LogForum

> Scientific Journal of Logistics <

2022, 18 (1), 97-107

http://doi.org/10.17270/J.LOG.2022.655

http://www.logforum.net p-ISSN 1895-2038 e-ISSN 1734-459X

ORIGINAL PAPER

FLIGHT DELAY PREDICTION BASED WITH MACHINE LEARNING

Irmak Hatıpoğlu[1], Ömür Tosun[2], Nedret Tosun[3]

1) Department of International Trade and Logistics, Faculty of Applied Sciences, Antalya, **Turkey**
2) Department of Management Information Systems, Faculty of Applied Sciences, Akdeniz University, Antalya, **Turkey**.
3) West Mediterranean Exportaters Association, Antalya, **Turkey**

ABSTRACT. Background: The delay of a planned flight causes many undesirable situations such as cost, customer satisfaction, environmental pollution. There is only one way to prevent these problems before they occur, and that is to know which flights will be delayed. The aim of this study is to predict delayed flights. For this, the use of machine learning techniques, which have become widespread with the development of computer capacities and data storage systems, is preferred.

Methods: Estimations are made with three up-to-date techniques XGBoost, LightGBM, and CatBoost techniques based on Gradient Boosting from machine learning techniques. The bayesian technique is used for hyper-parameter settings. In addition, the Synthetic Minority Over-Sampling Technique (SMOTE) technique is also used, as the majority of flights are on time and delayed flights, which constitute a minority class, may adversely affect the results. The results are analyzed and shared with and without SMOTE.

Results: As a consequence of the application, which was run on a data set containing all of an international airline's flights [18148 flights] for a year, it was discovered that flights may be predicted with high accuracy.

Conclusions: The application of machine learning techniques to anticipate flight delays is new, but it has a lot of potential. Companies will be able to avert problems before they develop if delays are correctly estimated, which can generate plenty of issues. As a result, concrete advantages such as lower costs and higher customer satisfaction will emerge. Improvements will be made at the most vulnerable place in the aviation business.

Keywords: GBDT, XGBoost, LightGBM, Catboost, delay prediction

INTRODUCTION

Except for the extraordinary pandemic situation in 2020, the number of global air traffic passengers has increased since 2006 [IATA, 2019]. While there was no case of an emergency in the first three months of 2020, a similar rise in demand was seen [Mazareanu, 2020]. According to NEXTOR [2010], the cost of delayed flights to airlines is $ 8.3 billion, and the cost of passengers due to delayed, canceled, or missed connecting flights is calculated as $ 9.4 billion in total.

Flight delays cause economic losses in the US of approximately 600 million $ per year, which has been estimated from passenger time and fuel consumption [Nakornsri, Apivatanagul, & Pisitkasem, 2020]. Environmental damage caused by delays is another negative impact that cannot be assessed as clearly as the cost. However, the additional fuel used during the delay and the resulting gas emissions have a negative impact on the environment [Simić & Babić, 2015; Dray, Antony, Vera-Morales, Reynolds, & Schafer, 2008]. Moreover, airline schedule success on time is a crucial factor in sustaining existing customer loyalty and attracting new customers [Abdelghany, Shah, Raina, & Abdelghany, 2004; Efthymiou, Njoya, Lo, Papatheodorou, & Randall, 2019].

Detailed information on flight timings is kept, as punctuality is of great importance in many different aspects of the entire airline industry. The aim of this study is to determine

Citation: Hatıpoğlu I., Tosun Ö., Tosun N., 2022. Flight delay prediction based with machine learning. LogForum 18 (1), 97-107, http://doi.org/10.17270/J.LOG.2022.655

Received: 10.09.2021, Accepted: 30.11.2021, on-line: 01.03.2022

the flights that will be delayed by examining the big data analysis methods by using various machine learning approaches in order to prevent the aforementioned problems. The big data used belongs to a private airline company and consists of all flights throughout the year. In addition to the flight data, the weather data of the time period closest to the flight was also added to the data set. After the initial tests, to prevent the imbalanced distribution of the dataset, the obtained data set was balanced with SMOTE and then the status of the flights was estimated using LightGBM, XGBoost, and Catboost methods.

The study is motivated by the need for information for commercial airline companies to overcome mentioned negative effects of delayed flights. In addition to this need, studies on flight delay topic are mostly built on the Bureau of Transportation Statistics. This study also showed that the data of individual airline companies is also used well. Another contribution is to find the most suitable solution using the three different gradient-based machine learning methods introduced by the creators as the most high-performance models.

