Big Data Analytics for Airline Delay Prediction

Li Guan, Robin Sah, Hengchuan Zhang

December 4, 2023

Introduction

Trip interruptions are common and unavoidable for frequent travelers. Airline delays, affected by the most sophisticated reasons including adverse weather conditions, mechanical repairs needed, air traffic congestion, crew availability issues, and security check delays. To avoid the huge financial costs of airline delays, potent prediction tools are needed. In this research, we are committed to analyze the incidence of airline delays using the big data and machine learning model. The model we proposed is capable of successful airline delay prediction with legitimate data visualization, and shows the feasibility to provide the users with accurate real-time airline delay predictions in the future.

Related Work

Our project draws inspiration from several key studies in the domain of airline delay prediction. The first study, using deep learning and the Levenberg-Marquardt algorithm, highlighted the challenges of data imbalance in flight delay prediction [3]. Another significant work employed big data analytics to understand the ripple effects of flight delays, particularly on connecting flights, using regression techniques [4]. Tang (2021) and Gui (2020) provided crucial insights into data collection and relevant factors affecting delays, emphasizing the effectiveness of decision tree models in prediction accuracy. Additional studies by Ehsan Esmaeilzadeh et al. and Khaksar et al. expanded the scope of machine learning applications in this field, demonstrating the versatility of various algorithms including SVM and hybrid methods for delay prediction. These works collectively underscore the importance of data quality, algorithm selection, and the potential of machine learning in addressing the complex problem of flight delay prediction.

The paper "Airline Flight Delay Prediction Using Machine Learning Models" provides us a solution on where and how to collect historical data sets. The author narrows down the scale to JFK airport and takes the historical data from 2018 - 2019 on Kaggle. (Tang 2021) In addition to that, the author also points out which information is valuable for airline delay prediction, such as day of the week, departure time, arrival time, distance of the fight, wind direction, temperature, etc. (Tang 2021) By knowing that information, we can clean our data set and prepare to train our machine learning

model. Tang is not the only person interested in predicting airline delay. In the paper "Flight Delay Prediction Based on Aviation Big Data and Machine Learning", Gui and his team extend that information in more detail. Gui and his team add the destination airport weather information as another factor to consider because sometimes airplanes wait at the airport due to the destination weather condition. (Gui 2020) With the information from these two papers, our team can focus on collecting data from the time of day, travel distance, air route and weather conditions.

Choosing which machine learning algorithm to use is vital in our project. Fortunately, both Tang and Gui's team tested different machine learning algorithms in their paper. And their results show that the decision tree model is the most accurate model compared to others. In Tang's paper, the decision tree model's accuracy rate is 97%. (Tang 2021) In Gui and his team's paper, the decision tree model's accuracy rate is 90.2%. (Gui 2020) Their discovery is valuable to us, which means we can focus on using the decision tree model for our project.

We further examined the paper *Machine Learning Approach for Flight Departure Delay Prediction and Analysis* by Ehsan Esmaeilzadeh et. al. This paper discussed the types of delay and associated possibilities. By collecting the data from EWR, JFK, and LGA in 2018, the paper collected a total of 33,548 flights including both domestic and international destinations. Taking advantage of the MATLAB, they implemented the SVM models and achieved 79.95% accuracy. Compared with the decision tree model provided by Tang's paper, the SVM model is less accurate but provides its ability to include a larger scale of factors. Given the fact that flight delays and reroutes could be affected by hundreds of factors, the SVM model shows its flexibility. This paper provides a comprehensive insightful data collection and data processing techniques. In the research article about "*Flight delay prediction based on deep learning and Levenberg-Marquart algorithm*" researchers looked closely at using a deep learning model, helped by the Levenberg-Marquardt (LM) algorithm, to guess flight delays accurately [3]. They made a model using something called stacked denoising autoencoders and checked how well it worked, focusing on finding a balance between how long the computer takes and the number of denoising autoencoders to make sure the model was just right. They had a big challenge with their data because it was imbalanced, meaning there were a lot more examples of non-delayed flights than delayed ones. They tried two different methods to balance the data in order to correct this, but ultimately decided to utilize under-sampling because up-sampling resulted in overfitting and took too long.

