



www.turkishstudies.net/appliedsciences

Turkish Studies - Information Technologies and Applied Sciences

eISSN: 2667-5633

Research Article / Araştırma Makalesi



Sponsored by IBU

Yapay Zekâ Teknolojisi ile Uçuş Fiyatı Tahmin Modeli Geliştirme

Developing Flight Price Prediction Models with Artificial Intelligence Technology

Mustafa Berk Keleş* - Aytürk Keleş** - Ali Keleş***

Abstract: Today, passengers can compare flights of different airline companies, find the most convenient flight in a certain time period, reserve and purchase, thanks to advances in internet technologies. Revenue policies of airlines and competition between them, seasons, holidays, time left to flight, number of seats available, tax policies of countries, travel policies between countries, etc. Many factors are effective in determining flight prices. At the same time, even the price of a ticket for the same flight can change within hours. Therefore, it is very important for airline companies and customers to be able to predict the ticket prices, which are so variable and dynamic. Today, many online travel agencies and airline companies collaborate on AI-focused Research&Development studies on dynamic price prediction. This study is part of the AI-supported "Flight Prices Predictor" project conducted by the Enuygun.com R&D Center. In this part of the the project, machine learning algorithms, which are one of the AI technologies, have been researched and two different prediction models have been developed by using Gradient Boosting (GB-Gradient Boosting) and Random Forest (RF-Random Forest) algorithms. The accuracy performances of "RF-FPPredictor" and "GB-FPPredictor" models, which were developed and tested using Gradient Boosting and Random Forest algorithms, are 90% and 92% respectively. The average absolute percentage errors (MAPE) of the models are 2.49% and 2.26%. These models are quite successful when compared with other models developed before in terms of prediction performance.

Structured Abstract: With the successful improvements of the transportation industry, airline has become a popular travel method. Today, thanks to internet technologies, customers can compare book and purchase flight tickets online. Pricing policies of airlines, flight distances, categories they serve, airlines, global population movements and many different dynamic factors are effective factors in the ticket pricing. Therefore, customers have to buy even the same flight at different prices at different times. Despite this, customers always want to buy their airline tickets at the lowest price. Today, thanks to the tools developed by

* Makine Öğrenmesi Mühendisi, Enuygun.com İnternet Bilgi Hizmetleri Teknoloji ve Ticaret A.Ş., AR-GE Merkezi
Machine Learning Engineer, Enuygun.com Internet Information Services Technology and Trade Inc., R&D Center
ORCID 0000-0002-4571-8126
mustafa.keles@enuygun.com

** Doç. Dr., Ağrı İbrahim Çeçen Üniversitesi, Eğitim Fakültesi, Bilgisayar ve Öğretim Teknolojileri Eğitimi Bölümü
Assoc. Prof. Dr., Agri Ibrahim Cecen University, Faculty of Education, Department of Computer and Education
ORCID 0000-0001-9755-295X
ayturkk@hotmail.com

*** Doç. Dr., Ağrı İbrahim Çeçen Üniversitesi, Eğitim Fakültesi, Bilgisayar ve Öğretim Teknolojileri Eğitimi Bölümü
Assoc. Prof. Dr., Agri Ibrahim Cecen University, Faculty of Education, Department of Computer and Education
ORCID 0000-0002-0785-9593
alakmus@hotmail.com

Cite as/ Atıf: Keleş, M. B., Keleş, A., & Keleş, A. (2020). Yapay zekâ teknolojisi ile uçuş fiyatı tahmin modeli geliştirme. *Turkish Studies - Applied Sciences*, 15(4), 511-520. <https://dx.doi.org/10.29228/TurkishStudies.45993>

Received/Geliş: 27 August/Ağustos 2020

Checked by plagiarism software

Accepted/Kabul: 20 December/Aralık 2020

Published/Yayın: 25 December/Aralık 2020

Copyright © INTAC LTD, Turkey

CC BY-NC 4.0

online travel agencies, customers can buy tickets with lowest prices by comparing compare ticket prices over time. Accurate prediction of ticket prices will help passengers make the right purchasing decisions and save money. This study seeks an answer to the question of whether we can successfully predict flight prices, which have a highly variable and dynamic structure under the influence of many factors, with machine learning techniques, which is a sub-field of AI technology.

Today, many techniques of machine learning are used successfully in classification and prediction. Ensemble learning algorithms, which are especially newly developed (An et al. 2016; Chen et al. 2015) and give very successful results, are used to solve many problems. Ensemble learning produces more successful results by combining the strengths (compensating for the weaknesses) of more than one sub models. There are different studies (Rajankar & Sakharkar 2019, Vu et al. 2018, Tziridis et al. 2017, Chen et al. 2015) on flight price using ensemble learning algorithms. One of the successful models developed so far was developed by Tsiridis. They were used different eight machine learning algorithms and compared their performances. Their dataset consists of 1,814 flights for a single international route and 8 features (departure time, arrival time, number of free luggage, days before departure, number of intermediate stops, holiday, time of day and day of week). All models were trained based on this dataset. The results revealed that the “Bagging Regression Tree” model outperforms the other models with accuracy of 87.42%, followed by Random Forest Regression Tree which achieved 85.91%.

In this study, Random Forest (RF) and Gradient Boosting (GB), which are ensemble learning algorithms were used for price estimation. Two different data sets were used to obtain the main data set. After pre-processing of the main data set, it was used to develop models. The data set consists of 200,000 flight records consisting of many different routes, national / international, connecting / non-stop, obtained from the online travel site enuygun.com.

