Büyük Veri Analitiği Final Ödevi

Ramazan Erduran

21821809

İstatistik
Hacettepe Üniversitesi

30.05.2023

Contents

1 Giriş **2**

2 Büyük Veri ve İstatistik **3**

3 Veri Kümesi **4**

4 Uygulama **5**
4.1 Verilerin Keşfi 5
4.2 İlişkisel Veri Analizi 5
4.3 Tahmin Yapma Aşaması 6

5 Sonuç ve Tartışma **8**

6 Kaynakça **9**

1 Giriş

Her geçen gün üretilen bilgi miktarı bir önceki güne kıyasla daha hızlı bir şekilde artmış ve bu nedenle bu büyük verileri analiz edip anlamlı içgörüler çıkarmak giderek daha önemli hale gelmiştir.

Büyük veri analitiği alanı, bu devasa veri havuzlarından anlamak, sınıflandırmak, analiz etmek, tahminlemeler yapmak ve değerli bilgiler çıkarmak gibi pek çok yöntemi kapsar. Ayrıca, bu alanda, araştırmacıların sofistike büyük veri analiz faaliyetleri uygulamalarını sağlamak için pek çok büyük veri işleme teknolojilerini kullanılır. KNIME'da bu teknolojilerden biridir.

Büyük veri faaliyetleri, ön işleme fonksiyonları, modelleme aktiviteleri ve öngörü işlemleri vb. şeyleri içerir.

Bu raporlamada büyük veri analitiğiyle ilişkili temel kavramları aydınlatmayı ve aynı zamanda bir büyük veri analitiği uygulamasında KNIME'ın nasıl kullanılabileceğini göstermeyi amaçlıyorum.

2 Büyük Veri ve İstatistik

Açıkcası bu kısımda büyük veri ile istatistiği birbirinden ayırmak biraz tuhaf olacaktır. Zira büyük veri analitiğinde kullanılan tekniklerin belki de %99'u istatistik temelli tekniklerdir.

Günümüzde karmaşık bilgilerin muazzam bir hacimle biriktirilmesi ve işlenmeye çalışılması, geleneksel hesaplama sistemlerinin etkin bir şekilde başa çıkamadığı bir seviyededir. İşte tam da burada büyük veri işlemede kullanılan teknolojiler devreye girer. Büyük veri, geleneksel veri işleme teknikleriyle kolayca yönetilemeyen yapılandırılmış ve yapılandırılmamış verilerdir. Öte yandan istatistik, veri toplama, analiz etme, yorumlama ve iç görüler üretmeye odaklı matematiksel bir daldır.

Büyük veri ve istatistik benzer konuları ele alsa da, veri ölçeğinde, veri türlerinde ve uygulanan metodolojide farklılıklar gösterir. Büyük veri, gelişmiş hesaplama çerçeveleri aracılığıyla yüksek ölçekte verilerin analizi ve yönetimine odaklanırken, istatistik daha küçük çaplı veri setleri ile alakadardır. Zaten istatistiğin temelinde olasılıkları ve kitaleyi en iyi temsil eden örneklemi ele alarak daha küçük veri setleriyle daha büyük veri setleri hakkında yargılara varmaya çalışmak vardır. İstatistikte örnekleme yöntemleri ve hipotez testleri kullanılırken, büyük veri analitiğinde zaten kitlenin kendisi analiz edilmeye çalışıldığından hipotez kurmaya veya örneklem çekmeye gerek yoktur.

Ancak bu farklılıklara rağmen büyük veri analitiği ve istatistik iç içe geçmiş alanlardır. Zira büyük veri analitiği ile elde edilen çıkarımlar istatistiksel teknikler ile yoğrularak daha doğru ve iyi kararlar almayı, veri kümelerini derinlemesine anlamayı sağlar.

3 Veri Kümesi

Uygulamada kullanılan veri kümesi Avrupa'daki bir bankadan (ismi belirtilmiyor) alınmıştır. Veri seti bankaya kayıtlı müşterilerin churn edip etmeyeceğine ilişkin verileri taşımaktadır.

Veri setinde bulunan değişkenlere ilişkin açıklamaya Tablo 1'te görüntülenmektedir.