Project Design

Our team focused on predicting airline delays with the use of data of historical airline information to train the machine learning model. Then, we used the prepared model to predict airline delays based on the real time data and store those data into our data warehouse. When the user types the start, destination airport and time, our user would access the data warehouse showing which flights are more likely to delay than others. The goal of our project is to help users avoid flights that may be delayed so that users can arrive at their destination on time.

The diagram below shows the data flow diagram for our project. The starting point is the historical data set which contains the history data of airline delay. In the data ingestion part, we will design a new schema from the old data set. The new schema should only contain data, time, start airport, destination airport, weather, travel distance and time of delay. Those data will be stored in Amazon S3(D1) as our data lake.

With the help of Amazon Athena, our team can use Amazon S3 to do machine learning and export the prediction result to another Amazon S3(D2) as a data warehouse. After training our model, we will send our model to the real-time data processing.

Our team will use the Airport API to get real time data from specific airports. For example, we can get the data from JFK airport in the following week and store it in another Amazon S3(D3) as a data lake. Next, we use the trained model to predict the real time airline and store those results in the final data warehouse(D4). Finally, the application will query the data warehouse(D4) and replay users the result.

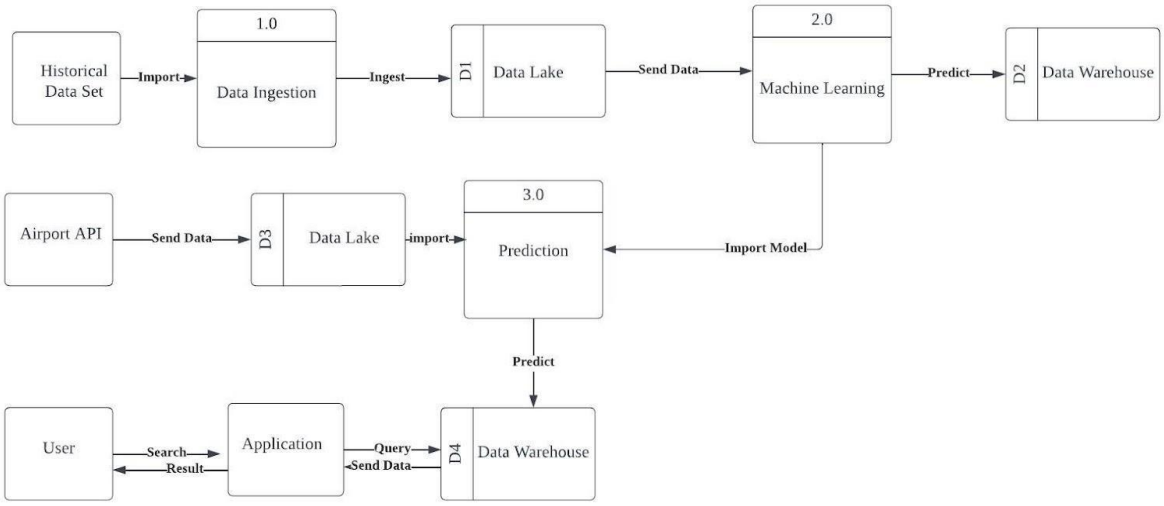

Fig.1 Data Flow Diagram for Airline Delay Prediction

Therefore, the system boasts a meticulous divide between historical data processing, real-time data assimilation, and predictive operations. By leveraging cloud services like Amazon S3 and Athena, it ensures scalability, real-time processing, and efficient data storage. The architecture seamlessly integrates data transformation, machine learning, and real-time prediction to offer users accurate delay predictions, incorporating the strengths of archival data with the dynamism of live flight information.

Data Collection and Ingestion - This phase encompasses the gathering of historical flight data and integrating real-time data using the Airport API. The data is then transformed and stored in a dedicated Data Lake.

Model Development and Training - Features are selected and optimized, and suitable machine learning models are identified and trained. This phase also includes the refinement of the model through hyperparameter tuning.

Real-time Data Integration and Prediction - Here, the trained model is integrated with the real-time data pipeline, and large-scale predictions are conducted, with results stored in the final Data Warehouse.

