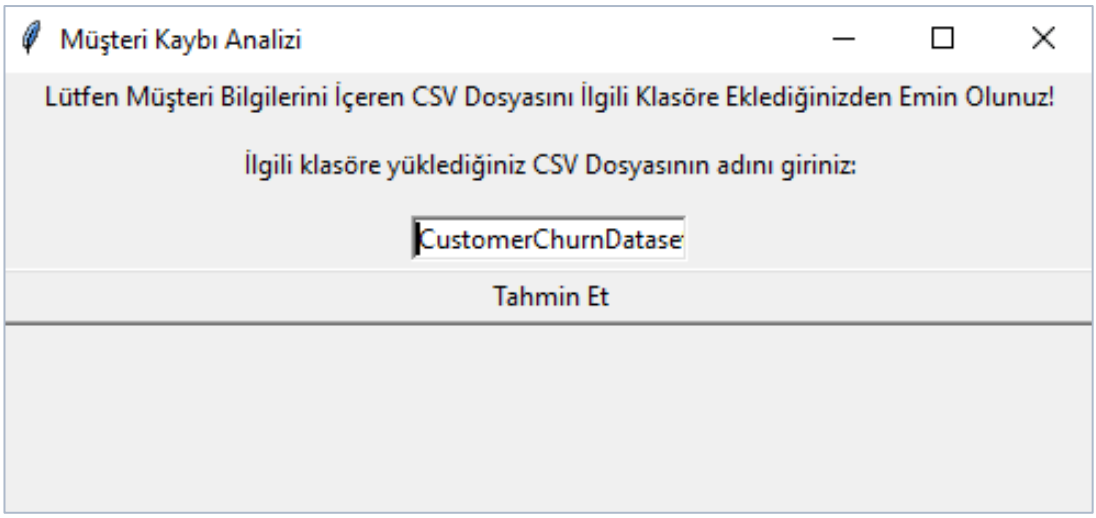
Tablo 5.7. KNN Algoritması ile oluşturulan modelin sonuçları

	Kesinlik	Duyarlılık	F-Ölçütü
S=0	0.69	0.70	0.70
S=1	0.80	0.79	0.79

5.3.5. Müşteri kayıp analizi uygulaması

Müşteri bilgilerini ve davranışlarını inceleyerek, bu müşteriler arasından ayrılma ihtimali olan müşterilere ait özellikleri çıkarıp, terk etme olasılığı yüksek olan müşterileri önceden tahmin etme amacını devam ettirmek için tüm bu adımlar bir sınıfa dahil edilmiştir.

Müşteri Kaybı Analizi Uygulaması önyüzündeki hesapla butonuna basıldığında bu sınıf çağrılarak içerisindeki en yüksek başarı oranını veren karar ağacı algoritması ile müşterilerin terk edip, etmeyeceği analiz edip arayüze yansıtılmaktadır. Şekil 5.2.’de Müşteri Kaybı Analizi arayüzü verilmiştir.