Literature Review on Machine Learning and Flight Delays

Machine learning and flights delay topic are started to be studied not long ago. One of the reasons for this is the development of machine learning methods and the fact that big data operations can be done easily with machine learning. Since the topic has vital importance in air traffic control, airline decision-making processes and ground operations have been studied from various perspectives. The very first study is done by Choi et al. [2016]. Flight delays that are caused due to weather are forecasted with the domestic flight data and the weather data is used from 2005 to 2015. Kim et al. [2016] also worked on a two-step machine learning model with Recurrent Neural Networks and Long Short Term Memory. The model predicts whether the aircraft will be delayed, and then gives results on how long they will be late within 15 minutes of categorical time periods. Data of ten major airports in the U.S. have been collected and the model's accuracy is between 85% and 92%. Another 2-step model is built by Thiagarajan et al. [2017]. Their model predicts whether the delay will happen or not at the first step and then the delay's time duration is determined. For the first step Random Forests, Gradient Boosting, AdaBoost, Extra-Trees, LOYCVKi, and K-Fold CV are tried and Gradient Boosting is preferred and for the next step Gradient Boosting, MLP, Random Forests, Extra-Trees are tried and Extra-trees are preferred. The US Domestic Airline On-Time Performance data and weather data [World Weather Online API] from the year 2012 to 2016 are used. Manna et al. [2017] build a model to determine the delay time of aircraft. They used Gradient Boosted Decision Tree and find out that the method gives outstanding accuracy results with Coefficient of Determination of 92.31% for arrival delays and 94.85% for departure delays. The busiest 70 airports in the USA within April-November 2013 time period of the US domestic airline data used. Similarly, a comparison of methods is done by Kuhn and Jamadagi [2017]. Decision tree, logistic regression, and artificial neural network models are developed. The results of the models are compared and it is shown that they are almost the same. The data set is collected from Kaggle. Another study which is again based on the Kaggle dataset is done by Venkatesh et al. [2017]. Artificial neural network and deep belief network are used for forecasting whether the aircraft is on time or late. The model gives 92% accuracy. Modammed et al. [2018] are also compare models, but their study focuses on decision tree classifiers which are REPTree, Forecast, Stump, and J48. The data used is gathered from Egypt Airlines and the best results are found with REPTree model. Yu et al. [2019] focused on average delays between Beijing Capital Airport and Hangzhou Xiaoshan International Airport. They used Novel Deep Belief Network model to find a mean of delays. McCarthy et al. are focused on flight delays with a different perspective. They analyze the two European low-cost airlines. Delays that are less than 15 minutes are predicted with Long Short-Term Memory (LSTM). Chen and Li [2019] are focused on delays for connected flights. Bureau of Transportation Statistics is used with Multi-Label Random Forest and approximated delay propagation model.

METHODOLOGY

In this section, proposed approaches are introduced. Firstly, the schematic flowchart is provided in Fig. 1 which presents the whole process flow of the study. As this figure depicts, the dataset consists of flight and weather datasets that are preprocessed, cleaned, and merged. Since the data set was unbalanced, the observations allocated for training are also passed through a preliminary stage using SMOTE, and then three different methods are implemented.

