https://doi.org/10.26552/ pas.Z.2024.1.20

UNIVERSITY OF ŽILINA

FORECASTING FLIGHT DELAYS USING MACHINE LEARNING

Mykhailo Soloviov
Air Transport Department
University of Žilina
Univerzitná 8215/1
010 26 Žilina

Benedikt Badánik
Air Transport Department
University of Žilina
Univerzitná 8215/1
010 26 Žilina

Abstract

This article considers machine learning and its utilization in the domain of air transportation. The first part of this research aims to define machine learning, describe its historical development and helps delineate machine learning in a broader framework of other data analysis approaches such as artificial intelligence, deep learning and data science. In the second part, the research is meant to explain what machine learning is, tell more about the types of machine learning and how it is used in different scenarios in the aviation industry. This part of the thesis discusses certain areas and real examples of how machine learning is used by multiple companies (aircraft manufacturers) as well as examines the available conclusions of researches already undertaken in the area and determines their connection to the current one. Finally, the practical part of the thesis uses the collected real-time data about departures from two American airports, analyses it with the help of statistical Python-based tools, describes the developed machine-learning algorithm to predict delays, runs experiments on data and discusses results.

Keywords

machine learning, data science, artificial intelligence, flight delay, prediction, classifier, estimator, data

1. Introduction

Machine learning is using an algorithm or computer programme to learn about different patterns in data and then taking that algorithm and what it has learnt to make predictions about the future using similar data.

For the past several years, much has been written about data collection systems onboard modern airplanes: GE jet engines collect information at 5,000 data points per second; a Boeing 787 generates an average of 500GB of system data a flight; an Airbus A380 is fitted with as many as 25,000 sensors. Much of this data is transmitted or downloaded to plan maintenance, where it can be decided on when to position spare parts or anticipate component failure [1].

Machine learning lets computers make decisions about data, it lets computers learn from data and then make predictions and decisions. What product does YouTube recommend? Does this person have heart disease? Is this email spam? All use machine learning. So we, essentially, let computers decide for us. Although computers only understand numbers, in the end of the day only ones and zeros, using machine learning we have found ways for computers to decide for us and answer hard questions that in the past only humans could answer. But machine learning is after all just a general term for when computers learn from data. It allows computers to do tasks that in the past required humans, and make our lives, hopefully, easier.

The main goal of this article is to show how machine learning and advanced data analysis is used in airport operations and to develop a flight delay prediction model that uses machine learning.

2. Methodology and methods of research

Now let us explore what flight delay is and what methods of predicting flight delays scholars have developed to address the issue.

Civil aviation authorities in different countries have different definitions of flight delay.

The United States Federal Aviation Administration (FAA) considers a flight to be delayed if its actual departure is 15 minutes later than scheduled departure. There are many reasons for flight delays including weather conditions, engineering maintenance, air traffic flow, military activity, incidents like runway incursion, runway/taxiway surface damage, passengers being late for boarding, unruly non-cooperative passengers, public health incidents, departure system failures etc.

As explained in the paper by Wang and colleagues *A Review of Flight Delay Prediction Methods* [2] the general process of flight delay prediction consists of:

1) Analysing flight delay prediction objects: determining flights, airports and timeframe as well as relevant variables such as weather, air traffic flow and military aviation activity

2) Data collection and pre-processing: collecting flight operation data, pre-processing the data, i.e. performing cleaning and transforming

3) Selecting a flight delay prediction method: selecting an appropriate and efficient method (machine learning, traditional statistical analysis, queuing theory)

4) Building a flight delay prediction model: building a prediction model, inputting collected and processed data and performing numerical simulation test

5) Model evaluation: using evaluation indicators such as RMSE (Root Mean Square Error), MSE (Mean Squared Error) and MAE (Mean Absolute Error) to evaluate the prediction effect

6) Model output and saving: analysing the model output and saving the model for future use