Each flight record consists of departure point, destination, price, airline, departure date, arrival date, stop count and route data. All features that may be effective in ticket prediction have been tried to be included in the data set. In addition, the prices of flights between different countries with the same flight distance are very different from each other due to various reasons. Therefore, the "meanPriceOf" feature was created by considering the flights between two countries as a category and by taking the average of the prices of the flights belonging to this category. As a result, 60% of this data set was used for training, 20% for testing and the remaining 20% for verification.

Other studies and the results obtained from this study can be examined in detail in Table.1 in the "Discussion and Conclusion" section. As seen in “Table 1”, the studies have some limitations in terms of data set size, number of routes, and generally including non-stop flights. When the studies evaluated in terms of the features used, it is seen that the departure time, arrival time, time to flight, and price features are used in common in all study. However, in this study, 200,000 flight records consisting of many different routes, national / international, connecting / non-stop, were used. However, when the models developed using the same algorithms are compared with the RF-FPPredictor and BG-FPPredictor models which were developed in this study, it is seen that they are much more successful than the others with high accuracy of 90% and above and an average absolute percentage error of less than 3%.

The reasons why the community learning models developed in this study have higher accuracy and lower mean absolute percentage error when compared to the models in other studies can be explained as follows.

- A comprehensive research has been done for this study.
- Algorithms with the potential to create successful models were preferred.
- A large data set was used in the study.
- Unlike other studies, in this study, models were trained using data with a wider range of features
- Feature engineering and data preprocessing were carried out more comprehensively and with more care than other studies.
- The optimization of models was done very well.

In the light of these findings, it is planned that the newly developed RF-FPPredictor and BG-FPPredictor models are integrated into the system by the enuygun.com R&D Center to successfully forecast variable flight prices. As a result, being able to predict flight prices, which has dynamic and variable structure

and are affected by many factors, will make an important contribution to the airline industry in both economic and strategic terms. In this way, airlines and online travel agents can determine the necessary strategy for revenue maximization and customers can save money by purchasing cheaper tickets.

Keywords: Computer Science, Artificial Intelligence, Gradient Boosting, Random Forest, Flight Price

Öz: Bugün internet teknolojilerindeki gelişmeler sayesinde yolcular farklı havayolu şirketlerine ait uçuşları karşılaştırabilir, belirli bir zaman periyodunda en uygun fiyatlı uçuşu bulabilir, rezerve edebilir ve satın alabilir. Havayolu şirketlerinin gelir politikaları ve aralarındaki rekabet, mevsimler, tatiller, uçuş süresine kalan süre, mevcut koltuk sayısı, ülkelerin uyguladıkları vergi politikaları, ülkeler arası seyahat politikaları vb pek çok faktör uçuş fiyatlarının belirlenmesinde etkilidir. Aynı zamanda, aynı uçuşa ait bir biletin fiyatı bile saatler içinde değişebilmektedir. Buyüzden bu kadar değişken ve dinamik olan bilet fiyatlarının önceden kestirebilmek hem hava yolu şirketleri açısından hem de müşteriler açısından çok önemlidir. Bugün pek çok çevrimiçi seyahat acentası ve havayolu şirketi işbirliği yaparak dinamik fiyat tahmini üzerine Yapay Zekâ odaklı Ar-Ge çalışmaları yürütmektedir. Bu çalışma da Enuygun.com Ar-Ge Merkezi tarafından yürütülmekte olan YZ destekli “Flight Prices Predictor” adlı projesinin bir parçasıdır. İlgili projenin bu kısmında YZ teknolojilerinden makine öğrenmesi algoritmaları araştırılmış ve topluluk öğrenme algoritmalarından Gradyan Artırım (GB-Gradient Boosting) ve Rastsal Orman (RF-Random Forest) algoritmaları kullanılarak iki farklı tahmin modeli geliştirilmiştir. Bu algoritmalar kullanarak geliştirilen ve test edilen “**RF-FPPredictor**” ve “**GB-FPPredictor**” modellerinin doğruluk performansları sırasıyla %90 ve %92’dir. Modellerin ortalama mutlak yüzde hataları (MAPE) ise %2.49 ve %2.26’dır. Bu modeller daha önce geliştirilen diğer modeller ile tahmin performansı açısından karşılaştırıldığında oldukça başarılı olduğu görülmektedir.

Anahtar Kelimeler: Bilgisayar Bilimleri, Yapay Zekâ, Gradyan Artırım, Rastsal Orman, Uçuş Fiyatı

Giriş

Ulaşım sektörünün gelişmesiyle havayolu daha popüler bir seyahat yöntemi haline geldi. Bugün internet teknolojileri sayesinde müşteriler uçak biletlerini çevrimiçi olarak karşılaştırabilir, rezervasyon yapabilir ve satın alabilir hale gelmiştir. Havayolu şirketleri gelir politikaları gereği bilet fiyatlarını mevcut uçuş koltuğuna, mevsimlere, tatil dönemlerine ve daha pek çok faktöre göre farklılaştırmaktadır. Aynı uçuşu farklı zamanlarda alış bile fiyatı değiştirmektedir. Buna karşın çevrimiçi seyahat acentalarının geliştirdiği araçlar sayesinde müşteriler bilet fiyatlarını zaman içinde etkili bir şekilde karşılaştırarak takip edebilmektedir. Müşteriler genel olarak uçak biletlerini en düşük fiyattan satın almak isterler. Ancak havayollarının fiyatlandırma politikaları (Smith vd.,1992), uçuş mesafeleri, hizmet verdikleri kategoriler, havayolları, küresel nüfus hareketleri ve pek çok farklı dinamik faktör bilet fiyatlarının oluşmasında etkindir. Buyüzden bilet fiyatlarının doğru tahmin edilmesi, yolcuların uygun satın alma kararları vermelerine ve masraflardan tasarruf etmelerine yardımcı olacaktır.