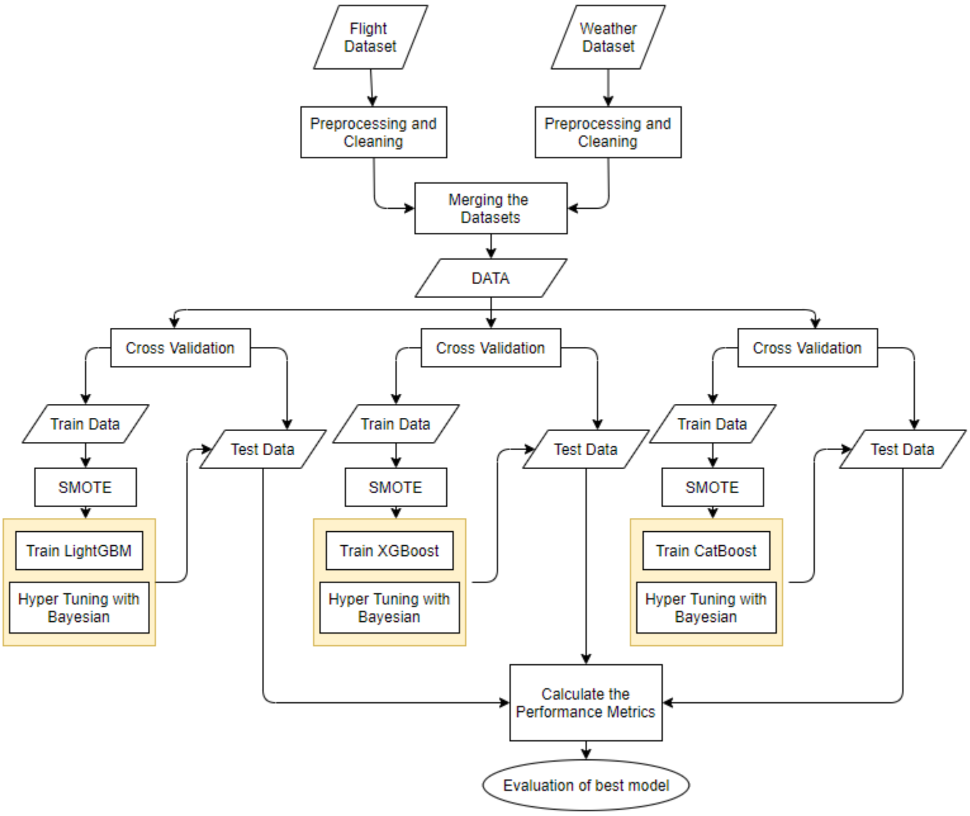

Fig. 1 Flowcharts of the proposed approaches

Imbalanced data handling

Even though machine learning techniques are well-built models to be applied to classification or regression problems, it is still very difficult to classify imbalanced data sets [Haixiang, Yijing, Shang, Mingyun, & Yuanyue, 2017], because imbalanced data sets could result in lower performance of the learners [He & Garcia, 2009]. Imbalanced distribution of the classes not only lowers the performance of the given model but also the model could focus on majority class to accurately predict or categorize, while the minority class is overlooked [López, Fernández, García, Palade, & Herrera, 2013]. Since minority class is ignored, performance metrics could result in misleading results [Loyola-González, Martínez-Trinidad, Carrasco-Ochoa, & García-Borroto, 2016]. The flight data set is also imbalanced since the delayed flights are 0.46 times lower than the on-time flights. In order to overcome this problem different suggestions are given in the literature. Two main categories of them are cost-sensitive learning approach and data preprocessing techniques [Hsixiang, Yijing, Shang, Mingyun, & Yuanyue, 2017]. Cost-sensitive models are aimed to solve this problem by giving higher penalties for the misclassifications of minority observations. The preprocessing techniques are often applied before the learning process and with the help of this better learning is aimed. Since the resampling methods are more useful and popular for this study Synthetic Minority Over-Sampling Technique [SMOTE] is preferred. The method is a useful way to generate synthetic minority and also SMOTE method is practiced in various problems with good results [Chawla, Bowyer, Hall, & Kegelmeyer, 2002].

Gradient boosting decision tree

Gradient boosting decision tree (GBDT) is a machine learning method that is applied frequently [Friedman, 2001] for different problems and performs well. The method basically works with the idea of building a strong classifier from a combination series of weak ones.

Given a training set $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, x represents the data samples and y represents the class labels. $F[x]$ is used to represent the estimated function. GBDT aims to minimize loss function which is $\hat{F}[y, F(x)]$:

$$\hat{F} = arg^{min}_{F} E_{x,y}\left[L\left(y, F(x)\right)\right] \qquad [2]$$

Model is updated with [3]:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \qquad [3]$$

Where $\gamma_m = \arg{}^{min}_{\gamma} \sum_{i=1}^{n} L[y_i, F_{m-1}(x_i) + \gamma h_m[x_i]]$, m is the iteration number, $h_m[x]$ represents the base decision tree.

Extreme Gradient Boosting (XGBoost)

The method is developed by Chen and Guestrin [2016] and it is implied that the algorithm is able to run ten times faster than the known ones. XGBoost works a set of classification and regression trees which are called CART. Differently, CART evaluates each leaf with a decision value and this enables the model to make better assumptions rather than doing just simple classification. Mathematically, it can be shown as where K is the number of trees, f is a function in the functional space of F where F represents all possible CARTs. For a data set with n observations and m features $D = \{(x_i, y_i), x_i \in R^m, y_i, \in R\}$ $[|D| = n]$ a tree ensemble model uses K to get better predictions with additive training strategy.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \qquad f_k \in F \qquad [4]$$

$F = \{f(x) = w_{q[x]}, q : R^m \rightarrow T, w \in R^T\}$ represents the set of all possible CARTs, w is vector of scores on leaves, q is function of each data assignment to corresponding leaf, and T is the number of leaves. Objective function can be given as

$$obj^{[t]} = \sum_{i=1}^{n}\left[l\left(y_i, \hat{y}_i^{t-1}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2[x_i]\right] + \Omega[f_t] \qquad [5]$$

where

$$g_i = \partial_{\hat{y}_i(t-1)} l\left(y_i, \hat{y}_i^{(t-1)}\right), h_i = \partial^2_{\hat{y}(t-1)} l\left(y_i, \hat{y}_i^{(t-1)}\right) \qquad [6]$$

Regularization, pruning, ability the work with missing values are the main differences of the method against GBDT.

