YBS Ansiklopedi

www.YBSAnsiklopedi.com

Cilt 9 Sayı 1, Haziran 2021

Müşteriye Özel Fiyat Tahmin Çalışması

Berkay AKAR[1]

1. Düzce Üniversitesi, Bilgisayar Mühendisliği Bölümü

Özet

Teknolojinin gelişmesi ile performanslarının ve kullanışlılığının artmasıyla hayatımızda vazgeçilmez bir yer almaya başlayan bilgisayar teknolojilerinin beraberinde getirmiş olduğu yapay zekâ teknolojilerinin günlük hayat üzerinden uygulamalarının, yoğunlukla çalışma alanlarının incelenmesi ve analizlerinin yapılarak sektörel bazlı işleyişinin incelenmesi amaçlanmıştır. Beraberinde yapılan bu çalışmanın amacı yapay zekâ teknolojilerinin bir pazar sürecinde müşterilere özgü indirim tahmininde bulunulması amaçlanmış, bu amaçlar doğrultusunda, farklı platformlarda geliştirilen bu çalışmadan çıkan makine öğrenmesi modelleri ve değerleri yorumlanmıştır. Çıkan sonuçlar tartışılarak uçtan uca bir veri bilimi süreci içerisinde müşteri analizleri gerçekleştirilmiş ve müşteriye özgü bir fiyatlandırmanın yapay zekâ kullanılarak nasıl gerçekleştirilebileceği açıklanmış olup, yapılan çalışmanın tutarlılığı için sektörel açıdan var olan bir AutoML teknolojisi ile karşılaştırılarak geçerliliği sorgulanmıştır.

Anahtar Kavramlar: *Otomatik Makine Öğrenmesi (AutoML), Knime, Sklearn, Müşteri Davranışları , İndirim Tahminleme*

Abstract

It is the objective of this research to examine the daily applications, most commonly used areas, and sector-based functioning of artificial intelligence technologies through analysis, which were brought into our lives with computer technologies and have begun to take an imperative place in our lives as their performance and usefulness have increased. Furthermore, this research aims to show how artificial intelligence technologies can be used to estimate customer-specific discounts in a market process, and hence, machine learning models and values developed on various platforms and deduced from this research have been interpreted for these purposes. Customer analyses have been conducted in an end-to-end data science process by presenting the findings, and it was demonstrated how artificial intelligence could be used to achieve customer-specific pricing. The study's validity was checked by comparison via AutoML technology available in the sector for consistency.

Keywords: *Automatic Machine Learning (AutoML), Knime Sklearn Costumer Behaivoral, Discount Prediction*

1.Giriş

Teknolojinin gün geçtikçe gelişmesi, mikroişlemci ve mikro denetleyici modüllerin üretim maliyetlerinin azalması, ham maddelerin ucuzlaması ve bu teknolojilerin kullanışlılığının artması, insanların yapabilecekleri iş hızını aşmaları sebebiyle bilgisayar ve mikro denetleyici aygıtlar hayatımızın vazgeçilmez bir bütününü oluşturmaktadır. Bu

durum bilgisayar bilimlerinin gelişmesine, beraberinde başka bilimler ile birleştirilerek insanların hayattaki problemlerine çözüm odağı olması için hızlı bir gelişme ve çözüm süreci getirmiş bulunmaktadır. Bu gelişmeler hayatımıza farklı sorunlara farklı çözümler sunulması ile çok büyük bir talep ve pazar payı oluşturmaktadır.

İşleyiş açısından hayatımızı her ne kadar kolaylaştırsa ve kullanışlılığı ile birçok sektörde kullanılsa bile sektörel ve kişisel kullanımları düşünüldüğünde artan veri büyüklüğü göz önünde bulundurulduğunda her geçen zaman bir önceki zaman diliminden daha fazla bilgi üretilmekte ve tüketilmektedir. International Data Corporation (IDC) bir rapora göre 2011'de genel olarak oluşturulmuş ve dünyadaki kopyalanan veri hacmi yaklaşık olarak 1.8 Zettabyte olarak hesaplanmıştır [3]. IDC'nin bir açıklamasına göre tahmini olarak 2025'te bu sayı 175 zettabytes'a ulaşabileceği düşünülmektedir[4]. IDC'nin düzenli olarak paylaştığı senelik üretilen veri grafiklerine bakılacak olursa bu açıklanan değerler baz alınarak her geçen sene bir önceki seneye göre daha fazla bir artış gerçekleşmektedir.

Bu üretim ve tüketimin en başlıca yaygınlaşma sebepleri olarak teknolojilerin yaygınlaşması, gerek ticari gerekse kişisel istek veya amaç doğrultusunda yeni isterlerin oluşması, bireysel kullanıcıların sosyal medya işleyişleri ve yoğunlaşmaların artması, kişisel tüketim elektroniği kullanımlarındaki artış, ve eski zamanlarda manuel çalışan cihazların yönetilebilirliğini kolaylaştırabilmek adına akılcılaştırılması ve bu süreç içerisinde gelişen cihazların işlemlerini yerine getirebilmeleri için sensörler vasıtasıyla düzenli olarak veri üretebilmeleri örnek verilebilmektedir[1]. Günümüzde üretilen verilerden uygulamalar vasıtası ile tahminlemelerde ve tavsiyelerde bulunma, yönetilebilirlik, öngörü istekleri günümüzde en sık karşılaşılan veri kullanım örneklerinin başlıcaları olmak üzere sürekli olarak her yerde bir veri üretimi ve tüketimin mevcut olduğunun göstergesidir.

Bu gelişmeler ve istemler neticesinde düzenli olarak yükseliş gösteren ve geleceğin teknolojisi olan veri bilimi sektörü birçok alanda çalışmalar sürdürülebilmeleri ve aynı zamanda verileri gerekli şekilde işleyerek birçok alanda insan işleyişinin kolaylaştırılması sebebi ile iş süreçlerinde daha yüksek başarılar elde edilebilmesinde öncü olmuştur. Sık kullanılan iş süreçlerine örnekler vermek gerekirse, sektörde müşteri kayıp analizi [5] olarak bilinen ve müşterilerin davranışları üzerinden ayrılma ihtimali olan abonelerin ayrılma öncesinde tespit edilmesi ve kaybın engellenmesi için gerekliliklerin yerine getirilmesi, veya yapay zeka endüstrisinin bir alt kolu olan görüntü işleme kullanılarak bir insanın vücudundaki anormallik tespiti yapılarak potansiyel olarak olabilecek hastalıkların önceden tespitinde ve tedavisinin erkan başlamasında ve bu nedenle tedavide başarının sağlanması [6], veya yazı işleme kullanılarak yazılmış bir içerik içerisinden yazan kişinin duygu analizleri [7], eldeki verileri kullanarak borsanı endeksi tahmini [8] gibi çalışma alanları örnek olara verilebilir.