Bilet fiyatlarının tahmin edilmesi üzerine yapılan farklı makine öğrenmesi çalışmaları mevcuttur. Bunlardan biri Boruah ve diğerlerinin (2018) “Kalman” filtresi olarak bilinen ünlü bir Bayesçi tahmin tekniği kullanılarak bilet fiyatları tahmin etmek için yaptıkları çalışmadır. Lantseva ve diğerleri (2015) ise bilet tahmini için deneysel veri odaklı bir Regresyon modeli öne sürmüşlerdir. Model, kalkış tarihinden önceki 90 gün içinde belirli bir uçuş için kilometre başına fiyatı tahmin etmektedir. 2015 baharında iki bağımsız çevrimiçi seyahat acentasından (AviaSales.com ve Sabre.com) toplanan verilere dayalı olarak yapılan çalışmada ulusal ve uluslararası uçuşlar kullanılmıştır. Bunlar Rusya’nın Moskova ve SaintPetersburg şehirlerinden 50 farklı şehre iç hat uçuşları ve yine bu iki şehirden 40 farklı uluslararası şehre yapılan dış hat uçuşları içermektedir. AviaSales’den 75 gün ve Sabre’den 90 gün boyunca günlük her uçuş için fiyat bilgisi toplandı. Modeli oluşturmak için kalkış şehri, varış noktası, bilet satın alma tarihi, kalkış tarihi, fiyatla birlikte bilet seçeneği bilgileri özellik olarak kullanıldı. Araştırmacılar

önerdikleri model ile erken bilet satın almanın ulusal ve uluslararası bilet fiyatları üzerindeki etkisini karşılaştırdılar. Buna göre erken bilet satın almanın uluslararası uçuşlarda bir avantaj olduğu, ulusal uçuşlarda ise somut sonuçlara varmak için ek araştırmalar gerektiği sonucuna vardılar. Bununla birlikte çalışmada kullanılan modelin performans değerlendirmesine yönelik bir çalışma yapılmadığı görülmektedir.

Li ve Li (2018) çalışmalarında bilet fiyatını tahmin etmek için iki farklı model kullandılar. İlk model ARMA ve Rastsal Orman algoritmasının kombünasyonudur. İkinci model ise sadece Rastsal Orman algoritmasına dayalıdır. Kombine edilen modelde ARMA algoritması bilet fiyatı için bir ön tahmin oluşturur. Daha sonra bu tahmin ile gerçek fiyat arasındaki hatayı tahmin ederek yeni bir tahmin oluşturmak için ise Rastsal orman algoritması çalıştırılır. Veri seti olarak 9 şehre ait 18 iç hat rotasının fiyat verilerini içerir. Bu veriler uçuştan 60 gün önceki fiyatları içermekte olup 12 saat aralıklarla düzenli olarak toplanmıştır. Oluşturdukları modelleri Akaike ve Bayesian (AIC-BIC) bilgi kriterlerine göre değerlendirdiler. Buna göre en küçük değerlere sahip model ARMA ve RF algortimaları kombünasyonundan oluşan modeldir.

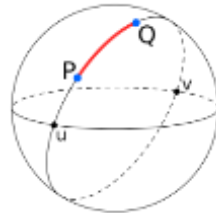
Bugün makine öğrenmesi modellerinin oluşturduğu topluluklar birden fazla alt modelin güçlü yönlerini (zayıf yönlerini telafi ederek) birleştirerek daha güçlü sonuçlar üretebilmektedir. Farklı yöntemleri bir arada kullanarak tahminleri iyileştirmek mümkün olabilmektedir. Bilet tahmini üzerine yapılan çalışmalar değerlendirildiğinde özellikle topluluk öğrenme algoritmaları ile geliştirilen modellerin diğerlerine göre daha başarılı oldukları gözlenmektedir (Keles vd., 2020).

Buyüzden bu çalışmada başarılı bir bilet tahmin modeli geliştirmek için bir çeşit topluluk öğrenme algoritması olan Gradyan Artırım (Gradient Boosting) ve Rastsal Orman (Random Forest) algortimaları tercih edilmiştir. Bu algoritmaları kullanarak geliştirdiğimiz modeller %90 ve üzeri yüksek doğruluk ve %3 den küçük ortalama mutlak yüzde hatasına sahiptir. Buna göre bu çalışmada geliştirdiğimiz modeller bilet tahmini için geliştirilen diğer modellerden çok daha başarılıdır.

Material ve Yöntem

Veri Seti ve Veri Önışleme

Çalışmada iki farklı veri seti kullanılmıştır ve bu veri setlerinin işlenip birleştirilmesi ile modellerin eğitiminde kullanılacak asıl veri seti elde edilmiştir. İlk veri seti uçuşlara ait bilgileri içermektedir. İkinci veri seti ise uçuşların gerçekleştiği hava limanlarının coğrafi koordinatlarını içermektedir. İlk veri setindeki “*origin*” ve “*destination*” özellikleri uçuşun gerçekleştiği havalimanlarına ait kodları içermektedir. Bu havalimanı kodları kullanarak ikinci veri setindeki coğrafi koordinatlara ulaşılmış, uçuş mesafeleri “Haversine” formülünden¹ faydalanılarak hesaplanmış ve elde edilen sonuçlar birinci veri setinde “*distance*” özelliğinde depolanmıştır (Şeki 1).