However, GBDT results could not be satisfactory according to their efficiency and accuracy when the data is big. In other words, if the data set contains a large number of samples or features a trade-off between efficiency and accuracy would emerge. A traditional GBDT requires scanning all data samples for every feature. In this way, it only gains the information to estimate all the best split points. In short, these computational complexities could require more time when big data is handled.

In this study, binary classification is used to make predictions if the airplane will be departure on time or not according to the input features. Three different methods based on GBDT are used. First, their definitions are given as follows and then their results are compared.

LightGBM

As mentioned before handling big data with GBDT could result in problems. In order to solve these, Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) are proposed by Ke et al. [2017] and this new method is named LightGBM. The GBDT uses the information to split each node whereas LightGBM uses GOSS in order to determine the split point via calculating variance gain. At first, GOSS sorts the data samples according to the absolute value and the top a x 100% data samples of gradient values are selected and called A. Afterwards, subset B which size b x 100% is obtained from the remaining data, whose size is b x $|A^c|$. In the end, the samples are split via the estimated variance $\widetilde{V_j}[d]$ on $A \cup B$.

$$\widetilde{V_j}(d) = \frac{1}{n}\left(\frac{\left(\sum_{x_{,}\epsilon A_l} g_i + \frac{1-a}{b}\sum_{x_i \epsilon B_i} g_i\right)^2}{n_l^j[d]} + \frac{\left(\sum_{x_{,}\epsilon A_i} g_i + \frac{1-a}{b}\sum_{x_i \epsilon B_r} g_i\right)^2}{n_r^j[d]}\right) \quad [7]$$

Where, $A_l = \{x_i \epsilon\, A : \ x_{ij} \ \leq d\}$, $B_r = \{x_i\ \epsilon\ B : \ x_{ij}\ > d\}$, g_i stands for the negative gradient of the loss function, $\frac{1-a}{b}$ is employed in order to normalize the sum of gradients as a constant.

Without changing the original data distribution by much, GOSS boosts the sample data with small gradients. EFB algorithm leads to the speedup of GDBT with the help of the ability to bundle many sparse features to the fewer dense features. The method is also based on decision tree, but the difference is the fitting operation of negative gradients of loss function one by one. The LightGBM equation $F_M[x]$ can be get through a weighted combination scheme.

$$F_M = \sum_{m-1}^{M} \gamma_m h_m[x] \qquad [8]$$

Categorical Boost

Categorical Boost (CatBoost) is another version of GDBT algorithm. It is developed by Yandex in April 2017 [Yandex, 2017]. Differently, the CatBoost allows the usage of the whole data set while the model is training. Firstly, a random permutation of the dataset is performed and then the average label value is calculated for each example with the same category value placed in the permutation before the given one. If a permutation is $\sigma = [\,\sigma_1, \dots, \sigma_n]$, then $x_{\sigma_p,k}$ is substituted with,

$$\frac{\sum_{j=1}^{p-1}\left[x_{\sigma_j,k} = x_{\sigma_p,k}\right] Y_{\sigma_j} + a\,.\,P}{\sum_{j=1}^{p-1}[x_{\sigma_j,k} = x_{\sigma_p,k}] + a} \qquad [9]$$

Where P is a prior value and a is the weight of the prior [Dorogush, Ershov, & Gulin, 2018]. Another different feature of the method is feature combinations. When a new split is going to build any combination is not considered for the first split in the tree, but for the following splits, all combinations are presented with all categorical features in the data set. All selected splits in the tree are perceived as a two-value classification and combined [Huang et al., 2019]. Moreover, unlike GBDT, Cat Boosting uses target statistics and thus deviation of the solution would not occur [Yandex, 2017].

Bayesian hyperparameter optimization

The supervised machine learning process is based on the input and output data and the learning process of the model, apart from this, hyper-tuning is what makes learning perfect. This process is not learned directly from the inputs. Selecting hyper-parameters manually is time-consuming, repetitive and requires ad-hoc decisions by the practitioner [Feurer, Springenberg, & Hutter, 2015]. For some, it is a "black art" since tuning requires expert knowledge and some luck [Snoek, Larochelle, & Adams, 2012]. Common methods are grid search, random search, and automatic hyper-parameter tuning. The Bayesian hyper-parameter tuning differs itself from them by using a different method, which is downscaling the search space according to past evaluations. Sequential Model-Based Global Optimization (SMBO) is a formalization of Bayesian optimization, it predicts hyper-parameters and sequentially updates the probability model to get better results [Hutter, Hoos, & Leyton-Brown, 2011]. In order to find local optima Expected Improvement function is used. The function is popularized by Jones et al., [1998] which is defined as follows:

$$EI_{y^*}(x) = \int_{-\infty}^{y^*} (y^* - y)p[y|x]dy \qquad [10]$$

In the Eq. [10] y^* represents the objective function's threshold value, *x* is the suggested set of hyper-parameters, the actual value of an objective function, which is calculated with hyper-parameters x, is represented by y, and $p[y|x]$ is surrogate probability model.