Bu makalede amacımız ve çalışmamız veri bilimi ve istatistiksel modellerin kullanılmasının değerlendirilmesinin satış, pazar tahminlemesi, müşteriye özel fiyatlandırma olacağı için ağırlıklı olarak bu konuların üzerinde durulacaktır.

1.2. Müşteri Davranışlarının Analizi ve Satış İlişkisi:

Bu makalenin de ana konusu olan ve üzerinde incelemelerin yapılacağı sorun üretilen bu verilerden gerekli istatiksel modeller uygulanarak satış işlemlerinin müşteri ve ürün arasındaki ilişkisinin incelenmesi ve bu ilişki sonucu müşterilerimize ürünleri ne kadar indirim ile satılabileceğinin tahminidir. Literatürde, "müşteri davranışları" (Customer Behaviour) incelemesi olarak adlandırılan müşterilerin hareketlerinden, ürünlere verilen tepkilerden, harcamalarından analiz yaparak daha iyi satış yapılabilir mi sorusunun cevaplanması konusu ile ilişkilidir. Bu kavram için gerekli analizlerin nasıl toplanıldığından bahsetmek gerekirse, müşterilerin arama geçmişlerinin analizleri, müşterilerin markalar üzerindeki davranışları, müşterilerin ürünlere ulaşmak için kullandıkları reklamlar ve metotlar, reklam ve metotların kullanıcılara sunuluş biçimleri (bir e ticaret sitesi ise uygulama ve sitenin kullanıcı dostluğu, reklamların kullanıcılara nasıl ulaştırıldığı, ürün dizilimleri ve kategorilendirmeleri, sunuş biçimindeki görsellik kalitesi), müşterilerin ürünler üzerindeki geçmiş akışı, platform kalitesi ve bilinirliği, sistemde kalma süreleri,

geçmişteki yapılan alışverişler ve alışveriş kategorileri, sistemi kullandıkları zaman aralıkları (günün hangi saatlerinde, hangi günlerde ve bu günün özel bir gün olup olmadığının kontrolü), müşterilerin demografik bilgilerinin analiz edilmesi bu analizlerden tahminlenmesinde kullanılan değişkenlerdir[9]. Bu değişkenlerin ve davranışların doğru analiz edilmesi sonucu gerekli yorumların yapılmasıyla birlikte potansiyel olarak alma ihtimali yüksek kullanıcılara daha verimli ve etkili bir şekilde satışlarımızı gerçekleştirebilir. Müşterileri ve ürünleri doğru kategorize edebilmek satışımızda ve müşterinin ihtiyacı olan ürüne erişiminde büyük bir etken olmaktadır. Uygun bir gelir dağılımına ya da üründen farklı tipteki bir ürünü o müşteriye reklam bildirim veya herhangi bir yola başvurarak müşteriye ilerletmemiz ilgisini çekmeyecektir. Bu yüzden müşteriler ve ürünler iyi analiz edip doğru bir kategorilendirme vasıtası ile ürünleri müşterilere daha kolay ve daha etkili sunulabilmesi gerekmektedir.

Yapılan bir literatür taramasında, Paulo Rita, ve arkadaşlarının yaptığı bir çalışmaya [10] göre bilgi teknolojilerinin hızlı gelişmesi ve sağlanan kolaylıklar sebebi ile müşteriler fiziksel mağazalar yerine elektronik ortamdaki mağazaları tercih etmeye başlamış ve bu durum, fiziksel mağazaların kapanmaya başlamasına ve e ticaretin yükselerek Endonezya için çok büyük bir pazar payı oluşturmasına, beraberinde müşterilerin alışveriş kültürlerinin değişmesine sebebiyet vermiştir. Değişen kültürle müşterilerin satın alma alışkanlıklarında hizmet kalitesinin artışının aynı müşterinin tekrar aynı platformu tercih ederek alışveriş yapma, müşteri aktifliği ve müşteri sadakati değişkenlerindeki artış ile doğru orantılı olduğu belirtilmiştir. Hizmet kalitesi olarak adlandırılan parametrelere bakıldığında versiyonlara göre değişiklik göstermekle birlikte gizlilik, güvenilebilirlik, kullanım kolaylığı, platform hızı, anlaşılabilirlik, yenilikçilik ve görsel çekicilik ve bütünlüğün genel olarak bahsedilen değişkenler olup hizmet kalitesini belirleyen ölçütler olarak tanımlanmıştır.

Fiziksel satış ortamlarında yapılan bir çalışmada ise, Kaneko ve Yada[11] bir süper market içerisinde konumlandırılan ürünler ile müşterilerin alışveriş boyunca kullandıkları yolları baz alarak çeşitli takip cihazları ile müşterilerin davranışlarını incelemiştir. İncelemeler sonucunda ürünlerin market içerisindeki konumlandırmanın müşterilerin satın alma eylemi üzerindeki davranışlarına etkileri incelenmiştir. Elde edilen sonuçlar ile müşterilerin kullandıkları yolların karmaşıklığı ve satın alma davranışları üzerinden bir modelleme gerçekleştirilmiştir. Incelemeler sonucunda elde edilen fraktallara neticesinde düşük ve yüksek müşteri gruplarına göre ayrılmalar gerçekleşmiş ve bu ayrımlar ziyaret edilen bölge sayısı, sepet büyüklüğü, mağazada kalma süresi ve satış tutarının parametreleri üzerinde ağırlıklı olarak farklılıklar göstermiştir. Aynı zamanda müşteri hareketlerinin karmaşıklığı arttıkça ve alışveriş süresi uzadıkça alınan ürün sayısı ile doğrusal bir ilişkide yakalanmıştır.

Müşteri alışkanlıklarını etkileyen değişkenleri çalışan William Applebaum [12], incelemesinde ürün türlerine ve müşterilerin bu alımların adeti ve sürekliliğinde aynı özellikleri gösteren guruplar olabildiğini gözlemlenmiştir. Beraberinde aynı çalışmada bir müşterinin o an gerçekleşmese bile zaman içerisinde alışveriş karakteristiğinin değişebileceği veya bir ürün bir müşteriye uygun olmasa dahil çevresel yönlendirmeler ile ürünü alabileceği ve bu değişimden kaynaklı olarak farklı guruptaki ürünleride tercih edebileceğini belirtmektedir. Bu şartlar neticesinde alışveriş davranışlarını satın alma yeri, satın alınan ürün, satın alma zamanı ve sıklığı, satın alma yöntemi gibi unsurlar ağırlıklı olarak etkilemiş bulunmaktadır.

