VERİ BİLİMİ DERGİSİ
www.dergipark.gov.tr/veri

An Approach for Airfare Prices Analysis with Penalized Regression Methods

Selim BUYRUKOĞLU [1]*, Yıldıran YILMAZ[2]

[1]*Çankırı Karatekin University, Computer Engineering, Çankırı*
[2]*Recep Tayyip Erdogan University, Computer Engineering, Rize*

Abstract

At present, the number of passengers preferring to use the airline is increasing with each passing day. Thus, correctly analysing the airfare prices is essential to raise awareness of passengers. Some researchers have applied different kinds of Machine Learning (ML) algorithms to predict the airfare prices. However, to the best of our knowledge, penalized regression methods have not been used to analyse the airfare prices. Ridge, Lasso, and Elastic Net regressions are penalized regression methods. The dataset used in this study consists of 1814 one-way flights from Greece to Germany. The developed Ridge, Lasso, and Elastic Net methods were achieved to provide convincing results for airfare prices analysis based on Mean Squared Error-MSE values (Ridge:160103, Lasso:159280, Elastic Net:174203) and Mean Absolute Error-MAE values (Ridge:147.74, Lasso:146.43, Elastic Net:346.86). MSE and MAE explain how close a regression line is to a set of points. They take the distances from the points to the regression line (refers to "errors"). MSE takes the squares of the errors, MAE takes the absolutes of the errors. The lower they are, the better the prediction is. Thus, in our case, Lasso regression can be considered better than the ridge and elastic net due to the lowest MSE and MAE values. In other words, the results and findings reveal that the proposed Lasso method is potentially better than the others in the analysis of datasets consisting of one-way flights.

Article Info

Received
05/07/2021
Accepted
27/07/2021

Keywords: *Airfare Price, Analysis Model, Penalised Regression Methods*

[1] Corresponding author: sbuyrukoglu@karatekin.edu.tr

** Part of this work was presented orally at the IV. International Conference on Data Science and Applications 2021.

1 Introduction

Airline companies aim to maximize their profitability, and therefore the companies prefer to use optimisation modeling systems [1]. This means that airfare prices are determined dynamically, and it is difficult to find a ticket at the lowest price. In addition, airfare companies desire to maximize the profitability of one passenger while lowering the per capita costs. The lowest fare per person for an airline company is when all seats on a flight are occupied [2]. As a result, the companies are trying to sell all the seats and not give any more discounts than they should be given.

If a passenger is going to travel at a busy time like holidays, he/she has to buy or book the ticket early, otherwise, his/her chances of finding cheap flights will be very low [1]. For this reason, it is not easy to buy low-cost flights. In literature, several studies have focused on detecting the right time to buy low-cost flights **[3, 4, 5].** Machine Learning (ML) algorithms have been used in the majority of these studies.

Studies about the airfare price prediction have highlighted the importance of features while buying cheap airfare tickets. Several feature selection techniques and ML algorithms have been applied in the airfare price prediction. More detailed information about the related studies is presented in Section 2. To the best of our knowledge, penalized regression methods (Ridge, Lasso, and Elastic Net) have never been used in the airfare price prediction. One of the advantages of the penalized regression methods is that they set coefficients for input features. It means that the importance of the features is presented based on the prediction of the target feature [6]. For instance, lasso regression can be considered as a feature selection approach since it sets the coefficients of unrelated features to zero. In contrast to lasso regression, ridge regression does not subtract unrelated variables while bringing the coefficients closer to zero [7]. Thus, the goal of this study is to compare the performance of penalized regression methods for airfare price prediction.

The structure of the paper is as follows: Section 2 presents related works in the field. Section 3 provides information about the dataset. Section 4 gives information about the background of linear regression and penalized regression methods. Section 5 includes the evaluation of the penalized regression methods. Section 6 concludes the overall study and outlines further work.