APPLICATION

Data description

In this study, flight data of a Turkish airline company is used. The data set consists of the daily flights of the company from 2018. This data is merged with weather condition information that is matched with the flights, which includes instant weather information that occurs in the closest time zone to the flight departure time.

The aim of the study is to predict whether a planned flight will be delayed or not. According to the international rules, if the time difference between actual departure and scheduled departure is greater than 15 minutes the flight is labeled as delayed. There are 18148 observations in which 5717 are delayed and 12431 are on-time flights. Based on the literature survey and expert decisions Table I shows selected variables.

Table 1. Attributes integrated into the dataset

Variable name	Definition	Data type
Day	Day of the month	Categorical
Month	Month of the year	Categorical
Weekday	Day of the week	Categorical
Scheduled departure	Scheduled departure time of the flight	Continuous
Taxi out	The time duration elapsed between departure from the airport gate and wheels off	Continuous
Wheels off	The time point that the aircraft's wheels leave the ground	Continuous
Time on runway	Time spent on the runway	Continuous
Scheduled arrival	Scheduled arrival time of the flight	Continuous
Wheels on	The time point that the aircraft's wheels touch the ground	Continuous
Taxi in	The time duration elapsed between wheels-on and gate arrival at the destination airport	Continuous
Air time	Flight duration	Continuous
Temperature	Air temperature of the airport at or near the flight time	Continuous
Po	Atmospheric pressure measured at the weather station of the airport at or near the flight time	Continuous
U	Relative humidity of the airport at or near the flight time	Continuous
DD	Mean wind direction measured of the airport at or near the flight time	Categorical
FF	Mean wind speed measurement of the airport at or near the flight time	Continuous
VV	Horizontal visibility measured at the airport [km]	Continuous
Td	Dew point temperature of the airport at or near the flight time	Continuous
C	Total cloud cover of the airport	Categorical

Except for the date, wind direction and cloud cover feature all the variables are collected as continuous numbers. The categorical variables are combined to numerical values using the “one-hot-encoding” transformation technique, in which each unique observation of the variables is transformed to binary variables. There are 19 unique values in wind direction, and 547 in cloud cover. Also 7 for the day of the week, 12 for the month, and 31 for the day of the month variables. Therefore, after the transformation, the data set has 630 variables. Also, 70% of the data is used for the training phase and the remaining part for the test data. All the codes are written in Python.

RESULTS AND DISCUSSION

Model selection

From different machine learning approaches, based on both popularity and their good performance, CatBoost algorithm, XGBoost and LightGBM are selected in this study. The model’s performance can be evaluated by various performance criteria. The main objective is the prediction of delayed flights thus developed models were evaluated by various performance measures such as accuracy, recall, receiver operating characteristic (ROC) score, and Cohen’s Kappa score. All performance indicators are based on the confusion matrix.

In binary classification problems, such as this problem, observations can be classified as positive or negative. According to this information, the results of the classification problem can be classified as true positive (TP), true negative (TN), false positive (FP), and false negative (FN). TP and TN represent correct classification. FP is a false alarm also called Type I error and FN represents miss-classified ones also called Type II error. The accuracy rate is obtained by dividing the correctly classified observations into all observations. Usage of this ratio is suitable when the classified classes are equal, otherwise, results will yield misleading outcomes. A recall is the proportion of real positive cases that are correctly predicted as positive. This ratio is suitable when minimum false negative classification is more important. ROC graph is formed with TP (Y axis) and FP rates (X axis). The score is calculated by measuring the area under the ROC curve, and a higher score indicates a better model since it indicates that model's capability of distinguishing the classes. Cohen's kappa score measures the inter-rater reliability [Cohen, 1960]. For classification problems, it takes into account random success as a norm, such as reality, and the observations are evaluated as their degree of agreement. There is no definite way to interpret the result but, Landis et al. [1977] provided a scale which is -1 to 0 indicates no agreement, 0-0.20 slight agreement, 0.21-0.40 as fair agreement, 0.41-0.60 as moderate agreement, 0.61-0.80 as substantial agreement, and 0.81-1 as almost perfect agreement. Finally, when the class labels are predicted and false negatives are more costly, F2 score is advised to consider [Fernández et al., 2018]. It is actually Fbeta-measure with a *beta* value of 2, so the recall score becomes more important. Again it is based on confusion matrix and calculated as follows [11]:

$$F2 - measure = \frac{\left(5 * \frac{TP}{TP + FP} * \frac{TP}{TP + FN}\right)}{\left(4 * \frac{TP}{TP + FP} + \frac{TP}{TP + FN}\right)} \qquad [11]$$

Cross-validation is conducted in order to evaluate the model. In other words, the dataset is randomly divided into 10 sets, where each set has approximately the same imbalance ratio. Furthermore, these sets are trained with proposed algorithms separately.

