DelayWiz

IEORE4523_001_2023_3 - DATA ANALYTICS

Vidhi Agrawal, Leonel Guevara, Leda Kyrialli, Yi Zhang, Runkai Zhou

17th December, 2023

1. Introduction

The core issue we aim to address in our project revolves around the prevalent challenge of flight delays experienced by travelers. It's a pervasive concern that can disrupt plans and schedules. These delays can result in missed or canceled activities, causing inconveniences that resonate throughout the entire travel experience. We wanted to create a model that could help customers predict if a flight would be delayed so they may be able to make any itinerary changes to account for a delay in their flight. We decided to use a dataset from the ***Bureau of Transportation Statistics**** with a date range of ***2021-05-01 to 2023-04-30***, this data is from actual flights in the United States for different airports and airlines. We decided to use ***John F. Kennedy International Airport (JFK)*** as our choice of airport to analyze with a total of just over ***240,000*** domestic flights. This dataset contains features like airline, departure_time, arrival_destination, distance etc. Prior to building our models, we came up with an assumption about what could cause a delay, "Weather will have a high impact on delays," and we also thought that it would be 'hard to predict extreme delays as they occur less often.' Our concern was that the data we had acquired did not have weather information, we had to go to the ***National Centers For Environmental Information**** to obtain this information which contains attributes on temperature, fog, wind speed etc. which we merged with the flights' dataset. As we progressed with our model of predicting if a flight would be delayed or not, we also decided if we could create a model to find what the delay time would actually be. This report will walk through our process, from cleaning the data we acquired to creating the models, how we addressed any issues in our datasets or models, describing the models, and how successful we were with the 2 models.

2. Data Preprocessing

Upon retrieving the dataset, our team initiated basic data cleaning procedures. To facilitate subsequent analyses, we reformatted the time-related attributes by delineating distinct variables for date, time, and day of the week. Additionally, we segmented each day into four distinct periods: morning, afternoon, evening, and night, for more granular analysis.

Our dataset looks balanced (59%:41%). During the cleaning process, we identified blank cells in the dataset, attributable to canceled flights. These constituted a mere 2.8% of the entire dataset, leading to the exclusion of 7,039 data rows encompassing these blank cells. In the weather dataset, instances of "NaN" were converted to "0" to quantitatively represent the absence of specific weather conditions. Furthermore, we introduced a dummy variable, "Delay", to signify instances of flight departure delays. This was determined by comparing the scheduled departure time (CSR_DEP_TIME) with the actual departure time (DEP_TIME).

In the final phase of data preparation, we amalgamated the monthly flight dataset with the weather data, using the "DATE" extracted from the time attributes. Given that the weather dataset's precision is limited to daily data and weather conditions vary throughout the day, this integration potentially introduces certain limitations or biases in our analysis.

Upon a more detailed examination of our data, the team opted to refine the dataset to enhance the accuracy of our model. We identified a subset of exceptionally prolonged delay cases (1,192 records, representing 0.48% of the entire dataset) and classified them as "outliers" following a graphical analysis. The exclusion of these outliers from our dataset was deemed essential for improving the model's accuracy. Consequently, we resolved to eliminate these records of extreme delays from our analysis. As a last step, we ensured all the numerical

columns are standardized to avoid introducing any biases as models are sensitive to feature scales and all categorical variables are encoded for the model to be able to interpret them as categories.

3. Exploratory Data Analysis

In our exploratory data analysis phase, we delved into our dataset to unveil patterns, trends, and relationships. To enhance our understanding, we extracted temporal details, such as day, month, and year from the FL_DATE variable, and the hour from the DEP_time variable. Additionally, we introduced a binary variable named DELAY, assigned a value of 1 if the flight experienced a delay (DEP_DELAY >= 0) and 0 otherwise.

Our initial visualization aimed at portraying the proportion of delayed flights by airline carriers, revealing a conspicuously high percentage. This observation substantiates the motivation behind our project, offering compelling evidence to potential stakeholders and airline operators of the substantial value that can be harnessed through this predictive model (see Appendix Figure 1).

