T.C. Yalova Üniversitesi
Mühendislik Fakültesi
Bilgisayar Mühendisliği Bölümü
Yapay Sinir Ağları Dersi
Proje Raporu

2. EL ARABA FİYAT TAHMİNİ

HAZIRLAYAN

Elif ARDA - 140101026

OCAK-2021

ÖZET

Günümüzde tüketicilerin en çok ilgi gösterdiği ürünler arasında ikinci el otomobiller yer almaktadır. Ekonomik koşullar göz önüne alındığında ikinci el otomobillerin geniş bir müşteri kitlesi bulunmaktadır. Geniş bir pazara sahip olan 2.el araba sektöründe doğru fiyatlandırma yapmak büyük önem taşımaktadır.

Bu çalışmada doğrusal olmayan zaman serilerinde Yapay Sini Ağları yönteminin geleneksel makine öğrenmesi yöntemlerinden daha üstün sonuç göstermesi sebebiyle Yapay Sinir Ağları yöntemi kullanılarak ikinci el araba fiyat tahmini yapılmıştır. Veriler için ABD'de 1990-2018 yılları arasında piyasa fiyatı ve bazı özelliklerle satılan yaklaşık 12.000 araba modelinin bulunduğu veri seti kullanılmıştır. Elde edilen sonuçlar incelenmiştir.

1.GİRİŞ

1.1. Proje Tanımı

2.el araba fiyatlarının yeni araba fiyatlarına göre daha değişken olduğu bilinmektedir. Bu çalışmanın amacı Yapay Sinir Ağı kullanarak 2.el araba fiyatlarında tahmin yapmaktır.

1.2. Veri Seti

Araba fiyatlarının birçok parametreye göre değişmesi sebebiyle mümkün olduğunca giriş sayısının fazla olduğu bir veri seti seçilmiştir. Seçilen veri seti içerisinde toplam 16 sütun ve 11914 satır bulunmaktadır.

Sütun	Açıklaması
Make	Markası
Model	Modeli
Year	Üretim Yılı
Engine Fuel Type	Motor Yakıt Tipi
Engine HP	Beygir Gücü
Engine Cylinders	Silindir
Transmission Type	Şanzıman Tipi
Driven Wheels	Çekiş Sistemi
Number of Doors	Kapı Sayısı
Market Category	Pazar Kategorisi
Vehicle Size	Araç Boyutu
Vehicle Style	Araç Stili
highway MPG	Şehir Dışı Yakıt Tüketimi
city MPG	Şehir İçi Yakıt Tüketimi
Popularity	Popülerliği
MSRP	Fiyatı

Giriş için kullanmak üzere 15 sütun seçilmiştir. Marka, motor yakıt tipi, şanzıman tipi, çekiş sistemi, araç boyutu ve araç stili kategorik verilerin olduğu sütunları modelin daha doğru çalışabilmesi için sayısal hale dönüştürmek için One-Hot Encoding işlemi yapıldı. Bu işlemle kategorik türde olan özniteliklere ait tüm değerler yeni bir öznitelik haline getirildi. Eğer örnek aynı kategorideki yeni özniteliğe sahipse değeri "1" aynı kategorideki diğer değişkenlerin ise değeri "0" oldu. Kategorik öznitelikleri ayırdıktan sonra veri seti toplam 89 giriş verisi ve 1 çıkış verisi olarak kullanıldı. Eksik öznitelik değerleri olan veri satırları, veri seti içinden temizlenerek son halinde toplam 8084 satır bulunmaktadır. Veri setine https://www.kaggle.com/CooperUnion/cardataset web adresinden ulaşılabilinir.

1.3. Veri Setinin Kullanıldığı Başka Projeler

Veri setinin alındığı web adresinden elde edilen bilgilere göre bu veri seti ile daha önce farklı makine öğrenmesi yöntemleri kullanılarak birçok fiyat tahmin için model oluşturulmuştur. Bunların dışında veri seti üzerinde analiz çalışmaları da bulunmaktadır.