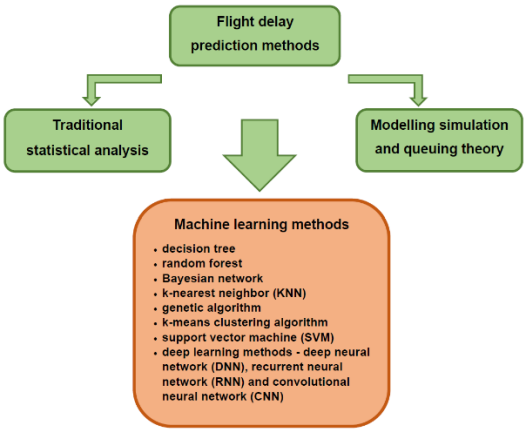


Figure 1. *Flight delay prediction methods framework. [Source: Author]*

Machine learning, a subset of artificial intelligence, leverages AI techniques to extract crucial features and construct machine learning models using extensive flight data. The principles and methodologies of machine learning have found extensive applications within the realm of civil aviation. In particular, the domain of flight delay prediction employs a variety of machine learning approaches, encompassing decision trees, random forests, Bayesian networks, k-nearest neighbors, k-means clustering algorithms, support vector machines, deep learning, and more.

Decision tree is a tree-like model where each node represents a decision based on the value of a specific feature. The leaves of the tree represent the outcomes or class labels. The commonly used decision tree algorithms include ID3 (Iterative Dichotomiser 3), C4.5, CHAID (Chi-squared Automatic Interaction Detector) and CART (Classification and Regression Trees). Cheng [3], Zhou [4] and Liu [5] used different approaches and the experimental results showed that the moel's accuracy is close to 80%.

Random forest is a classifier containing multiple decision trees, which randomly generates multiple independent decision trees from historical data. Certain research has been done while fusing historical flight data with meteorological data and it showed that the recall and accuracy are improved after the integration of meteorological data [6].

Bayesian network consists of a graph and a probability table, Xu proposed the use of Bayesian networks to investigate and visualize delay propagation between airports [7].

The k-Nearest Neighbors (k-NN) algorithm is a simple, yet powerful, supervised machine learning algorithm used for both classification and regression tasks. The primary idea behind k-NN is to predict the label or value of a new data point based on the majority class or average of the k-nearest data points in the feature space.

Support vector machine is a supervised learning algorithm for both classification and regression, SVM is particularly effective in high-dimensional spaces, but can be computationally intensive for large datasets. Esmaeilzadeh [8] employed SVM model to explore the nonlinear relationship between flight delay results.

Deep learning is a new research direction in the field of machine learning. Deep neural networks can automatically learn hierarchical representations of features from raw data such as departure times, weather conditions, airport information, and historical flight data. Deep learning models can integrate information from various sources, such as structured tabular data and unstructured data like text or images. This is beneficial, because one can predict flight delays based on diverse factors like weather reports, social media sentiment, etc. It's important to note that while deep learning methods can offer powerful capabilities, their success depends on factors such as the quality and quantity of data, appropriate model architecture, hyperparameter tuning, and careful validation.

For this project the very popular and widely-used tool for ML will be used: Scikit-Learn [9]. There is an extensive documentation available at their official website [10] about different methods and estimators used in ML and the settings for their usage from a point of problem's stating to the data available and implementation.

For our purposes the following 'map' will be used.