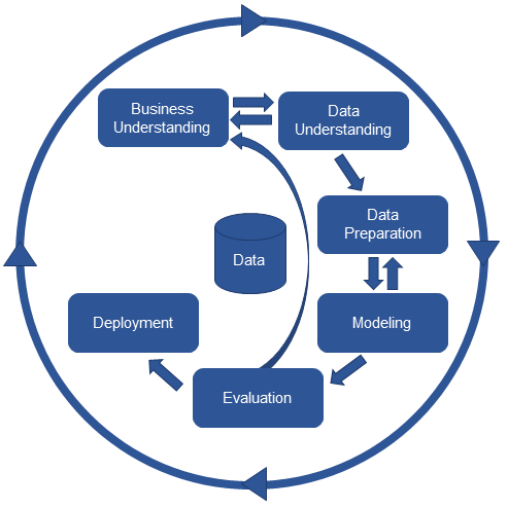
Müsteri hareketlerini hedef alarak bir pazar çalışması yapan Mu-Chen Chena, Ai-Lun Chiub , Hsu-Hwa Changc [13] müşteri davranışlarında müşterileri ürünlere göre eşleştirmek ve servis kalitesinin yanında müşteri davranışlarının zaman içerisinde değişebileceğini ve bu değişimlerin tespit edilerek pazarlama stratejilerinde kullanılabileceğini savunur. müşteri davranışlarını bu zamana kadar anket gibi metotlar ile veri toplayıp tahminlemelerde bulunmak yerine tavsiye algoritmalarında sıklıkla başvurulan birliktelik kural çıkarımı algoritmalarından birisi olan apriori algoritması kullanarak müşterilerin davranışlarının kalıplar halinde tanımlamış, farklı dönemlerdeki kalıplar ile karşılaştırılarak farklılıkları tespit etmiştir. Yapılan bu literatür taraması William Applebaum'ün [12] yaptığı çalışmada bahsettiği üzere müşterilerin zamana ve çevresel etkilere bağlı olarak davranışlarında değişiklikler gösterebileceğinin bir analizidir.

Müsteri hareketlerini metinler üzerinden algılayarak isterleri belirlemek ve pazarlama sürecinde kullanılabilmesini hedef alan bir çalışmaya göre Jian jin, Ying Liu, Ping Ji ve Honngguang Liu [14] müşterilerin isterlerini ve duygularını analiz ederek Pazar odaklı ürün tasarımı üzerine bir çalışma gerçekleştirmiştir. Pazarda başarı açısından diğer çalışmalardaki bahsedilenleri desteklemekle ürün üretiminde müşterilerin piyasada var olan ürünlere karşı davranışlarını analiz ederek daha başarılı üretim için bir çalışma gerçekleştirmiştir. Çalışmada sektörde bilindik olan e-ticaret platformları üzerindeki ürünleri bularak hem ürünlerin özelliklerini analiz etmiş hem de müşterilerin ürün üzerindeki geri bildirimlerini dikkate alarak bir model geliştirmiştir. Bu çalışma gerçekleştirilirken 661 adet ürün hakkında 113.467 adet inceleme toplanmıştır. Elde edilen müşteri değerlendirmeleri yorumlanarak müşterilerin bu davranışları üzerinden üretilecek olan yeni ürünlerde istekleri belirlenmiştir.

Yapılan literatür değerlendirmelerini özetlemek gerekirse gelişen zamanla pazarlama stratejilerinde eski yöntemler yerine zamanın getirmiş olduğu sosyal medyalar ve e-ticaret platformları ile müşterilerin alışveriş kültürlerinde bir değişim gerçekleşmiştir. Beraberinde müşterilerin sıklıkla kullandıkları bu ortamlarda her geçen gün veri sayısında bir atışla birlikte büyük veri kavramı doğmuştur. Pazarlamada müşterilere özel ürünlerin bu büyük verilerden belirli istatiksel modeller çevresinde uygulanması ile müşterilere karakteristik özelliklere göre daha uygun ürünler önerilebilmiş, fiyatlandırmalarda kolaylıklar sağlamış ve süreklilik sağlanmaya çalışılmış ve beraberinde büyük bir sektörü meydana getirmiştir. Müşteri davranışları ise ürünün sunuluş biçimine, ürünün özelliklerine, tüketim sıklığına, müşterilerin demografik özelliklerine, pazarlanma biçimine, zamana, platformun kullanılabilirliğine ve güvenilebilirliğine göre sınıflandırmalar yapılabileceğini göstermiştir.

1.3. Metodolojinin Belirlenmesi

Veri bilimi sürecinde, bir metodoloji belirlenerek ve bu metodoloji adımları izlenerek iş geliştirme sürecine başlanmıştır. Metodolojilerin ama amacı bir veri bilimi projesine nasıl başlanacağı hangi adımların sırası ile devam edileceği ve sonucunda neler yapmamız gerektiğini belirten ve proje işleyişini yönlendiren yöntemlerdir. SEMMA, KDD ve CRISP-DM [15] olmak üzere bilinen ve kullanılan üç farklı metodoloji bulunmaktadır. Güncel hayatta kullanışlılığı ve endüstriyel süreçlere daha rahat uyarlanabilmesi ve birçok endüstriyel süreç içerisinde uyumluluğu sebebiyle bu çalışmamız boyunca tercih ettiğimiz metodoloji CRSİP-DM (Cross-Industry Standard Process for Data Mining) metodolojisidir. Bu metodoloji detaylı bir şekilde açıklamak gerekirse 6 evreden oluşur ve bazı evreler arasında dönüşüm mevcuttur. Bu dönüşüm aşamaları ve aralarındaki ilişkiler, Şekil 1'de görselleştirilmiştir.



Şekil 1 CRISP-DM evreleri ve adım yönleri

1.3.1. İşin Anlaşılması (Business Understanding)

Bu aşamada iş probleminin ve istenilenlerin anlaşılması evresidir. Problemin ne olduğunu, neden çıktığını, problem çözümü için iş işleyişinin anlaşılması, iş sürecinde ilgili verilerin tespiti, şayet problem çözülmesi olasılığını alarak hedeflerimizin belirlenmesi ve çözümü sonrası için başarı kriterleri nelerdir bunların analizi yapılır. Metodolojinin 1. Evresi olan bu evre en önemli ve hata yapılması durumunda en baştan başlanılması gereken evredir.

1.3.2. Verinin Anlaşılması (Data Understanting)

Metodolojinin 2. evresi olan bu evrede iş tanımı ve hedeflerimiz doğrultusunda istenilen verilerin anlaşılmasına, gerekli görselleştirmelerin yapılarak verinin anlaşılmasını, eğer ki elimizde veri yoksa bu problem çözümü için uygun tipte veriler bularak bu verilerin anlaşılmasına nicelik istatistikleri çerçevesinde değerlendirilmesine bir sonraki aşamada işlenecek veri üzere gerekli tespitlerin yapılmasına dair bir işleyiştir.