For the Bayesian optimization, hyper-parameters for each machine learning algorithm, range of the values, and the best-found values are given in the following Table II with and without using SMOTE. After the hyper-parameter tuning, the selected algorithms are applied. Table III shows the results of all three algorithms without SMOTE usage and Table 4 shows results with SMOTE. These results are given in order to clarify the contribution of SMOTE algorithm. Normally, it is expected to get better results with SMOTE, however for this dataset without SMOTE results are preferable. To overcome this situation Table III should be evaluated mostly with the F2 measure, as it is appropriate when the data is unbalanced and the minimization of false negatives is more important to this particular model. LightGBM is more suitable for our case, according to both accuracy, recall, Cohen's Kappa, and F2 score. XGBoost can be considered to have performed slightly worse. Although Catboost has the lowest results, it is worth noting that these results, which are considered to be bad, are around 0.90. When Table IV is examined by focusing on the recall score only, it can be said that LightGBM yielded better results again, but when we look at all the results, it will not be overlooked that it has yielded very close results with XGBoost. CatBoost shows little underperformance here, too. As a result, the recommended model was LightGBM, although all of them were actually available. It is possible to attribute good results without using SMOTE to the fact that the grouping is good at the cross-validation stage and the models are advanced.

Table 2. Hyper-parameters

Algorithm	Hyper-parameter	Range	Best value found	Best value found with SMOTE
XGBoost	Learning rate	[0.01, 1]	0.160	0.076
	Number of estimators	[100, 1000]	740	380
	Max depth	[3, 10]	9	9
	Subsample	[0, 1]	1	0.641
	Gamma	[0, 5]	0.32	0.779
	Minimum child weight	[0, 20]	0	1
LightGBM	Number of leaves	[25, 45]	25	42
	Max depth	[5, 35]	7	33
	Lambda L1	[0, 0.05]	0.011	0.047
	Lambda L2	[0, 0.05]	0.013	0.034
	Minimum child samples	[5, 100]	10	21
	Minimum data in leaf	[5, 100]	7	35
	Feature fraction	[0.1, 0.9]	0.839	0.64
	Bagging fraction	[0.8, 1]	0.852	0.821
CatBoost	Bagging temperature	[3, 10]	6	4.67
	L2 leaf regularization	[2, 5]	5	2
	Max depth	[5, 15]	14	12

Table 3. Algorithm results without SMOTE

Model	Accuracy		Recall		ROC Score		Cohen’s Kappa		F2 Measure	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
XGBoost	0.999	0.969	1	0.925	0.999	0.958	0.999	0.926	1	0.9662
LightGBM	0.999	0.967	1	0.929	0.999	0.957	0.999	0.924	1	0.9707
CatBoost	1	0.947	1	0.858	1	0.923	1	0.873	1	0.9379

Table 4. Algorithm results with SMOTE

Model	Accuracy		Recall		ROC Score		Cohen’s Kappa		F2 Measure	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
XGBoost	0.998	0.962	0.999	0.903	0.999	0.946	0.998	0.910	0.999	0.951
LightGBM	1	0.959	1	0.904	1	0.945	1	0.905	1	0.957
CatBoost	0.999	0.937	0.999	0.825	0.999	0.906	0.999	0.847	1	0.931

CONCLUDING REMARKS

Flight delays have become a regular phenomenon in the aviation industry with the ever-growing travel demand, restricted airport capacity, and increasing amount of aviation traffic. Therefore, the prediction of late flights is important for all parties affected by this situation.

This paper has developed a new approach for airline companies to detect delayed flights. In order to achieve this different approaches, which are XGBoost, LightGBM, and CatBoost, were used. In addition, the SMOTE method was used in order not to be affected by the instability of the data set, but according to the results, the proposed algorithms performed well with the available data without the need for synthetic data processing. The reason for this could be the implementation of cross-validation or the use of advanced models.

This study, which is carried out by examining the data of a particular airline, can be evaluated as an event study. However, the results obtained are extremely promising. This degree of accurate estimation of delayed flights by algorithms also paves the way for future studies. In the following studies, it will be possible to predict delays in connecting flights or to replace cargo flights that require urgent transportation with alternatives. According to all these estimates, it is thought that opportunities such as applying different pricing policies or making flight insurance by taking into account the estimates in insurance costs can be offered.