Maintenance and Monitoring - The project transitions into this stage emphasizing continuous model performance evaluation and system health checks.

Big Data Challenges

One of the biggest challenges overall is financial concerns. Most data pipeline service providers charge a considerable amount of money. Also, the price for querying our airline delay information is very expensive. For the purpose of this project, we only queried the free data available instead of the huge stream of live data. The estimated cost for the whole data pipeline and querying the data is about $1500 per day. This post challenges our projects that we are not able to query and process the real time data.

Another challenge is machine learning. The whole group is not familiar with machine learning, and we spent considerable time struggling in applying the existing machine learning models to our dataset.

Project Outcomes and Discussion

This project aims to revolutionize air travel planning by providing accurate flight delay predictions. Despite financial and technical hurdles, the advancements in model development are encouraging. Our work not only contributes to the field of delay prediction but also demonstrates the effective application of big data and machine learning in addressing complex real-world challenges, providing a valid demonstration of the application of big data pipelines in combination with machine learning to address the real-word issues.

Data Query

In terms of data query or data collection, we first need to find a good data source to train the mode learning model. Luckily, we find the historical data of airline delays from kaggle. The dataset contains all information we need including airline information, airports information, estimated time, actual time, weather information, etc. That valuable information is good enough to train an accurate model. However, the main challenge is to get that information in real life for prediction use.

We research many airport API on the Internet. Most of the APIs do not open for individuals. The rest of them are also expensive to use. In this situation, we decided to use the Airlab API because it offers 5000 queries for free. The downside of Airlab API is it can only give us the next 10 hours of the airline schedule with only 100 rows on each airport for the free tier. Airports API can give us the next 3 days airline schedule but only 50 queries on free tier. We think the volume of the dataset is the most important thing. Therefore, Airlab API is our choice.

We query the airline company, departure airport, arrival airport, estimate departure time, and estimate arrival time from the API. This is another trade-off. In the historical dataset from kaggle, it has information about wheel off and carriver delay, which makes our prediction accurate. However, there is no way to get that information on future airlines.

As a result, we query the top 31 airports in the USA for their future 10 hours of airline schedule and save that information as batch data. And we only train our model with the airline company, departure airport, arrival airport, estimate departure time, and estimate arrival time. After all, that is all information available for the historical dataset and Airlab API.

Machine Learning

Data Processing

In our project, the initial data processing involved examining the Data Frame's information to understand the structure and types of data we were dealing with. Our dataset contained over 5.8 million entries and 28 features, including various delay metrics, times, and identifiers associated with flights. We then addressed missing values, a crucial step in data preprocessing. Significant gaps were found in delays and time-related features, which required careful imputation or exclusion to ensure the integrity of our models.

Next, we conducted a correlation analysis to determine which features had the most significant relationship with our target variable, 'ARR_DELAY'. The correlation matrix revealed that 'DEP_DELAY' had the strongest correlation, followed by various other types of delays. This step was essential for feature selection, helping us narrow down the most influential predictors for the arrival delay.

Lastly, we prepared our data for model training. This involved handling categorical variables, normalizing numerical inputs, and splitting our data into training and testing sets. Ensuring our data was clean and appropriately formatted was vital for the subsequent training of our machine learning models, which included Random Forest, KNN, SVM, Decision Tree, and Neural Networks, each chosen for their unique approaches to learning from data and predicting outcomes.

Classification Models

In this project, a comparison of five machine learning algorithms was conducted to determine their efficacy in predicting flight arrival delays. The algorithms included Random Forest, k-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree, and a Neural Network Model. Each model was trained using a dataset featuring six predictive variables: departure delay, carrier delay, late aircraft delay, National Airspace System delay, weather delay, and taxi-out time. The target variable for prediction was the arrival delay.

Models Comparison

The Random Forest model, known for its robustness and accuracy in regression tasks, outperformed other models with an R-squared value of 0.9527. This indicates that the Random Forest model could explain approximately 95.27% of the variance in the arrival delays, which is a substantial portion, reflecting the model's strong predictive power. The Mean Squared Error (MSE) for this model was 70.906, which is a reasonable error rate given the complexities involved in predicting real-world flight delays.

