Bitcoin Fiyat Tahmini
BİL470 Proje Raporu

Kemal Bayık
TOBB ETÜ Bilgisayar Mühendisliği
Ankara, Türkiye
kbayik@etu.edu.tr

***Abstract*—Bu rapor TOBB Ekonomi ve Teknoloji Üniversitesi BİL470 dersi proje raporudur. Raporda, Bitcoin fiyat tahmini için oluşturulmuş 5 farklı modelden ve model eğitimi sonucunda elde edilen sonuçlardan bahsedilmiştir.**

I. Giriş

A. Motivasyon

Kripto varlıklar günümüzün en popüler konularından bir tanesidir. Yaklaşık 8 senedir var olmasına karşın son bir sene içerisinde kripto varlıklara olan ilgi dünyada ve ülkemizde bir hayli artmıştır. Bu ilgi ile beraber insanlar kripto varlıkları çok ciddi bir yatırım aracı olarak kullanmaya başladılar. Aynı zamanda da birçok firma kripto varlıklar ile ödeme almaya başladı. Bu gelişmelerin ışığında makine öğrenmesi ile Bitcoin fiyat tahmini yapmanın mantıklı ve kullanılabilir bir proje olduğunu düşünüyorum.

Bu projede kripto varlık birimi olarak Bitcoin seçmemin sebebi ise, diğer kripto varlık fiyatlarının Bitcoin fiyatına endeksli bir şekilde hareket ediyor olması. Bu şekilde diğer kripto varlıkların fiyat hareketleri hakkında bir bilgi sahibi olmakta mümkün oluyor.

B. Problem Tanımı (Sınıflandırma / Regresyon)

Bitcoin fiyat tahmini problemi bir zaman serisi regresyon problemidir.

C. Amaç / Hedef

Bu projenin amacı,

- Bitcoin fiyat tahmininde en iyi performans verebilecek 5 algoritmayı araştırmak
- Bulunan algoritmalar ile 5 farklı model oluşturmak
- Oluşturulan modelleri eğitmek
- Elde edilen sonuçları karşılaştırmak ve analiz etmektir.

D. Başarım Metrikleri

Bitcoin fiyatının günümüzde yaklaşık 38000 dolar olduğunu göz önüne aldığımda, mean absolute error olarak 200 dolar ve altında yapılan tahminleri başarılı olarak kabul ettim. Kullandığım metrikler ise MSE (Mean Squared Error), MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), MAX (Max Error) ve R2 Score.

II. Veri Seti, Veri Öznitelikleri, Öznitelikler

A. Veri Kaynağı

www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory

B. Veri Kümesi

Veri kümesi, 29 Nisan 2013 ve 6 Temmuz 2021 arasındaki günlük bitcoin fiyatının en düşük fiyatını, en yüksek fiyatını, açılış fiyatını, kapanış fiyatını, hacmini, market değerini, coin adını, sembolünü ve tarihi içeriyor.

C. Önişleme Aşamaları

- Date formatını " 4/29/2013 23:59" şeklinden "2013-04-29 23:59:00+00:00" şekline getirdim. Bazı modellerde ise date'i float64'e çevirerek kullandım.
- Train ve test set için sadece "Date" ve "Close" özniteliklerini kullandım. Diğer öznitelikleri işleme almadım.

D. Öznitelik Açıklamaları

- Veri kümesinde hatalı veya eksik veri içeren bir öznitelik bulunmuyor.
- Veri kümesi structured.

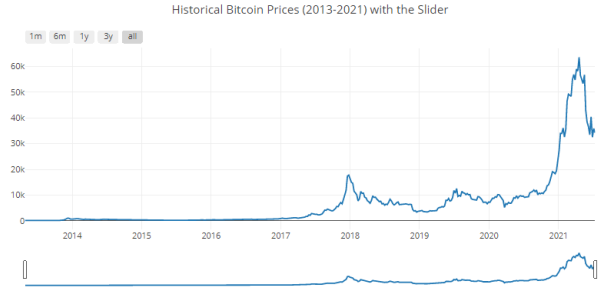
E. Sıralama, Kategorizasyon

- Veriler tarihe göre sıralı
- Veriler tarihe göre karşılaştırılabilir

F. Data normalizasyonu, Öznitelik Normalizasyonu

Projede herhangi bir normalizasyon tekniği kullanmadım.