1.4. Projenin Faydası

Pazar payının çok geniş olduğu bir sektörde doğru fiyatlandırma yapmak için ideal fiyatları hesaplamak büyük önem taşımaktadır. Alış verişin daha adil bir şekilde yapılması için araba özelliklerinin göz önünde bulundurulması gerekmektedir. Bu çalışmayla 2.el araba fiyat piyasasına katkıda bulunmak amaçlanmıştır.

2. YÖNTEMLER

2.1. Kullanılan Araçlar

Bu çalışmada Python (version 3.8.6) programlama dili ile Jupyter-Notebook kullanılmıştır. Kullanılan kütüphaneler ve versiyonları aşağıda belirtilmiştir.

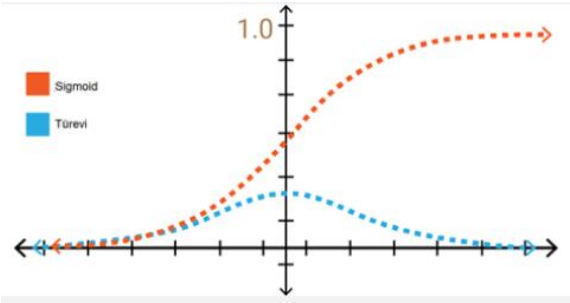
- Numpy 1.19
- Pandas 2.7.3
- Matplotlib 3.3.3
- Seaborn 0.11
- Scikit-learn 0.24
- Tensorflow 2.4.0
- Keras 2.4.3

2.2. Kullanılan Algoritma

Yapay Sinir Ağı modeli oluşturmak için Python'da yazılmış Keras kütüphanesi kullanılmıştır. Keras kütüphanesinde bulunan Sequential model ile yapay sinir ağı oluşturulmuştur. Keras, varsayılan olan geri besleme (Backpropagation) algoritması kullanır.

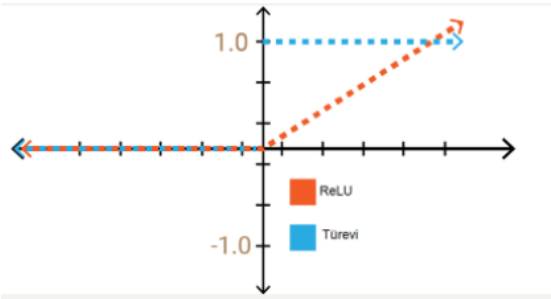
2.3. Kullanılan Aktivasyon Fonksiyonları

Aktivasyon fonksiyonları yapay sinir ağının doğrusal olmamasını sağlar. Aktivasyon fonksiyonları ileri besleme adımında fonksiyonun kendisi ile tahminler yapılmasını sağlarken geri besleme adımında fonksiyonun türevi ile öğrenmeye katkı sağlar. Aktivasyon fonksiyonları model üzerinde denenerek hangisinin daha iyi olduğu karar verilmiştir. Bu çalışmada ara katmanlarda "relu" ve çıkış katmanında "sigmoid" aktivasyon fonksiyonu kullanılmıştır.



Sigmoid Fonksiyonu (0,1)

$$f(x) = \sigma(x) = \frac{1}{1+e^{-x}}$$

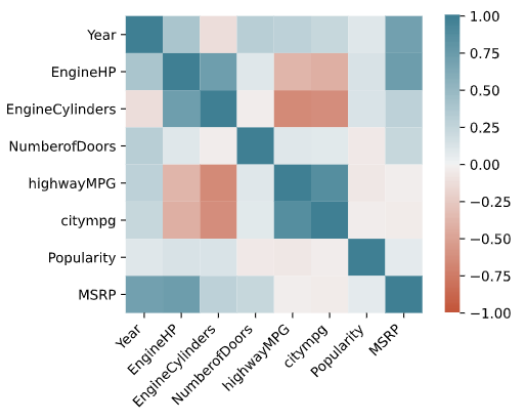


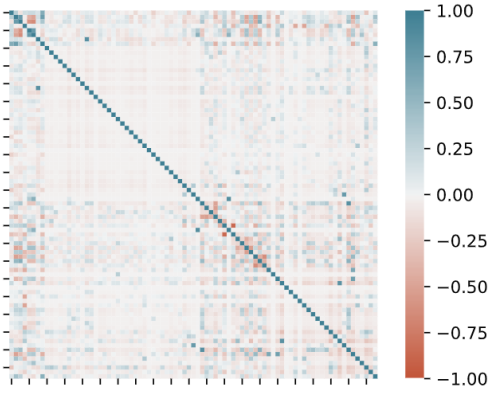
ReLU fonksiyonu [0,∞)