The Neural Network Model, leveraging its capacity to model complex non-linear relationships, yielded an MSE of 86.021 on the test set. Although higher than the Random Forest model, this error rate is still within acceptable limits, considering the intricacies of the dataset and the potential for overfitting with neural network architectures.

The KNN algorithm, which makes predictions based on the proximity of data points, showed an MSE of 0.7973. However, there is a possibility that this figure represents another performance metric, such as R-squared, which would normally range between 0 and 1 for regression problems. If this is indeed the correct MSE, it would suggest an unusually precise model, which may necessitate a review to ensure the validity of the model evaluation process.

Similarly, the Decision Tree Regressor reported an MSE of 0.8312, and the SVM model reported an MSE of 0.5299. These values, notably lower than those of the Random Forest and Neural Network models, raise questions about their accuracy or the possibility of a metric reporting error. These values, if correct, would indicate an exceptionally high level of predictive accuracy that is rare in the context of flight delay prediction.

Models	Mean Squared Error
Random Forest	95.27
Neural Network	86.021
k-Nearest Neighbor	79.73
Decision Tree	83.12
Support Vector Machine	52.99

Prediction on AIRLAB API data

In our project, we embarked on a comprehensive analysis of flight delay predictions, utilizing an extensive dataset spanning from 2009 to 2018. Initially, we amalgamated a subset of 100,000 records from each yearly dataset into a consolidated training dataset, ensuring a robust and diverse range of historical data points. The prediction models were trained on this rich historical dataset, which included the '2015.csv' file, with the aim to capture and learn from the patterns of delays over the years.

Upon training, our models were put to the test with real-world applications by integrating real-time flight data obtained from the AIRLAB API. This API supplied us with a limited set of features compared to our historical data, providing only 'DEP_TIME', 'OP_CARRIER', 'ORIGIN', and

'DEST". The constraints of receiving fewer variables and the inherent differences in the data structure presented a challenge for maintaining the accuracy levels achieved with the historical dataset. For instance, the 'DEP_TIME' from the API came in a more granular timestamp format, including year, month, day, hour, and minute, which necessitated conversion into a uniform HHMM format to align with our model's expectations. Similarly, the string data for 'OP_CARRIER', 'ORIGIN', and 'DEST" required preprocessing to transform them into a usable format for model input.

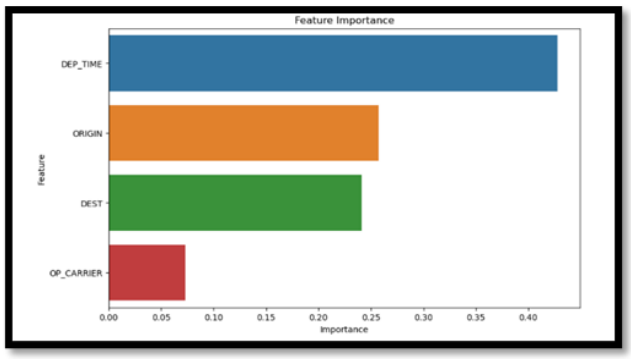

Fig. Feature Importance from API

Despite these challenges, we optimized our data processing pipeline to accommodate the real-time data's format and content, enabling us to feed it effectively into our predictive models. Our focus shifted towards adapting the models to provide reliable predictions based on the real-time data features available, understanding that this might lead to a trade-off with the level of accuracy we had previously achieved. The goal was to harness the predictive power of our models to offer timely and practical insights into flight delays, prioritizing the application of our analysis to the dynamic and often unpredictable nature of live flight data.

Data Visualization

We analyzed the total delayed time per airport, and analyzed the diagram that the airports delay the most as Figure 2. We also generated a diagram comparing the airports that delays the least.