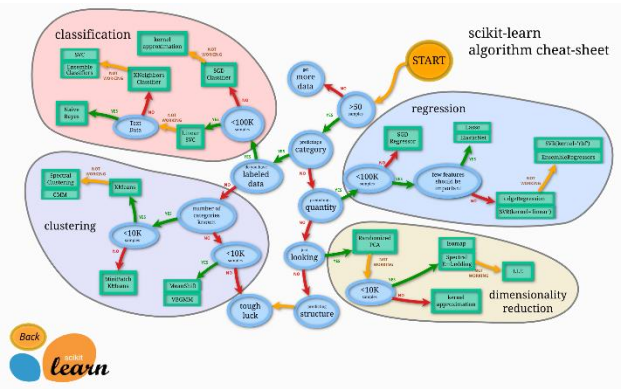


Figure 2. *Scikit-Learn estimators map [Source: [10]]*

For our classification problem the SVC, KNNeighbors and Ensemble Classifier (Random Forest) will be used. For the regression problem the Ridge and Ensemble Regressor (Random Forest) will be used according to the map above.

Data for this project has been collected from the official website of the United States Government Department of Transportation [11], which is an official website of the United States Government Department of Transportation and it has the search engine capable of looking for on-time real-world statistics from many airlines, specifically those that have at least 0.5 percent of total domestic scheduled-service passenger traffic.

For the purpose of this project two airports were chosen: ATL Atlanta Hartsfield Jackson International Airport and PHX Phoenix Sky Harbor International Airport. The time frame chosen is full calendar year – 2019. It has not been affected in any way by the COVID-19 pandemic, so the operation can be deemed normal. The airline, whose flights were selected is American Airlines (AA). Also both these airports are focal points of the aviation traffic in the United States, serving two large urban areas in the south-eastern and south-western USA respectively.

The Jupyter Notebook software for data analysis is used in this project, being the best and the most popular tool for data analysis projects.

At first some EDA (Exploratory Data Analysis) was done to show in more detail the features of the dataset we were working with.

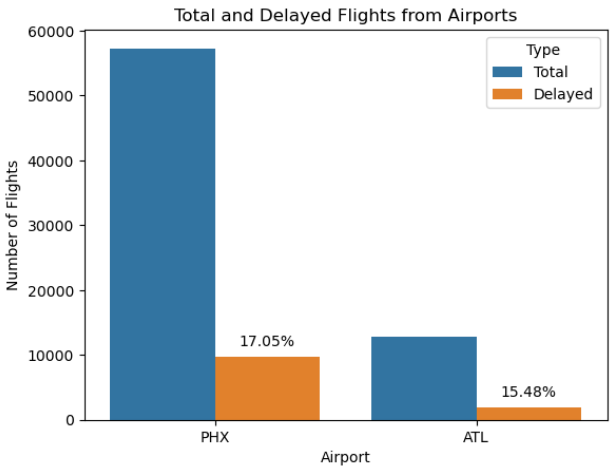


Figure 3. *Delayed and on-time flight at PHX and ATL [Source: Author]*

We see that delayed flights make up a total of ~16.76% of 70 036 recorded flights.

Now let's see the distribution of delays and what delays are the most common.