Table 1: Meta

Değişken	**Açıklaması**
RowNumber	Kayıt (sıra) numarasına karşılık gelir.
CustomerId	Müşteriye ait benzersiz bir değerdir. Her müşterininki farklıdır.
Surname	Müşterinin soyadı
CreditScore	Müşterinin kredi puanı
Geography	Müşterinin yaşadığı yeri belirtir.
Gender	Müşterinin cinsiyetini belirtir.
Age	Müşterinin yaşını belirtir.
Tenure	Müşterinin bankanın müşterisi olduğu yıl sayısını ifade eder.
Balance	Müşterinin hesap dengesini belirtir.
NumOfProducts	Bir müşterinin banka aracılığıyla satın aldığı ürün sayısını ifade eder.
HasCrCard	Müşterinin kredi kartı olup olmadığını gösterir.
IsActiveMember	Müşterinin aktif olup olmadığını gösterir.
EstimatedSalary	Müşterinin tahmini maaşını gösterir.
Complain	Müşterinin şikayetinin olup olmadığını gösterir.
SatisfactionScore	Müşterinin tatminiyet skorunu gösterir.
CardType	Müşterinin sahip olduğu kart tipini gösterir.
PointsEarned	Müşterinin kredi kartıyla topladığı puanı gösterir.
Exited	Müşterinin churn edip etmediğini gösterir (hedef değişken).

Bilindiği üzere bankaya yeni müşteri katmaktansa eldeki müşterilerin bankadan ayrılmasını engellemek daha az maaliyetli ve daha az yorucu bir iştir. Ayrıca eldeki müşterilerin memnuniyeti sağlandığında bankayı önerme olasılıkları artar ve dolaylı yoldan yeni müşteriler de elde etmiş oluruz. Bu amaçla hangi müşterilerin bankadan ayrılacağına ilişkin bir makine öğrenmesi projesi ile tahminlerde bulunacağız. Bankadan ayrılacak müşterilere ilişkin promosyon önerileri gibi önerilerde bulunulabilir...

4 Uygulama

4.1 Verilerin Keşfi

Verileri keşfetme sürecinde çeşitli görselleştirilmeler ile veri seti içerisinde gizli ilişkiler ve hedef değişkene olan etkiler araştırıldı.

Çıkarımlar şu şekilde elde edildi:

1. Bankadan ayrılan müşterilerin çoğunluğunu kadınlar oluşturdu.
2. Bankadan ayrılan müşterilerin çoğunluğunu kadınlar oluşturdu.
3. Bankadan ayrılan müşterilerin konumlarına bakıldığında Almanya ve Fransa topluluğun %40'ını oluştururken churn eden müşterilerin en azı İspanya'da bulunanlar oldu.
4. İspanya'daki müşterilere ayrıcalık mı tanınıyor şeklinde bir inceleme yapıldı ve üç ülkenin de ortalama tatmin skoru eşit bulundu.
5. Bankadan ayrılan müşterilerin tatmin skorlarında bir ilişki bulunamadı. Ayrılan ve ayrılmayan müşterilerin skorlarının ortalamaları 3 olarak bulundu. Demek ki müşterilerin gitmesine sebep olan farklı bir durum var, zira tatmin skorları 3/5 olarak bulundu.

4.2 İlişkisel Veri Analizi

Değişkenlerin birbiri ile olan ilişkileri incelendi ve sonuçlar şu şekilde elde edildi:

1. Bağımsız değişkenlerin kendi aralarındaki ilişkileri incelendiğinde sadece *ürün sayısı* ile *hesap bakiyesi* değişkenleri arasında orta düzeyde doğrusal ilişki bulundu. Diğer değişkenlerin birbiri ile ilişkisi çok zayıf ya da yok olarak belirlendi.
2. Müşterinin bankadan ayrılıp ayrılmadığını belirten, hedef değişken olan *churn* değişkeni üzerindeki ilişkileri incelendiğinde en çok etkisi olan değişken ***şikayet*** değişkeni old. Öyle ki karar ağaçları algoritmasında bu değişkenin veri setine eklenmesi ile başarı oranı %100'e ulaştı. Bu nedenle bu değişkeni veri setinden çıkarıp daha gerçek dünyaya uygulanabilir bir veri setiymişcesine işlemler yapılmaya devam edildi.
3. Hedef değişkenle olan ilişkilerin yanı sıra bağımsız değişkenler için PCA (Temel Bileşenler Analizi) yapılmak istendi ancak bağımsız değişkenlerin kendi içlerindeki çok zayıf/yok derecesindeki korelasyondan dolayı bu işleme devam edilemedi.

Yukarıdaki sonuçlardan hareketle veri setine dair tüm çıkarımların yapıldığı sonucuna varılıp makine öğrenmesi adımına geçildi.

4.3 Tahmin Yapma Aşaması

Tahmin aşamasına gelindiğinde sınıflandırma algoritmaları olarak:

- Karar Ağaçları Algoritması
- Lineer Regresyon Algoritması
- En Yakın Komşu Algoritması

Kullanıldı.

Algoritmaların sonuçları değerlendirilirken iki adet metrik kullanıldı. Bu metrikler yüzdelik doğruluk oranı (Accuracy) ve Kappa katsayısı oldu. Kullanılan metriklere göre en iyi sonucu veren algoritma lojistik regresyon oldu.