P (lat1,lon1)
Q (lat2,lon2)

$$distance(P, Q) = 6377.83 \times \arccos[(\sin(lat1) \sin(lat2) + \cos(lat1) \cos(lat2) \cos(lon2 - lon1))]$$

¹ https://en.wikipedia.org/wiki/Haversine_formula

Şekil 1. Bir uçuşa ait “kalkış-Origin” ve “Varış-destination” noktalarının enlem ve boylam bilgileri ile gösterilmesi

Hesaplama sırasında coğrafi koordinatların değerleri radyan cisine çevrilerek kullanılır. Koordinatlar (lat, long) radyan cinsinden verilmiştir. Aşağıda Esenboğa-ESB (0.700, 0.576) ve Dudlin-DUB (0.932, -0.109) havalimanı koordinatları kullanılarak mesafe yani “distance” özelliğinin nasıl hesap edildiği örnek olarak gösterilmiştir.

$$\begin{aligned} distance &= 6377.83 \\ &\quad * \arccos[(\sin(0.700) \sin(0.932) \\ &\quad + \cos(0.700) \cos(0.932) \cos(0.576 - (-0.109))] \end{aligned}$$

$$distance(ESB, DUB) = 279.368$$

Veri seti içerisinde bulunan “departure date” özeliğinden faydalanılarak uçuşa kalan günü ifade eden “remainingday”, uçuşun hafta sonuna denk gelip gelmediğini ifade eden “isweekend” ve uçuşun haftanın hangi günü olduğunu gösteren 6 farklı özellik (day1, day2,...,day6) oluşturulmuştur.

Ülkeler ve havayollarının uyguladıkları vergiler ve diğer birçok etmen uçuş fiyatlarına yansımaktadır. Bu sebeple uçuş mesafesi aynı olan farklı ülkeler arasındaki uçuşların fiyatları birbirinden oldukça farklıdır. Hatta bazı ülkeler arasında yapılan uçuşların fiyatları diğer ülkeler arasında yapılan uçuşlardan çok daha pahalı veya ucuz olduğunu söylemek mümkündür. Uçuşların gerçekleştiği ülkeler ve havalimanları düşünüldüğünde veri setine eklenmesi gereken yüzlerce yeni kategori oluşmaktadır. Bu kadar çok kategoriye veri setine eklemek hesaplama gücü açısından çok maliyetli olacağından hem de başarıyı olumsuz yönde etkileyebileceğinden kategorilere (örn: TR-DE Türkiye-Almanya) ait uçuşların fiyatlarının ortalaması alınarak “mean_of_ctg” özelliği oluşturulmuştur. Çok sayıdaki kategoriyle baş etmek için izlenen bu yöntem veri bilimi projelerinde ve yarışmalarında sıkça kullanılmaktadır.

Ayrıca veri setindeki aykırı veriler tespit edilerek temizlenmiştir. Veri setindeki numerik özelliklere de “Z-Score standardizasyonu” uygulanmıştır. Birçok makine öğrenmesi algoritmasının başarısını standardizasyon büyük ölçüde etkiler.

$$x_{standard} = \frac{x - \text{mean}(x)}{\text{standard deviation}(x)}$$

origin	destination	price	stops_count	isweekend	distance	remainingday	mean_of_ctg	day_1	day_2	day_3	day_4	day_5	day_6	is_holiday
HAJ	AYT	1142.9508	0.0	0	0.242518	0.515759	0.195599	0	0	1	0	0	0	0
IST	BOM	2127.9536	1.0	0	0.599911	0.100017	0.757346	0	0	0	0	1	0	0
CGK	KDI	712.1200	0.0	1	0.174983	0.289341	0.099878	0	0	0	0	0	0	1
SAW	MRY	640.9689	1.0	0	0.315586	0.223496	0.682610	0	0	0	1	0	0	0
AMS	ORY	258.2700	0.0	0	0.032067	0.123209	0.166673	0	0	0	1	0	0	1

Şekil.2 Ön işleme sonrası elde edilen veri tablosundan örnek kayıtlar

RF-FPPredictor ve GB-FPPredictor Modellerinin Geliştirilmesi

Öncelikle YZ ile uçuş fiyatı tahmini üzerine yapılan çalışmalar alanda tespit edilmiş ve detaylıca analiz edilmiştir. Elde ettiğimiz bulgular ışığında özellikle topluluk öğrenme algoritmalarının bireysel makine öğrenmesi algoritmalarından çok daha başarılı oldukları gözlenmiştir (Tziridis vd., 2017). Buyüzden bu çalışmada iyi sonuçlar verdiği gözlenen Rastsal Orman “Random Forest” ve Gradyan Artırım “Gradient Boosting” algoritmaları model geliştirmek için kullanılmıştır.