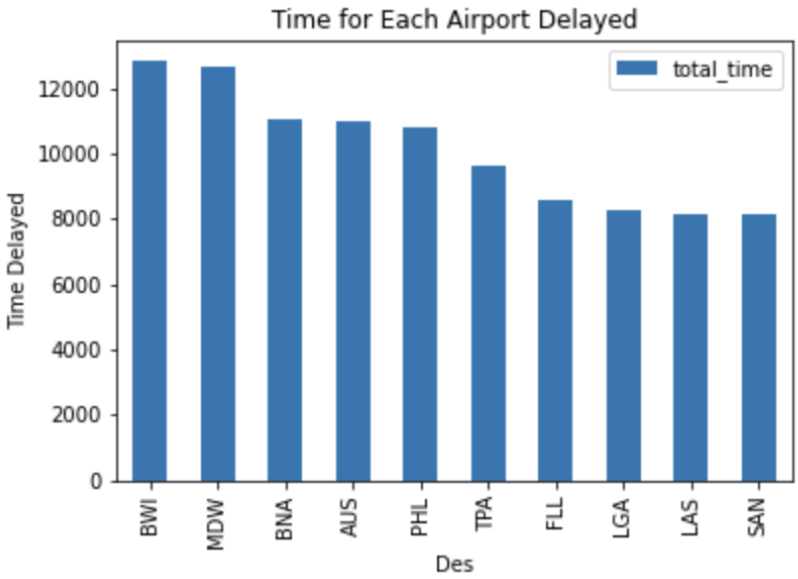

Figure 2. Total delayed time per destination of our sample dataset

Using our machine learning model, we present the projected delay time for all airlines in our dataset. We will take BWI and Southwest as an example to illustrate our data visualization. An example is provided in Table 1. The first column is the code for the airlines. For example, WN stands for Southwest Airlines, and NK stands for Spirit. The outlier of the Southwest delay is attributed to the well-known Southwest huge delay in 2023, resulting in the grand congestion in Baltimore-Washington International Airport (BWI) as shown in Figure 3. The general founding is that all low-cost carriers delay longer than full-service carriers. The carrier's distribution in the airport is provided as well as shown in Figure 4.

OP_CARRIER	Predicted_ARR_DELAY
WN	7764.639000
F9	649.640000
NK	488.627500
KL	283.190000
DL	283.190000
AA	259.830000
B6	250.300000
UA	224.180000
NZ	224.180000
MH	173.180000
BA	173.180000

Table 1. Expected delay time by airlines.

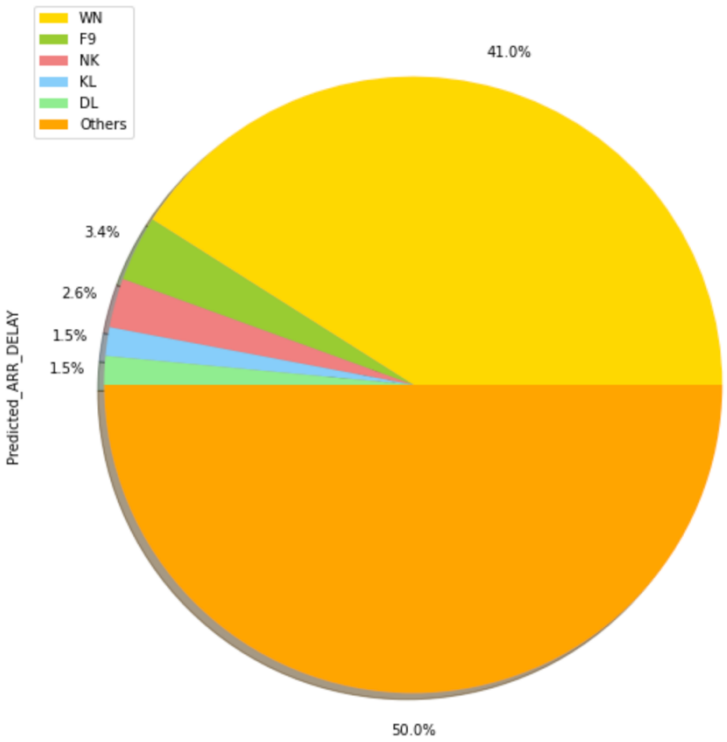

Figure 3. Carries delayed in BWI.