G. Verinin 2D gösterimi



H. Öznitelikler Arasındaki İlişkiler

Öznitelikler arasında belirli bir ilişki bulunmuyor. Genel olarak marketcap, close price ve volume, close price arasında doğrusal bir ilişki olduğu varsayılabilir fakat bu her zaman doğru olmuyor. Seçtiğim ve kullandığım öznitelikler olan date ve close price arasında ise bir ilişki bulunmuyor.

III. Kullanılan Modeller

Kullandığım modellerde çarpraz doğrulama kullanmadım. Bunun birinci sebebi veri zamana göre değiştiğinden dolayı ve doğru tahmin yapmak için bir trend yakalamak gerektiğinden dolayı çarpraz doğrulama için uygun bir veri bölümü bulmakta zorlandım. İkinci sebebi ise kullandığım modellerde ve aldığım sonuçlarda herhangi bir overfitting sorunu ile karşılaşmadım. Test veri setini 1 Ocak 2018 ile 1 Ocak 2020 arası olarak belirledim. Kalan kısmı training veri seti olarak kullandım. Yani yaklaşık olarak %75 training, %25 test olarak ayırdım.

A. LSTM

LSTM, uzun süreli bağımlılıkları öğrenebilen özel bir RNN türüdür. 1997 yılında Hochreiter Schmidhuber tarafından tanıtılmıştır ve sonraki çalışmalarda bir çok kişi tarafından geliştirilmiş ve popüler hale gelmiştir. Çok çeşitli problemler üzerinde iyi sonuç verir ve yaygın olarak kullanılmaktadır.

LSTM'ler, uzun vadeli bağımlılık sorununu önlemek için tasarlanmıştır. Bu sorunu önlemek için bilgileri uzun süreler boyunca hatırlama yolunu kullanır.

LSTM, zaman serilerini sınıflandırmak, işlemek ve öngörmek için oldukça uygundur. [1]

B. XGBOOST

XGBoost(eXtreme Gradient Boosting), Gradient Boosting algoritmasının çeşitli düzenlemeler ile optimize edilmiş yüksek performanslı halidir. Tianqi Chen ve Carlos Guestrin'in 2016 yılında yayınladıkları "XGBoost: A Scalable Tree Boosting System" adlı makale ile hayatımıza dahil olmuştur. Algortimanın en önemli özellikleri yüksek tahmin gücü elde edebilmesi, aşırı öğrenmenin önüne geçebilmesi, boş verileri yönetebilmesi ve bunları hızlı yapabilmesidir. Tianqi'ye göre XGBoost diğer popüler algoritmalardan 10 kat daha hızlı çalışmaktadır.

Daha az kaynak kullanarak üstün sonuçlar elde etmek için yazılım ve donanım optimizasyon tekniklerini uygulanmıştır. Karar ağacı tabanlı algoritmaların en iyisi olarak gösterilir. [2]

C. ARIMA

ARIMA zaman serisi modeli, t zamanındaki değerin geçmiş değerler ile modellenmesidir. AutoRegressive(AR),Integrated(I) ve Moving Average(MA) kısımlarından oluşur ve p,d,q parametre değerlerini alır. AR seviyesini "p", I seviyesini yani kaç fark alındığını "d" ve MA seviyesini "q" gösterir. [3]

D. Linear Regression

Doğrusal regresyon analizi, bir değişkenin değerini başka bir değişkenin değerine göre tahmin etmek için kullanılır. Tahmin etmek istediğiniz değişken, bağımlı değişken olarak adlandırılır. Diğer değişkenin değerini tahmin etmek için kullandığınız değişken ise bağımsız değişken olarak adlandırılır.