ACKNOWLEDGMENTS AND FUNDING SOURCE DECLARATION

This work was supported by The Scientific Research Projects Coordination Unit of Akdeniz University. Project Number: SBA-2019-4841

REFERENCES

Abdelghany, K. F., Shah, S. S., Raina, S., & Abdelghany, A. F., 2004. A model for projecting flight delays during irregular operation conditions. Journal of Air Transport Management, 10, 385-394. http://doi.org/10.1016/j.jairtraman.2004.06.008

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P., 2002. SMOTE: Synthetic Minority Over-Sampling Technique. Journal of artificial intelligence research, 16: 321-357. https://dl.acm.org/doi/10.5555/1622407.1622416

Chen, J., & Li, M., 2019. Chained Predictions of Flight Delay Using Machine Learning. AIAA SciTech Forum, San Diego: American Institute of Aeronautics and Astronautics, Inc. 1-25. https://doi.org/10.2514/6.2019-1661

Chen, T., & Guestrin, C., 2016]. Xgboost: a scalable tree boosting system. Proceedings of the 22nd acm sigkdd International Conference on Knowledge Discovery and Data Mining, 785-794. https://dl.acm.org/doi/10.1145/2939672.2939785

Choi, S., Kim, Y. K., Briceno, S., & Mavris, D., 2016. Prediction of Weather-induced Airline Delays Based on Machine Learning Algorithms. 35th Digital Avionics Systems Conference, 1-6, IEEE. https://doi.org/10.1109/DASC.2016.7777956

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and psychological measurement, 20(1): 37-46. https://doi.org/10.1177%2F001316446002000104

Dorogush, A. V., Ershov, V., & Gulin, A., 2018. CatBoost: gradient boosting with categorical features support. arXiv preprint, 1-7. https://arxiv.org/abs/1810.11363

Dray, L. M., Antony, E., Vera-Morales, M., Reynolds, T. G., & Schafer, A., 2008. Network and Environmental Impacts of Passenger and Airline Response to Cost and Delay. 8th AIAA Aviation Technology, Integration and Operations Conference, 8890-8901. Anchorage. https://doi.org/10.2514/6.2008-8890

Efthymiou, M., Njoya, E. T., Lo, P. L., Papatheodorou, A., & Randall, D., 2019. The Impact of Delays on Customers' Satisfaction: an Empirical Analysis of the British Airways On-Time Performance at Heathrow Airport. Journal of Aerospace Technology and Management, 11. http://dx.doi.org/10.5028/jatm.v11.977

Fernández , A., García , S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F., 2018. Learning from imbalanced data sets. Switzerland: Springer. https://doi.org/10.1007/978-3-319-98074-4

Feurer, M., Springenberg, J. T., & Hutter, F., 2015. Initializing bayesian hyperparameter optimization via meta-learning. In Twenty-Ninth AAAI Conference on Artificial Intelligence. https://dl.acm.org/doi/10.5555/2887007.2887164

Friedman, J. H., 2001]. Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232. http://doi.org/10.1214/aos/1013203451

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., & Yuanyue, H., 2017. Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications, 73: 220-239. https://doi.org/10.1016/j.eswa.2016.12.035

He, H., & Garcia, E., 2009]. Learning from imbalanced data. IEEE Transactions on knowledge and data engineering, 21(9): 1263-1284. https://doi.org/10.1109/TKDE.2008.239

Huang, G., Wu, L., Ma, X., Zhang, W., Fan, J., Yu, X., Zhou, H., 2019. Evaluation of CatBoost method for prediction of reference evapotransportation in humid regions. Journal of Hydrology, 1029-1041. https://doi.org/10.1016/j.jhydrol.2019.04.085

Hutter, F., Hoos, H., & Leyton-Brown, K., 2011. Sequential model-based optimization for general algorithm configuration. International conference on learning and intelligent optimization, 507-523. Berlin: Springer. https://doi.org/10.1007/978-3-642-25566-3_40

IATA, 2019. International Air Transport Association Annual Review . Seoul: IATA.

Jones, D. R., Schonlau, M., & Welch, W. J., 1998. Efficient global optimization of expensive black-box functions. Journal of Global optimization, 13(4): 455-492. https://doi.org/10.1023/A:1008306431147

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., & Liu, T.-Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems, 3146-3154. https://dl.acm.org/doi/10.5555/3294996.3295074

Kim, Y. J., Choi, S., Briceno, S., & Mavris, D., 2016. A Deep Learning Approach to Flight Delay Prediction. IEEE/AIAA 35th Digital Avionics Systems Conference [DASC] [s. 1-6]. IEEE. https://doi.org/10.1109/DASC.2016.7778092

Kuhn, N., & Jamadagni, N., 2017. Application of Machine Learning Algorithms to Predict Flight Arrival Delays., 1-6.