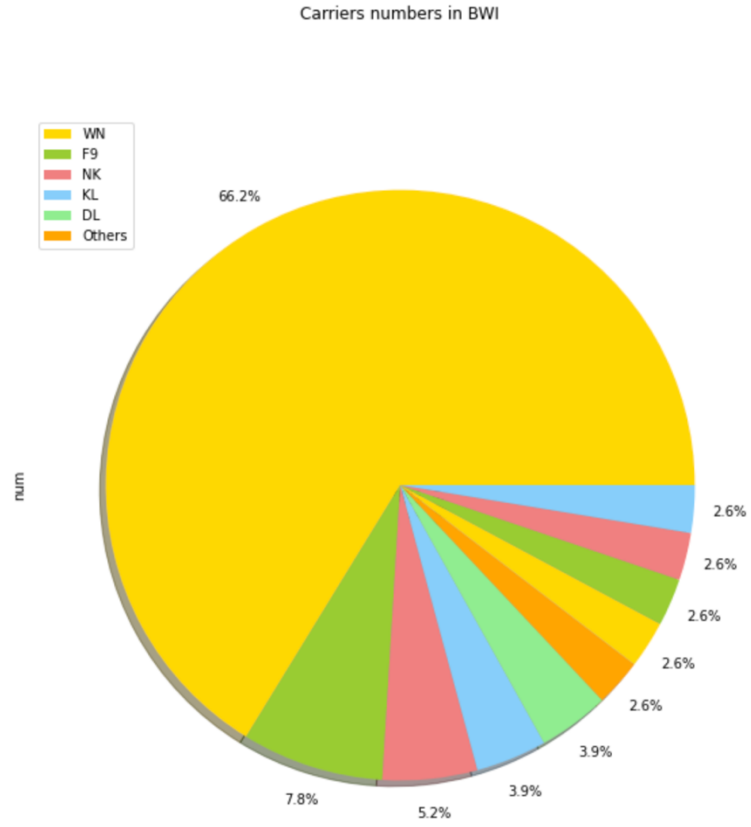

Figure 4. Carrier numbers in BWI.

Surprisingly, we found that the total time distribution at airports forms a normal distribution with an average of 7600 of total delayed time as Figure 5.

