Makine Öğrenmesi Ödevi

Semih Yazici - 19011087

Yıldız Teknik Üniversitesi
semihyazci@gmail.com

11 Aralık 2022

Özet

Bu çalışmada bir bankanın müşteri kayıp analizi için tahmin sistemi geliştirilmiştir. Bu proje kapsamında kullanılan veri kümesi Kaggle'dan sağlanmıştır. Veri kümesi genel olarak bayraklardan oluşmaktadır. Kullanılan modellerde elde edilen sonuçlar arasında çok büyük fark vardır. Test sonucu en başarılı olan modeller %88 ile Naive Bayes olup %85 ile arkasından derin öğrenme modelidir.

I. Giriş

Müşteri kayıp analizi uzun süredir üzerinde çalışılan bir konudur. Makine öğrenmesi kullanılmadan önce uzmanlar yardımıyla çözülmeye çalışılan bu problem bir müşterinin davranışlarından bulunduğu firmanın eşdeğeri bir firmaya tercih edilip edilmeyeceği anlaşılmaya çalışılır. Bu problemin zor olmasının temel sebeplerinden birisi çok fazla özelliğin bu problem kapsamında kullanabilir olmasıdır. Bir uzman bütün özelliklerin müşterinin davranışına etkisini hesaplaması zordur ancak makineler yüksek seviyeli özellikle işlemleri kolayca yapabilmektedir. Bu sebeple makine öğrenmesi tekniklerinin bu alanda kullanılmaya başlanması ile müşteri kayıp analizi problemi kolaylaşmıştır.

II. Veri Kümesi

Bu çalışma kapsamında kullanılan veri kümesi Kaggle'dan alınmış bir banka verisidir.Veri kümesi üşteri kayıp analizi veri kümelerinin hepsinde karşılaşılabilecek bir problem olan dengesiz veri dağılımı problemini içermektedir. Müşteriler genel olarak bulundukları firmayı ortabüyük bir problem olmadığı takdirde değiştirmeme eğiliminde olduklarından bu problemin gerçekleşme sebebi anlaşılabilir. Genel olarak müşteri kayıp analizi problemlerinde %3-%20 arasında müşteri firmayı terketmiş olur. Kullanılan veri kümesinde ise bu oran %20'dir

Veri kümesinde toplamda 10.000 örnek bulunmaktadır. Veri kümesinde toplamda 11 özellik bulunmaktadır ancak müşteri numarası ayırt edici bir özellik olmadığından kullanılmamıştır. Geri özellikler aşağıda açıklanmıştır.

- Kredi Puanı: Bir müşterinin kredi puanıdır. Bu özelliğin tipi tam sayıdır.
- Ülke: Bir müşterinin bulunduğu ülkeyi içerir. Bu özelliğin tipi stringdir.
- Cinsiyet: Müşterinin cinsiyet bilgisidir. Bu özelliğin tipi stringdir.
- Yaş: Müşterinin yaş bilgisini içeririr. Bu özelliğin tipi tam sayıdır.
- Bulunduğu Süre: Müşterinin bankada ne kadar süredir bulunduğu bilgisidir. Bu özelliğin tipi tam sayıdır.
- Bakiye: Müşterinin hesabında bulunan para miktarıdır. Bu özelliğin tipi floattır.
- Sahip Olunan Ürün Miktarı: Müşterinin bankada kullandığı ürün miktarıdır. Bu özelliğin tipi tam sayıdır.
- Kredi Kartı: Müşterinin kredi kartına sahip olup olmadığı ile ilgili tutulan bir bayraktır. Bu özelliğin tipi tam sayıdır.
- Aktif Kullanıcı: Müşterinin bankaya göre aktif bir kullanıcı olup olmadığı ile ilgili tutulan bir bayraktır. Bu özelliğin tipi tam sayıdır.
- Maaş: Müşterinin tahmin edilen yıllık gelir

bilgisidir. Bu özelliğin tipi floattır.

III. Sınıflandırma Modelleri

Müşteri analizi gerçekleştirilirken Naive Bayes, Karar Ağacı, Derin Öğrenme ve KNN yöntemleri kullanılmıştır.

i. Naive Bayes

Naive Bayes yöntemi koşullu olasılık formülünü temel alan bir yöntemdir. Bu yöntem çok basit bir temele sahip olmasına karşı kategorik verilere sahip problemlerde başarılıdır. Uzun süre boyunca spam email probleminde kullanılmış bir yöntemdir. Naive Bayes yönteminin müşteri kayıp analizinde kullanılmasının temel sebebi her bir özelliğin kategorik hale çevirilebilir olmasından kaynaklı olarak bu problemin başarılı olacağı düşünülmesidir. Test sonuçlarına bakıldığı durumda da görülebileceği üzere en yüksek başarıya sahip olan model Naive Bayes olmuştur.