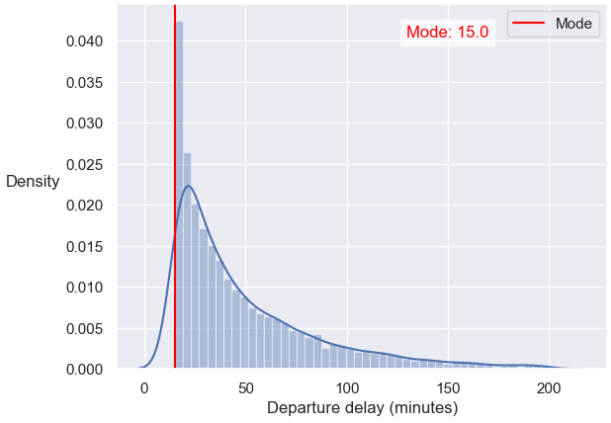


Figure 4. *Distribution of delays at PHX [Source: Author]*

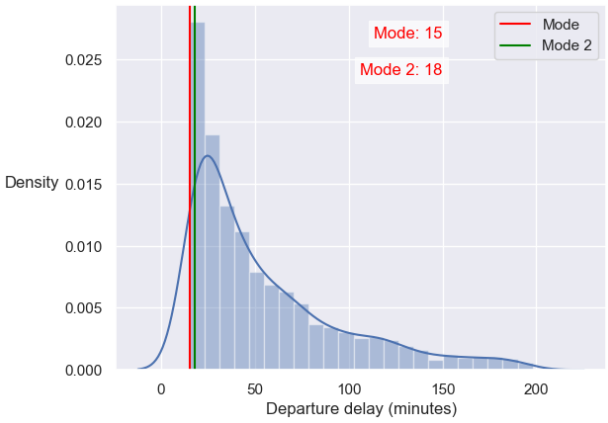


Figure 5. *Distribution of delays at ATL [Source: Author]*

Now let's see the distribution of delays and what delays are the most common.

We see that distribution is highly right-skewed, meaning that majority of flights' delays oscillate around 0, with the most common delay figure in the PHX dataset being 15 minutes and the most common in ATL dataset – 15 and 18 minutes.

By importing, identifying and removing the missing data points from PHX, this overview of dataset is obtained:

It contains a total of 57270 rows (flights) and these columns:

carrier_code, date_(mm/dd/yyyy), flight_number, tail_number, destination_airport, scheduled_departure_time, actual_departure_time, scheduled_elapsed_time_(minutes), actual_elapsed_time_(minutes), departure_delay_(minutes), wheels-off_time, taxi-out_time_(minutes).

The corresponding ATL dataset has 12766 rows (flights) and same columns.

2.1. Classification model

The first model we are going to implement here is a classification model. It will look at the training part of the data and try and find some correlation between features and target variable. Target variable takes a value of '1' or '0' depending on whether the flight is delayed or on time, '1' representing delay of over 15 minutes and '0' – no delay or a delay of less than 15 minutes.

Our original dataset would have all these columns (features) as displayed in the first step in the above figure, then some columns are dropped, because they would have made it very obvious for the algorithm to predict if the flight was delayed or on-time and furthermore they are not known at the time of prediction, which is usually some time before the flight's departure. The correlation matrix in Fig. 6 shows the columns that this algorithm is going to be working with. The ML model will try to predict whether or not the flight is going to be delayed based on the flight number, scheduled elapsed time (the estimated duration of the flight), destination airport (where the flight is headed), month and day of the flight, and the scheduled departure time of the flight.

Figure 5. *Correlation matrix [Source: Author]*

Three different estimators are used in the classification problem: KNN (k-Nearest Neighbor), SVC (Support Vector Classifier) and Ensemble Classifier (namely Random Forest Classifier) as shown in Figure 2.

2.2. Regression model

The goal of the regression model is to let the model learn from the data of the delayed flights and try and predict a number of minutes the flight is going to be delayed for. So, we are considering only delayed flights for this.

In this scenario, two regression estimators will be used: Ridge Regression and Ensemble Regressor (namely Random Forest Regressor).

After splitting the data into training and testing subsets, two models were applied to the data and immediately evaluated using a R^2 (R-squared) parameter or coefficient of determination. What this parameter shows us is how the model's predictions correspond to the mean of the targets. The score can range from negative infinity to 1. A score of 1 indicates perfect prediction, a score of 0 indicates that the model is no better than predicting the mean of the target variable for all observations, and a negative score indicates that the model is performing worse than predicting the mean.

For a Ridge regression model the R^2 score obtained was: 0.0021642457789885494. Seeing as the value is very close to 0 means that this model is as good at predicting the actual value in minutes by which the flight is delayed as just taking the mean delays in the test dataset and saying that the value obtained is the predicted delay for all flights. For a Random Forest regression model the R^2 score was even less: -0.03244153536392225. Meaning that the model performs even poorer than just predicting the delay as mean of all delays in the dataset.