1.3.3. Veri Hazırlanması (Data Preprosessing)

Verinin hazırlanması aşamasından gelen verilerimizi ve işin anlaşılması sürecinden gelen isterler doğrultusunda verimizi gerekli bir şekilde işlemeye başlanılan aşamadır. Veri biliminde her problem her algoritma ile çözülmemektedir. Aynı şekilde her algoritmada aynı tipten verilerle çalışmamakta ver kendilerine özgü özellikler göstermektedir. Aynı amaca hizmet edilen verilerde ve algoritmalarda dahi aynı problem çözümü için farklı ön işlemeler yapılabilmektedir. Algoritma isterine göre verilerimizi görselleştirme doğrultusunda gelen bilgiler ışığında işleyerek verilerimizi algoritmalarımıza veririz.

1.3.4. Modelleme:

Ön işlemeden gelen veriler için bu aşamada gerekli makine öğrenmesi algoritmalarına vererek modellerimizin çıkartılması işlemidir. Gerekli kütüphaneler veya ortam üzerinden veriler öncesinde bir train ve test olarak bölünme yaşar. Train verileri ile sistem eğitilerek test kısmında ise daha önce algoritmanın hiç görmediği verilerin testi yapılır.

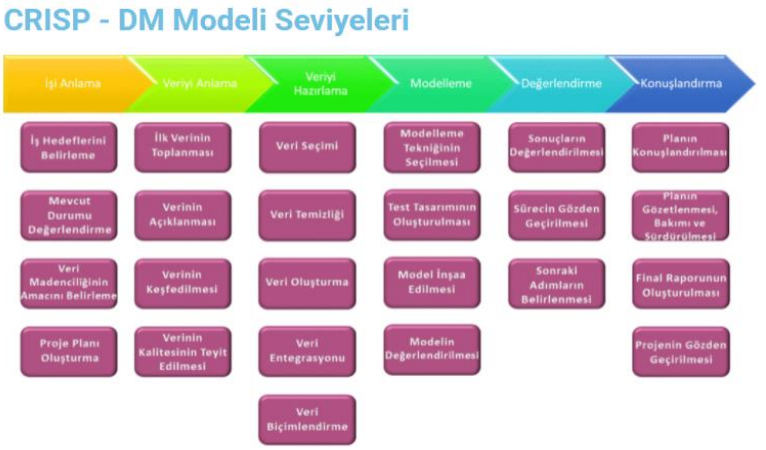
1.3.5. Değerlendirme (Evolution)

Modelleme aşamasında üretilen modelin yorumlanacağı aşamadır. Birçok algoritmaya göre bu yorumlamalar farklılık göstermekte ve her problem türüne göre farklı bir yorumlanma biçimi barındırmaktadır. Aynı problem ve algoritmalar için dahi yorumlanmada değişik birçok metrik mevcuttur. Yorumlamalardan geçen ve daha öncesinde iş analizi kısmında belirttiğimiz değerlere göre kabul edilebilir değerler elde edebilirinse bir sonraki aşamada bu modeller canlıya alma (deploy) olarak adlandırılan aşamaya alınarak gündelik hayatta kullanılabilir bir biçime dönüştürülecektir.

1.3.6. Canlıya Alma (Deploy)

Alınan modellerin kabul edilebilir olması durumunda canlı bir hayata uygulanması ve gerekli dönüşümler yapılarak sistemlere aktarılması işi olarak düşünülebilir. Bu bir web programı için ise gerekli backend işlemleri için

arka planda bağlanabilir, satış tahminlemeye dair bir otomasyonsa gerekli uygulamaya eklenerek kullanıcıların kullanımına sunulabilir.



Şekil 2 CRİSP-DM aşama bazlı açıklayıcı şablon

Şekil 2'de CRİSP-DM methodolojisinin evrelerini ve evrelerdeki işlemler detaylı bir biçimde gösterilmiştir.

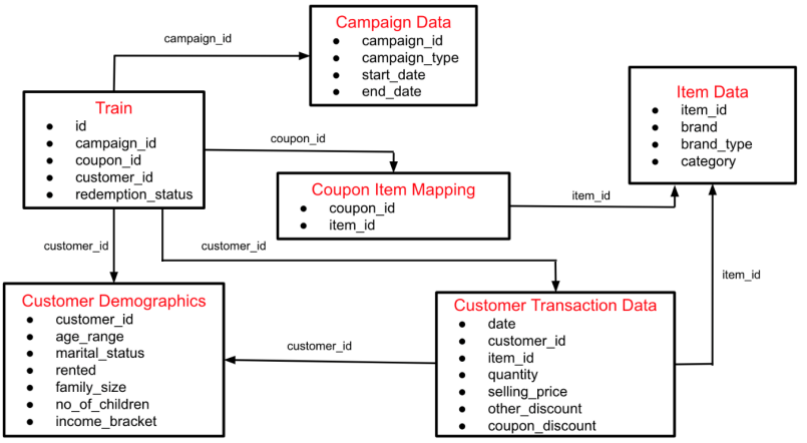
2. Uygulama

Bu başlık boyunca problemi çözebilmek için sırası ile hangi adımlarının gerçekleşeceği anlatılacaktır. Çalışma boyunca Python dilinin Sklearn kütüphanesi, grafikler için Matplotlib ve Seaborn kütüphaneleri, nocode yaklaşımı ile tasarlanmış olan Knime analytics platformu kullanılmıştır. Bu başlık altında CRİSP-DM methodolojisine bağlı kalınarak sırası ile verilerin anlaşılması, keşifçi veri analizi, ön işleme adımları incelenecektir

2.1. Verilerin Anlaşılması

CRISP-DM metodolojisinin 2. basamağı olarak bilinen bu aşamada işe başlanılabilmesi için bir veri kümesi toplanması gerekmektedir. Veri bilimi ve makine öğrenmeleri projelerinde kullanıcılara kişisel verilerinden olabildiğince arındırılmış ve var olan iş problemlerine uygun olarak bir sürü veri seti içerisinde barındıran Kaggle[19] platformu üzerinden bir veri seti tedariki gerçekleştirilmiştir.

Veri seti seçiminde makalenin 1.2. başlık altında belirtildiği üzere müşteri davranışlarının analizi ve satış ilişkisi olarak incelemek istediğimiz bu çalışmada veri seti kolonlarımızın müşterilerin demografik özelliklerini barındıran, eğer mümkünse ürünler üzerinde gerçekleştirdikleri davranışların olduğu niteliklere sahip olan, ürün özelliklerinin de içerisinde barındığı bir tablo tipi ile çalışılması gerekmektedir. Bu çalışma için Kaggle platformu üzerinde "Predicting_Cupon_Redemption_Pca"[16] isimli veri setinin çalışma amacının doğrultusunda uygun olacağı düşünülerek seçilmiş bulunmaktadır.