$$f(x) = \begin{cases} 0 \; için \; x < 0 \\ x \; için \; x \geq 0 \end{cases}$$

2.4. Korelasyon Matrisi

Korelasyon matrisi özniteliklerin birbiriyle olan ilişkilerini göstermektedir. Korelasyon matrisinde 2 özniteliğin arasındaki ilişki için -1 ile 1 arasında değer verilir. Veri seti üzerinde işlem yapılmadan önce 15 sütunlu ve veri setinin son halindeki 81 giriş değeri belirlenen öznitelikler arasındaki ilişkinin gösterildiği korelasyon matrisleri aşağıda verilmiştir. Grafik üzerinde mavi kareler 1'e yaklaştıkça pozitif korelasyonları, kırmızı kareler -1'e yaklaştıkça negatif korelasyonları göstermektedir. Eğer beyaz kareler yani korelasyon 0'a yakın ise veriler arasında lineer ilişki olmadığı sonucu görülmüştür.





2.5. Hata Matrisi (Confusion Matrix)

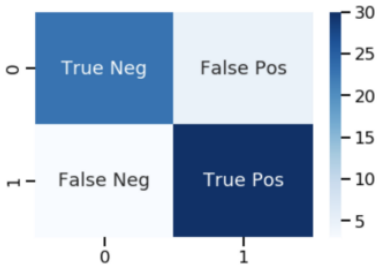
Sınıflandırma modellerinin performansını değerlendirmek için çıktı değerine ait tahminlerin ve gerçek değerlerin karşılaştırıldığı bir matristir.

TP (True Positive - Doğru Pozitif) : Pozitif tahmin doğru

FP (False Positive - Yanlış Pozitif) : Pozitif tahmin yanlış

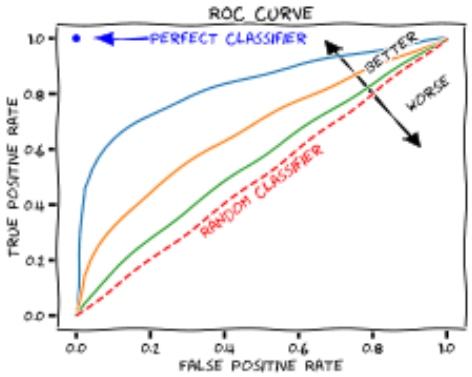
TN (True Negative - Doğru Negatif) : Negatif tahmin doğru

FN (False Negative - Yanlış Negatif) : Negatif tahmin yanlış



2.6. AUC - ROC Eğrisi

ROC eğrisine bakılarak model performansı hakkında değerlendirme yapılabilir. İşlem Karakteristik Eğrisi (Receiver Operating Curve), eşik değeri geliştirilerek Doğru Pozitif Oranı - Yanlış Pozitif Oranı grafiğidir. ROC eğrisi Yanlış Pozitife yaklaştıkça başarı seviyesi düşer. Sistem başarısı ROC eğrisinin altında kalan alan ifade edilir. Bu alan değeri ne kadar büyükse sistemin güvenirlik değeri de o kadar yüksek olur.