Table 1. *Evaluators of regression models [Source: Author]*

Evaluator	Ridge	Random Forest
R^2	0.00216	-0.03244
MAE	44.554 mins	47.508 mins
MSE	7813.765 mins2	8084.753 mins2

In the table above there are also two other common evaluators that are used when evaluating a regression model. MAE is the average of the absolute differences between predictions and actual values. It gives us an idea of how wrong our predictions were. Lower is better. In our case they are ~44 and ~47 mins respectively. Given as the mean delay is about 9 mins for both datasets which was easily found out from EDA, our MAE constitutes an error of about 500%, which makes this model unusable under these circumstances.

By finding that out, we can safely say that the dataset obtained and used in this thesis, is not suitable for a prediction of delay using regression ML algorithm. For purposes of improving this model a different approach would have needed to be applied: collecting more data and of different type and further feature engineering.

That's why regression model will not be further discussed in the next chapter.

3. Results

The first classification model that was run was KNN classifier and it achieved accuracy of 80.60%, but then some tuning was performed and by adjusting the number of neighbors to 20 a better result was achieved of 83.05%. The SVC classifier model achieved accuracy of 83.10%. The Random Forest model achieved accuracy of 77.72%.

By computing the accuracy we measure a fraction of all data points that were correctly predicted [12]. Only we have to take into account that accuracy alone might not be sufficient to evaluate a classification model, especially if the dataset is imbalanced. Other metrics take stage in these scenarios like precision, recall, F1 score and ROC-AUC and we will explore them next.

Table 2. *Classification report for KNN model [Source: Author]*

	Precision	Recall	f1-score	Support
0	0.83	1.00	0.91	11641
1	0.46	0.02	0.04	2367
Accuracy			0.83	14008

Table 3. *Classification report for SVC model [Source: Author]*

	Precision	Recall	f1-score	Support
0	0.83	1.00	0.91	11641
1	0.00	0.00	0.00	2367
Accuracy			0.83	14008

Table 4. *Classification report for Random Forest model [Source: Author]*

	Precision	Recall	f1-score	Support
0	0.84	0.90	0.87	11641
1	0.24	0.15	0.19	2367
Accuracy			0.78	14008

In the tables illustrated above we see that KNN model has the same accuracy as SVC model, but they differ in precision and recall. We see that SVC has zero values in both precision and recall in the 1 class, meaning that the model classified flights as being either not delayed and they'd turned out as not delayed (11641) or the model classified them as delayed, when they were indeed on time (2367). There are no flights in the TP (True Positive) part of the confusion matrix (see Fig. 7), meaning that the model didn't actually predict a single flight to be delayed and was right, so we can safely disregard this model from further evaluation.

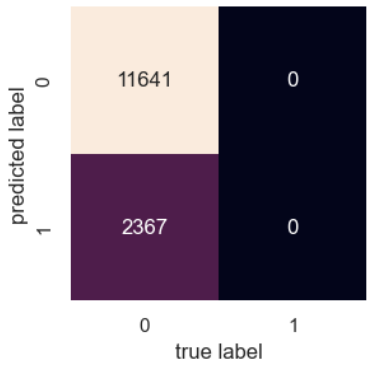


Figure 7. *Confusion matrix SVC [Source: Author]*

Now we are left with comparing the KNN model and Random Forest model. Let's see both their confusion matrices and also compare their classification reports.

Looking at the confusion matrix of the KNN model (see Fig. 8) we roughly get the idea of where the high accuracy of 83% is coming from. The model is pretty good at capturing and identifying non-delayed flights. After all out of 11641 non-delayed flights, 11589 were correctly classified as non-delayed. But at the same time 2323 flights (out of 2367 actually delayed flights) were predicted to be delayed when indeed they were not. Only 44 flights (about 2% of all delayed flights) were correctly identified as delayed, and 52 flights were misclassified as non-delayed.

Random Forest model (see Fig. 9) is better at identifying the delayed flights – 359 (out of total 2367, about 15%) flights were correctly classified as delayed. But it comes at a price of more False Positive identifications (flights identified as 1 – 'delayed' when they were indeed not) and less True Negative identifications (model is worse at capturing non-delayed flights).