Şekil 3 Tedarik edilen veri seti hakkında tablolar arası ilişkiler şablonu

Veri setimizin tablo analizlerini yapmamız gerekirse costumer_demograpich isimli 760 satır, 7 sütundan oluşan tabloda müşterilerin demografik özellikleri, costumer_transaction isimli 1324566 satır, 7 sütundan oluşan tabloda müşterilerimizin yaptıkları işlemlere ait özellikleri barındıran, item_data isimli 74066 satır ve 4 sütundan oluşan tabloda ürün özelliklerimizi barındıran, cupon_item_mapping isimli 92663 satır ve 2 sütundan oluşan tablosunda ise ürünler ile kuponların eşleşmelerini içeren ve campaing_data isimli 28 satır 4 sütundan oluşan tabloda kampanya süreleri ve bu kampanyaların özelliklerini belirten bir veri seti olarak karşımıza çıkmaktadır. Değişkenler, işlevleri ve veri tipleriyle beraber tablo 1 de belirtilmiştir.

Tablo 1 Veri Setindeki Değişkenler Ve Özellikleri

Değişken İsmi	Değişken Özelliği	Değişken tipi
Costumer_id	MÜŞTERİLERİ BENZERSİZ KILAN BİR ANAHTAR DEĞİŞKENDİR	İNT
Age_range	MÜŞTERİLERİN YAŞA ARALIĞINI BELİRTİR VE 5 ADET TİPE SAHİPTİR	KATEGORİK
Marital_status	MÜŞTERİNİN EVLİLİK DURUMUNU BELİRTİR (1 : EVLİ, 0 : BEKAR)	İNT
Rented	MÜŞTERİNİN EV SAHİPLİK DURUMUNU İÇERİR (1 : EV SAHİBİ, 0 : KİRA)	İNT
Family_size	AİLEDEKİ BİREY SAYISINI BELİRTİR.	KATEGORİK

No_of_children	AİLEDEKİ ÇOCUK SAYISINI BELİRTİR	KATEGORİK
İncome_bracket	MÜŞTERİNİN SOSYO-DEMOGRAFİK GELİR SEVİYESİNİ BELİRTİR	İNT
Date	İŞLEMLERİN TARİHİNİ VERİR (GÜN-AY-YIL FORMATINDA)	DATE64
İtem_id	ÜRÜNLERİN BENZERSİZ KILAN BİR ANAHTAR DEĞİŞKENDİR.	İNT
Quantity	MÜŞTERİNİN O ÜRÜNDEN KAÇ ADET ALDIĞINI GÖSTERİR	İNT
Selling_price	İŞLEM ÜCRETİDİR (İNDİRİMLERİN UYGULANMASI İLE ELDE EDİLEN FİYAT)	İNT
Other_discount	KUPONSUZ İNDİRİM TUTARINI VERİR	İNT
Cupone_discount	KUPON KULLANIMI İLE GERÇEKLEŞEN İNDİRİM TUTARINI VERİR	İNT
Brand	ÜRÜNLERİN MARKALARINI VERİR	İNT
Brand_type	ÜRÜNÜN MARKA TİPİNİ VERİR	KATEGORİK
Category	ÜRÜN KATEGORİSİNİ VERİR	KATEGORİK
Coupon_id	KULLANILAN KUPONUN İD BİLGİSİNİ VEREN BİR ANAHTAR DEĞİŞKENDİR	İNT
CampaIgn_id	KAMPANYA DEĞİŞKENİNİ BENZERSİZ KILAN BİR ANAHTAR DEĞİŞKENDİR	İNT
Campaign_type	KAMPANYA TİPİNİ VERİR	KATEGORİK
Start_date	KAMPANYA BAŞLAMA TARİHİN (GÜN-AY-YIL FORMATINDA)	KATEGORİK
End_date	KAMPANYA BİTİŞ TARİHİ (GÜN-AY-YIL FORMATINDA)	DATE64

2.2.) Keşifçi Veri Analizi (EDA)

Keşifçi veri analizi adımlarından bu başlık altında bahsedilmiştir. İlk olarak her bir tabloyu kendi içerisinde inceleyerek bir keşifçi veri analizi yapılmıştır. Yapılan keşifçi veri analizindeki çıkarımlar maddeler halinde şu şekilde belirtilmiştir.

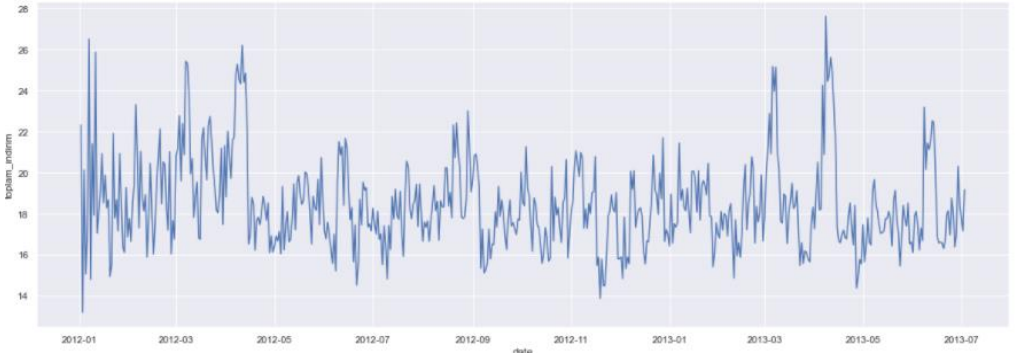
- Campaign_data isimli tabloda 22 adet y, 6 adet x kampanya tipinde kayıt bulunmuştur.
- Kampanya sürelerine bakıldığında 32 gün süren 10 adet kayıt olmakla birlikte geri kalan kampanyaların hepsinin geçerlilik süresi farklılık oluşturmaktadır.
- Bir ürün ile birden fazla kupon eşleşmiştir.
- Müşterilerin evlilik durumu, çocuk sayısı gibi değişkenlerde eksik verilerin olacağı ve belirli kurallar ile bu eksikliklerin doldurulabileceği keşfedilmiştir.
- Gerekli zaman tipleri dönüşümleri alınarak günlük yapılan toplam satış sayısı ve ortalama indirim tutarlarının zaman dilimlerine göre bir etkisinin olup olmadığı incelenmiştir.

Şekil 4 Zamana göre toplam gerçekleşen satış sayısı

Şekil 4' te günlük gerçekleşmiş olan toplam satış miktarı gözükmektedir. Yapılan incelemeler sonucunda Ocak 2021 ve Mart 2021 tarihleri arasında zaman ile toplam satış arasında doğrusal bir ilişki tespit edilmiştir. Bu zaman dilimlerinin dışında ise düzensiz bir değişim gözlemlenmiştir.