Bu analiz biçimi, bağımlı değişkenin değerini en iyi öngören bir ya da daha fazla bağımsız değişkeni kullanarak doğrusal denklemin katsayılarını tahmin eder. Doğrusal regresyon, öngörülen ve gerçek çıkış değerleri arasındaki uyumsuzlukları en aza indiren düz bir çizgi ya da yüzeye yerleşir. Bir çift eşleştirilmiş veri kümesi için en uygun satırı keşfetmek üzere "en küçük kareler" yöntemini kullanan basit doğrusal regresyon hesaplayıcılar vardır. Daha sonra, Y'den (bağımsız değişken) X'in (bağımlı değişken) değerini tahmin edersiniz. [4]

E. Decision Trees

Ağaç tabanlı öğrenme algoritmaları, en çok kullanılan ve denetimli öğrenme yöntemlerinden biri olarak düşünülmektedir. Ağaç tabanlı yöntemler, yüksek doğruluk, kararlılık ve yorumlanma kolaylığına sahiptir. Doğrusal modellerin aksine doğrusal olmayan ilişkileri de oldukça iyi eşleyebilirler. Sınıflandırma veya regresyon, elde edilen her türlü sorunun çözümünde uyarlanabilirler. Karar ağaçları, rastgele orman, gradyan güçlendirme gibi yöntemler, her türlü veri bilimi probleminde yaygın şekilde kullanılmaktadır [5]

Bu modellere yaptığım literatür araştırmasının sonucunda karar verdim. Problem zaman serisi regresyon problemi olduğundan dolayı literatür araştırmalarımı buna göre yaptım. Bu problemde en çok kullanılan modelleri analizledim ve seçimlerimi buna göre yaptım. Linear regression ve decision tree modellerini ise hem zaman serisi regresyon problemlerinde kullanıldıklarından dolayı hemde ders esnasında gördüğümüz ve anlaşılması diğer modellere oranla daha kolay olduğu için seçtim.

IV. Test Sonuçları ve Sonuçların Yorumları

Bu bölümde modellerin eğitim ve test sonucunda MSE (Mean Squared Error), MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), MAX (Max Error) ve R2 Score performans metrikleri ile alınan sonuçlar gösterilecek ve incelenecektir.

Modeller toplamda 10 kere çalıştırılmıştır ve grafiklerde bu sonuçlar gösterilmektedir. Bazı modellerde alınan sonuçlar farklı çalıştırma sonuçlarında değişmemiştir.

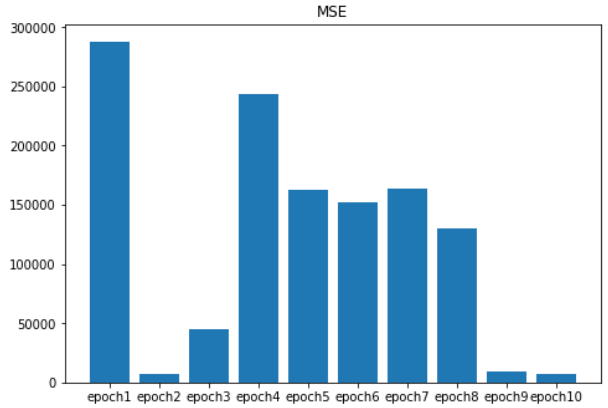
Oluşturduğum modellerde elde ettiğim başarı sırası ise şu şekilde :

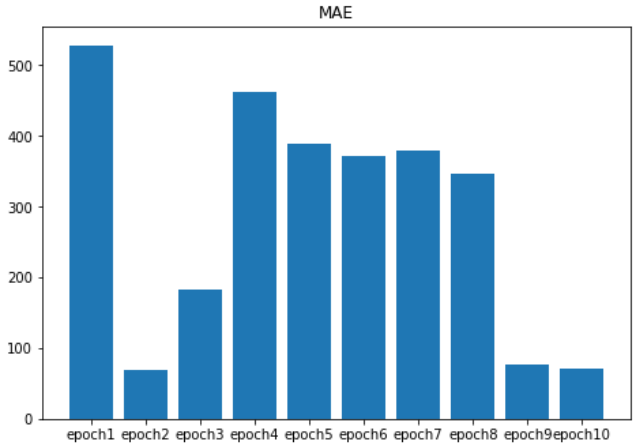
1. Linear Regression
2. Decision Trees
3. LSTM
4. ARIMA