Moving forward, we conducted a correlation analysis utilizing a heatmap to scrutinize multicollinearity (see Appendix Figure 2). This process enabled us to identify and omit variables that could potentially impact our model's performance, such as correlated weather conditions. Notably, we pinpointed variables with a pronounced correlation with our target variable, DEP_DELAY.Further insights were gained through visualizations of key variables. For instance, a bar chart depicted the distribution of binary variables indicating reasons for delays, revealing notable patterns, especially in carrier delays and issues with the arrival time of the aircraft from a previous flight (see Appendix Figure 3).To understand the temporal aspects of flight delays, we visualized the distribution of departure hours. Notably, a dense pattern emerged between 15:00 and 20:00, shedding light on periods of heightened activity (see Appendix Figure 4).One intriguing categorical variable omitted from the initial correlation analysis was the Destination (see Appendix Figure 5). Our exploration uncovered that certain destinations, notably in states like California, Boston, and Florida, exhibited a significant number of delayed flights.

In summary, our exploratory analysis provided valuable insights into the dataset, informing our feature engineering decisions and enhancing our understanding of the relationships and dynamics at play.

4. Modeling and Result Analysis

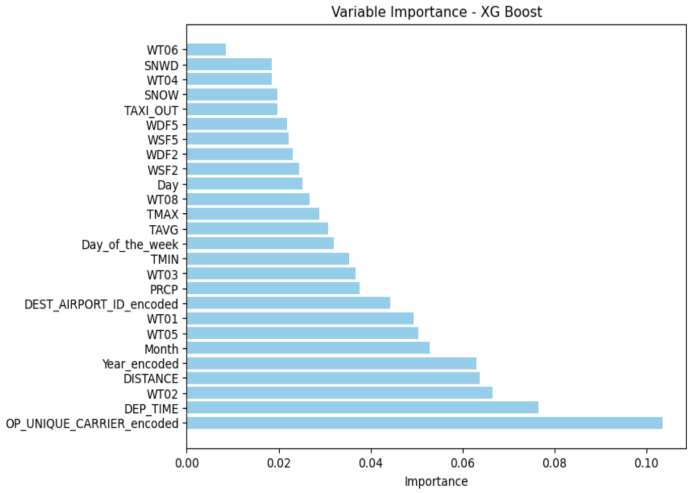
4.1 Predicting Delay Status (Binary Problem)

The first objective of the project was to build a binary classification model that would predict whether a flight will get delayed or not- target variable- ‘DELAY’. We build this model on the data that was prepared during the pre-processing step. We use sklearn’s train_test split functionality to split the entire dataset into two parts: Training(80%) and Testing(20%). Subsequently, we experimented with multiple classification algorithms like LogisticRegression, Decision Trees, RandomForest, XGBoost, LightGBM and HistGradientBoosting and NNs.

Model	Accuracy	F1 score	Recall
HistGradientBoosting	75.9%	0.74	0.73
Logistic Regression	65%	0.60	0.60
XG Boost	82.7%	0.83	0.83
Random Forest	61%	0.40	0.51
LightGBM	82.4%	0.81	0.78

The comparative analysis revealed Gradient Boosting algorithms' superior performance over Bagging Classifiers, attributed to their adeptness in capturing intricate relationships. Further optimization via GridSearchCV for hyperparameter tuning notably enhanced model efficacy. Employing k-fold

cross-validation (achieving 82% accuracy) ensured consistent model performance, with **Recall** being the pivotal metric due to the substantial cost implications of missing delayed flights. XGBoost emerged as the optimal choice, boasting the highest Recall, accuracy, and f1-score.



The variable importance plot for XG Boost revealed significant influencers in delay predictions, such as specific airlines, Departure_time, Distance, and select weather variables like fog, heavy fog and hail (e.g., WT05, WT01). The challenge of encoding categorical data was met with surprise as LabelEncoding outperformed One Hot Encoding, despite initial concerns regarding potential ordinality. Extensive research concluded that Gradient Boosting algorithms, reliant on minimizing entropy as tree-based methods, were unaffected by the chosen encoding method. This comprehensive exploration underscores the critical role of feature importance assessment and encoding methodologies in refining model performance. We integrated SHAP values, a game theory-based model explainability tool, providing an alternative feature importance plot (Appendix Figure 6) for the same model. Although resembling the sklearn plot, the SHAP-based plot differs due to their distinct calculation methodologies—SHAP values employing game theory and sklearn using Gini index. While both highlight feature importance, SHAP values offer insights into magnitude and directionality of importance, setting them apart.

