İkinci El Otomobil Fiyat Tahmini

Buğra Soysal, Oğuzhan Erten
Bilgisayar Mühendisliği
Kocaeli Üniversitesi
İzmit, Türkiye
bugrasoysal@outlook.com, ouerten@yahoo.com

Bu projede, ikinci el otomobil verisi, decision tree, random forest, neural network gibi makine öğrenmesi modelleri ile eğitilmiş, test verisi için fiyat tahmini yapılmış ve kullanılan modellerin hata oranları karşılaştırılmıştır.

Anahtar Kelimeler – Otomobil, eğitim, test, regresyon, tahmin.

1. TANIM

İkinci el otomobil pazarında, arabaların fiyat tahmini önemli bir problemdir. Bir satıcı, aracı her zaman en yüksek fiyata satmaya çalışır ve otomobilin fiyatı, müşterilerin satın alma seçiminde çok önemli bir faktördür. Bu nedenle, fiyatların etkin bir şekilde tahmin edilmesi, ikinci el otomobil sektöründe hem satıcı hem de alıcılar için çok kullanışlı bir kaynak olacaktır. Genel olarak, kullanılmış bir otomobilin fiyatının belirlenmesi markasına, modeline, üretim yılına ve diğer birçok faktöre bağlıdır. Bu, ikinci el bir arabanın fiyatının tespitini çok karmaşık hale getirir. O yüzden, makine öğrenme yöntemleri, belirlenen faktörler ve ikinci el otomobillerin fiyatları arasındaki gizli ilişkiyi belirlemek için uygun olacaktır. Model oluşturulduktan sonra, diğer benzer sorunlara da bazı referanslar sağlayabilir. Örneğin, mevcut işlenmiş bir arabanın yıpranmasının değerlendirilmesi veya bir arabanın değiştirilmesi gerekip gerekmediğinin tahmini.

1.1. PROBLEM TANIMI

Bu problemin çözümü, Ebay'den alınan veri kümesi için oluşturulmuştur. Veri seti, Mart 2016-Nisan 2016 tarihleri arasında Ebay'de yayınlanan kullanılmış araba ilanlarının tüm detaylarını sunmaktadır. Detayların birçoğu fiyat tespiti ile bağlantılıdır. Ayrıca, bu detaylar, nitelendirilebilir veya işlenebilirdir. Fiyat bir sayı biçiminde olduğu için, tahmin ve gerçek değeri karşılaştırmanın yolu, bu iki miktar arasındaki farkın mutlak değerini, yani mutlak hatayı almak olabilir. Bu bilgilere dayanarak makine öğrenmesi yöntemiyle ikinci el otomobil fiyatlarının tahmininin yapılması beklenmektedir.

1.2. HATANIN BULUNMASI

Arabanın fiyatı sayı olarak belirtilmiştir. İlanda verilen gerçek fiyat, tahmin fiyatıyla doğrudan karşılaştırılabilir. Model, iki fiyat arasındaki ortalama mutlak hata kullanılarak değerlendirilebilir, bunlar:

$$E_{ae} = \frac{\sum_{i=1}^{N} |P_p - P_a|}{N}$$

Eae ortalama mutlak hata ise, Pp otomobil fiyatının tahmini, Pa ilanı verilen otomobilin fiyatıdır.

2. ANALİZ

2.1 VERİNİN AÇIKLANMASI

Veri kümesi, kaggle'daki ikinci el otomobil veritabanından çıkarılmıştır. 371528 satırndan oluşan bu verinin her satırı, ikinci el otomobil ilanı hakkında aşağıdaki bilgileri içerir;