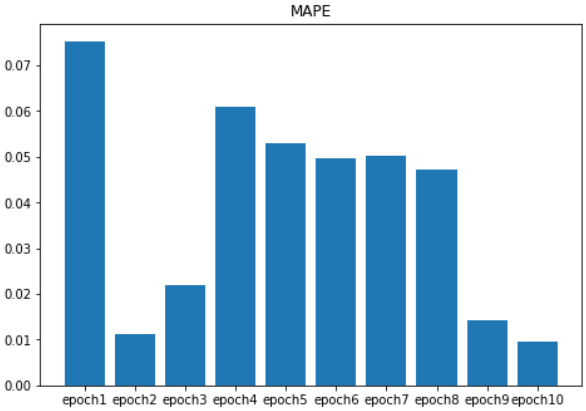
5. XGBOOST

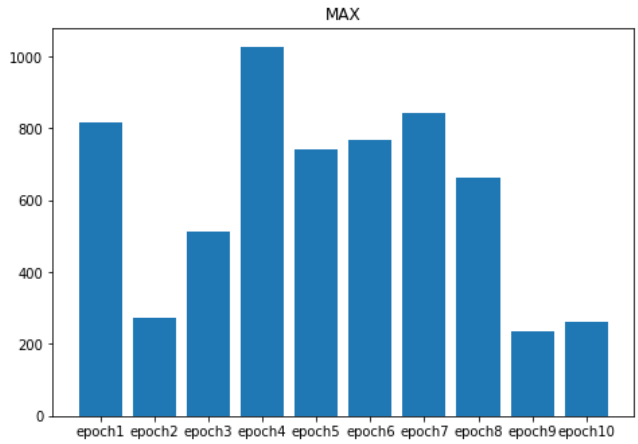
A. *LSTM*

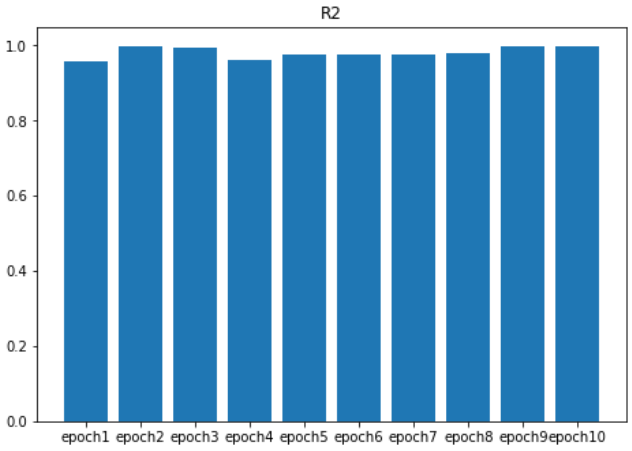
LSTM ile oluşturduğum modeller sonucunda iyi bir sonuç aldığımı düşünüyorum. 10 çalıştırma sonucunda yaklaşık 300 dolar yakın bir tahmin sonucuna ulaştım. En düşük 90 dolar en fazla ise 550 dolar yakın bir tahmin elde ettim.











B. *XGBOOST*

Oluşturduğum XGBOOST modeli oluşturduğum beş model arasında en kötü sonuç aldığım model oldu. Oluşturduğum model sonucunda aldığım sonuçlar şu şekilde :

- MSE : 10486355.035
- MAE : 2638
- MAPE : 0.40
- MAX : 11312
- R2 : -0.627

XGBOOST ile de 10 ayrı model oluşturdum ve hepsinde aynı sonucu aldığımdan dolayı grafiksel olarak göstermedim.

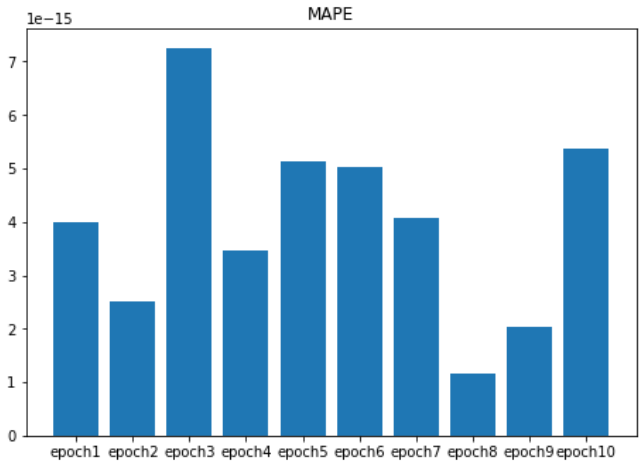
C. *ARIMA*

Oluşturduğum ARIMA modeli oluşturduğum beş model arasında en kötü sonuç aldığım ikinci model oldu. Oluşturduğum model sonucunda aldığım sonuçlar şu şekilde :