ii. Karar Ağacı

Karar ağacı uzun süredir bilinen güçlü bir makine öğrenmesi tekniğidir. Karar ağaçları Naive Bayes yönteminin aksine kategorik veri ile birlikte sayısal verilerle de çalışabilir. Karar ağaçları probleme uyum sağlama potansiyeli en fazla olan makine öğrenmesi yöntemidir. Bunu aşırı derecede küçük gruplara ayırarak yapabilmektedir. Bu sebeple overfit problemi sık karşılaşılan bir sorundur.

iii. Derin Öğrenme

Derin öğrenme yöntemi günümüzde en çok kullanılan makine öğrenmesi tekniklerden birisidir. Karar ağaçları gibi hem kategorik veri ile hem de sayısal verilerle çalışabilmektedir. Karar ağaçları kadar probleme uyum yeteneğine sahip olmamasına karşı overfit probleminden daha az etkilenirler.

Derin öğrenme yöntemi bir fonksiyon yakınsaması problemi olduğundan daha derin bir mimari daha yüksek öğrenme kapasitesi anlamına gelmektedir. Ancak bu durum iki uçlu bir bıçak olarak düşünülebilir. Çünkü artan kapasite modelin öğrenme potansiyelini arttırmakla birlikte yeterli miktarda veri yoksa overfit riskini de arttırmaktadır. Özellik müşteri analizi gibi dengesiz dağılıma sahip problemlerde derin öğrenme modellerinin overfit etmesi çok sık karşılaşılabilecek bir problemdir. Ancak derin öğrenme alanında yapılan son 20 yıllık çalışmalar sonucunda overfit problemi ciddi düzeylerde iyileştirilmiş ve daha stabil bir model üretilmiştir. Bu çalışma kapsamında da derin öğrenme modelinde iyileştirilmeler yapılmıştır.

iv. KNN

KNN(K En Yakın Komşuluk) bir örneğin hangi örnek kümesine ait olduğunu bulmak amacıyla kullanılan bir yöntemdir. Bu yöntem doğru ayarlamalar yapıldığı takdir az miktarda veri içerilen durumlarda derin öğrenme gibi çok fazla veriye ihtiyaç duyan modellerin aksine stabilliğini koruyan bir yöntemdir.

KNN yöntemi doğru K seçimi ile stabil model oluşturulabilmesine karşı, K yüksek bir değer seçilmesi durumunda model overfit edebilir. Özellikle veri kümesi müşteri analizi gibi dengesiz veri dağılımına sahip ise overfit riski daha da artmaktadır.

IV. Deneysel Analiz

i. Veri Hazırlık Aşaması

Kullanılan sınıflandırma yöntemlerinin gereksinimleri düşünüldüğünde ve daha yüksek başarıya sahip olmaları amacıyla veri kümesinin kategorik hale getirilmesine karar verilmiştir. Ancak bakiye, yaş, maaş gibi geniş aralığa sahip olan özellikler için direkt olarak kategorizasyon işlemi gerçekleştirilememektedir. Bu durumda yapılabilecek iki işlem bulunmaktadır; ilk yaklaşımda bu veriler normalize edilip kategorik olmayan şekilde verilebilir. Ya da ikinci yaklaşımla bu veriler aralıklara bölünüp

gruplanabilir. Örneğin maaş aralıkları yaratılıp bunlara göre insanlar düşük, ortalamanın aşağısı, ortalamanın üstü ve yüksek gelir gruplarına ayrılabilir. Bu sayede kümeleme işlemleri ve kıyaslama işlemleri kolaylaşır. Bu çalışma dahilinde ikinci yöntem kullanılmıştır. Bütün veri grupları kategorik hale getirildikten sonra bütün örnekler bir vektör haline getirilmiştir.

ii. Elenen Özellikler

Projede elenen tek özellik müşteri numarasıdır. Çünkü müşteri numarası her müşteriye özel olduğundan ötürü bir anlama sahip değildir.

iii. Model Özellikleri

iii.1 Naive Bayes

Naive Bayes yönteminde doğrulama aşamasında en iyi model sonucu %81 F1 skorudur.

iii.2 Karar Ağacı

Karar ağacı yöntemi kullanılırken farklı ağaç derinlikleri denenmiştir. Ancak en iyi F1 skoru 3 derinlikle %77 elde edilmiştir. Ağaç derinliği 3'ten az olduğu durumda model underfit ediyorken 3'ü aştığı durumda ise overfit etmeye başlamaktadır. Özellikle ağaç derinliği 7'ye geldiği durumda eğitim ve denetleme verisi arasındali fark yaklaşık %20 olmaktadır.