Gradyan Artırma (GB-Gradient Boosting) Algoritması

Gradyan Artırma, makine öğrenmesinde kullanılan bir tür topluluk algoritmasıdır. Bu algoritmanın temelleri Friedman ve arkadaşlarının yaptığı çalışmalara dayanmaktadır (Freund ve Schapire, 1997; Friedman vd., 2000; Friedman, 2001). Karar ağaçlarının en büyük dezavantajı basit ağaçlarda büyük bir “bias”, karmaşık ağaçlarda ise büyük bir varyans oluşturmalarıdır. “Bootstrap” yani torbalama bir veri kümesinden Rastsal sayıda veri seçme yöntemidir. Çoğunlukla ağacın varyansını azaltmak için kullanılır. Bu algoritmada genellikle karar ağaçları gibi zayıf öğrencilerin birbirlerinin zayıf yönlerini telafi ederek daha güçlü bir öğrencinin ortaya çıkarıldığı topluluk oluşturulur. Bu algoritma ile Makine öğrenimi ve veri madenciliği uygulamalarında oldukça başarılı sonuçlar elde edilmektedir (Bissacco vd., 2007; Hutchinson vd., 2011; Pittman ve Brown, 2011; Johnson ve Zhang, 2012).

Gradyan Artırma algoritmasına göre öncelikle ilk iterasyonda bir tahmin fonksiyonu oluşturulur. Tahminler ile gözlemler arasındaki fark hesaplanarak bu farklardan bir kayıp fonksiyonunu elde edilir. İkinci iterasyonda tahmin ve kayıp fonksiyonları birleştirilerek tekrar tahminler ile gözlemler arasındaki fark hesaplanır. Böylece sürekli üstüne ekleyerek tahmin fonksiyonunun başarısı artırılmaya çalışılır. Elde edilmek istenen tahminler ile gözlemler arasındaki farkın mümkün olduğunca sıfıra yaklaşmasıdır.

$$Loss (Kayıp Fonksiyonu) = MSE = \frac{1}{N} \sum (y_i - y_i^p)^2 \quad (1)$$

Burada y_i i. gözlenen değeri, y_i^p ise i. tahmin değerini, $L(y_i, y_i^p)$ kayıp fonksiyonunu ve α değeri ise öğrenme oranını temsil eder.

$$y_i^p = y_i^p + \alpha * \delta \sum \frac{(y_i - y_i^p)^2}{\delta y_i^p} \quad (2)$$

$$y_i^p = y_i^p - \alpha * 2 * \sum (y_i - y_i^p) \quad (3)$$

Rastsal Orman (RF-Random Forest) Algoritması

Rastsal Orman, karar ağaçlarına dayanan bir çeşit topluluk öğrenme algoritmasıdır. Makine öğreniminde karar ağaçları, tahmin modelleri oluşturan gözetimli öğrenme tekniğidir. Birden fazla karar ağacı ile karar ormanı oluşturur. Rastgele bir ormanda, her düğüm, o düğümde rastgele seçilen bir tahmin alt kümesi arasından en iyisi kullanılarak bölünür (Breiman, 2001). Algoritma hem sınıflandırma hem de regresyon problemlerinin çözümünde kullanılmaktadır. Regresyon problemlerinde veri setini daha küçük veri setlerine ayırarak birçok karar ağacı oluşturur. Regresyon probleminde verilen bir girdinin çıktısı bu karar ağaçlarının verdiği çıktılarının ortalamasıdır.

Modellerin Optimizasyonu ve Hiper Parametrelerin Belirlenmesi

Birçok hiper parametrenin alabileceği değer aralığı sonsuz veya çok geniştir. Ancak daha önce benzer problemlerden edinilen tecrübelerle dayanarak da hiper parametrelerin alabileceği değerler için makul bir aralık belirlemek mümkündür. Bu aralıkta bazı ana noktalar belirlenir ve bir hiper parametre için bir değer listesi oluşturulur. “Grid Search” metodu, belirlenen aralıktaki değerlerin kombünasyonlarını hiper parametrelere atayarak modeli eğitip sonuçlarını değerlendirir ve en iyi sonucu veren kombünasyonu bulur.

Bu çalışmada Gradient Boosting ve Random Forest modellerinin hiper parametreleri Grid Search metodu kullanılarak belirlenmiştir. GB-FPPredictor modelinin hiper parametrelerinin optimizasyonu için yazılan kodlarda kullanılan tüm değerler görülmektedir (Şekil 4). GridSearch metodunun optimizasyon için tercih edilmesinin en önemli sebeplerinden biri, bu metodun paralel işlemler koşturarak zamandan tasarruf sağlamasıdır.