- MSE : 7950945.368216312
- MAE : 1463.4171594609902
- MAPE : 0.2899441675798125
- MAX : 11539.584205465391

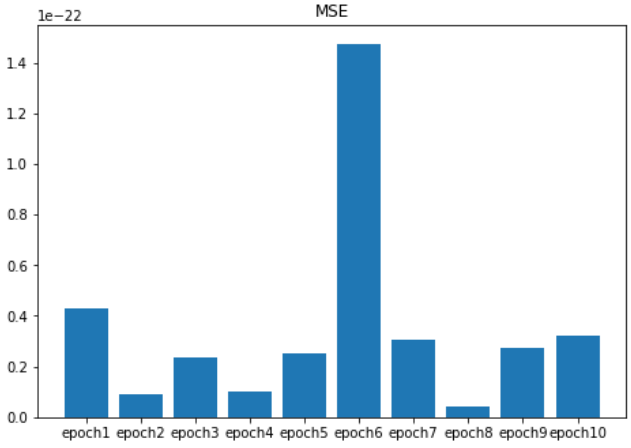
- R2 : 0.9366307170321426

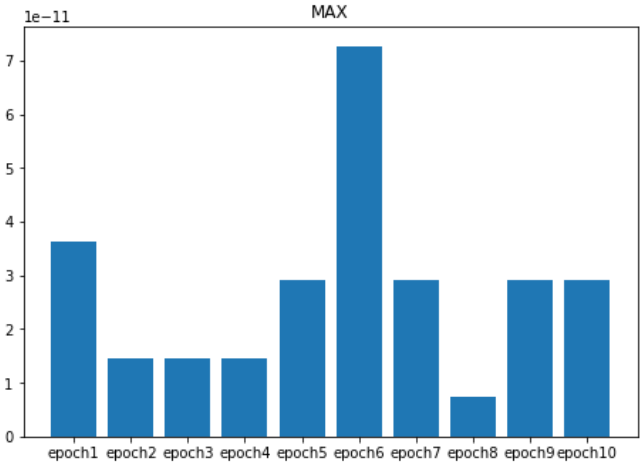
ARIMA ile de 10 ayrı model oluşturdum ve hepsinde aynı sonucu aldığımdan dolayı grafiksel olarak göstermedim