4.2 Predicting Delay Time in Minutes (Continuous Problem)

In addition to estimating the likelihood of a flight being delayed, it is also beneficial to employ machine learning models to predict the duration of the delay. To this end, we selected 28 independent variables, encompassing both flight-related factors and weather conditions, as features to forecast the dependent variable (DEP_DELAY), which represents the delay duration.

We employed two models for this analysis: Linear Regression (LR) and Support Vector Regression (SVR). Standard data preprocessing techniques were applied. This included data standardization to ensure uniformity across all numeric variables and the removal of rows with invalid inputs. Additionally, flights with excessively long delays (≥300 minutes) were excluded, as they were deemed outliers.

Upon preparing the data, we divided it into a training set and a test set. The training set, comprising 75% of the total flights, was used to train the model, while the remaining 25% formed the test set, used to assess the model's predictive performance. The models' predictions were then compared against the actual delay times, evaluated using the Mean Absolute Error (MAE) metric.

	Linear Regression	Support Vector Regression
Mean Absolute Error	22.24	17.91
Root Mean Squared Error	37.84	40.08

Although the Root Mean Squared Error (RMSE) was also calculated for reference, we focused on MAE as it is less sensitive to extreme delay times. Despite the exclusion of outliers (delays ≥300 minutes), some flights still experienced significant delays (up to 200+ minutes), which could disproportionately increase RMSE. Consequently, MAE, with values of approximately 22.24 for LR and 17.91 for SVR, offered a more relevant measure, indicating that, on average, the absolute difference between the predicted and actual delay times was about 22 minutes for LR and 18 minutes for SVR. In terms of MAE, SVR demonstrated marginally superior predictive accuracy compared to LR.

Further analysis of the distribution of residuals (the differences between predicted and actual delay times) for each model provided additional insights (see Appendix Figure 7). For LR, the residuals were most frequently observed in the 5-10 minute range, suggesting a tendency of this model to overestimate delay times. In contrast, the residuals for SVR were most commonly around zero, indicating a more accurate general prediction of delay times.

However, it is not straightforward to conclude that SVR is categorically superior to LR for predicting delay times. A detailed examination of the data, where actual and predicted delay times are juxtaposed, reveals that although SVR generally offers greater accuracy, it underperforms in predicting longer delays. In contrast, LR tends to be more accurate for longer delays ranging from 30 to 60 minutes (see Appendix Figure 8).

This discrepancy can be attributed to the distribution of each model's predictions. LR is capable of predicting a broader range of delay durations, whereas SVR is restricted to a narrower range, primarily shorter delays. Contextually, LR is more adept at capturing longer flight delays, while SVR achieves higher accuracy for shorter delays. Consequently, the selection of an appropriate model depends on the specific scenario. A prudent strategy might involve utilizing LR for flights historically prone to longer delays and resorting to SVR in other cases.

The key takeaway from this analysis is that the model with the most favorable evaluation metrics may not always be the optimal choice for every situation. Each model has its strengths and weaknesses, necessitating careful consideration under various circumstances. Understanding the underlying reasons for a model's performance is crucial in determining the most appropriate application for each.

Conclusion

DelayWiz is capable of predicting a majority of flight delay scenarios, employing a linear model for long-duration delays and a Support Vector Machine (SVM) model for shorter delays. However, the accuracy of our model may diminish in predicting extreme circumstances, often attributable to unforeseen, uncontrollable events. Moving forward, the integration of a more extensive dataset enriched with additional explanatory variables holds the potential to enhance the precision of our predictions and possibly broaden their applicability.