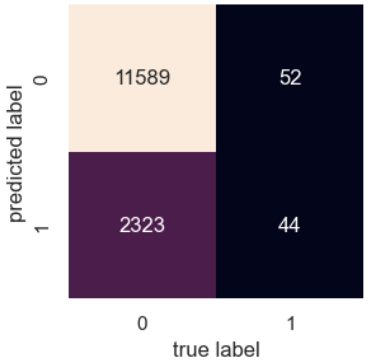


Figure 8. *Confusion matrix KNN [Source: Author]*

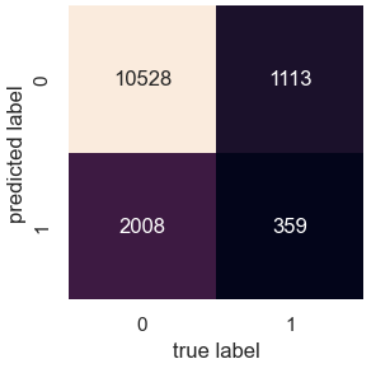
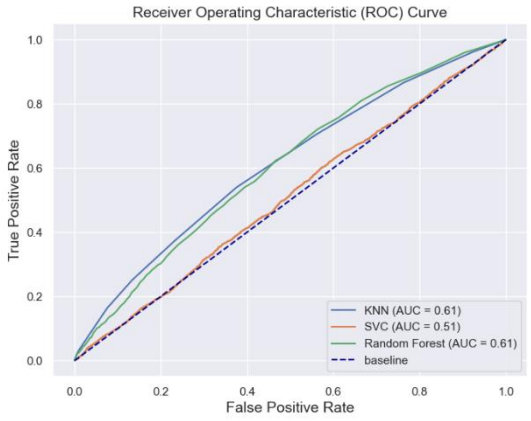


Figure 9. *Confusion matrix RF [Source: Author]*

Certain attention should be given to the ROC-AUC plot. It is a plot used to evaluate the performance of binary classification models. It illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate (specificity) at various thresholds. AUC (Area Under Curve) can be maximum 1.0, indicating perfect prediction. A higher AUC value indicates better overall model performance in distinguishing between the two classes.

Figure 10. *ROC-AUC plot [Source: Author]*



Clearly we see two leaders here: KNN and Random Forest. We can make an assumption going forward that with implementation of more relevant data those two algorithms have more promise in providing better results.

4. Conclusion

In this thesis, the aim was to introduce the reader to ML, specifically in aviation and to develop a flight delay ML algorithm. For those purposes a brief introduction to ML, including its history and classification was done, then the thesis focused on ML applications in the aviation industry. Relatable scientific papers were reviewed and discussed.

Then we embarked on a journey to develop a flight prediction machine learning model focusing on the timely operation of flights. The project began with data collection from reliable sources, cleaning, preprocessing, and culminated in the evaluation of classification and regression models. For the purpose of making the work accessible to public, all the Jupyter Notebook files (.ipynb extension) along with the raw datasets were uploaded to GitHub at this link [13].

The building of the models were split in two ways: classification and regression.

The classification model aimed to predict whether a flight would experience a delay exceeding 15 minutes. Features such as flight number, scheduled elapsed time, destination airport, month, day, and scheduled departure time were utilized for training. Three classifiers - KNN, SVC, and Random Forest - were employed, with SVC demonstrating the highest accuracy of 83.10%, meaning that the model was able to predict if the flight would be delayed 15 mins or more in 83 out of 100 flights. However, a deeper analysis revealed limitations (see Fig. 7-9, Table 2-4), particularly in precision and recall metrics, highlighting the need for comprehensive evaluation beyond accuracy.