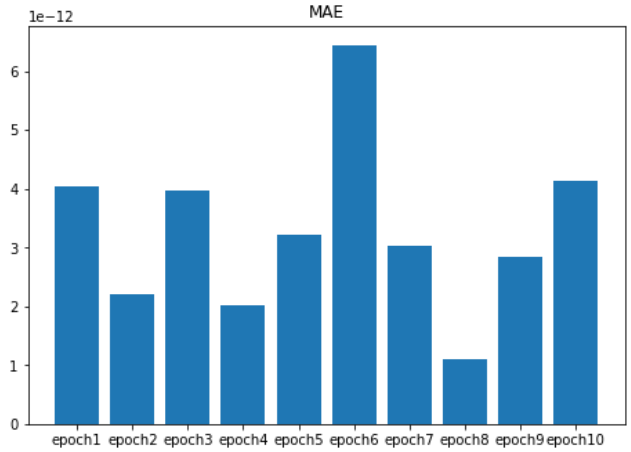


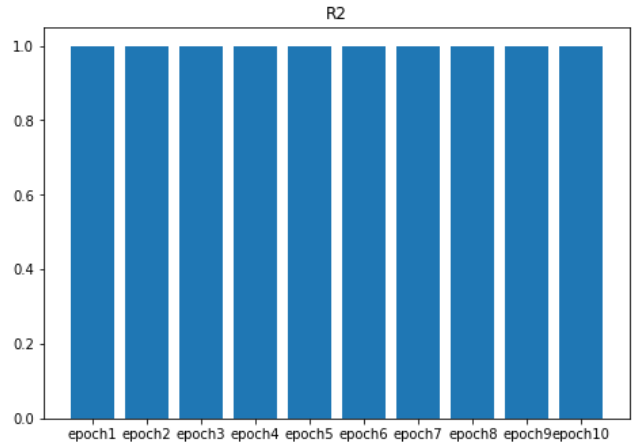
D. Linear Regression

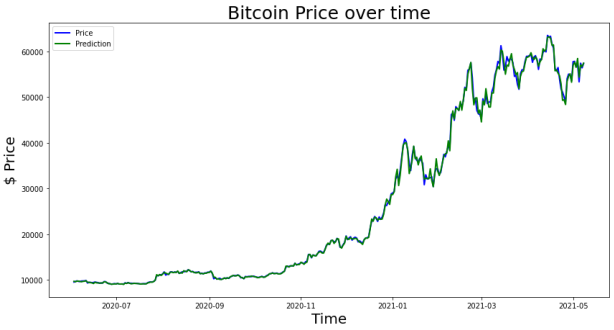
Linear Regression ile oluşturduğum modeller sonucunda iyi bir sonuç aldığımı düşünüyorum. 10 çalıştırma sonucunda en yüksek doğruluğu elde ettiğim model oldu.





