Bitcoin fiyatının machine learning ile tahminlenmesi
Predict Bitcoin price with machine learning

Murat Özkul

Özetçe—Bu çalışmada makine öğrenmesi ile bitcoin fiyatının tahminlenmesine çalışılmıştır. Henüz yeni yeni yaygınlaşan ve olgunlaşmamış bir piyasaya sahip olan ve dinamikleri normal para piyasalarından farklı olan bu piyasada anlık ve beklenmedik durumlar söz konusu olabilmektedir. Beklenmemiş hareketleri ve fiyatı tahminlemek bu nedenle diğer piyasalara göre daha zordur. Çalışmada bunları tahminleyebilmek için Python dilinde yazılmış scikit-learn kütüphanesi aracılığı ile karar ağaçları, linear regression, neural network algoritmaları kullanılmıştır. Modellerin doğruluk analizi içinde "Coeffecion of determinants" yöntemi kullanılmıştır.

Anahtar Kelimeler — bitcoin; tahminleme; machine learning

I. BİTCOİN NEDİR?

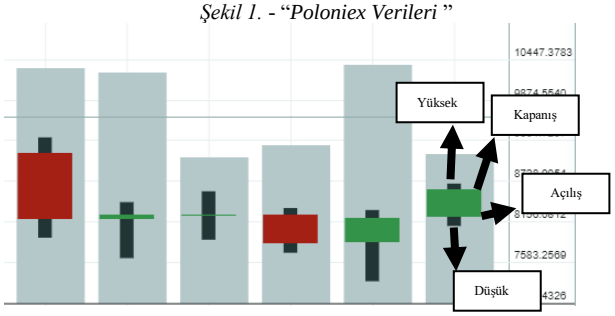
Bir grup gönüllü yazılımcı tarafından geliştirilen merkeziyetçi olmayan, sınırlı sayıda bulunan şifrelenmiş para birimidir. Blok zinciri tabanlı yapı ile yapılan her işlem değiştirilemez şekilde kaydedilir. Sistemde bulunan herkes hangi cüzdanda ne kadar bitcoin olduğunu ve hangi işlemleri yaptığını bilir.

II. KULLANILAN BORSA VERİLERİ NELERDİR?

Fiyat tahminlememizi Poloniex adlı bitcoin borsasının verileri üzerinden yapılmıştır. Şekil 1 de örnek bir borsa verisi gözükmektedir. Şekil 2 de ise bu görselin verilerinin Poloniex borsasından çekildiğindeki hali bulunmaktadır.

Örnek olarak bir kayıt;

Şekil 1. - "Poloniex Verileri "



Şekil 2 "API verileri"


```
{
"date": 1521763200,
"high": 8907.7,
"low": 8270,
"open": 8700,
"close": 8907.7,
"volume": 23230776.098734,
"quoteVolume": 2737.51850322,
"weightedAverage": 8486.07089647
}
```


Poloniex 2015-02-20(YYYY-MM-DD) tarihinden itibaren bizlere datasını API ile vermektedir.

- Date : Linux timestamp
- High : Günün en yüksek fiyatı
- Low : Günün en düşük fiyatı
- Open : Günün açılış(ilk) fiyatı
- Close :Günün kapanış(son) fiyatı
- Volume : Hacim günlük yapılan işlemlerin toplamı
- Quote Volume : Alım/Satım emirinde bekleyen hacim
- Weighted Average : Gerçekleşen işlemlerin ağırlıklı ortalaması

III. KULLANILAN TEKNOLOJİLER

Bu araştırmada kodlama dili olarak Python kullanılmıştır. Machine learning algoritmaları sıfırdan oluşturulmamıştır mevcutta open-source olarak geliştirilen scikit-learn(sklearn) kütüphanesi kullanılmıştır. Bu kütüphane kendi içinde bir çok machine learning algoritmasını barındırmakta ve sizlerin modellerinizde kullanabilmenizi sağlamaktadır.

Araştırma esnasında scikit-learn kütüphanesinin desteklediği aşağıdaki algoritmalar kullanılmıştır;

1. Lasso Regression
2. Bayesian Ridge Regression

3. Elastic-Net Regularization
4. Neural Networks
5. Decision Tree

Lasso, Bayesian Ridge ve Elastic-Net algoritmaları linear regression algoritmalarıdır.

1. Lasso Regression

Açılımı "Least Absolute Shrinkage and Selection Operator" olan linear regression algoritmasıdır.