- dateCrawled: İlanın ilk kez tarandığı tarih. Diğer tüm alanlar bu tarihteki ilandan gelir.
- name: Otomobilin ismi, marka, model vb. bilgileri içerebilir.
- seller: Satıcı.
- offerType: Teklif türü tüm veriler için aynıdır, bu nedenle bu alan kullanışsızdır.
- price: Arabanın ilandaki fiyatı. Bu tahmin edilmesi gereken veridir. Bu yuzden bu alan verilerden kaldırılacak.
- abtest: Bir ebay-intern değişkeni.
- vehicleType: Sekiz araç kategorisinden biri.
- yearOfRegistration: Aracın ilk kez kayıtlı olduğu yıl.
- gearbox: Otomobilin vites kutusunun tipi, manuel veya otomatik.
- powerPS: Otomobilin PS cinsinden gücü.
- model: Otomobilin modeli.
- kilometer: Otomobilin kat ettiği kilometre sayısı.
- monthOfRegistration: Aracın ilk kaydedildiği ay.
- fuelType: Bir arabanın yedi yakıt kategorisinden biri.
- brand: Bir araba markası.
- notRepairedDamage: Henüz onarılmayan bir hasar varsa.
- dateCreated: Ebay'daki ilanın oluşturulduğu veriler.
- nrOfPictures: Tarayıcıda bir hata olduğundan, bu alandaki tüm sayılar 0'dır.
- alan işe yaramaz.
- postalCode: Almanya'da arabanın bulunduğu yer.
- lastSeenOnline: Tarayıcının bu reklamı en son çevrimiçi olarak gördüğü süre.

Veri kümesi, verilen csv dosyası kullanılarak oluşturuldu, bilgi olmayan alanlar boş bir string değeri ile dolduruldu. Veri setindeki 7 örnek aşağıda gösterilmiştir:

	0	1
dateCrawled	2016-03-24 11:52:17	2016-03-24 10:58:45
name	Golf_3_1.6	A5_Sportback_2.7_Tdi
seller	privat	privat
offerType	Angebot	Angebot
price	480	18300
abtest	test	test
vehicleType		coupe
yearOfRegistration	1993	2011
gearbox	manuell	manuell
powerPS	0	190
model	golf	
kilometer	150000	125000
monthOfRegistration	0	5
fuelType	benzin	diesel
brand	volkswagen	audi
notRepairedDamage		ja
dateCreated	2016-03-24 00:00:00	2016-03-24 00:00:00
nrOfPictures	0	0
postalCode	70435	66954
lastSeen	2016-04-07 03:16:57	2016-04-07 01:46:50

Şekil. 1 Veri kümesinde ilk dört örnek.

İlk gözlem için sonuçları:

1. 'Abtest' alanının anlamını söylemek zor. Bu nedenle, bu alan ve fiyat tespiti arasındaki ilişkiyi bulmak da zorlaşıyor.
2. 'İsim' alanı temel olarak bir otomobilin marka ve model bilgilerini içerir, böylece bu alan 'marka' ve 'model' alanı ile ilgili olarak kapatılır.
3. 'seller', 'offerType' ve 'nrOfPictures' alanı tüm kayıtlar için aynı değerde olabilir. Bu doğruysa, o zaman bu alan fiyat tespiti için işe yaramaz.
4. 'vehicleType', 'model', 'gearbox', 'feulType' ve 'notRepairedDamage' alanları içinde değeri olmayan bazı kayıtlara sahip, yani değer olarak boş bir string var. Bu nedenle, tüm potansiyel faydalı alanlar, alanda boş değerleri olan kayıtların olup olmadığını görmek için kontrol edilecektir.
5. 'powerPS' ve 'monthOfRegistration' alanlarında makul olmayan değerler var. Yani sıfır olan değerler. Bu nedenle, tüm sayısal alanlar makul olmayan değerler olup olmadığını görmek için kontrol edilecektir.
6. 'monthOfRegistration' ve 'postalCode' sayısal biçimlerde görünmesine rağmen, fiyat belirlemesi ile ilgili olarak belirleyici anlamlar taşımamaktadır. Bu yüzden metin alanlarıyla aynı tutumda bulunulması gerekir.

7. 'formCrawled', 'dateCreated' ve 'lastSeen' tarih alanları metin biçiminde görünür. Ancak, değer ile sabit seviye tarihi arasındaki gün sayısını hesaplayarak sayısal formlara dönüştürülebilirler.

Daha önce belirtildiği gibi, 'model' alanındaki eksik değer, 'name' alanından çıkarılan bilgilerle doldurulur. 'VehicleType', 'gearbox', 'fuelType' ve 'notRepairedDamage' gibi diğer alanlardaki eksik değerler için, mevcut bilgileri kullanarak doldurmanın açıkcası bir yolu yoktur. Böylece bu dört alandaki boş değerleri olan kayıtlar atılacaktır.