López, V., Fernández, A., García, S., Palade, V., & Herrera, F., 2013. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information sciences, 250: 113-141. doi: https://doi.org/10.1016/j.ins.2013.07.007

Loyola-González, O., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A., & García-Borroto, M., 2016. Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases. Neurocomputing, 175: 935-947. https://doi.org/10.1016/j.neucom.2015.04.120

Manna, S., Biswas, S., Kundu, R., Rakshit, S., Gupta, P., & Barman, S., 2017. A Statistical Approach to Predict Flight Delay Using Gradient Boosted Decision Tree. International Conference on Computational Intelligence in Data Science, 1-5. IEEE. https://doi.org/10.1109/ICCIDS.2017.8272656

Mazareanu, E., 2020, Global air traffic - annual growth of passenger demand 2006-2021. 09 25, 2020 statistica: https://www.statista.com/statistics/193533/growth-of-global-air-traffic-passenger-demand/

McCarthy, N., Karzand, M., & Lecue, F., 2019. Amsterdam to Dublin Eventually Delayed? LSTM and Transfer Learning for Predicting Delays of Low Cost Airlines. The Thirty-First AAAI Conference on Innovative Applications of Artificial Intelligence. 33: 9541-9546. Proceedings of the AAAI Conference on Artificial Intelligence. https://doi.org/10.1609/aaai.v33i01.33019541

Mohamed, H. M., Al-Tabbakh, S. M., & El-Zahed, H., 2018. Machine Learning Techniques for analysis of Egyptian Flight Delay. J. Sci. Res. Sci., 35: 390-399. https://dx.doi.org/10.21608/jsrs.2018.25523

Nakornsri, P., Apivatanagul, P., & Pisitkasem, P., 2020. Density Analysis Based Flight Delay Prediction withGenetic Algorithm Hyperparameter Tuning. Rangsit Graduate Research Conference: RGRC, 15: 2324-2337.

NEXTOR, 2010. Total Delay Impact Study . Federal Aviation Administration Air Traffic Organization Strategy and Performance Business Unit.

Simić, T. K., & Babić, O., 2015]. Airport traffic complexity and environment efficiency metrics for evaluation of ATM measures. Journal of Air Transport Management, 42: 260-271. https://doi.org/10.1016/j.jairtraman.2014.11.008

Snoek, J., Larochelle, H., & Adams, R. P., 2012. Practical bayesian optimization of machine learning algorithms. Advances in neural information processing systems, 2951-2959. https://dl.acm.org/doi/10.5555/2999325.2999464

Thiagarajan, B., Srinivasan, L., Sharma, A. V., Sreekanthan, D., & Vijayaraghavan, V., 2017]. A Machine Learning Approach for Prediction of On-time Performance of Flights. 36th Digital Avionics Systems Conference (DASC), 1-6, IEEE. https://doi.org/10.1109/DASC.2017.8102138

Venkatesh, V., Arya, A., Agarwal, P., S, L., & Balana, S., 2017, Iterative Machine and Deep Learning Approach for Aviation Delay Prediction. 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics [UPCON] [562-567]. IEEE. https://doi.org/10.1109/UPCON.2017.8251111

Yandex, 2017, CatBoost Now Available in Open Source. 10 07, 2020, catboost: https://catboost.ai/news/catboost-now-available-in-open-source

Yandex, 2017. Feature importance. 06 03, 2020, catboost.ai: https://catboost.ai/docs/concepts/fstr.html#fstr__regular-feature-importance

Yu, B., Guo, Z., Asian, S., Wang, H., & Chen, G., 2019. Flight delay prediction for commercial air transport: A deep learning approach. Transportation Research Part E, 125: 203-221. https://doi.org/10.1016/j.tre.2019.03.013

Irmak Hatıpoğlu   ORCID ID: https://orcid.org/0000-0001-5244-9115
Department of International Trade and Logistics,
Faculty of Applied Sciences,
Antalya, **Turkey**
e-mail: mailto:irmakdaldir@akdeniz.edu.tr

Ömür Tosun   ORCID ID: https://orcid.org/0000-0003-1571-7373
Department of Management Information Systems,
Faculty of Applied Sciences, Akdeniz University,
Antalya, **Turkey.**
e-mail: omurtosun@akdeniz.edu.tr

Nedret Tosun   ORCID ID: https://orcid.org/0000-0003-4566-6693
West Mediterranean Exportaters Association,
Antalya, **Turkey**
e-mail: tosunn@baib.gov.tr