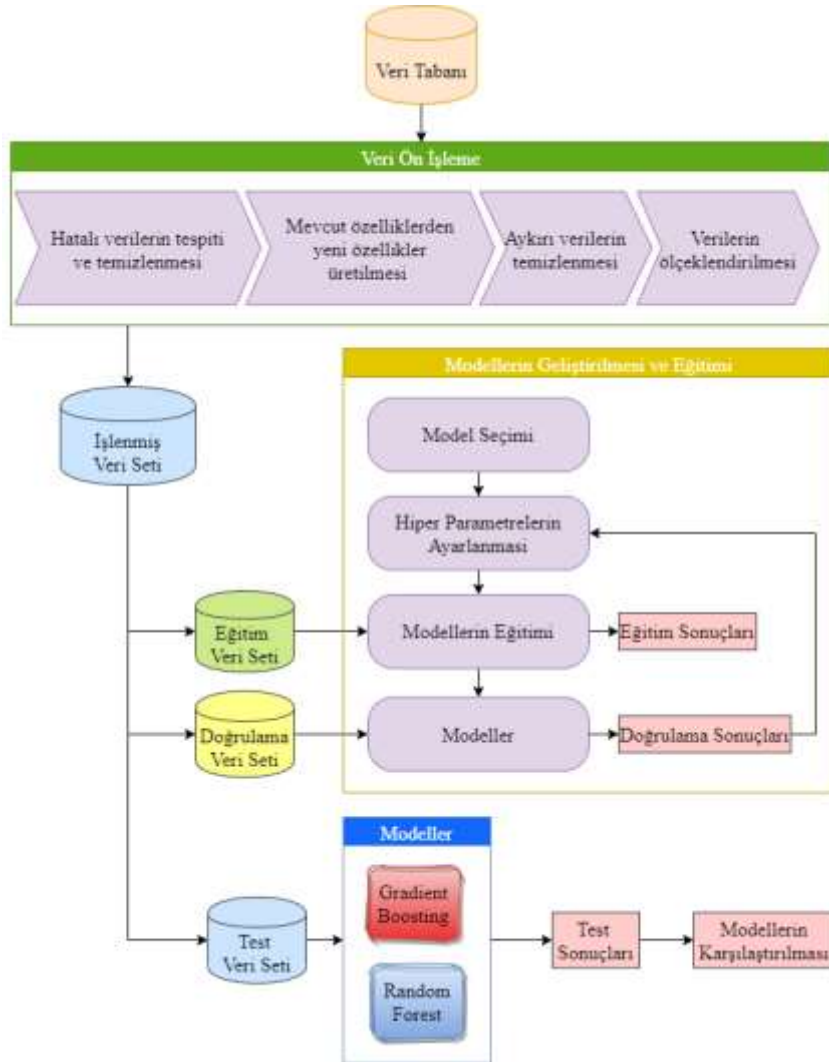
```

1 import matplotlib.pyplot as plt
2 import numpy as np
3 from sklearn.model_selection import shuffle_split, train_test_split
4 from sklearn.ensemble import GradientBoostingRegressor
5 from sklearn.model_selection import GridSearchCV
6 def GridSearchCVGradientBoostingRegressor(X,y):
7     X_train, X_test, y_train, y_test = train_test_split(X.values, y.values, random_state=0, test_size=0.3)
8     estimator = GradientBoostingRegressor(random_state=42, n_estimators=100, learning_rate=0.05, max_depth=8)
9     cv = shuffle_split(n_splits=5, random_state=42, test_size=0.3)
10    param_grid={
11        'subsample':[1.0]
12    }
13    n_jobs=8
14    regressor = GridSearchCV(estimator=estimator, cv=cv, param_grid=param_grid, n_jobs=n_jobs)
15    regressor.fit(X_train, y_train)
16    y_pred = regressor.best_estimator_.predict(X_test)
17    print(regressor.best_params_)
18    print(print_accuracy_report(y_test, y_pred, X_test, regressor.best_estimator_))
19    return y_test, y_pred

```

Şekil 4. GB-FPPredictor modelinin hiper parametrelerinin optimizasyonu için yazılan kodlar

Aşağıdaki diyagram (Şekil 3) bu çalışmanın metodolojisini ve RF-FPPredictor ve GB-FPPredictor modellerinin geliştirilme sürecini ve optimizasyonunu göstermektedir.



Şekil 3. RF-FPPredictor ve GB-FPPredictor modellerinin geliştirilme süreci

Tartışma ve Sonuç

Uçuş fiyatı tahmini üzerine yapılan çalışmalar dikkatle incelendiğinde pek çok makine öğrenmesi modelinin hem ayrı hem de farklı kombibnasyonlarla birlikte kullanıldığı gözlemlenmektedir. Bu çalışmalarda, başarılı performansları ile topluluk öğrenme algoritmaları özellikle öne çıkmaktadır. Bu bölümde, bu çalışmaların nasıl yapıldığı izah edilerek bir birleri ile karşılaştırmalı olarak değerlendirilmiştir (Table 1).

Chen ve diğerleri (2015) belirli bir seyahat rotası için en düşük fiyatı tahmin eden bir çeşit topluluk algoritması olan Learn++.NSE ile bir model geliştirdiler. Bu model mevcut gün-d1, rota-r ve kalkış tarihi- dn ile tanımlanan belirli bir seyahat programı (r, dn) verildiğinde, model ardışık (d2, d3 ...dn-1) günler için en düşük fiyatları tahmin edebilmektedir. Model sadece aktarmasız uçuşları dikkate alır. Learn ++. NSE, fiyat değişikliklerini geçmiş uçuş örüntülerinden aşamalı olarak öğrenmek ve gelecekteki fiyatları tahmin etmek için eğitilmiştir. Gelecekteki birçok fiyatı tahmin edebilmek için yinelemeli bir strateji kullanır. Modelde, rotanın günlük fiyatı, rotanın kalkış gününden önceki son fiyatları, rotanın haftanın aynı gününe ait fiyatları, rotanın ayın aynı gününe ait fiyatları özellik olarak kullanılmıştır. Model, 5 farklı uluslararası rotanın günlük fiyatlarını içeren bir veri seti ile test edilmiştir. Bu veri seti Çin'deki çevrimiçi bir seyahat acentasından toplanmıştır. Veri seti 11 Şubat - 01 Haziran 2015 tarihlerini kapsayan toplamda 110 günlük veriyi içermektedir. Her bir gün için sonraki 60 günde gerçekleşecek seyahatlerin en düşük fiyatları kaydedilmiştir. Her bir rota için 110x60 kayıt oluşturulmuştur. Learn++.NSE ile oluşturulan modelin ortalama mutlak yüzde hatası (MAPE) %10,7'dir.