Sayısal alanlar için bazı istatistiksel bilgiler aşağıda verilmiştir:

	price	yearOfRegistration	powerPS	kilometer
count	3.715280e+05	371528.000000	371528.000000	371528.000000
mean	1.729514e+04	2004.577997	115.549477	125618.688228
std	3.587954e+06	92.866598	192.139578	40112.337051
min	0.000000e+00	1000.000000	0.000000	5000.000000
25%	1.150000e+03	1999.000000	70.000000	125000.000000
50%	2.950000e+03	2003.000000	105.000000	150000.000000
75%	7.200000e+03	2008.000000	150.000000	150000.000000
max	2.147484e+09	9999.000000	20000.000000	150000.000000

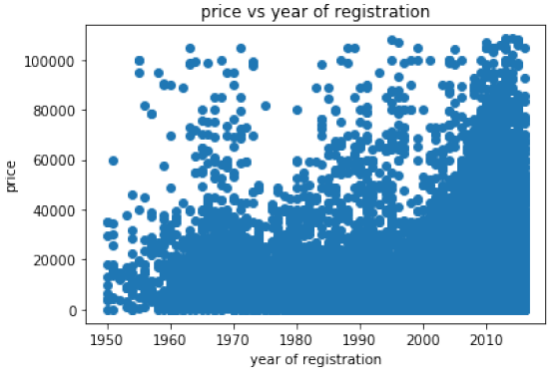
Şekil. 2 Sayısal alanların istatistikleri.

1. 'price', 'yearOfRegistration', 'powerPS' ve 'monthOfRegistration' alanında geçerli olmayan, 0 değerleri görünüyor.
2. 'yearOfRegistration' alanında, '1000' ve '9999' gibi geçerli olmayan erken ve geç yıllar var.
3. Fiyat verilerinin çoğu binlerce dolar aralığında yer alırken, 2.1e9 gibi geçerli olmayan yüksek fiyatlar vardır.
4. 'powerPS' alanında '20000' gibi geçerli olmayan yüksek değerler görünmektedir.
5. 'kilometre' alanındaki arabaların yarısından fazlası 150000 değerine sahip. Bu, kilometredeki verinin yanlış olabileceğini anlamına gelir. Çünkü ebay'da, kullanılmış bir otomobilin kilometresi için doldurulabilecek en büyük değer 150000'dir.

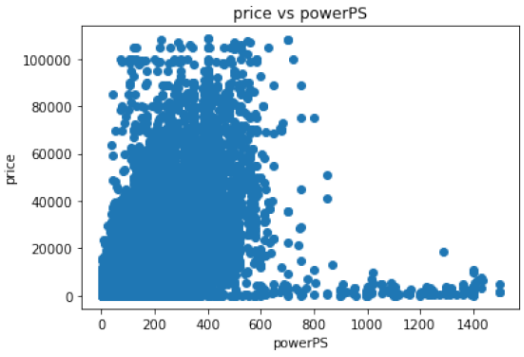
Yukarıdaki gözlemlere dayanarak, geçerli olmayan değerlere sahip kayıtlar atılacaktır. 'yearOfRegistration' ve 'powerPS' alanlarındaki değerler geçerli değerleri sınıflandırmayı sağlar. 'yearOfRegistration', değeri araba ilanının yeni bir araç yerine eski bir araç olması için [1950,2016] aralığında olması gerekir. 'PowerPS' sayısı, bir otomobilin bilinen en yüksek powerPS'si olan 1500'den az olmalıdır. Fiyat verilerinin çok fazla aykırı olduğunu kısa bir incleme sonrası anlayabiliyoruz, bu nedenle en yüksek %0,1 fiyatlar verilerden çıkartılacaktır. Kalan kayıtlar daha makul bir maksimum değere sahiptir ve öncekinden daha az aykırı değer içermektedir.