2 Related Work

There are many kinds of research that have been developed to predict airfare prices and to reveal the most effective features in the prediction of airfare prices. In a study [5], machine learning (ML) algorithms have been used to analyze airfare prices. The ML algorithms employed in [5] are Multilayer Perceptron (MLP), Random Forest (RF), Linear Regression (LR), Support Vector Machines (SVM), Bagging Regression Tree. The best accuracy rate (87.93%) has been obtained through the bagging regression tree. Five hundred decision trees were used in the development process of the model, and also 10-fold cross-validation was applied in the evaluation process. In addition, different feature combinations have been tried to determine the optimal features. In the end, the result without the overnight flight feature provided the best result compared to other combinations. It means that the overnight flight feature can be considered the least effective feature. In a different study, different ML algorithms have been applied for airfare price prediction [8]. Algorithms' performances were compared with and without a feature selection approach. Root Mean Squared Error (RMSE) and R^2 adjusted evaluation metrics were used in the study. The best result is obtained applying the RF with feature selection approach (RMSE = 62.75, R^2 adjusted = 0.869). It is noted that the RF algorithm again provided the best performance compared to other ML models without feature selection (RMSE = 66.58, R^2 adjusted = 0.858). Lu, J. [9] proposed a method to provide solutions for time series problems applying ML algorithms for the prediction of airfare prices. AdaBoost-Decision Tree Classification provided the best performance compared to other ML algorithms. In a study [10], a model was proposed to determine the best time in terms of booking a flight. SVM, k-Nearest Neighbor (kNN), RF, and Logistic Regression algorithms were implemented in the study. The results show that the RF algorithm provided the best performance compared to others with a 95.2% test accuracy rate. The RF model was created with eighteen trees. On the other hand, kNN provided the second-best performance (91.3% test accuracy rate) which is fairly similar to the performance result of RF. In another study, the PLS regression model was proposed to predict airfare prices, and this model achieved 75.3% accuracy rate [11]. Also, Papadakis

[12] developed a model based on ML algorithms to predict the ticket price that is likely to drop in the future. Ripple Down Rule Learner presented the best performance (74.5% accuracy) when compared to the other ML algorithms used in the study. Furthermore, performance comparison of LR, Naïve Bayes, Softmax Regression, and SVM models are carried out in the prediction of airfare prices. The best performance is obtained through SVM with an 80.6% accuracy rate.

All the aforementioned studies proposed different ML models for the prediction of airfare prices. Even if they achieved promising results, to the best of our knowledge, penalized regression methods have never been applied to predict airfare prices, and the prediction problem is still unexplored. Therefore, the contribution of this study is the performance comparison of penalized regression models in airfare price prediction.

3 Dataset

The dataset used in this research consists of nine features for each flight. The features are listed in Table 1. It consists of 1814 one-way flights of Aegean Airlines from Thessaloniki (Greece) to Stuttgart (Germany). This dataset is obtained from GitHub [13].

Table 1. Information about features

Features	Feature Type
Departure time	Numeric
Arrival Time	Numeric
Number of free luggage	Numeric
Days left until departure	Numeric
Number of intermediate stops	Numeric
Holiday day	Categorical
Overnight flight	Categorical
Day of week	Numeric
Price	Numeric

4 Penalized Regression Methods

This section describes the proposed method for the airfare price prediction problem.

4.1 Linear Regression

There are two types of linear regressions, namely simple and multiple linear regression. Simple linear regression can be considered as an effective model to predict a response based on a single predictor. In contrast to simple linear regression, multiple linear regression is an effective model if a response is predicted based on more than one predictor [14]. This study is aimed to predict airfare prices with more than one predictor, and so the multiple linear regression function is explained as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \qquad (1)$$

where the jth predictor is outlined by X_j, and also an association between a variable and the response is measured by β_j.

4.2 Penalized Regression Methods

4.2.1 Ridge Regression

Hoerl and Kennard proposed the ridge regression in 1970, and it aims to find the coefficients that minimise the error sum of squares by applying a penalty to these coefficients. Ridge regression function is explained as follows:

$$PR_{Ridge} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{P}\beta_j^2 \qquad (2)$$

where $\lambda \geq 0$ is a tuning parameter, and the term $\lambda\sum_{j=1}^{P}\beta_j^2$ is called a shrinkage penalty. It is resistant to overfitting and provides a solution to multidimensionality. It establishes a model using all parameters, and it does not subtract unrelated variables while bringing the coefficients closer to zero. Thus, it is necessary to find a good value for alpha (penalty) in setting up the model process [7]. Also, using ridge regression is not an advantage if $\lambda = 0$.

4.2.2 Lasso Regression

It makes both variable selection and regularization to increase the accuracy and interpretability of the produced statistical model. It aims to find the coefficients that minimise the sum of squared error by applying penalties to the coefficients [7]. Lasso regression function is explained as follows:

$$PR_{Lasso} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{P}|\beta_j| \qquad (3)$$

In contrast to ridge regression, it sets the coefficients of unrelated variables to zero. In other words, the term β_j^2 in Eq. (2) has been replaced by $|\beta j|$ in Eq. (3), lasso regression.