iii.3 Derin Öğrenme

Derin öğrenme yöntemi kullanılırken veri miktarı az olduğu için derin bir mimari tasarlanmamaya özen gösterilmiştir. Modelin overfit etmesi engellenmesi ve stabil olmasını sağlamak amacıyla dropout kullanılmıştır. Ancak bu yöntemin en büyük problemi hata fonksiyonudur. Çünkü hata fonksiyonları veri dağılımı eşitken çok başarılı iken dengesiz dağılıma sahip verilerde modelin tek tarafa odaklanmasına sebep olur. Bunu engellemek amaçlı hata fonksiyonu iyileştirilmiştir. Bu sayede hata fonksiyonu yüksek orana sahip örneklerden daha az şey öğrenirken daha düşük orana sahip örneklerden daha fazla şey öğrenmektedir. Bu tamamen olarak başarıyı dengelemese de modelin daha yüksek başarıya sahip olması sağlanmıştır. Ayrıca hata fonksiyonununda direkt olarak bir örneğin 1 veya 0 olarak etiketlenmesi yerine 0.9-0.1 şeklinde etiketlenmesi modelin aşırı özgüvenli olmasını engelleyerek daha stabil bir model ve daha yüksek başarıya sahip olmasını sağlamaktadır. Açılan iyleştirilmeler yapılarak overfit eden derin öğrenme modeli %78'lik bir F1 skoruna sahiptir. Bu kadar iyileştirmeye karşı F1 skorunun diğer yöntemlere kıyasla benzer olmasının temel sebebi, diğer yöntemlerin derin öğrenme modeline kıyasla düşük orana sahip örneklerle daha az ilgilenmesinden kaynaklıdır.

iii.4 KNN

KNN yönteminde en yüksek başarı K= 3 iken elde edilmiştir. Karar ağaçlarında yaşananın benzeri yaşanmıştır. Model K < 3 için underfit ederken K > 3 için overfit etmiştir. Elde edilen en yüksek başarı %74'lük F1 skor olmuştur.

iv. Konfigürasyon Sonuç Yorumu

iv.1 Naive Bayes Karmaşıklık Matrisi

		Tahmin	
		Kayıp Değil	Kayıp
Gerçek	Kayıp Değil	237	3
	Kayıp	30	30

iv.2 Karar Ağacı Karmaşıklık Matrisi

		Tahmin	
		Kayıp Değil	Kayıp
Gerçek	Kayıp Değil	238	2
	Kayıp	46	14

iv.3 Derin Öğrenme Karmaşıklık Matrisi

		Tahmin	
		Kayıp Değil	Kayıp
Gerçek	Kayıp Değil	175	65
	Kayıp	41	19

iv.4 KNN Karmaşıklık Matrisi

		Tahmin	
		Kayıp Değil	Kayıp
Gerçek	Kayıp Değil	226	14
	Kayıp	51	9

Karmaşıklık matrislerinde de görülebileceği üzere modeller kayıp değil sınıfını bir sıkıntı tahmin etmekte sıkıntı çekmemektedir. Bunun temel sebebi sınıfın örneklerinin çok fazla olmasıdır. Derin öğrenme modelinin sonuçlarına bakıldığında ilk dikkat çeken şey diğer modellere kıyasla kayıp değil sınıfını bulmakta zorlandığıdır. Bunun temel sebebi hata fonksiyonunda yapılan geliştirmelerdir. Bu sayede model daha fazla tahmininde kayıp sınıfıyla etiketleme eğilimde oluyor. Karmaşıklık matrislerine bakıldığı zaman görülen şey Naive Bayes modelinin diğer modellere kıyasla çok yüksek başarısıdır.

v. Test Sonuçları Yorumu

Yöntem	Accuracy	Recall	Precision	F1
Naive Bayes	0.89	0.89	0.89	0.88
Karar Ağacı	0.84	0.85	0.84	0.80
Derin Öğrenme	0.84	0.87	0.84	0.85
KNN	0.82	0.80	0.82	0.79

vi. En Başarılı Öğrenme Modeli

Sonuçlar tablosunda da görüldüğü üzere Naive Bayes bütün başarı metriklerinde en başarılır model olmuştur.

V. Sonuç

Kullanılan modeller içerisinde derin öğrenme ve karar ağaçları uyum yetenekleri sayesinde en başarılı iki model olacağını düşünülüyordu. Derin öğrenme için doğru karar verilmiş olsa da karar ağaçları beklenen başarıyı elde edemedi. Naive Bayes yöntemi ise test verisinde beklenmeyen bir başarıyı elde etmiştir. Özellikle doğrulama aşamasında diğer modellere kıyasla çok düşük bir skor elde ettiği düşünüldüğünde şaşırtıcı bir başarıdır. KNN yöntemi probleme en uygun yaklaşım olduğu düşünülmemiştir bu yüzden sonuncu olacağı tahmin edilmiştir. Bu problem özelinde bakıldığı takdirde Naive Bayes hem başarı hem de hız ve basitlik açısından diğer modellere kıyasla çok önde olduğundan ötürü bu probleme en uygun model olarak gözükmektedir.