2.2. GÖRSELLEŞTİRME

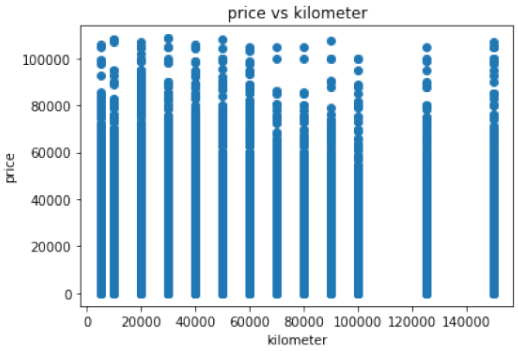
Geçerli değerlere sahip veriler için, yani sırasıyla 'yearofRegistration', 'powerPS' ve 'kilometer' verilerini ‘price’ verisine göre dağılımlarını yapalım. Aşağıdaki grafiklerden, sayısal özellikler ile fiyat verileri arasında belirgin bir doğrusal ilişki olmadığı görülmektedir. Dolayısıyla, basit bir doğrusal modelin, bu veri setindeki kullanılmış otomobiller için kesin fiyat tahminleri yapma olasılığı daha düşüktür.



Şekil. 3 Fiyat ve kayıt yılı karşılaştırılması



Şekil. 4 Fiyat ve beygir gücünün karşılaştırılması.



Şekil. 5 Fiyat ve kilometre karşılaştırılması.

2.3. ALGORİTMA VE ÇÖZÜM TEKNİKLERİ

Bu problem için, naive bayes, decision tree, linear regression, random forest modeli ve çok katmanlı bir neural network modeli, veri setinden işlenmiş verilere dayanarak eğitilecektir. Giriş özellikleri ile hedef değişken arasındaki doğrusal olmayan bir ilişki vardır. Her katmana eklenen aktivasyon fonksiyonları sayesinde, çok katmanlı bir neural network modeli, girişler ve çıkışlar arasındaki doğrusal olmayan ilişkiyi temsil edebilir.

Orijinal veri setinde, girdiye özelliklerini üretmek için ön işlemler yapılacak ve fiyat verileri hedef değişken olarak alınacaktır. Girdi özellikleri ve fiyat verileri hazır olduklarında, işlenen veri setinden rastgele seçilerek train seti, validation seti ve test seti 6:2:2 oranında bölünecektir.

İlk önce, basit bir ortalama değer tahmini için naive bayes modeli oluşturulacaktır. Daha sonra, decision tree, linear ve random forest gibi regresyon modelleri veri setimiz için test edilecektir. Doğrusal bir modelin bu problem üzerinde nasıl bir performans sergilediğini görmek için iç katman ve aktivasyon fonksiyonları olmayan bir neural network modeli eğitilecektir. Doğrusal olmayan bir modelin performansı iyileştirip iyileştirmediğini görmek için aktivasyon işlevine sahip tek katmanlı bir neural network modeli test edilecektir. Sonra son bir model eğitilecek ve parametreleri ayarlanacaktır. Son neural network modelinde ayarlanması gerekenler katman sayısı, her katmandaki düğüm sayısı, aktivasyon işlevlerinin türü ve dropout oranı’dır.

3. YÖNTEM

3.1 VERİ ÖN İŞLEME

Orijinal veri setine aşağıdaki işlemler yapılacaktır:

1. 'nrOfPictures' alanındaki tüm değerler 0 olduğu için atılacak.
2. 'seller' veya 'offerType' alanlarındaki azınlık değerine sahip kayıtlar atılacak.

3. 'Gearbox', 'vehicleType', 'fuelType' ve 'notRepairedDamage' alanlarındaki boş değerli kayıtlar silinecek.
4. 'Model' alanındaki boş değeri olan kayıtlar, 'name' alanından bilgi çıkarılarak yeniden düzenlenecek. Daha sonra sabit olmayan kayıtlar silinecek ve 'name' alanı atılacak. 'brand' ve 'model' alanlarındaki stringler, farklı durumlardaki kopyaları ortadan kaldırmak için büyük harfe dönüştürülecek.
5. 'price', 'powerPS', 'yearOfRegistration' ve 'monthOfRegistration' alanlarındaki geçerli olmayan değerlere sahip kayıtlar ve %0,1'lik fiyatların en yüksek olduğu kayıtlar silinecek.
6. 'dateCrawled', 'dateCreated' ve 'lastSeen' alanları, ilk önce datetime verilerine, ardından datetime verileri tarih içindeki günleri hesaplayarak sayısal verilere dönüştürülecek.
7. 'postalCode' alanındaki değerlerin son üç hanesi silinecek, ardından kalan bir veya iki hane metin olarak değerlendirilecektir.
8. 'price' alanı hedef değişken verileri olarak gösterilecektir.
9. 'abtest', 'vehicleType', 'gearbox', 'model', 'fuelType', 'brand', 'notRepairedDamage', 'monthOfRegistration' ve üzerinde değişiklik yapılmış olan 'postalCode' alanları metin alanları olarak değerlendirilecektir.
10. 'powerPS', 'yearOfRegistration' ve 'kilometer' alanları sayısal alanlar olarak ele alınacak ve min-max ölçekleyici tarafından normalizasyon uygulanacaktır.