4.2.3 Elastic Net Regression

The goal of the elastic net is the same as with lasso and ridge regression. Elastic net combines the ridge regression and lasso. In the combination process,

the punishment style of the elastic net is the same as ridge regression while the variable selection style of it is the same as lasso regression [15]. Besides, it has considerable computational advantages over ridge and lasso. The elastic regression function is explained as follows:

$$PR_{Elastic} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda_1\sum_{j=1}^{P}|\beta_j| + \lambda_2\sum_{j=1}^{P}\beta_j^2 \quad (4)$$

4.3 Performance Evaluation

4.3.1 Mean Squared Error (MSE)

The dataset used in this study consists of continuous variables. Mean Squared Error (MSE) measures the average of the squares of the errors. In other words, it refers to the average squared difference between the estimated and the actual value [16]. MSE function is explained as follows.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(P_i - \widehat{P}_i\right)^2 \quad (5)$$

where P_i is the actual value, $\widehat{P}_i$ is the predicted value from the model and n is the number of observations.

4.3.2 K-fold Cross-Validation

The reason behind using the k-fold cross-validation (CV) is that it allows us to see whether the high performance of the model is random [17]. Dataset is randomly divided into k folds which are the nearly same size. k-1 subsets are used to train the data, and the remaining last subset is used as test data. The average error value obtained as a result of k experiments indicates the validity of the model. The k value is usually chosen as 3 or 5. In addition, the k value can be chosen as 10 or 15, but this may cause a very expensive calculation and waste of time. In this study, 10-fold cross-validation is used.

5 Results and Discussion

This section provides the experimental results and discusses the performance of three regression methods such as Ridge, Lasso and Elastic Net Regression considering Mean Squared Error as an evaluation metric.

5.1 Results on the Ridge Regression

In the case of the large alpha value used in the ridge regression, the coefficient is expected to be much smaller compared to the usage of a small alpha value. As can be seen from Figure 1, the coefficient decreases as alpha increases.

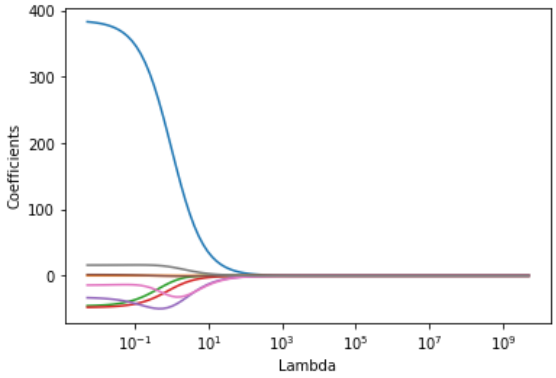

Figure 1. Ridge Plot

5.2 Results on the Lasso Regression

Figure 2 shows the lasso plot and some of the coefficients exactly equal to zero as alpha increases. Two features (Overnight and Arrival Time) are subtracted based on the lasso regression which means that their coefficients are exactly zero.

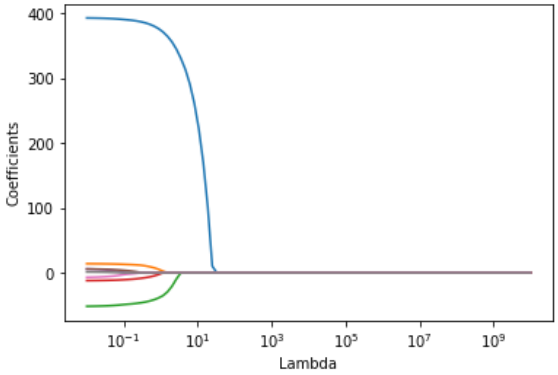

Figure 2. Lasso Plot

5.3 Results on the Elastic Net Regression

Figure 3 shows the Elastic Net plot. Four coefficients are estimated as very close to zero including 'Days until departure', 'Intermediate Stops', 'Day of Week', and 'Arrival Time'.

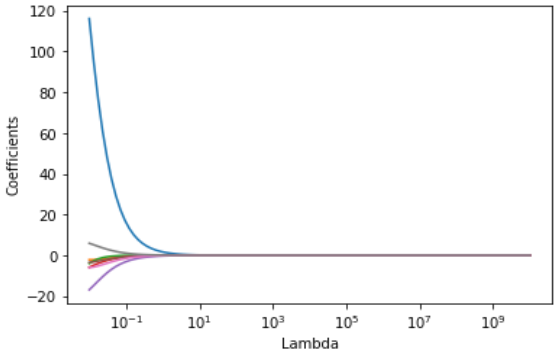

Figure 3. Elastic Net Plot

5.4 Overall Interpretation of the Results

Performance comparison of penalized regression methods is presented based on Mean Squared Error (MSE) and Mean Absolute Error MAE in the evaluation process of the methods. The comparison is presented in Table 2. Lasso regression method achieved to provide the best MSE rate. On the other hand, the worst MSE rate was obtained by the elastic net. Furthermore, MAE Lasso regression has the best MAE rate in predicting airfare price.