Şekil 5 Zamana göre toplam indirim incelemesi

Şekil 5' te günlük gerçekleşmiş olan toplam indirimler gözükmektedir. Yapılan incelemeler sonucunda Ocak 2021 ve Mart 2021 tarihleri arasında zaman ile toplam indirim arasında doğrusal bir ilişki tespit edilmiştir. Bu zaman dilimlerinin dışında ise düzensiz bir değişim gözlemlenmiştir.



Şekil 6 Zamana göre toplam indirim incelemesi

Şekil 6' te günlük gerçekleşmiş olan Ortalama indirimler gözükmektedir. tarihlerin tiplerinde ve zaman çizelgesinde yapılan incelemeler sonucunda kısmi olarak doğrusallık mevcut olsa dahi bu durum tüm zaman çizelgesi için aynı ilişkiyi göstermemektedir. Bu sebepten ötürü zaman değişkenlerinin bu veri kümesi için bir hedef değişken olmadığı belirlenmiştir.

2.3 Ön İşleme:

Analizlerin gerçekleştirilmesiyle birlikte algoritmalarda kullanılmak üzere ön işlemeler bu aşamada gerçekleşecektir.

Eksik verilerin doldurulması aşamasında birçok doldurma yöntemi olmakla birlikte [17] kural tabanlı doldurmanın daha doğru olacağı düşünülerek bu yönteme başvurulmuştur. Kurallar şu şekilde belirlenmiştir :

- Aile sayısı 1 olan müştrilerin hepsi bekardır ve çocukları yoktur.
- Aile sayısı - çocuk sayısı = 1 koşulunu sağlayan müştrilerin hepsi bekardır.
- Bir kişi evli ve aile sayısı iki ise çocuk sayısı 0 dır.
- Evli insanların aile sayısı iki ise ve ailede iki kişi varsa çocuk yoktur.
- Aile sayısı 1 olan müşteriler ise çocuk sahibi olmayacagından ötürü çocuk sayısı 0 dır.

Eksik veriler tamamlanamadığı için marital_status ve no_of_children sistemden çıkartılmıştır.İşlem sonuçlarında literatüre bağlı kalınarak müşteri, ürün ve işlem özelliklerini içeren tablolar birleştirilerek tahmin algoritmalarına verilmiştir.

3.) Değerlendirme

Elde edilen son değişkenler ile birlikte modellere bağlı değişken seçimleri gerçekleştirilmiş, modellemeler gerçekleştirilmiş, hiperparametre optimizasyonları da yapılarak kullanılan tüm algoritmalar ve değeleri sırası ile tablo 2 de verilmiştir. Model Doğrulaması için cross-validaton, değişken seçiminde feature importance ve hiperparametre optimizasyonunda ise gridsearch yöntemleri kullanılmıştır.

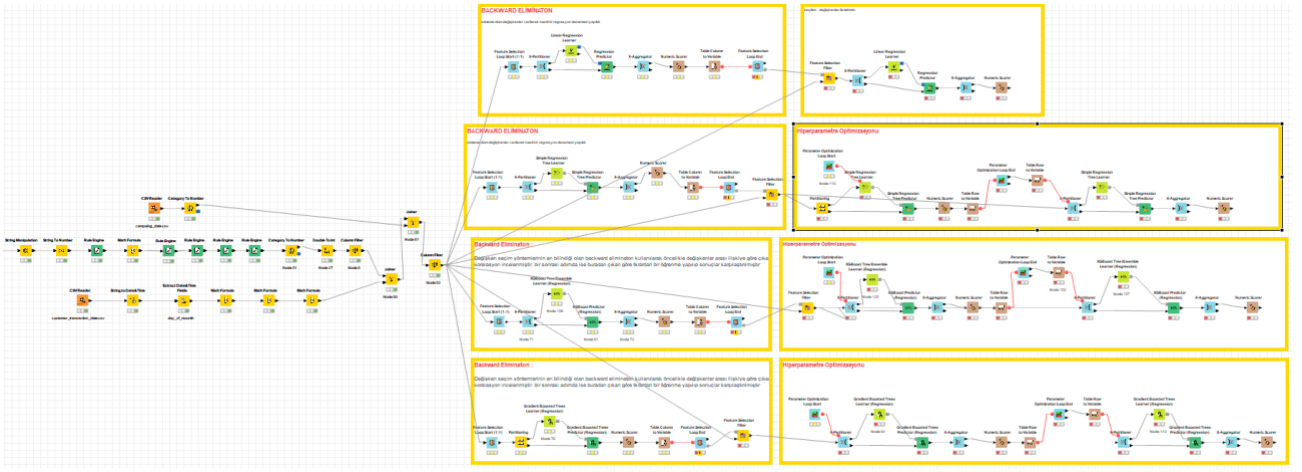
Tablo 2 Algoritma ve sonuçları tablosu

Algoritmalar	Başarım Oranları	Model Eğitim Süresi (Saniye)
Simple Linear Regression (SLR)	-0.015	0.046
Multiple Linear Regression (MLR)	0.0029	0.28
Desicion Tree Regression (DTR)	0.279	1.68
Random Forest Regression (RFR)	0.414	191
Extreme Gradient Boosting (XGBoost)	0.317	42
Gradient Boosting Tree (GBR)	0.232	40
Light GBM (LGBM)	0.300	5.43

Elde edilen değerler sonucunda en yüksek doğruluk değeri Random Forest Regresyon ile elde edilmiştir. Fakat performans kriterleri göz önüne alındığında doğruluk değeri, model eğitim süresi gibi avantajları sebebi ile modellemede en uygun algoritma Light GBM olarak gözlemlenmiştir.

4. Knime Platformu Üzerinde Çalışmanın Bir Uygulaması

Knime üzerinde uçtan uca veri bilimi işi yürütebilmemizi sağlayan, açık kaynak kodlu, data mining suit olarak bilinen bir yazılımdır. Geliştirme süreci boyunca 4.3.2 sürümü kullanılmıştır.



Şekil 7 Uygulamanın Knime Platformu Üzerindeki İş Akışı

Ön işleme işlemleri sklearn kısmındaki akış ile aynı olup algoritmalar kısmında değişken seçim yöntemi olarak sarmal (wrapped) method olarak bilinen Backward-Eliminaton yöntemi kullanılmış olup Knime içerisinde yer alan ve hazır olarak bulunan Backward Feature Elimination Metanode'u kullanılacak algoritmaya göre öğrenme algoritmaları değiştirilmiş ve tüm veri setine uyarlanabilirlik açısından iste partitioner düğümleri, X-partitioner düğümleri ile değiştirilmiştir. Elde sonuçların birebir olarak sklearn ile kullanılan modeller ile kıyaslanarak tablo 3'te verilmiştir.