3.2. UYGULAMA

Problem çözümü için ilk önce naïve bayes modeli uygulanmıştır. Naïve bayes modeli tahmin için yaklaşık 5439,84 puan almıştır, bu durum öngörülen fiyat ile gerçek fiyat arasındaki farkın ortalama 5439,84 $ olduğu anlamına gelmektedir. Yani naïve modelinin pek iyi performans gösterdiği söylenemez. Daha sonra, bir linear regression modeli oluşturulmuştur, oluşturulan bu model ile tahmin yapıldı ancak ortalama mutlak değer 460000, skor ise -2 gibi anlamsız bir değer ortaya çıkmıştır. Bu, modelin problemimiz için uygun olmadığı anlamına gelmektedir.

Bir sonraki modelimiz ise Decision Tree (Karar ağaçları). Bu model genelde non-linear veriler üzerinde daha iyi bir sonuç veriyor diyebiliriz. Kullandığımız veri seti üzerinde denediğimizde ise eğitim sürecinin çok hızlı olduğunu söyleyebiliriz. Modelimizin, test verisi için tahmin aşamasında ortalama mutlak hatası 1586,30, skoru ise 0,80'dir. Naive bayes ve linear regression modellerine göre çok daha güzel puanlar aldığını görebiliyoruz.

Oluşuturduğumuz üçüncü model bir ensemble modeli olan Random Forest. Gözlemlerimize, göre bu model doğrusal olmayan verileri linear regression modeline göre verileri daha kolay ve hızlı bir şekilde eğitiyor. Eğitim sonrasında test verisi için tahmine geçildiğinde ortalama mutlak hata yaklaşık 1279,90, skor ise 0,87 çıkmıştır. Elde ettiğimiz bu değerler şu ana kadar denediğimiz diğer modellere göre veri setimize en uygun modelin Random Forest olduğunu gösteriyor. %87'lik bir başarı sağlasak da diğer modelleri denemeye devam ediyoruz.

Deneyeceğimiz son model neural network. Bu neural network modelini geliştirmek için keras paketi kullanılmıştır. İlk olarak, lineer bir neural network modelinin bu problem karşısında çözüm sunup sunamayacağını test etmek için eğitilmiştir. Bu model, sadece 128 düğümden ve lineer aktivasyon fonksiyonuna sahip tek bir iç katman içermektedir. Yani modelde toplam 54529 parametre vardır. Optimizer Adam, kayıp fonksiyonu mean absoulte error'dur. Eğitim modeli 20 epoch'ludur ve batch boyutu 500'dür. 11 test kümesi için 3569.53 skor almaktadır.

Yani lineer neural network modeli, naive bayes modeline göre çok daha iyi fakat Random Forest modeline göre hala düşük fiyat tahminleri yapmaktadır. Bu da bu problem için neural network yöntemlerinin kullanılabileceğini ama daha çok geliştirilmesi gerektiğini gösterir.

Ancak, belirtilen araç özelliklerinin işlenmeden kullanılması durumlarında lineer model, problemi çözmekte çok başarılı olmayacaktır.

Belirtilen özellikler arasındaki lineer olmayan ilişkileri doğru bir şekilde ele almak zordur, tecrübe ve çok sayıda deneme gerektirir. Daha pratik bir yol, lineer olmayan aktivasyon fonksiyonu içeren birçok katmanlı nöron ağı kullanmaktır. Bu model, belirtilen özellikler ve fiyat arasındaki lineer olmayan ilişkileri otomatik olarak bulacaktır.