Appendix

1. **Bureau of Transportation & Statistics- https://www.transtats.bts.gov/Fields.asp?gnoyr_VQ=FGK**
2. ***National Centers For Environmental Information-*** **https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00094789/detail**

Figure 1

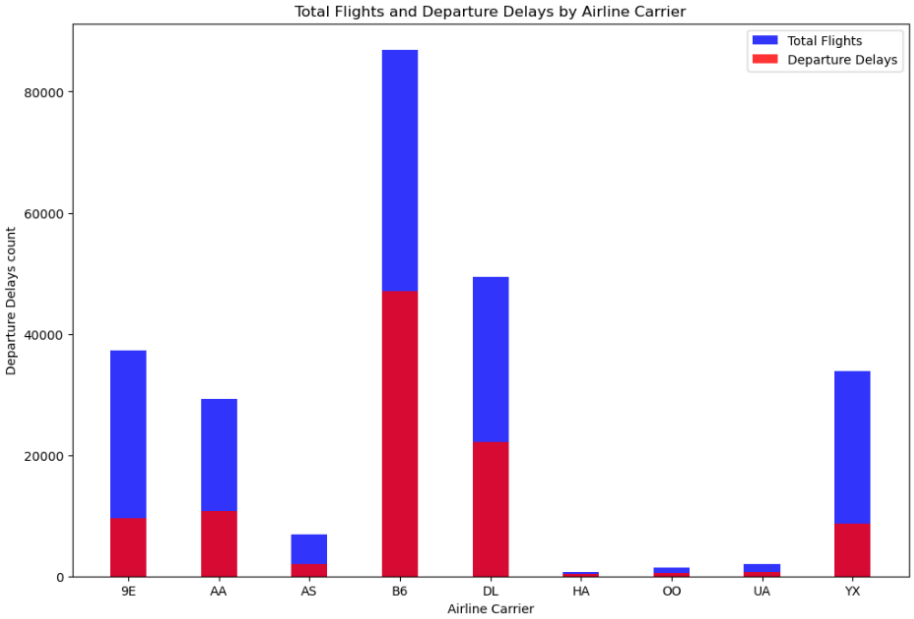


Figure 2

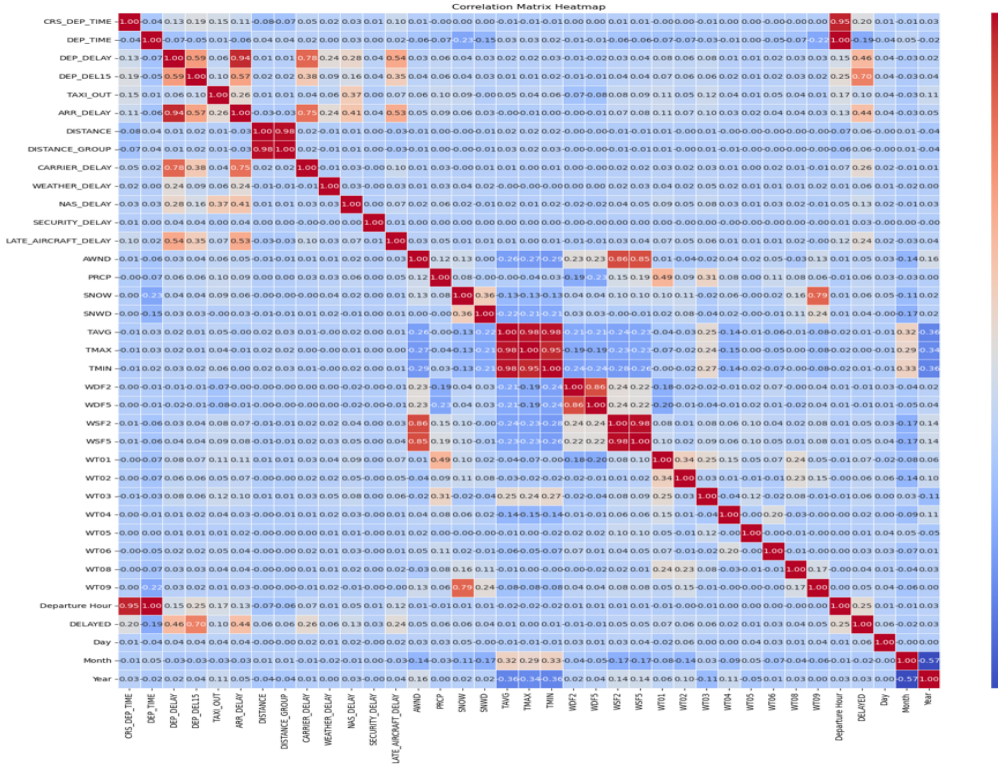


Figure 3

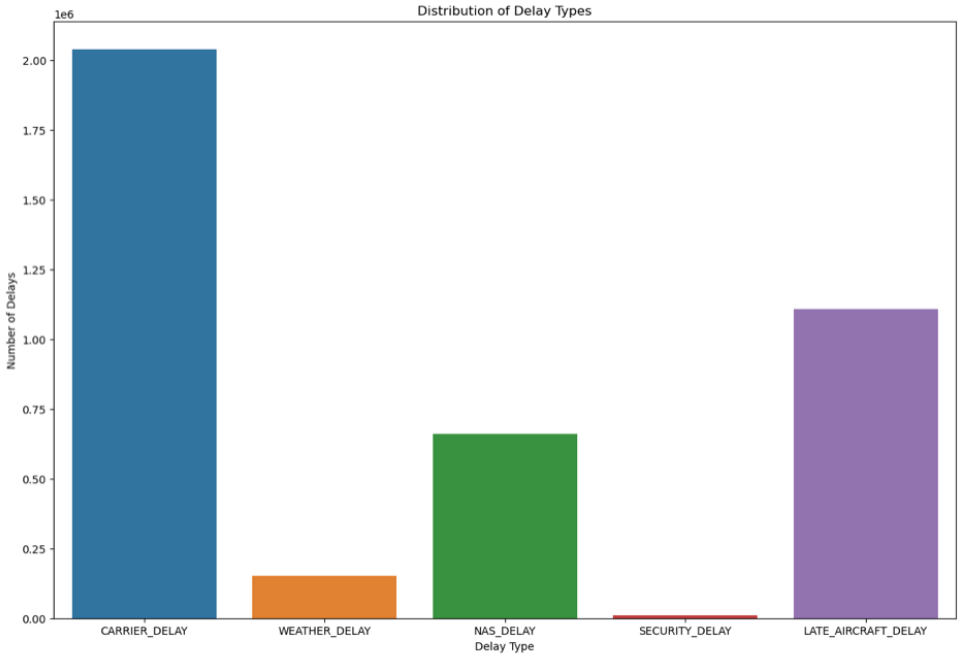