Table 2. Performance Results of Methods

Method	MSE	MAE
Ridge Regression	160103	147.74
Lasso Regression	159280	146.43
Elastic Regression	174203	346.81

The results indicate that lasso regression can be considered as the more effective method than the ridge and elastic net in the prediction of airfare prices.

6 Conclusion and Further Work

Performance of penalized regression methods is compared in the prediction of airfare price. The dataset has 1814 one-way flights of Aegean Airlines from Thessaloniki to Stuttgart. The results show that ridge and lasso regression methods provided almost similar performance based on MSE. However, the best performance about the prediction of airfare price prediction is obtained through the lasso regression. Two unrelated features are subtracted based on the lasso regression namely, overnight and arrival time. In the future, airfare prices can be predicted based on larger airfare datasets. Besides, deep learning algorithms can be applied to predict the airfare price prediction [18].

References

[1] Abdella, J. A., Zaki, N., Shuaib, K., & Khan, F.. "Airline ticket price and demand prediction: A survey." *Journal of King Saud University-Computer and Information Sciences*. (2019)

[2] Szabo, S., Mako, S., Tobisova, A., Hanak, P., & Pilat, M. "Effect of the load factor on the ticket price". *Transport problems*, 13. (2018).

[3] Groves, W., & Gini, M. "A regression model for predicting optimal purchase timing for airline tickets," Technical Report 11-025, University of Minnesota, Minneapolis, 2011.

[4] Groves, W., & Gini, M. "An agent for optimizing airline ticket purchasing," *12th International Conference on Autonomous Agents and Multiagent Systems*, St. Paul, MN, pp. 1341-1342, May 06 - 10, 2013.

[5] Tziridis, K., Kalampokas, T., Papakostas, G. A., & Diamantaras, K. I. Airfare prices prediction using machine learning techniques. *25th European Signal Processing Conference* (EUSIPCO), IEEE, 1036-1039, 2017.

[6] Ajana, S., Acar, N., Bretillon, L., Hejblum, B. P., Jacqmin-Gadda, H., & Delcourt, C. "Benefits of dimension reduction in penalized regression methods for high-dimensional grouped data: a case study in low sample size". *Bioinformatics*, 35(19), 3628-3634, 2019.

[7] James, G., Witten, D., Hastie, T., & Tibshirani, R. "An introduction to statistical learning", Vol. 112, p. 18,. New York: springer, 2013.

[8] Wang, T., Pouyanfar, S., Tian, H., Tao, Y., Alonso, M., Luis, S., & Chen, S. C.. "A framework for airfare price prediction: A machine learning approach". *20th International Conference on Information Reuse and Integration for Data Science (IRI)*, IEEE, pp. 200-207, 2019.

[9] Lu, J. "Machine learning modeling for time series problem: Predicting flight ticket prices". *arXiv preprint arXiv:1705.07205*, 2017.

[10] Ren, Q. When to Book: Predicting Flight Pricing. *Standford university*.

[11] Groves, W., & Gini, M. "An agent for optimizing airline ticket purchasing". In Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems, 1341-1342, 2013

[12] M. Papadakis, "Predicting Airfare Prices," 2012.

[13] https://github.com/humain-lab/airfare prediction. (14.04.2021).

[14] Uzut O.G., Buyrukoglu S. "Prediction of real estate prices with data mining algorithms". *Euroasia Journal of Mathematics, Engineering, Natural and Medical Sciences*. 2020;77-84. 2020, https://doi.org/10.38065/euroasiaorg.81

[15] Zhang, Z., Lai, Z., Xu, Y., Shao, L., Wu, J., & Xie, G. S. "Discriminative elastic-net regularized linear regression". *IEEE Transactions on Image Processing,* 26(3), 1466-1481, 2017.

[16] Muniz, G., & Kibria, B. G. "On some ridge regression estimators: An empirical comparisons". *Communications in Statistics—Simulation and Computation*, 38(3), 621-630, 2009.

[17] Fushiki, T. (2011). "Estimation of prediction error by using K-fold cross-validation". *Statistics and Computing,* 21(2), 137-146, 2011

[18] Zerman, M., Bulut, F., (2021). Büyük Verilerde Birliktelik Kurali İle Satiş Yönetimi: Havaalani Örneği, Mühendislik Alaninda Araştirma Ve Değerlendirmeler, Cilt 2, Sayfalar 113-146, Gece Kitaplığı, ISBN: 978-625-7342-73-5