Bu nedenle, problemi çözmek için iki iç katmanlı ve lineer olmayan aktivasyon fonksiyonlu bir nöron ağı oluşturulup eğitim ve test için kullanılmıştır. Modelin ilk katmanında 128 düğüm, ikinci katmanında 32 düğüm, her iki katmanında da aktivasyon fonksiyonu olarak relu fonksiyonuı vardır. Modelin eğitilmesi için 58561 parametre vardır. Optimizer Adam fonksiyonudur, kayıp fonksiyonu mean absolute error'dur, model 20 epoch ile eğitilmiştir ve batch boyutu 500 dür. Test kümesi için 1784.62 skoru almaktadır. Tek katmanlı ve relu aktivasyon fonksiyonlu basit bir nöron ağı, tahmin fiyatlarını hatayı neredeyse yarı yarıya düşürmektedir. Bu durum, lineer olmayan özelliklerin sayısını arttırdıkça performansın arttığını gösterir. Fakat hala diğer modellere göre daha düşük tahmin skoru vermektedir. Bu yüzden daha çok katmanlı neural network denendi.

Bu neural network modelinin parametrelerini ayarlamak için aşağıda verilen deneyler yapılmıştır:

1. Ne kadar katmanın en iyi puanı verdiğini belirlemek için modele 2,3 ve 4 katman sırasıyla eklendi.
2. Her katmanda kaç tane düğüm olması gerektiğini belirlemek için {1024, 512, 256,128, 64, 32, 16, 8} kümesinde testler yapıldı.
3. Etkinleştirme işlevlerinin türü için, {'relu', 'sigmoid', 'softmax', 'tanh'} kümesinden biri seçildı.
4. Her katmanın içindeki dropout oranı da ayarlandı. {0, 0.2, 0.4, 0.5} değerleri bırakma oranı olarak test edildi ve en uygun değer seçildi.

Son model, kabul edilebilir bir sonuç sağlayan aşağıdaki dört iç katmana sahiptir:

1. 128 düğümlü ilk iç katmanda, aktivasyon fonksiyonu relu'dur ve dropout yokur.
2. 64 düğümlü ikinci doğrusal katmanda, aktivasyon fonksiyonu linear'dir ve dropout yok.
3. 32 düğümlü üçüncü iç katmanda, aktivasyon fonksiyonu relu'dur ve dropout yoktur.
4. 8 düğümlü dördüncü iç katmanda, aktivasyon fonksiyonu linear'dir ve dropout yoktur.

Model, eğitilecek 65009 parametreye sahiptir. Optimizer'ı Adam ve loss function'ı mean absolute error fonksiyonudur. Model, 20 epoch ve 500 batch boyutu ile eğitilmiştir. Son model test seti için 1361,26 puan almıştır. Bu puan her çalıştırma için farklı değerler verebilir. Fakat yine de Random Forest modelinden daha düşük bir puan almıştır. Son modelimizin elimizdeki en iyi ikinci model olduğunu söyleyebiliriz.