The regression model sought to predict the duration of flight delays, focusing exclusively on delayed flights. Two regressors - Ridge and Random Forest - were applied, but both models exhibited poor performance, as indicated by the low R-squared scores and high Mean Absolute Errors (MAE). This outcome suggested that the dataset, in its current form, lacked the predictive power required for regression-based delay prediction. To predict the exact number of minutes the flight is going to be delayed for more features in the dataset are required – weather information, causes of delays in the training dataset when and if the flight was delayed (whether it was caused by weather, operations of airline, airport, force majeure factors, aircraft maintenance, delay propagated from previous flight, military incidents etc.).

In conclusion, classification models showed promise in identifying delayed flights – all of them were able to predict whether the flight would be delayed or not in at least 78% of cases. The regression model's performance was inadequate for practical application. The practical application of the results reached in this thesis is understanding that for a complex task of predicting flight delays Ensemble methods (such as Random Forest) should be used and further built upon to develop a more robust flight delay prediction engine.

Future research should focus on enhancing the dataset quality, incorporating additional features, and exploring advanced machine learning techniques to improve predictive accuracy. Additionally, deploying ensemble methods or exploring neural networks may offer avenues for enhanced performance. Moreover, addressing the issue of imbalanced datasets and fine-tuning models could further refine prediction capabilities. Despite the challenges encountered, this thesis lays a foundational framework for developing more robust and accurate flight prediction models, essential for optimizing airline operations and passenger experiences.

References

[1] SHAH, D. 2014. How Big Data could improve commercial aviation safety. In: Aerospace Manufacturing & Design. 2014. Available at: https://www.aerospacemanufacturinganddesign.com/news/millions-of-data-points-flying-part2-121914/ [cit. 2023-11-06]

[2] T. Wang, Y. Zheng and H. Xu, "A Review of Flight Delay Prediction Methods," 2022 2nd International Conference on Big Data Engineering and Education (BDEE), Chengdu, China, 2022, pp. 135-141, doi: 10.1109/BDEE55929.2022.00029.

[3] H. Cheng, Y. M. Li, Q. Luo, et al., "Study on flight delay with C4.5 decision tree based prediction method," Systems Engineering-Theory & Practice, vol. 34, pp. 239-247, 2014.

[4] T. Zhou, Q. Gao, N. Ma, et al., "Flight delay prediction based on clustering analysis and CHAID decision tree algorithm," Journal of Wuhan University of Technology, pp. 32-40, 2017.

[5] F. Liu, J. Sun, M. Liu, et al., "Generalized flight delay prediction method using gradient boosting decision tree," IEEE Veh. Technol. Conf., pp. 1-5, May 2020.

[6] R. B. Wu, J. Y. Li, and J. Y. Qu, "Parallel flight delay prediction model based on fusion of meteorological data," Journal of Signal Processing, vol.34, pp. 505-512, 2018.

[7] N. Xu, G. Donohue, K. B. Laskey, et al., "Estimation of delay propagation in the national aviation system using Bayesian networks," Proc. USA/Europe Air Traffic Manag. Res. Dev. Semin., ATM, pp. 353-363, June 2005.

[8] E. Esmaeilzadeh, S. Mokhtarimousavi, et al., "Machine learning approach for flight departure delay prediction and analysis," Transp. Res. Rec., vol. 2674, pp. 145-159, 2020.

[9] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[10] Scikit-Learn Official Website [online]. Available at: https://scikit-learn.org

[11] Bureau of Transportation Statistics, United States Department of Transportation [online]. Available at: https://www.transtats.bts.gov/ontime/ [cit. 2023-08-15]

[12] MÜLLER, ANDREAS C., GUIDO, SARAH. 2017. Introduction to Machine Learning with Python. 1st edition. O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472, 2017. 398 pages. ISBN: 978-1-449-36941-5

[13] PROJECT CODE. SOLOVIOV, MYKHAILO. 2024. Available at: https://github.com/michaelatt/flight_delay_prediction