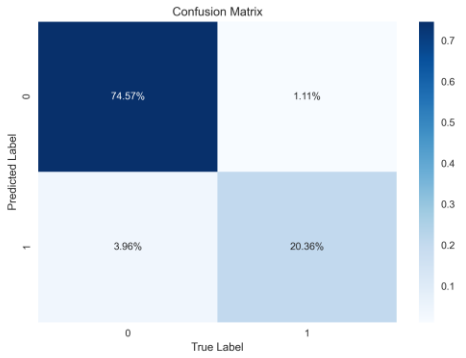
3. UYGULAMA DETAYLARI

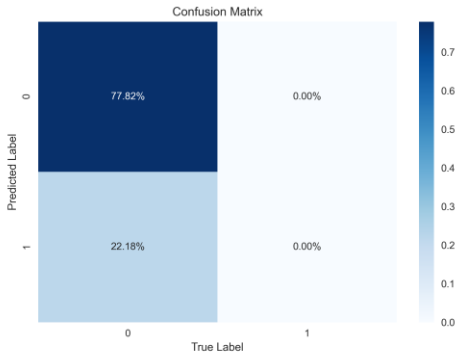
Yapay Sinir Ağı kullanarak 2.el araba fiyat tahmin modeli için önce csv formatındaki veri seti okundu. Veri seti üzerinde boş verilerin temizlenmesi, sütun ayırma ve tip dönüşüm işlemleri yapıldı. Bu modelin sınıflandırma algoritması olabilmesi ve düzgün bir başarı elde edilebilmesi için çıkış değeri olan fiyat sütunu sınırlandırıldı. Veri seti giriş ve çıkış değerlerine ayrıldı. Giriş ve çıkış değerlerini normal hale getirerek performansı arttırmak için MinMax Scaling ile verilerin 0 - 1 arasında değer alması sağlandı. Veri seti %70 eğitim verisi ve %30 test verisi olarak ayrıldı. Yapay Sinir Ağı modeli Keras kütüphanesinin Sequential modeli ile oluşturuldu. Giriş, ara katmanlar ve Overfitting'i engellemek için Dropout katmanı modele eklendi. Giriş ve ara katmanlarda düğüm yani nöron kullanıldı. Ara katmanlarda relu ve çıkış katmanında sigmoid aktivasyon fonksiyonu kullanıldı. Öğrenme oranını kontrol etmek için optimizer olarak "adam" ve "stochastic gradient descent" algoritmaları kullanıldı. Farklı öğrenme kat sayıları denendi. Hata fonksiyonu olarak "mean absolute error" seçildi. Modeli eğitim sürecinde farklı epoch,batch_size ve momentum değerleri denendi.

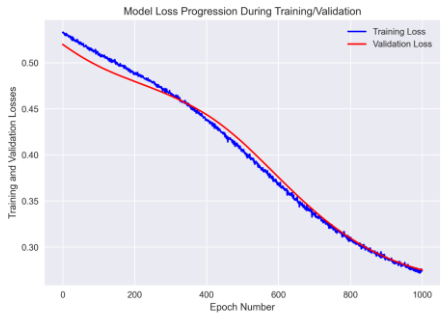
4. SONUÇLAR

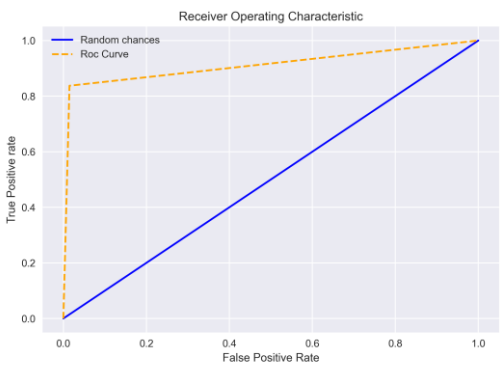
Optimizer	Adam
Learning Rate	0.001
Epochs	1000
Batch Size	1024
Auc Score	%91
Test Accuracy Score	%94

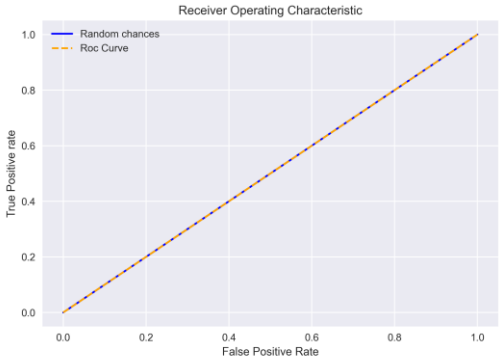
Optimizer	Adam
Learning Rate	0.00001
Epochs	1000
Batch Size	1024
Auc Score	%50
Test Accuracy Score	%77