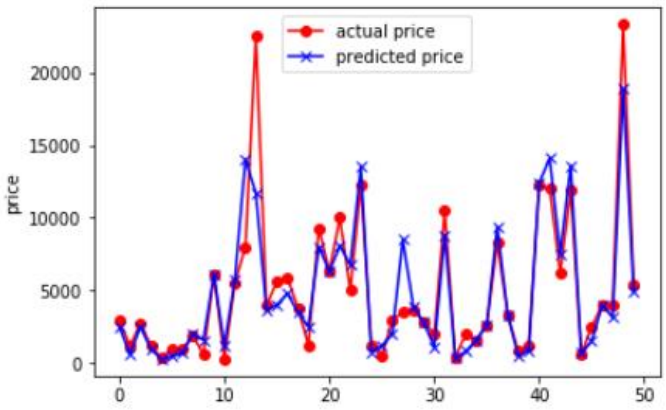
4. SONUÇ

4.1 MODEL DEĞERLENDİRME

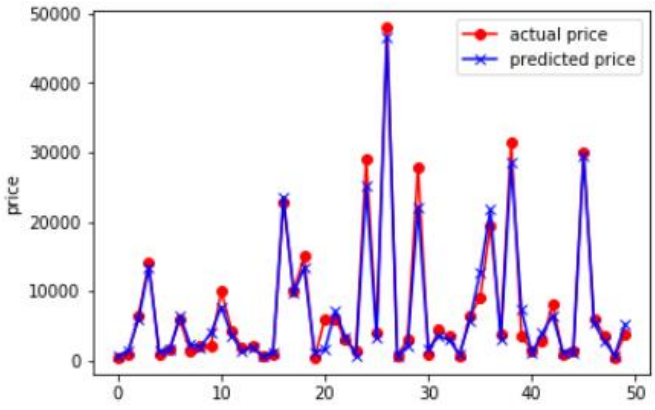
Oluşturduğumuz son model her bir katmanda yeterli sayıda düğüme ve dört iç katmana sahip neural network, ondan önce oluşturduğumuz diğer iki model ise decision tree ve random forest'tır. Bu modeller ikinci el otomobil ilanı için işlenmiş verileri alır ve otomobilin fiyatını tahmin eder. Test setindeki verilerin performansına dayanarak, tahmini fiyat ile gerçek fiyat karşılaştırıldığında yapay sinir ağı için yaklaşık 1360, decision tree için 1586, random forest için 1278 dolarlık ortalama mutlak hata ortaya çıkmaktadır. Bu değerler kabul edilebilir değerlerdir.

Ortalama mutlak hatayı metrik olarak kullanan naive bayes modeli yaklaşık 5440 puan alırken, son model yaklaşık 1360, random forest modeli 1278, decision modeli ise 1586 puan alımıştır. En iyi sonuç veren model Random Forest olmuştur. Bu modeller, naive bayes modeline göre çok daha az mutlak hata ile tahminlerde bulunur. Bu modellerin sonucunu doğrulamak için, rastgele test setinden 50 örnek alırız, gerçek fiyat ve fiyat tahminini karşılaştırırız, örneklerin çoğu için model iyi tahminler yapmıştır. Fakat, fiyatların 10000 ABD dolarından yüksek olduğu kayıtlar için, tahmin fiyatı ile gerçek fiyat arasında büyük hata yapan bazı örnekler vardır. Bu olaya iki olası faktör neden olabilir:

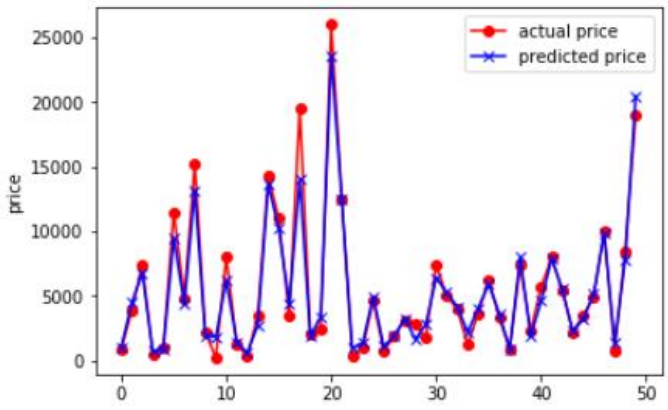
- Verilerin %78'inden fazlasının fiyatı 10000 ABD dolarından daha azdır. Bu yüzden fiyatları 10000 ABD dolarından yüksek olan otomobiller için eğitim verileri küçük boyutlardadır. Gerçek fiyatların yüksek olduğu otomobillere ilişkin yanlış tahmin, modelin eğitimi için yetersiz verilerden kaynaklanıyor olabilir.
- İlanı verilen ikinci el otomobilin fiyatı o arabanın tam değeri değildir. İlanı veren kişi, kişisel faktörlerinden etkilendiğinden, bu fiyat yalnızca ikinci el otomobilin ayrıntılarıyla belirlenmez, aynı zamanda ilanı veren kişi ile de ilgilidir. Ancak, veri kümesindeki ilanı veren kişi hakkında hiçbir bilgi bulunmamaktadır. Bu nedenle, arabalara ilişkin yanlış tahminler veri setinden gelen yetersiz bilgilerden de kaynaklanıyor olabilir.