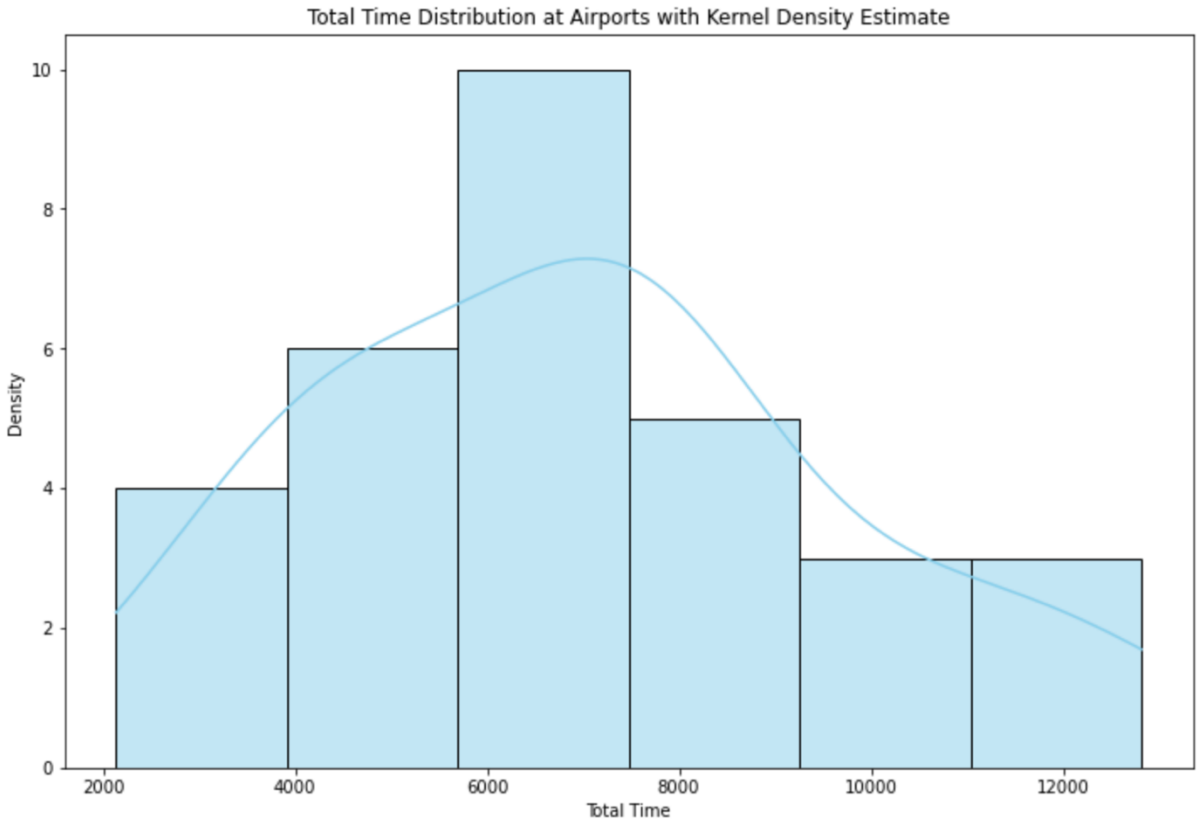

Figure 5. Total time distribution at airports with kernel density estimate.

The above visualizations could be easily deployed to websites or mobile applications for future work.

Conclusion

The project's findings suggest that while all models have their merits, the Random Forest algorithm provided the most reliable predictions for this particular task. The significant difference in error rates between the Random Forest model and the other models emphasizes the importance of algorithm selection in predictive modeling. Nevertheless, the possibility of a reporting error for the KNN, Decision Tree, and SVM models warrants a re-examination of the evaluation metrics to confirm these results. The final model selection would ideally be based not only on statistical

measures but also on considerations such as interpretability, computational efficiency, and the specific operational requirements of the predictive task at hand.

In conclusion, our project's journey through the complexities of flight delay prediction culminated in the development of various machine learning models, each with their own strengths and adaptability to the nuances of aviation data. By leveraging a decade's worth of historical data, we built a robust foundation for our predictive analysis. However, the real test of our models' utility came with their application to real-time data from the AirLab API. Despite facing challenges such as feature limitations and data formatting discrepancies; we innovatively adapted our models to provide actionable insights. While the precision of predictions may have been compromised due to the constrained feature set, our project has laid the groundwork for real-time predictive analytics in air travel, balancing historical knowledge with the agility required for live data application. This endeavor not only showcased the potential of machine learning in practical scenarios but also highlighted the importance of continuous model refinement to address the evolving landscape of data-driven decision-making in the airline industry.

Photo of our team member:

Li Guan:

References

[1] Yuemin Tang. 2021. Airline Flight Delay Prediction Using Machine Learning Models. In *2021 5th International Conference on E-Business and Internet (ICEBI 2021), October 15-17, 2021, Singapore, Singapore*. ACM, New York, NY, USA, 7 Pages. https://doi.org/10.1145/3497701.3497725

[2] G. Gui, F. Liu, J. Sun, J. Yang, Z. Zhou and D. Zhao, "Flight Delay Prediction Based on Aviation Big Data and Machine Learning," in IEEE Transactions on Vehicular Technology, vol. 69, no. 1, pp. 140-150, Jan. 2020, doi: 10.1109/TVT.2019.2954094

[3] Yazdi, M. F., Kamel, S. R., Chabok, S. J. M., & Kheirabadi, M. (2020). Flight delay prediction based on deep learning and Levenberg-Marquart algorithm. *Journal of Big Data*, *7*, 1-28.

[4] "Flight Disruption Insights with Big Data Analytics.", Krishna Bathula, https://activityinsight.pace.edu/ctappert/intellcont/KrishnaBathula-1.pdf

[5] Esmaeilzadeh, Ehsan & Mokhtarimousavi, Sajad. (2020). Machine Learning Approach for Flight Departure Delay Prediction and Analysis. Transportation Research Record: Journal of the Transportation Research Board. 2674. 036119812093001. 10.1177/0361198120930014.

[6] H. Khaksar; A. Sheikholeslami. "Airline delay prediction by machine learning algorithms". *Scientia Iranica*, 26, 5, 2019, 2689-2702. doi: 10.24200/sci.2017.20020