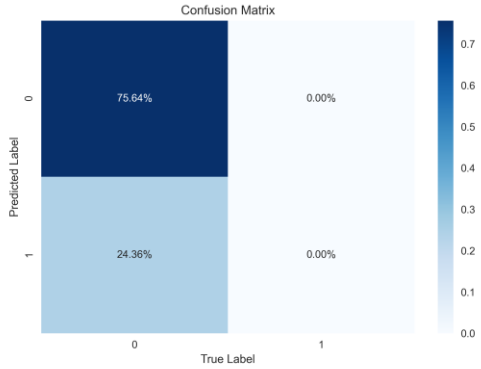


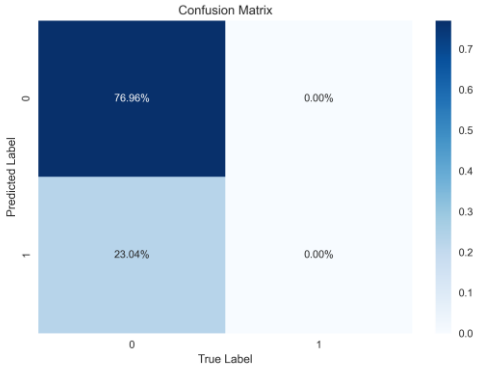


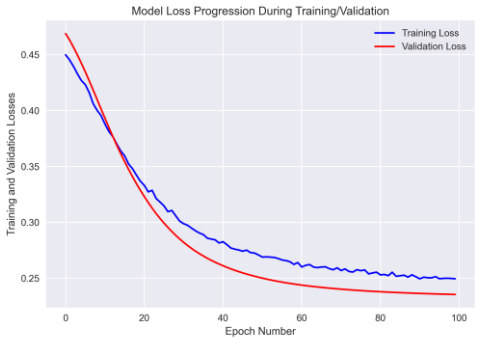
Öğrenme oranını 0.001'den 0.0001'e düşürülmesi tahmin başarasını düşürmüştür. Hedeften uzaklaşılmıştır. Öğrenme oranın 0.001 olduğu modelde roc eğrisi 1'e yakındır ve model hassasiyeti %91'dir. Testin ayırt etme gücü ve performansı yüksektir.

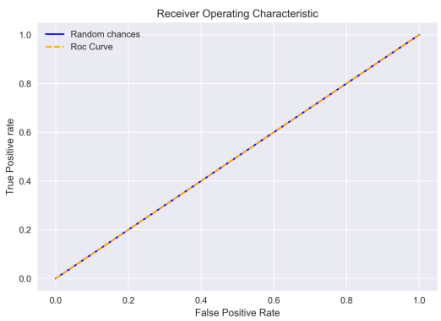
Optimizer	SGD
Learning Rate	0.001
Momentum	0.9
Epochs	100
Batch Size	512
Auc Score	%50
Test Accuracy Score	%75

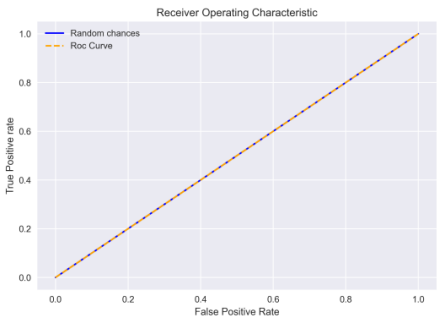
Optimizer	SGD
Learning Rate	0.001
Momentum	0.5
Epochs	100
Batch Size	512
Auc Score	%50
Test Accuracy Score	%76