Tablo 3 Knime Ve Sklearn modellerinin karşılaştırılması

Modeller	**R^2**
Simple Linear Regression (Python)	- 0.015
Simple Linear Regression (Knime)	0.018
Multiple Linear Regression (Python)	0.027
Multiple Linear Regression (Knime)	0.038
Desicion Tree Regression (Python)	0.280
Simple Regression Tree (Knime)	0.469
Random Forest Regression (Python)	0.414
XGBoost (Python)	0.318

XGBoost (Knime)	0.460
Gradient Boosting Regerssion (Python)	0.232
Gradient Boosting Regerssion (Knime)	0.192
LightGBM (Python)	0.300

5.) AutoML Yapay Zeka Yaklaşımı ile Karşılaştırma

AutoML yaklaşımı veri bilimi faaliyetlerini veri bilimcisine ihtiyaç duymadan veya daha az ihtiyaç duyarak yapabilmeyi amaçlayan verileri belirli yazılımlar vasıtası ile otomatik olarak analiz edebilen ve analizler beraberinde değerler üretebilen bir yöntemdir. Optiwisdom [18] firmasının çıkartmış olduğu Optiscorer isimli yapay zeka motoru gerekli her türlü işleyişi ve skorlamayı sağlayabilmektedir.

İşleyiş için yapay zeka motorunun işleyebilmesi için bir örnek olarak eğitim ve test kümeleri temin edilmelidir. Örnek alım süreci için df.sample(n) methodu kullanılarak bir örneklem alınmıştır. Beraberinde bu motorun 5000 satır limiti bulunduğundan ötürü 5000 satırlık bir örneklem ile test gerçekleşmiştir.

Alınan örnekler sonucu R^2 değerimiz -0.179 olarak bulunmuştur. Kendi hiperparametreli modelimiz ise bu aşamada 0.101 olarak bulmuştur. Bu farklılık veri setindeki satırların birbirleri ile aynı sıra ve şekilde dağılmamasından kaynaklı olarak 5000 satırlık bir veri seti üzerinde iki algoritma sırası ile test edilecektir. Test süresi boyunca hız, doğruluk, ve performanslar sebebi ile Light GBM algoritması kullanılarak karşılaştırılacaktır. Aynı zamanda bu algoritma Optiscorer Motoru'nun varsayılan makine öğrenmesi algoritmasıdır.

Yapılan doğrulama yönteminde bir test ve train işleminden geçmiş bulunmaktadır. Bu işlem için önceden aldığımız 5000 satırlık veri setini doğrusal bir sıralama ile 1000 er satır oluşturacak bir şekilde test ve 4000 er den oluşacak şekilde eğtim kümeleri olarak ayrılmıştır. Bu işlemler ile iki algoritma karşılaştırılmış ve tutarlılık sağlanmıştır.

1. parça ile yapılan regreson sonucu 0.124 gibi bir değer elde etmektedir. Kendi modelimde gerçekleşen bu değer 0.188 olarak gerçekleşmiştir. Beraberinde varsayılan parametreli bir model oluşturulduğunda 0.188 olarak gerçekleştirmiştir.

2. parça ile elde edilen değerler için Optsicorer kendisi 0.131 değerini elde ederken hiperparametreli modelim 0.089 olarak bir değer elde etmiştir. Beraberinde varsayılan parametreli bir model oluşturulduğunda 0.136 olarak gerçekleştirmiştir.

3. parça ile gerçekleşen işlemler ise optsicorer için -0.162, Hiperparametreli modelimiz için 0.101, varsayılan model için 0.155 olarak hesaplanmıştır

4. parça ile ise işleyiş optsicorer için -0.162, hiperparametreli model için 0.448, varsayılan parametreli model için -0.258 olarak hesaplanmıştır/

5. parça ile işleyiş optsicorer için 0.101, varsayılan parametreli model için 0.144 ve hiperparametreli model için 0.109 olarak gerçekleşmiştir.

Cross validaton sonuçlarının 3 model üzerinden tartışmamız gerekirse

Tablo 4 Sonuçların Karşılaştırılması

	Varsayılan Parametreli	**Optiscorer**	**Hiperparametreli Model**
1. parça	0.188	0.124	0.172
2. parça	0.138	0.131	0.089
3.parça	0.155	-0.162	0.104
4.parça	-0.256	-0.162	-0.448
5. parça	0.144	0.101	0.109
Toplam:	0.0738	0.0064	0.0052

Elde edilen sonuçların daha önce elde edilmiş olan çalışmamız sonuçları ile uyumsuzluğu gözlemlenmiştir. Bu farklılığın daha düşük bir veri seti üzerinde gerçekleştirilmesinden kaynaklı olarak düşünülmesi sebebi ile tutarlılığını sorgulayabilmek adına 1000 satırdan oluşan farklı bir veri kümesi örneği üzerinde 800 ve 200 oranlarında doğrusal bir şekilde sıralı olarak eğitim ve test uygulamaları gerçekleştirilerek bir deneme daha yapılmış ve sonuçlar karşılaştırılmıştır. Karşılaştırma sonuçları tablo 5 üzerinde verilmiştir.

Tablo 5 1000 satırlık veriler ile karşılaştırma

	Varsayılan Parametreli	**Optiscorer**	**Hiperparametreli Model**
1. parça	-0.052	-0.105	-0.284
2. parça	0.0236	-0.055	-0.214
3. parça	0.074	-0.034	-0.035
4. parça	-0.017	-0.196	-0.233
5. parça	0.098	-0.015	-0.0065
Toplam:	0.025	-0.08	-0.11

6. Sonuç

Yapılan testler sonucu bazı çıkarımlarda bulunulmuştur. Hem tüm makalenin özeti olarak sayılabilecek olan beraberinde modeller ve değerlerin sorunları ile ilgili tespitler bu aşamada gerçekleştirilmiştir.

- Elde edilen büyük veriye bakıldığında elde edilen değerlerin birçoğu modelleme için uygun olmayıp beraberinde küçük bir veri setinden elde edilen modeller büyük veriye uygulandığı zaman çok düşük doğruluklarda çıkmaktadır. Gerek demografik gerekse ürün özellikleri içerisindeki dengesiz dağılım durum sebebi ile makine modelleme yapamamakta ve veriler tutarsız olması sebebi ile iyi regresyon değerleri elde edilememektedir.

- İkinci aşamada yapılan tüm veri seti kullanılarak yapılan regresyon tipi en sağlıklı regresyon tipi olup küçük veriler üzerinde uygulandığında ortaya çıkan durum göre daha iyi sonuçlar elde edilebilmektedir. Büyük veri setlerinde yeteri kadar eğitim ve test kümeleri için veri bulunması bu durumun açıklamasıdır.