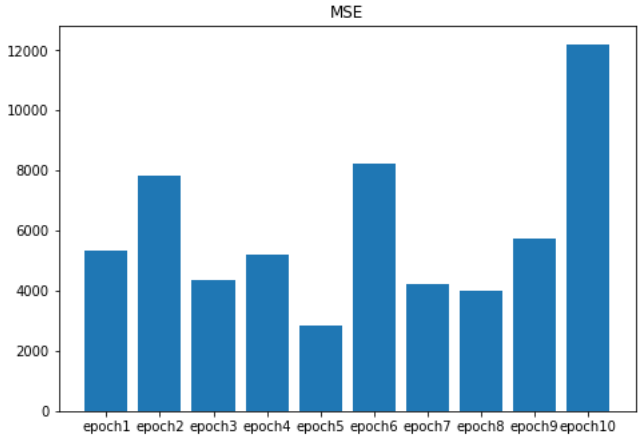


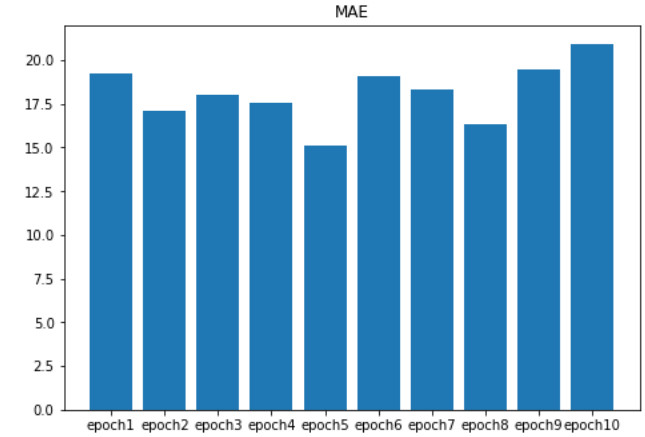


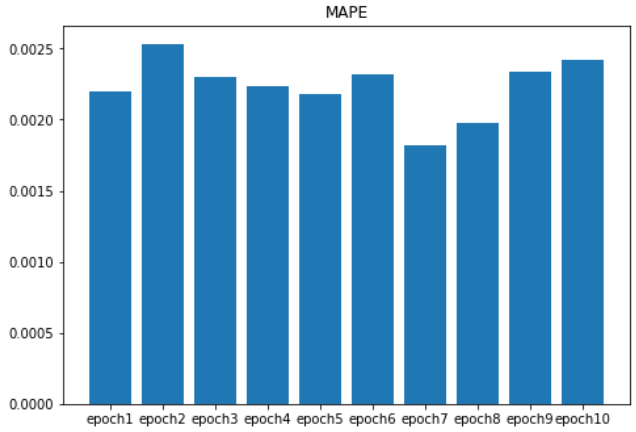


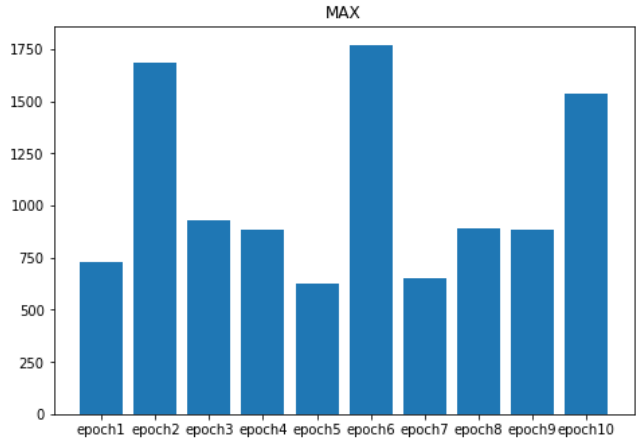
E. Decision Trees

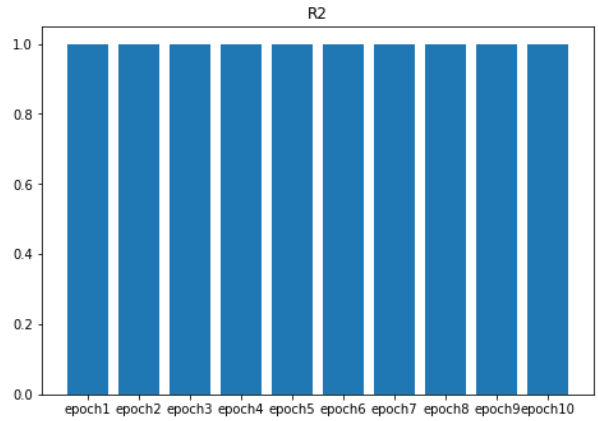
Decision Tree ile oluşturduğum modeller sonucunda iyi bir sonuç aldığımı düşünüyorum. 10 çalıştırma sonucunda en ikinci yüksek doğruluğu elde ettiğim model oldu. Oluşturduğum modeller sonucunda yaklaşık 18 dolar yakın tahmin elde ettim.











V. Sonuçlar

BIL 470 dersi kapsamında yapmış olduğum Bitcoin Price Prediction projemde beş farklı algoritma ile modeller oluşturdum ve en yakın Bitcoin fiyat tahmini yapan modeli elde etmeye çalıştım. Elde ettiğim sonuçları performans metrikleri ile kıyasladım.

Çalışma sonucunda bir time series forecasting problemi olan bitcoin price predictionda makine öğrenmesi ve doğru seçilmiş

modeller ile gayet yakın tahminler yapılabileceği görülüyor. Fakat bununla birlikte bu fiyatların genel olarak insanların düşünce ve davranışarına odaklı olduğu ve manipulasyona çok açık olduğu unutulmamalıdır.

Projede kullandığım modelleri seçerken daha önce bu konu üzerinde yapılmış projelerde kullanılan modelleri inceledim ve seçimlerimi buna göre yaptım fakat bazı modellerde beklediğim sonucu elde edemedim. Bu yüzden daha iyi sonuç veren modeller seçip projeyi daha verimli bir şekilde tamamlayabilirdim diye düşünüyorum.

Bir kripto varlık olan bitcoin hayatımıza yeni girmiş olmasına karşılık günümüzde oldukça popüler bir yatırım aracı olarak görülüyor. Bu sebepten dolayı makine öğrenmesi ile fiyat tahmininin ileride üzerinde çok daha fazla çalışılacak bir proje olduğunu düşünüyorum. Kullandığım veri setine o gün yaşanan önemli olaylar (Amerikan Merkez Bankası faiz açıklamaları, Elon Musk tweetleri gibi) eklenirse manipulasyona bu kadar açık olan bir fiyatın tahminin daha verimli olabileceğini düşünüyorum.

REFERENCES

[1] https://colah.github.io/posts/2015-08-Understanding-LSTMs/
[2] https://www.veribilimiokulu.com/xgboost-nasil-calisir/
[3] https://medium.com/icrypex/arima-ile-bitcoin-fiyat-tahmini-510dd7a94d97
[4] https://www.ibm.com/tr-tr/analytics/learn/linear-regression
[5] https://veribilimcisi.com/2018/02/23/karar-agaclari-decision-trees/
[6] https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory
[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

Sunum linki : https://youtu.be/v01RD7qpfME