Figure 4

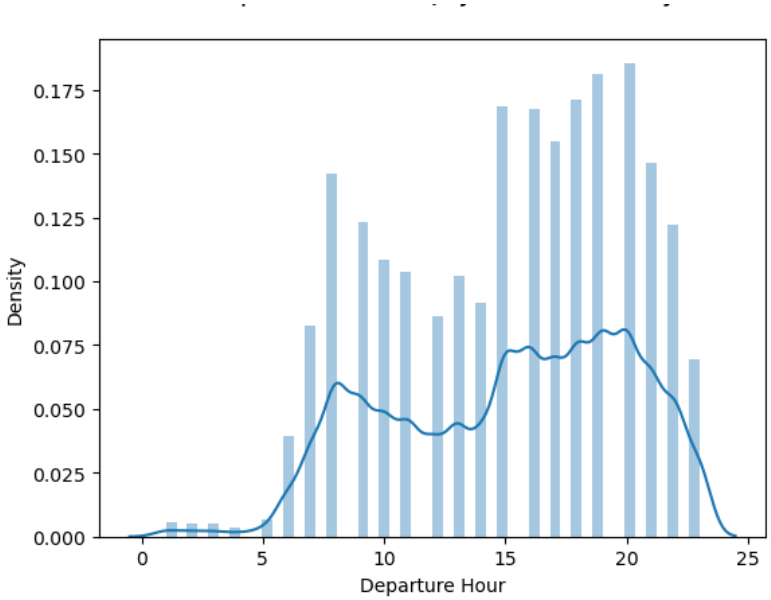


Figure 5

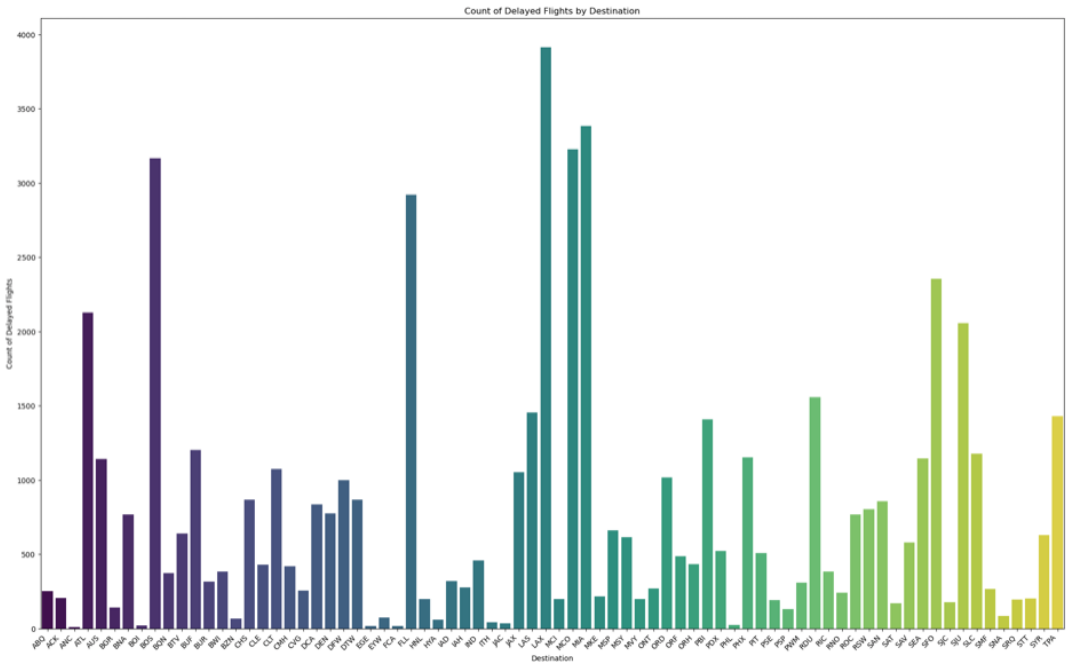


Figure 6

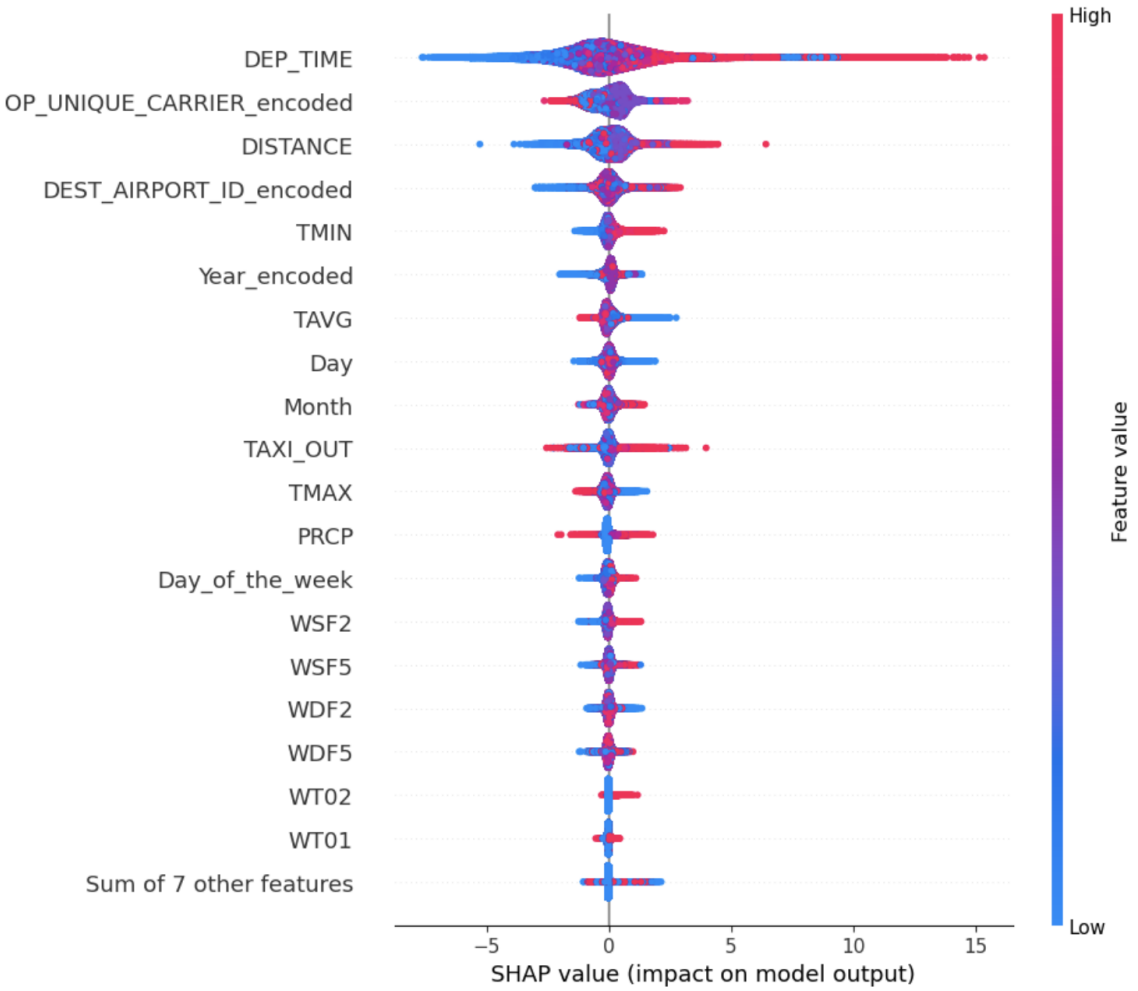
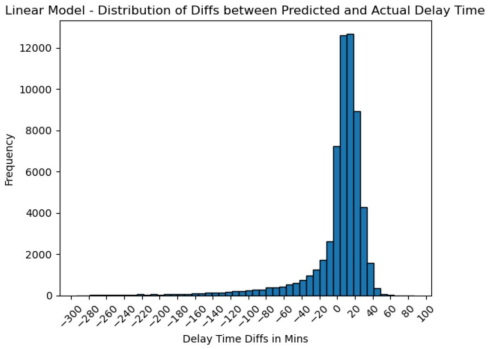