Tziridis ve diğerleri (2017) uçak bileti fiyatlarını tahmin etmek için 8 çeşit makine öğrenmesi algoritması ile farklı modeller oluşturdular ve bu modellerin performanslarını karşılaştırdılar. Bununla birlikte uçak bileti fiyat tahmini üzerinde yüksek etkiye sahip olan faktörleri de belirlemeye çalıştılar. İncelenen faktörler arasında kalkış saati, varış saati, boş bagaj sayısı, kalkışa kalan gün sayısı, aktarma sayısı, tatil, günün saati ve haftanın günü yer almaktadır. Modeller, tek bir uluslararası rota için 1.814 uçuş bilgisinden oluşan bir veri seti kullanılarak eğitilmiştir. Çalışmada kullandıkları sekiz model ve elde ettikleri performanslar sırasıyla şöyledir: Çok Katmanlı Algılayıcı (MLP- *Multi Layer Perceptron*: %80.28), Genelleştirilmiş Regresyon Sinir Ağı (GRNN- *Generalized Regression Neural Network*: %66.83), Aşırı Öğrenme Makinesi (ELM- *Extreme Learning Machine*: %68.68), Rastsal Orman Regresyonu (RGR- *Random Forest Regression*: %85.91), Regresyon Ağacı (RT- *Regression Tree*: %84.13), Torbalama Regresyon Ağacı (BRT- *Bagging Regression Tree* %87.42), Destek Vektör Regresyonu (SVR-*Support Vector Regression* Polinomal: %77 ve SVR- *Support Vector Regression* Doğrusal: %49.40) ve Doğrusal Regresyon (LR- *Linear Regression*: %57.25). Buna göre en iyi öğrenici modeller topluluk öğrenme algoritmaları olan Torbalama Regresyon Ağaçları ve Rastsal Orman Regresyonu ile oluşturulmuştur.

Bu çalışmada kullanılan ver seti ulusal/uluslararası, aktarmalı/aktarmasız birçok farklı rotadan oluşan 200.000 uçuş kaydını barındırmaktadır. Her bir uçuş kaydı kalkış noktası, varış noktası, fiyat, havayolu, kalkış tarihi, varış tarihi, aktarma sayısı ve rota verisinden oluşmaktadır. Bilet tahmini için topluluk algoritmalarından Rastsal Orman (RF) ve Gradyan Artırma (GB) kullanılmıştır. Bu algoritmalar kullanılarak yeni tahmin modelleri geliştirilmiştir. RF ve GB algoritmaları ile geliştirilen RF-FPPredictor ve GB-FPPredictor modellerin doğruluk performansları sırasıyla %90 ve %92'dir. Modellerin ortalama mutlak yüzde hataları (MAPE) ise %2.49 ve %2.26'dır. Bilet tahmini için farklı veri setleri ve özellikler kullanılarak geliştirilmiş modeller ile bu çalışmada geliştirilen yeni modeller Tablo 1'de karşılaştırmalı olarak sunulmuştur. Buna göre bu çalışmada geliştirilen yeni modeller çok yüksek doğruluk ve düşük MAPE değeri ile oldukça başarılı görülmektedir.

Ayrıca uluslararası uçuşlarda uçuşa kalan gün sayısı azaldıkça bilet fiyatlarının arttığı, uçuş mesafesi arttıkça bilet fiyatının arttığı bununla birlikte mesafe aynı olsa bile farklı ülkeler arasındaki uçuşlarda fiyatın değiştiği gözlemlenmiştir.

Tablo 1. Topluluk Öğrenme Algoritmaları Kullanılarak Geliştirilen Uçuş Tahmin Modelleri ve performansları

Referanslar	Veri Seti	Özellikler	Uygulama Performansı
Rajankar & Sakharkar (2019)	Aktarmasız ulusal tek rota için 45 günlük uçuş kaydı	Kalkış noktası, varış noktası, kalkış tarihi, kalkış zamanı, varış zamanı, fiyat, havayolu, bilet alışı tarihi	RF: %67 (Doğruluk) GB: %47 (Doğruluk)
Vu vd. (2018)	3 ulusal hava yoluna ait haftanın 7 günü için aktarmasız yapılan 51.000 uçuş kaydı	Havayolu, uçuş numarası, satın alma tarihi, kalkış tarihi, kalkış zamanı, varış zamanı, ücret sınıfı, aktarma sayısı, fiyat, kalkış havalimanı, varış havalimanı, varış tarihi	Stacked Predict Model (KNN&RF&Bayesian): %44 (Doğruluk) RFT&MLP: 77% (Doğruluk)
Tziridis vd. (2017)	Tek uluslararası rota için 814 uçuş kaydı	Kalkış zamanı, varış zamanı, bedava bagaj sayısı, uçuşa kalan gün sayısı, aktarma sayısı, tatil günü (Evet/hayır), gece uçuşu (Evet/hayır), haftanın günü	RF : %85.91 (Doğruluk) BR: %87.42 (Doğruluk)
Liu vd.(2017)	19 farklı rotadan oluşan 92 günlük uçuş kayıtları	Geçmiş bilet fiyatları, kalkış tarihi tatil günü mü (Evet/Hayır), kalkıştan önceli gün sayısı	KNN&RF&Bayesian: %7 (MAPE)
Chen vd.(2015)	5 farklı uluslararası rotanın 110 günlük verisi	Rotanın günlük fiyatı, rotanın uçuş günden önceki son fiyatları, rotanın haftanın aynı gününe ait fiyatları, rotanın ayın aynı gününe ait fiyatları	Learn++.NSE (MAPE): %10.7
Bu çalışmada	Ulusal/uluslararası, aktarmalı/aktarmasız birçok farklı rotadan oluşan 200.000 uçuş	Kalkış noktası, varış noktası, fiyat, havayolu, kalkış tarihi, varış tarihi, aktarma sayısı ve rota	RF (Doğruluk/MAPE): %92/ %2.49 GB (Doğruluk/MAPE): %90/ %2.26