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$
$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \le t.$$

Lasso regresyonun en önemli özelliği kendiliğinden feature selection uygulamasıdır. Önceliği ve modelde etkisi az olan feature ları algoritmada etkisi 0 a indirgenmektedir.[1][2]

2. Bayesian Ridge Regression

Ridge algoritmasını temel alıp geliştirilmiş bir linear olasılık tahminleme algoritmasıdır.[3]

$$p(w|\lambda) = \mathcal{N}(w|0, \lambda^{-1}\mathbf{I_p})$$

3. Elastic-Net Regularization

Ridge ve Lasso regresyonlarının beraber kullanımı sonucu oluşturulmuş bir algoritmadır.

$$P_\alpha(\beta) = \frac{(1-\alpha)}{2}\|\beta\|_2^2 + \alpha\|\beta\|_1 = \sum_{j=1}^{p}\left(\frac{(1-\alpha)}{2}\beta_j^2 + \alpha|\beta_j|\right).$$

$\alpha = 1$ olması durumunda Lasso regression ile aynı sonuç üretilecektir. [4][5]

4. Neural Networks

Canlıların beyinlerini örnek alarak geliştirilen machine learning algoritmalarıdır. Beyindeki nöronların birbirine bağlanması gibi feature ların karar verici nodelara bağlanarak örnekler üzerinden beyin gibi öğrenmesi amaçlanmaktadır. [6]

5. Decision Tree

Daha çok sınıflandırma amacıyla kullanılan bir algoritmadır. Geçmiş örnekler kökten yapraklara olmak üzere dağıtılıp daha sonra bu dağılımlar üzerinden kategori sınıflaması yapılmaktadır. Regression da ise önceki örneklerdeki değişimler kökten yapraklara kadar dağılıp daha sonra bu artışlar gruplanıp buna göre tahminleme yapılmaktadır.[7][8]

IV. MODELLER VE TAHMİNLEMELERİ

Yukarıda tanımlaması yapılan 5 farklı machine learning algoritması scikit-learn kütüphanesi ile Python üzerinde geliştirildi. Herbir algoritma farklı parametrelerle denendi her deneme için R^2 (coefficient of determination) skorları hesaplanıp en yüksek skora sahip olan versiyonları seçildi. Seçilen bu modeller hergün yeni gelen datalarla eğitilip baştan oluşturulup tahminlemelerde bulunuldu.

Oluşturulan modellerden alınan skorlar aşağıdaki gibidir.

Tablo 1. - "Algoritma Skorları"

Algoritma Tipi	Algoritma	Skor
Lineer	Lasso	0.867013659539
Lineer	Elastic-Net	0.867013531006
Lineer	Bayesian Ridge	0.869218514189
Neural Network	Neural Network	0.866293900411
Decision Tree	Decision Tree	0.89607374373

[a.] Skorlar -1 ile 1 arasındadır 1 en yüksek -1 en düşük

Tablo 1 de görüldüğü gibi en yüksek skoru decision tree algoritması almaktadır. Bu neden ile çalışmama decision tree algoritması ile devam edilmiştir.

Decision tree algoritması ile model geliştirildiğinde elimizde bulunan feature ların modeldeki etkisini çıkarttığımızda. Yani hangi feature bitcoinin fiyatını daha çok etkilediğine baktığımızda aşağıdaki gibi bir tablo karşımıza çıkmakmıştır.

Tablo 2. - "Modeli en çok etkileyen feature sıralaması"

High	1
Close	2
Low	3
Open	4
Volume	5
Weighted Average	6
Quote Volume	7

Decision tree algoritması içindeki maksimum derinlik değerleri ile ve piyasa koşullarının değişmesini göz önünde bulundurduğumuzda geçmiş verileri dışarıda tutup modelleri kontrol ettiğimizde ise skorlar aşağıdaki gibi oluşmaktadır.

Tablo 3. - "*Decision Tree değişkenlerle skor incelemesi* "

Tarih limiti	Derinlik	Skor
Evet	6	0.886838922205
Evet	7	0.898244635965
Evet	Yok	0.999940018413
Hayır	6	0.874058735942
Hayır	7	0.89607374373
Hayır	Yok	0.999991759984

[b.] Tarih limitli;
Evet: Datadan ilk 600 gün çıkartılmış
Hayır: Tüm data kullanılmış.

Tablo 3 de görüldüğü gibi tarih limiti verilmesi durumunda daha yüksek bir skor elde edilmektedir. Derinlik verilmemesi durumunda ise tarih limiti verilmeyen model daha başarılı çıkmaktadır. Fakat derinlik verilmediği durumlarda model aşırı öğrenmeye gittiği için tercih edilmemiştir.

Derinlik 7 olan modellerde tarih limitli ve limitsiz arasında büyük farklar olmadığından ve de tarih limitsizde daha fazla örneklem bulunduğundan ve ileride olası piyasa döngüsünün tekrarlanma durumları da düşünüldüğünde tarih limiti bulunmayan derinliği 7 olan model kullanılmıştır. Bu model ile her gün yeni günün verisini de ekleyerek her gün model geliştirilmiş ve 20.05.2018 tarihi ile 57 adet tahmin gerçekleşmiştir.