- Beraberinde elde edilen Knime ve Sklearn çalışmaları her ne kadar aynı veri setleri içerisinde modellemeler gerçekleşse ve yakın değerler elde edilse dahi tam manası ile aynı değerleri elde edilememesinin birkaç sebebi vardır. Veride eğitime ve teste giren veriler sıralama bazında aynı veriler olmamaktadır. Bu durumun bir örneği 5. Başlıktaki aynı veri setlerini vermemize rağmen Automl ile çalışmadaki verilerin aynı olması fakat sıralamaların farklı olması sebebi ile elde edilen farklılık değeri bir örnektir. beraberinde varsayılan model parametreleri sebebi ile aralarında fark elde edilmektedir. Seçilen değişkenler farklı olup hiper parametre iyileştirmelerinde (hyper parameter tuning) birbirleri ile aynı değerlere yaklaşılmıştır. Tam olarak aynısının yakalanamamasında hiperparametre optimizasyonu için belirli algoritmalarda iki taraf içinde aynı ayarlar olmaması ve seçilen ayarlar dışındaki bazı ayarların varsayılan değerlerinin diğerlerinden farklılıklar olmasıdır.

- Beraberinde elde edilen modeller ile ilgili olarak her ne kadar model isimleri aynı gibi gözükse de Knime ile sklearn arasındaki farklılıklar incelendiğinde gerek varsayılan model parametreleri gerekse model parametrelerinin aynı bir biçinde ayarlanamaması farklılıkların başlıca göstergesidir.

- Son aşamada optsicorer gibi bir motor ile yapılan teslterin düşük çıkması sebeplerine gelinecek olursa yapılan test ve incelemelerde elde edilen değerlerde ilk aşamada ki gibi verilerin karıştırılarak yapılan tahmin çok daha iyi bir değer iken son aşamada lineer bir bölüntüleme ile almamız regresyondaki kötülüğü arttırmaktadır.

- Bir diğer sebep ise eğitim ve test için yeterli sayıda veri kullanamamamız ve verilerin herhangi bir istatiksel yaklaşım ile değil rastgele seçilerek alınması başlıca göstergesidir.

Yapılan çalışma neticesinde literatürde müşteri davranışı (“costumer behaivor”) olarak geçmekte olan kavramın veri madenciliği teknikleri kullanılarak bir çalışması gerçekleştirilmiştir. Çalışma boyunca müşterilerin sosyo-demografik özellikleri, ürün özellikleri gibi verilerden faydalanılarak modellemeler gerçekleştirilmiştir. Müşterilerin ürünleri tahmini olarak satın alabilecekleri indirim oranları belirlenerek kullanıcılara sunulmuştur. Modelleme süreçleri doğruluk açısından en yüksek değeri Random Forest Regresyon algoritması ile yapılan modelleme verirken CRISP-DM döngüsünün son aşaması olan projenin canlıya alınması süreçleri içerisinde hız,

performans, ve yüksek doğruluk değerlerinin önemliliği sebebi ile modellemeye uygun olarak Light GBM algoritması seçilmiştir.

Aynı özellikleri gösteren müşteriler ve ürünler yorumlanmış, beraberinde aynı özelliklere sahip müşterilerin aynı davranışları gösterebileceği kanıtlanmıştır. Müşterilerin bu davranışlardan yararlanılarak potansiyel olarak ürünü alabilecekleri oranların belirlenmesi ile satışların arttırılması amaçlanmıştır. Satış, pazarlama ve hizmet gibi alanlarda kullanılması hedeflenen bu çalışmada veriler bağımsız platformlardan çok sayıda yöntem ile toplanabilir ve 1. Başlık altındaki süreçlere ve değişkenlere bağımlı kalınarak bir modelleme yapılabilir. Çalışmamız boyunca verilerin birbirleri ile kolerasyonarına bakıldığında sosyodemografik veriler ve ürün verileri dışında, başka değişkenlerin yeterli miktarda bir iyileştirmede bulunmadığı ve bu fazlalıkların performans kaybı gibi sebepler doğurmasından ötürü sadece sosyo-demografik verilere ve ürün özellikleri kararlaştırılarak bir modelleme yapılmış, çalışma gerçekleştirilmiştir.

6.) Kaynakça

[1] Dr. Ertuğrul AKTAN , Big Data: Application Areas, Analytics and Security Dimension

[3] Min Chen , Shiwen Mao, Yunhao Liu , Big Data: A Survey (2014)

[4] İnternet Kaynağı: https://www.networkworld.com/article/3325397/idc-expect-175-zettabytes-of-data-worldwide-by-2025.html

[5] Fahrulnisa Sema ÖZBAŞ, Müşteri Kayıp Tahmini (Customer Churn Prediction), YBS Ansiklopedi, Cilt 8, Sayı1, Haziran 2020

[6] İzmir Kâtip Çelebi Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 2020; 3(1):80 - 92

[7] Kader BASTEM, Veri Madenciliği Yöntemleri ile Twitter Üzerinden MBTI Kişilik Tipi Analizi, v. 4, is. 2, Haziran, 2017

[8] Doç. Dr. Birgül Kutlu Yrd. Doç. Dr. Bertan Badur , Yapay Sinir Ağları İle Borsa Endeksi Tahmini Boğaziçi Üniversitesi Yönetim Bilişim Sistemleri Bölümü

[9] ProTicaret e-ticaret akademisi konferansı (2017)

[10] The impact of e-service quality and customer satisfaction on customer behavior in online shopping Paulo Rita a,* , Tiago Oliveira a , Almira Farisa

[11] Fractal Dimension of Shopping Path: Influence on PurchaseBehavior in a SupermarketYuta Kanekoa,∗, Katsutoshi YadaaaData Science Laboratory, Kansai University, 3-3-35 Yamate, Suita, Osaka 564-8680, Japan

[12] Studyıng Customer Behavıor In Retaıl Stores Wıllıam Applebaum Stop Shop, Inc.

[13] Mu-Chen Chena, Ai-Lun Chiub , Hsu-Hwa Changc , Mining changes in customer behavior in retail marketing Expert Systems with Applications 28 (2005) 773–781

[14] Jian jin, Ying Liu, Ping Ji ve Honngguang Liu, Understanding Big Costumer Opinion Data For Market-driven Product Desing

[15] A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA) Umair Shafique and Haseeb Qaiser Department of Information Technology, University of Gujrat, Gujrat, Pakistans

[16] İnternet Kaynağı : https://www.kaggle.com/vasudeva009/predicting-coupon-redemption-pca

[17] Şadi Evren ŞEKER, Eda Eşmekaya, Eksik Verilerin Tamamlanması (Imputation) Cilt 4, Sayı 3, Eylül 2017

[18] OptiScorer: Otomatik Makine Öğrenmesi ile Skorlama Şadi Evren ŞEKER Cilt 8 Sayı 1, Ocak 2020

[19] İnternet Kaynağı : https://www.kaggle.com/