Şekil. 6 Decision Tree kullanılarak gerçek ve tahmini fiyatların karşılaştırılması.



Şekil. 7 Neural Network kullanılarak gerçek ve tahmini fiyatların karşılaştırılması.



Şekil. 8 Random Forest kullanılarak gerçek ve tahmini fiyatların karşılaştırılması.

4.2 ÇÖZÜM AŞAMALARI

Bu projenin sorun çözüm süreci aşağıdaki gibidir:

1. Problem ile ilgili veri seti bulundu ve tanımlandı.
2. Veri seti gözlem, istatistiksel ve görselleştirme ile incelenmiştir.
3. Model tasarımları ve metrikleri araştırıldı ve tartışıldı.
4. Veri seti işlendi ve daha sonra train, validation ve test için girdi verilerine dönüştürüldü.
5. Naive bayes modeli oluşturuldu ve test edildi.
6. Decision Tree modeli oluşturuldu ve test edildi.
7. Linear Regression modeli oluşturuldu ve test edildi.
8. Random Forest modeli oluşturuldu ve test edildi.
9. Yapay sinir ağı modeli basit bir lineer modelden karmaşık bir lineer olmayan modele dönüştürüldü.
10. Son modelin parametreleri ayarlandı.
11. Modelin sonucu gözlemlendi. Modeli daha da geliştirmek için deneyler yapıldı.

10. aşamada parametrelerin ayarlanması çok zaman alır. İlk önce, mümkün olduğu kadar çok katman ve düğüm denendi, daha sonra overfitting durumunu azaltmak için daha az katman ve düğüm olarak değiştirildi.

Veri setindeki geçerli olmayan kayıtları temizlemek ve kalan verileri önceden işlemek için çok çaba sarf etmek gerekiyor. Geçerli olmayan verileri filtrelemek için bazı alanlar sayısal alanlara dönüştürülmüştür. Bu ayarlamalar modelin performansını etkileyecektir.

Veri setinde onarılamayacak bazı veriler olabilir. Bu yüzden, kapsamlı bilgiyle kesin ve doğru veri toplama süreci, gerçek bir problemi başarıyla çözmenin en önemli faktörlerindendir.

KAYNAKLAR

[1] Hüseyin Daştan (2016), "Türkiyedeki ikinci el otomobil fiyatlaırnı etkileyen faktörlerin hedonik fiyat modeli ile belirlenmesi", Gazi Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, Ankara, s.15-250.

[2] Ecer, F. (2013), "Türkiye'de 2. El Otomobil Fiyatlarının Tahmini ve Fiyat Belirleyicilerinin Tespiti", Anadolu Üniversitesi Sosyal Bilimler Dergisi, 13(4), s. 101-112.

[3] Boyel, S.E., Hogarty, T. F. (1975), "Pricing Behavior in the American Automobile Industry, 1957-71", The Journal of Industrial Economics, 24, s. 81-95.

[4] Will Koehrsen, "Improving the Random Forest in Python", Erişim Tarihi: 02.05.2019, https://towardsdatascience.com/improving-random-forest-in-python-part-1-893916666cd

[5] Alper, E., & Mumcu, A. S. (2000). "Türkiye'de otomobil talebinin tahmini". İstanbul: Boğaziçi Üniversitesi, Ekonomi Bölümü, Ekonomi ve Ekonometri Merkezi.

[6] Avinash Navlani, "Decision Tree Classification in Python", Erişim Tarihi: 01.05.2019, https://www.datacamp.com/community/tutorials/decision-tree-classification-python

[7] Asilkan, Ö., & Sezgin, I. (2009). "İkinci el otomobillerin gelecekteki fiyatlarının yapay sinir ağları ile tahmin edilmesi". Süleyman Demirel Üniversitesi İktisadi İdari Bilimler Fakültesi Dergisi, 14(2), 375-391.

[8] Hamza Erol, Pelin İyi (2008), "Çoklu lineer regresyonda en iyi model seçimi", Ç. Ü. Fen Bilimleri Enstitüsü İstatistik s. 50-55.

[9] Jason Brownlee, "Develop Your First Neural Network in Python With Keras Step-By-Step", Erişim Tarihi: 05.05.2019, https://machinelearningmastery.com/tutorial-first-neural-network-python-keras/