Tablo 4. - "*Tahminler* "

Tarih	Tahmin	Gerçekleşen	Fark	Sapma Oranı
24.03.2018	8934,660979	8907,7	-26,96097946	0,30
25.03.2018	8493,148073	8454,814447	-38,33362605	0,45
26.03.2018	9469,999983	8116,523046	-1353,476937	16,68
27.03.2018	8326,875008	7782,716713	-544,1582952	6,99
28.03.2018	7951	7943,583816	-7,41618365	0,09
29.03.2018	8009,373293	7080,402768	-928,970525	13,12
30.03.2018	7659,969799	6837,725325	-822,2444735	12,03
31.03.2018	7127,338161	6933,879205	-193,4589557	2,79
1.04.2018	7256,693464	6823,020401	-433,6730627	6,36
2.04.2018	6945,494074	7058,894148	113,4000738	1,61
3.04.2018	6931,885888	7412,270339	480,3844506	6,48
4.04.2018	7052,834962	6769,945088	-282,8898733	4,18
5.04.2018	7078,508917	6775	-303,508917	4,48
6.04.2018	7057,937995	6613,545638	-444,3923576	6,72
7.04.2018	7040,25437	6890	-150,2543704	2,18
8.04.2018	6880,957851	7030	149,0421489	2,12
9.04.2018	6881,962534	6761,537785	-120,4247498	1,78
10.04.2018	6896,766281	6834	-62,76628092	0,92
11.04.2018	6901,172837	6966,460062	65,28722596	0,94
12.04.2018	6896,005695	7929,999999	1033,994303	13,04
13.04.2018	7649,478221	7875	225,5217793	2,86
14.04.2018	7999,974796	7996	-3,974796597	0,05
15.04.2018	8255,423452	8356	100,5765483	1,20
16.04.2018	8241,011038	8060	-181,0110377	2,25
17.04.2018	8247,063088	7891,1	-355,9630883	4,51
18.04.2018	8237,709934	8164,190618	-73,51931639	0,90
19.04.2018	8221,204699	8286,999998	65,79529896	0,79
20.04.2018	8218,61315	8860	641,3868502	7,24
21.04.2018	8597,650954	8931,999998	334,3490449	3,74
22.04.2018	8597,650954	8791,999999	194,3490455	2,21
23.04.2018	8488,35637	8934,977783	446,6214126	5,00
24.04.2018	8635,415103	9633,630952	998,2158487	10,36
25.04.2018	9792,444341	8860,024074	-932,420267	10,52
26.04.2018	9120,702345	9286	165,2976553	1,78
27.04.2018	9857,500000	8920,01	-937,4900001	10,51
28.04.2018	9341,98048	9330,761257	-11,21922267	0,12
29.04.2018	8960,446625	9423,799996	463,3533717	4,92
30.04.2018	8942,240653	9218,792931	276,552278	3,00
1.05.2018	9022,085907	9218,792931	196,7070238	2,13
2.05.2018	8778,332575	9245	466,6674246	5,05
3.05.2018	9090,563238	9.746	655,5196241	6,73
4.05.2018	9004,851370	9719,999999	715,1486287	7,36
5.05.2018	10156,181298	9878,367954	-277,8133438	2,81
6.05.2018	10083,484415	9670	-413,4844151	4,28
7.05.2018	9638,091161	9390	-248,0911614	2,64
8.05.2018	9211,669277	9020	-191,6692772	2,12
9.05.2018	8901,250038	9300,4	399,149962	4,29
10.05.2018	9210,145095	9020	-190,145095	2,11
11.05.2018	9320,340540	8385,231466	-935,1090737	11,15
12.05.2018	8248,187054	8471,752489	223,5654355	2,64
13.05.2018	8248,187054	8701,613284	453,4262299	5,21
14.05.2018	8760,848245	8671	-89,84824455	1,04
15.05.2018	8757,557413	8449,740002	-307,8174116	3,64
16.05.2018	8257,129671	8336,803816	79,67414442	0,96
17.05.2018	8257,129671	8050	-207,1296713	2,57
18.05.2018	8260,194061	8240	-20,19406148	0,25
19.05.2018	8252,409096	8219,5	-32,909096	0,40

Tablo 4 de gösterilen 57 adet tahminin ortalama sapma değeri:4.26 dır. En yüksek 5 ve en düşük 5 adet sapma dışarıda bırakıldığında ise 3.74, En yüksek ve en düşük 10 adet sapmayı dışarıda bırakırsak ise ortalama sapma 3.41 olarak hesaplanmaktadır.