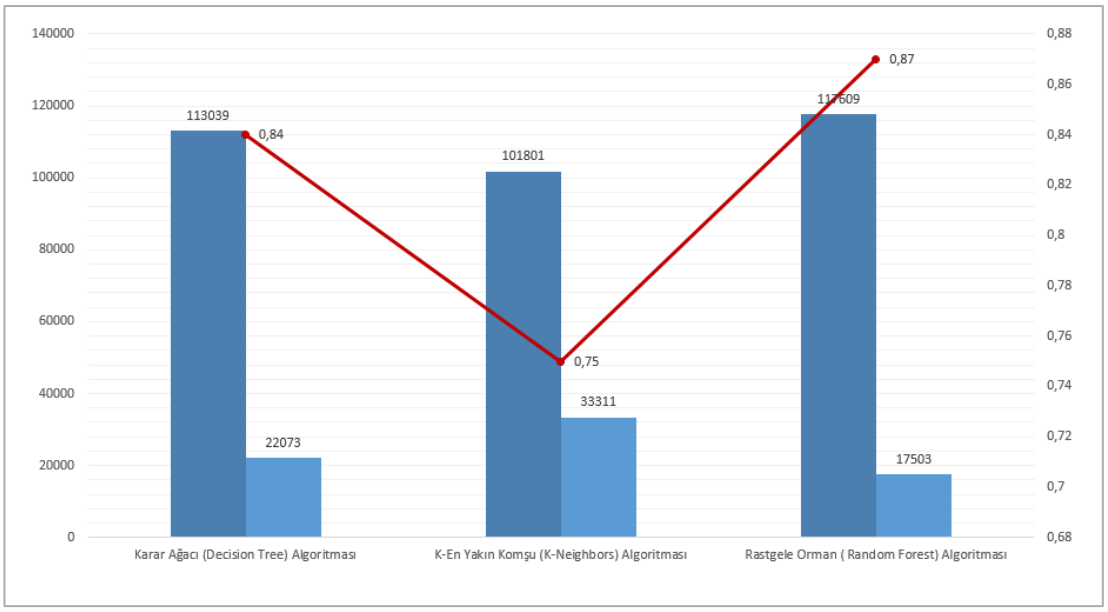


Şekil 5.2. Müşteri kaybı analizi arayüzü

BÖLÜM 6. SONUÇLAR VE ÖNERİLER

Günümüz şartlarında şirketler arası rekabet ortamının artması, pazarlama stratejilerinin gelişmesi ve müşterilerin daha bilinçli hale gelmesi ile müşteri sadakati önem kazanmıştır. Bir şirketin varlığını devam ettirebilmesi için mevcut müşterilerini elde tutmaları ve yeni müşteri edinmeleri oldukça önemlidir. Sigortacılık sektöründe yeni bir müşteri edinmek, mevcutta ki müşterinin ayrılmasını önlemekten çok fazla maliyetlidir. Mevcut müşterilerin profilleri ve davranışları incelenerek şirketi bırakma ihtimali olan riskli müşterileri bulma işlemine müşteri kaybı analizi denir. Bu sayede şirketler rekabet açısından üstünlük kazanır ve çeşitli stratejiler geliştirerek bu kayıpların önüne geçmeye çalışır.

Bu çalışmada sigortacılık sektöründe, aldığı sigorta poliçesini iptal etme olasılığı olan müşteleri, makine öğrenmesi yöntemlerinden olan 3 farklı sınıflandırma yöntemi ile tespit edilmeye çalışılmıştır. Bu 3 farklı sınıflandırma yöntemlerinden 117609 müşterinin poliçesini yenileyip yenilemeyeceğini doğru tahmin edip, en yüksek başarı oranını %87 oranla veren Rastgele Orman (Random Forest) Algoritması, peşinde 113039 adet müşteriyi doğru tahmin edip, %84 başarı oranı veren Karar Ağacı (Decision Tree) Algoritması ve 101801 doğru müşteri tahmini ile %80 başarı oranı veren K-En Yakın Komşu (K Neighbors) Algoritması olduğu gözlenlenmiştir. Şekil 6.1.'de bu üç algoritmanın doğru tahmin ettiği müşteri sayıları ve bu tahminlerin başarı oranları verilmektedir.



Şekil 6.1. Uygulanan algoritmaların karşılaştırması

%87 doğruluk oranını veren Rastgele Orman (Random Forest) Algoritması ile aynı makine öğrenmesi modeli bir sınıfa dahil edilmiştir. Pyhton programlama dili kullanılarak yapılan önyüze tahmin edilecek müşterilerin tutulduğu dosya dizini ismi girilerek tahmin et butonuna basılmasının ardından yapılan makine öğrenmesi sınıfı tetiklenerek, excelde bulunan müşteriler sınıfta öğretilen model yardımıyla tek tek tahmin edilip ekrana yazdırılmaktadır.

Yapılan çalışmamızın en kritik ve önemli süreci modellerde kullanılacak verinin kaynak sistemlerden üretilip bu üretilen verinin veri işleme yöntemleri ile temizlenerek modele uygun hale getirilmesidir. Çalışmadaki tüm süreçler tamamen ücretsiz Spyder programı üzerinde Pyhton programlama dili ile yapılmıştır. Makine öğrenmesi sürecinin ardından oluştulan arayüz de yine pyhton programlama dili ile oluşturulmuştur. Pyhton programlama dili ve kütüphanelerinin makine öğrenmesi uygulamalarında başarılı sonuç verdiği gözlemlenmiştir.

Bu çalışmada ortaya çıkan model ve arayüz, şirketin aktif müşterileri üzerinden çalıştırılarak , poliçesini iptal etme olasılığı olan müşteriler tespit edilebilmektedir. Oluşturulan arayüzün testleri çalışmamız kapsamında yapılmış olup, kullanımı ve

çıktıları KVKK kapsamında şirket içinde yapılmakta olması sebebiyle çalışmamızda yer almamıştır.

Yapılan çalışmada farklı modellere ait sonuçların performans metriklerini sonuçları karşılaştırılmış ve başarılı olan yöntemle kodlama veya veri analizi bilmeyen kullanıcıların da bu analizden fayda sağlamalarına olanak tanınmıştır. Başarılı model seçiminde doğruluk oranının yanında başka hangi kriterlerin göz önünde bulundurulması konusunda öneriler sunulmuştur. Veri setinde sınıf dağılımı eşit olduğu için bu konuda ekstra bir çalışma yapılmamıştır. Analiz sonucuna göre Rastgele Orman Algoritması'nın duyarlılık, kesinlik ve f-ölçütü ile değerlendirilmeye alınıp en yüksek sonuç verdiği söylenebilmektedir. Çalışmanın devamına rastgele orman algoritması ile devam edilip veri işleme, model oluşturulma aşamaları bir sınıfa dahil edilip, oluşturulan ön yüz sayesinde bu sınıfın tekrar tekrar çağrılıp, girilen her veri seti için ayrıca çalışmasına olanak sağlanmıştır.