Figure 7



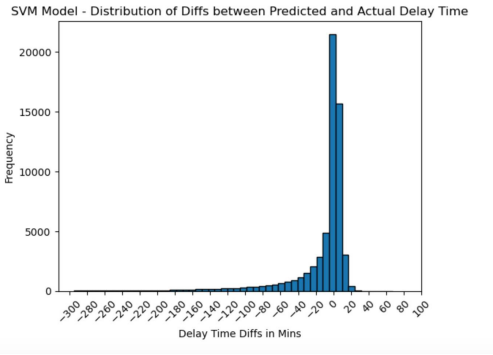


Figure 8

	DEP_DELAY	Prediction_linear	Diff_linear	DEP_DELAY	Prediction_SVM	Diff_SVM
103727	-3.0	24.171453	27.171453	-3.0	3.274146	6.274146
110212	13.0	-6.109699	-19.109699	13.0	-4.772322	-17.772322
154174	-2.0	-4.162542	-2.162542	-2.0	-3.799438	-1.799438
239240	-2.0	-6.327225	-4.327225	-2.0	-4.701268	-2.701268
111335	0.0	5.216472	5.216472	0.0	-6.623870	-6.623870
165199	106.0	15.461103	-90.538897	106.0	0.382808	-105.617192
123828	242.0	5.097101	-236.902899	242.0	-1.911146	-243.911146
31363	28.0	26.737896	-1.262104	28.0	7.003469	-20.996531
27177	-3.0	24.451621	27.451621	-3.0	7.830660	10.830660
166386	26.0	22.527182	-3.472818	26.0	2.369565	-23.630435
107151	4.0	12.481456	8.481456	4.0	-2.358735	-6.358735
127891	193.0	25.393353	-167.606647	193.0	2.466868	-190.533132
117056	-4.0	-1.223968	2.776032	-4.0	-6.045593	-2.045593
131181	-7.0	2.401674	9.401674	-7.0	-3.496723	3.503277
165986	53.0	35.329986	-17.670014	53.0	6.445616	-46.554384