Tablo 5. - "*Sapma Oranları* "

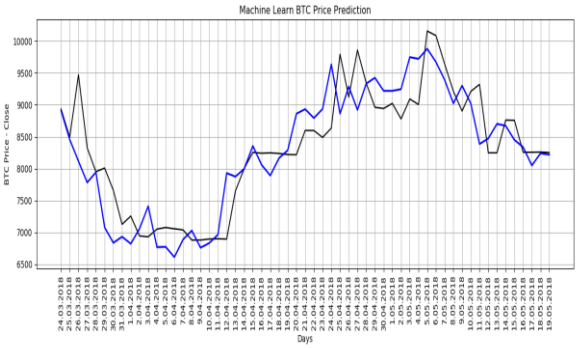
Ortalama Sapma	4,26
En Yüksek/Düşük 5 Dışarıda	3,74
En Yüksek/Düşük 10 Dışarıda	3,41

Sapma oranlarının yüksek olduğu tarihler incelendiğinde piyasayı etkileyen haberlerin bulunduğu gözlemlenmiştir. Örnek olarak Twitter, Facebook, Google'ın cryptocurrencies reklamlarını engellemesi, MT.Gox adlı eski borsanın kayyumunun elindeki yüklü miktardaki bitcoin leri satması vb. bu nedenle en yüksek ve en düşük sapma değerlerini

çıkardığımız zaman ortalama sapma değerlerinde düşüş gözlemlenmektedir.

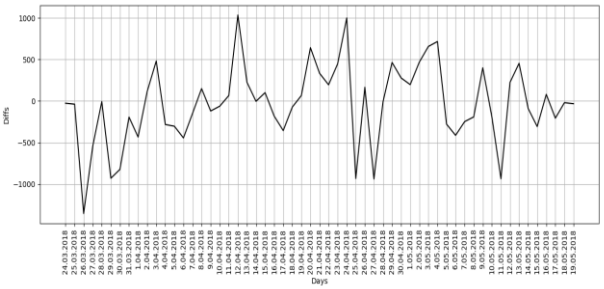
Bu verilere grafiksel olarak bakacak olursak;

Şekil 2. - "Fiyat-Tahmin grafiği"

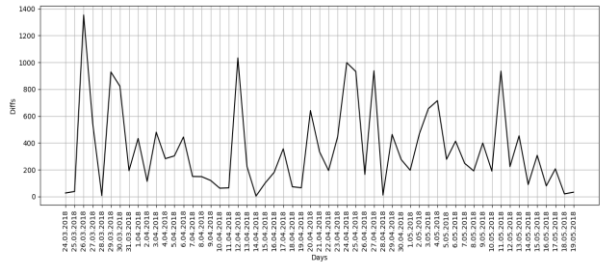


Şekil 2 de siyah çizgi o gün için gerçekleşen fiyat iken mavi çizgi ise modelin tahminlediği rakamlardır.

Şekil 3. - "Fiyat-Tahmin arasındaki fark"

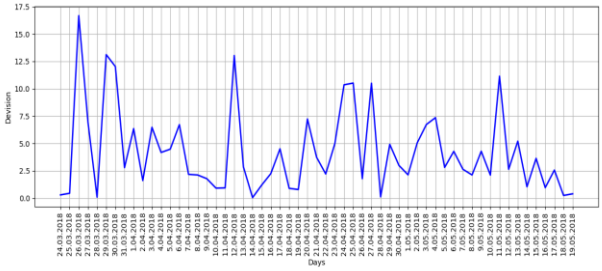


Şekil 4. - "Mutlak değerli Fiyat-Tahmin arasındaki fark"



Şekil 3 de gerçekleşen fiyat ile tahmin arasındaki fark gösterilmekteyken Şekil 4 de ise bu farkın mutlak değeri bulunmaktadır.

Şekil 5. - "Sapma grafiği"



Şekil 5 de ise sapma değerinin günlük değişimi gözükmektedir.

KAYNAKLAR

[1] Trevor Hastie, Robert Tibshirani, Jerome Friedman *The Elements of Statistical Learning*, Springer, Corrected 12th printing - Jan 13, 2017.

[2] E. Alpaydın, Yapay öğrenme. İstanbul, Boğaziçi Üniversitesi Yayınevi, 2011

[3] http://scikit-learn.org/stable/modules/linear_model.html#bayesian-ridge-regression

[4] Zou, H. and T. Hastie. *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society, Series B, Vol. 67, No. 2, pp. 301–320, 2005.

[5] Friedman, J., R. Tibshirani, and T. Hastie. *Regularization paths for generalized linear models via coordinate descen*, Journal of Statistical Software, Vol 33, No. 1, 2010

[6]https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html#What%20is%20a%20Neural%20Network

[7] J.R. Quinlan ,*Induction of Decision Trees, 1986 Kluwer Academic Publishers, Boston*

[8] Brendan Kitts , *Regression Trees, Reading, MA. 01867. USA*