Bu bulgular ışığında geliştirilen güçlü tahmin modelleri (RF-FPPredictor ve BG-FPPredictor) enuygun.com Ar-Ge Merkezi çalışmaları kapsamında sisteme entegre edilerek değişken uçuş fiyatlarının başarılı bir şekilde tahmin edilmesi mümkündür. Sonuç olarak pek çok faktörün üzerinde etkin olduğu dinamik ve bir o kadar değişken yapıya sahip uçuş fiyatlarının tahmin edilebilmesi hem ekonomik hem de stratejik anlamda havayolu sektörüne önemli bir katkı sağlayacaktır. Bu tür YZ destekli çalışmaların hava yolu şirketleri ve çevrim içi seyahat acentalarında hayata geçirilmesi şirketlerin gelir maksimizasyonuna büyük fayda sağlayacaktır. Diğer taraftan müşterilerin de çok daha uygun fiyata bilet satın alarak tasarruf etmelerine yardımcı olacaktır.

Kaynakça

- An, B., Chen, H., Park, N. & Subrahmanian, V. S. (2016). MAP: Frequency-based maximization of airline profits based on an ensemble forecasting approach. In *KDD 2016 - Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 421-430. <http://doi.org/10.1145/2939672.2939726>
- Bissacco, A., Yang, M.-H. & Soatto, S. (2007). Fast human pose estimation using appearance and motion via multi-dimensional boosting regression, *IEEE Conference on Computer Vision and Pattern Recognition, CVPR'07*. <http://doi.org/10.1109/CVPR.2007.383129>

-
- Breiman, L.(2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Boruah, A., Baruah, K., Das, B., Das, M.J. & Gohain, N.B. (2018). A Bayesian Approach for Flight Fare Prediction Based on Kalman Filter. In: Progress in Advanced Computing and Intelligent Engineering. *Advances in Intelligent Systems and Computing*, Springer, 714, 191-203.
- Chen, Y., Cao, J., Feng, S. & Tan, Y. (2015). An ensemble learning based approach for building airfare forecast service. In: *2015 IEEE International Conference on Big Data (Big Data)*, Santa Clara, CA, 2015, 964-969.
- Freund, Y., & Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst.* 55, 119–139.
- Friedman, J., Hastie, T. & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. *Ann. Stat.* 28, 337-407. <http://doi: 10.1214/aos/1016218222>
- Friedman, J. (2001). Greedy boosting approximation: a gradient boosting machine. *Ann. Stat.* 29, 1189–1232. <https://doi: 10.1214/aos/1013203451>
- Hutchinson, R. A., Liu, L.-P. & Dietterich, T. G. (2011). Incorporating boosted regression trees into ecological latent variable models, *AAAI'11,(San Francisco, CA)*, 1343-1348, <http://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/view/3711>
- Johnson, R. & Zhang, T. (2012). Learning Nonlinear Functions Using Regularized Greedy Forest. Technical Report. arXiv:1109.0887. <http://doi: 10.2172/1052139>
- Keles M.B., Keles A. & Keles A. (2020). Makine öğrenmesi yöntemleri ile uçuş fiyatlarının tahmini, *Euroasia Journal of Mathematics, Engineering, Natural & Medical Sciences*, 7(11), 72-78. <https://doi.org/10.38065/euroasiaorg.237>
- Lantseva, A., Mukhina, K., Nikishova, A., Ivanov, S. & Knyazkov, K. (2015). Data-driven Modeling of Airlines Pricing. *Procedia Comput. Sci.* 66, 267–276.
- Li, Y. & Li, Z. (2018). Design and Implementation of Ticket Price Forecasting System, *AIP Conference Proceedings*, 1967, 040009. <http://doi.org/10.1063/1.5039083>
- Liu, J., Liu, B., Liu, Y., Chen, H., Feng, L., Xiong, H. & Huang, Y. (2017). Personalized Air Travel Prediction. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9, 1-26, <http://doi.org/10.1145/3078845>
- Pittman, S. J. & Brown, K. A. (2011). Multi-scale approach for predicting fish species distributions across coral reef seascapes. *PLoS ONE* 6:e20583. <http://doi: 10.1371/journal.pone.0020583>
- Rajankar, S. & Sakharkar, N. (2019). A Survey on Flight Pricing Prediction using Machine Learning, *International Journal Of Engineering Research & Technology (Ijert)*, 8(6), 1281-1284.
- Smith, B.C., Leimkuhler, J.F. & Darrow, R.M. (1992). Yield management at american airlines, *INFORMS Journal on Applied Analytics*, 22(1), 8–31. <http://doi.org/10.1287/inte.22.1.8>
- Tziridis, K., Kalampokas, T., Papakostas, G.A. & Diamantaras, K.I. (2017). Airfare prices prediction using machine learning techniques, *25th European Signal Processing Conference (EUSIPCO)*. Kos 2017, 1036–1039.
- Vu, V.H., Minh, Q.T. & Phung, P.H. (2018). An airfare prediction model for developing markets. In: *2018 International Conference on Information Networking (ICOIN)*, Chiang Mai, 765-770.
-