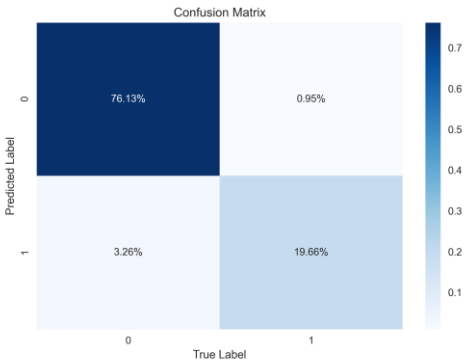


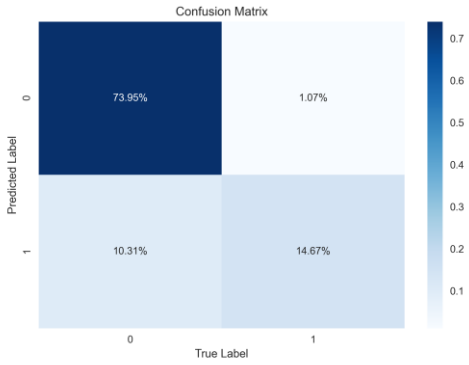


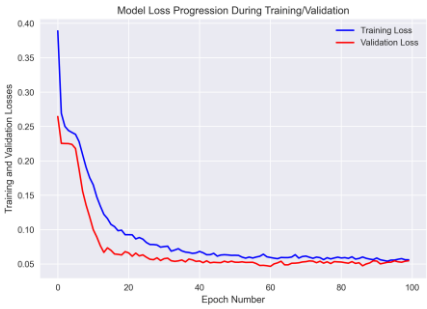
Stochastic gradient descent algoritması Adam algoritmasına göre kötü sonuçlar verdi. Ancak kendi içinde değerlendirecek olursak SGD algoritmasında momentum değeri değiştirilmesine rağmen başarı oranında büyük bir değişiklik gözlenmememiştir. Roc eğri grafiğine bakıldığında auc değerinin yani eğri altında kalan alanın 0.5 olduğu görülmektedir bu modellerin başarısız olduğunu gösterir.

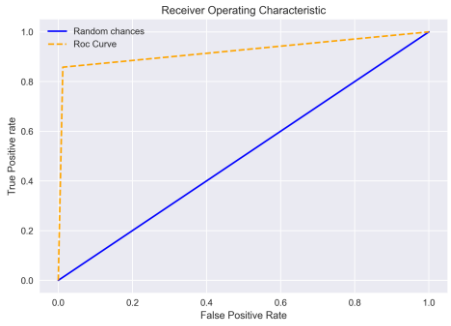
Optimizer	Adam
Learning Rate	0.001
Epochs	100
Batch Size	64
Auc Score	%92
Test Accuracy Score	%95

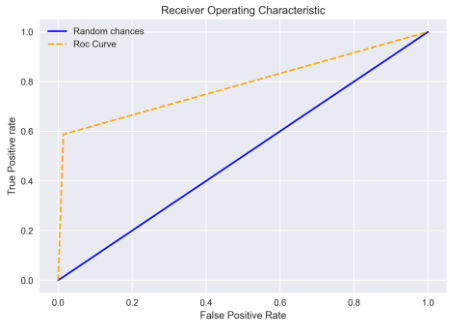
Optimizer	Adam
Learning Rate	0.001
Epochs	10
Batch Size	64
Auc Score	%78
Test Accuracy Score	%88











Öğrenme oranı ve batch size sabit tutulup epochs değeri 100 ve 10 denendiğinde model başarısı düşmüştür. Devir sayısını düşürmek model hassasiyetini de olumsuz etkileşmiştir. Training ve validation loss grafiklerine bakıldığında epoch değerinin artmasıyla modelin underfitting ya da overfitting olmadan eğitildiği görülmektedir.

Training loss ile validation loss arasında düzgün korelasyon olduğu görülmektedir. İkiside azalıp sabit bir değerde kalmıştır bu modelin iyi eğitildiği ve hem eğitim verilerinde ve test verilerinde eşit derecede iyi olduğu anlamına gelir.

Yukarıdaki örneklerde gizli katman sayısı 2 ve gizli katmanlardaki nöron sayıları 45, 25 sabit tutulmuştur. Yapılan denemelerde adam algoritmasının en optimize olduğuna karar verilmiştir. Eğitim oranı 0.001 olduğunda en iyi sonuç alınmıştır. Epochs değeri arttıkça model performansı artmaktadır.

Çalışmanın kaynak kodlarına "github.com/eaarda/Used_Car_Price_Prediction" web adresinden erişim sağlanmaktadır.