Şikayet değişkeni modeldeyken %100 doğruluk ile en iyi sonucu veren model karar ağaçları algoritmasıyken bu değişkeni veri setinden çıkardığımızda en iyi sonucu veren algoritma lojistik regresyon oldu.

Değişkenin modelden çıkarılması hususunda:

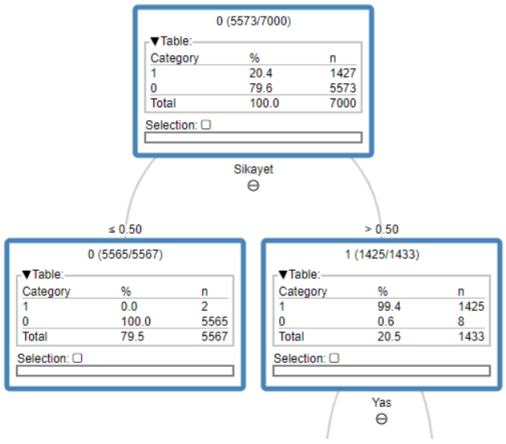

Figure 1: Şikayet değişkeni modeldeyken

1 Numaralı resimde de görüldüğü üzere şikayeti 0.50'nin altındaki kişiler gerçekten de churn etmemiş ve buradaki oran %79.5 olarak belirtilmiş. Geriye kalan %21.5'luk veriyi de kolay bir şekilde dallandırarak %100'lük bir başarıya ulaşmış. Ancak bu durum gerçek hayatı yansıtmayacağından ilgili değişken veri setinden çıkarılıp tekrar denendiğinde ise 2'de bulunan sonuçlar elde edildi.

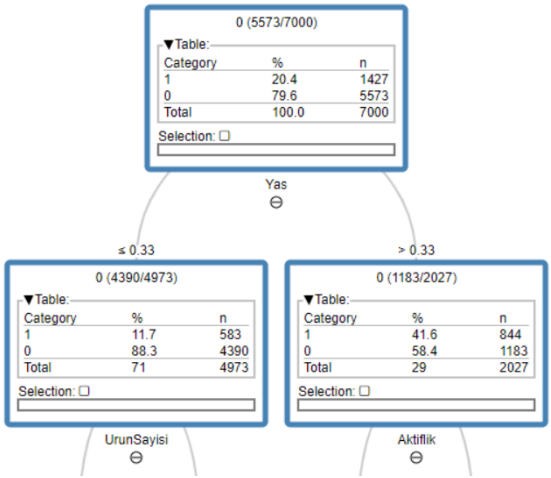

Figure 2: Şikayet değişkeni modelden çıkarıldığında

2 Numaralı resimde de görüldüğü üzere artık model %100'lük bir doğruluktan ziyade daha normal, kabul edilebilir, gerçek dünyaya uygun şekilde tahmin yapmakta.

Tüm bu yorumlamalardan sonra en nihayetinde *Şikayet* değişkeni veri setinden çıkarıldı ve 3 farklı model denendi. En iyi model ise lojistik regresyon modeli oldu.

Modele hiper parametre ayarı da yapıldı ve en iyi modelin hiper parametreleri;

1. Solver: Iteratively reweighted least squares

Şeklinde elde edildi.

5 Sonuç ve Tartışma

Sonuç olarak veri ön işleme, veri görselleştirme ve bu görselleştirmelerden içgörü elde etme, keşifsel ve ilişkisel veri analizi yapma, makine öğrenmesi ile tahminlerde bulunma, modelleri çeşitli metriklerle değerlendirme ve sonucunda hipoer parametre ayarı ile en iyi modeli bulma adımları tamamlandı.

Bu adımların sonucunda müşterilerin bankadan ayrılmasını en çok etkileyen değişken **şikayet** değişkeni olurken en iyi tahmin modeli (**şikayet** değişkeni veri setinde değilken) lojistik regresyon modeli oldu.

Modellerin değerlendirme süreçlerine ilişkin sonuçlar Tablo 2'te görülebilir

Table 2: Değerlendirme Sonuçları

Model	**Metrikler**	
	Accuracy	**Kappa**
Decision Tree	%80.167	0.373
Lojistik Regresyon	%80.733	0.206
KNN	%78.267	0.187

Değerlendirme sonucu en iyi olan lojistik regresyon modelinin hipoer parametre ayarı yapıldıktan sonra %81.467 accuracy ve 0.245 kappa sayısına sahip oldu.

6 Kaynakça

1. Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., & Sieb, C. (2008). KNIME: The Konstanz Information Miner. Data Mining and Knowledge Discovery Handbook, 1-17.

2. Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. MIS quarterly, 36(4), 1165-1188.

3. KNIME Hub: https://hub.knime.com/

4. Statology: https://www.statology.org/