Çalışmada en başarılı olan algoritmanın hesaplanması gibi manuel kalmış tüm süreçler otomatikleştirilip, son kullanıcının tetiklediği sınıfa o başarı oranı ile devam edilmesi bir sonraki çalışmalara örnek olarak gösterilmektedir. Ek olarak uygulamanın arayüzünün tasarımı kullanıcıların veya şirketlerin tercihine göre geliştirilebilmektedir.

KAYNAKLAR

Aydoğan E., G. A. (2009). Veri Madenciliği Tekikleri İle Bir Kozmetik Markanın Ayrılan Müşteri Analizi Ve Müşteri Bölümlenmesi. Mühendislik ve Fen Bilimleri Dergisi.

Bagheri F, T. M. (2015). Customer behavior mining based on RFM model. Journal of Industrial Engineering and Management Studies.

Baykal, A., & Coşkun, C. (2011). Veri madenciliğinde sınıflandırma algoritmalarının bir örnek üzerinde karşılaştırılması. Akademik Bilişim Konferansı.

Burez, J. V. (2009). Handling class imbalance in customer churn. Elsevier, 4626-4636.

Buttle, F. M. (2015). Customer Relationship Management. Concepts and Technologies.

Huigevoort, C. (2015). Customer churn prediction for an insurance company. Eindhoven Teknoloji Üniversitesi.

Karaağaç, Ş. (2015). Churn Analysis And Churn Prediction In A Private Bank. Marmara Ünüversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi.

Kaynar, O., Tuna, M., Görmez, Y., & Deveci, M. (2017). Makine Öğrenmesi Yöntemleriyle Müşteri Kaybı Analizi. C. Ü. İktisadi ve İdari Bilimler Dergisi.

Kim, K., Jun, C., & Lee, J. (2014). Improved churn prediction in telecommunication industry by analyzing a large network. Elsevier.

Koçtürk, Y. (2010). Veri Madenciliğine Bağlılık. İstanbul Teknik Üniversitesi, Yüksek Lisans Tezi.

Kotler, P., & Keller, K. (2015). Marketing Management. 5. Baskı.

Kunt, M. (2019). Telekomünikasyon Sektöründe Müşteri Kaybı Analizi. Ankara Üniversitesi Fen Bilimleri Enstitüsü.

Moore, A. (2001). Decision Trees. Mellon University.

Nettleton, D. (2014). Commetcial Data Mining. Analysis and Modeling for Predictive Analtyics Projects.

Özmen, M. (2006). Churn Modelling In Telecommunications Sector. İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü.

Ravi, V., Raju, S., & Farquad, M. (2014). Churn prediction using Comprehensible support vector machine: An analytical CRM application. Applied Soft Compting, 31-40.

Scheel, I., Aldrin, M., Glad, I., Sorum, R., Lyng, H., & Frigessi, A. (2005). The influence of missing value imputation on detection of differentially expressed genes from microarray data. Bionformatics, 4272-4279.

Sharma, A., & Panigrahi, D. (2011). A Neural Network Based Approach for Predicting Customer Churn in Cellular Network Services. International Journal of Computer Applications.

Tosun, T. (2006). Veri Madenciliği Teknikleriyle Kredi Kartlarında Müşteri Kaybetme Analizi. İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi.

Tsiptsis, K., & Chorianopoulos, A. (2009). Data Mining Tecniques in CRM: Inside Customer Segmentation. John Wiley & Sons.

Vafeiadis, T., Diamantaras, K., Sarigiannidir, G., & Chatzisavvas, K. (2015). A comparison of machine learning techniques for customer churn prediction. Elsevier.

Yılmaz, H. (2014). Random Forests Yönteminde Kayıp Veri Probleminin İncelenmesi ve Sağlık Alanında Bir Uygulama.

ÖZGEÇMİŞ

Adı Soyadı : Hande Esin AKYİĞİT

ÖĞRENİM DURUMU

Derece	**Eğitim Birimi**	**Mezuniyet Yılı**
Yüksek Lisans	Sakarya Üniversitesi / Fen Bilimleri Enstitüsü / Bilişim Sistemleri Mühendisliği	Devam ediyor
Lisans	Sakarya Üniversitesi / Bilgisayar ve Bilişim Bilimleri Fakültesi / Bilişim Sistemleri Mühendisliği	2017
Lise	Gülizar Zeki Obdan Anadolu Lisesi	2012

İŞ DENEYİMİ

Yıl	**Yer**	**Görev**

2019-Halen	Allianz	Veri Mühendisi
2017-2019	Obase	İş Zekası Danışmanı

YABANCI DİL

İngilizce

ESERLER (makale, bildiri, proje vb.)

1. Yaygın Kurumsal İş Zekası Uygulamalarının Çok Yönlü Karşılaştırmalı Analizi, Uluslararası Marmara Fen ve Sosyal Bilimler Kongresinde sunulan bildiri, Erişim Adresi: http://imascon.com/dosyalar/imascon2018/imascon2018_tam_metin.pdf

HOBİLER

Müzik Dinlemek, Seyahat Etmek, Kitap Okuma