**Predicting Airline Flight Delay:**

**An Analysis of Bureau of Transportation Flight Data**

Ned Kost, Brian Morris, Anova Youngers

Shiley-Marcos School of Engineering

MSAAI-501: Introduction to Artificial Intelligence

Prof. Andrew Van Benschoten & Dave Friesen

University of San Diego

**Introduction**

Flight delays are injurious to both passengers and airlines, costing a respective $16.7 billion and $8.3 billion annually (Anupkumar, 2023). The environmental impact for delayed flights can also be measured as delayed flights burn an additional 5529 tons (Dissanayaka et al, 2018) to try and make up time. Beyond these measurable factors, the frustration and dissatisfaction felt by travelers tarnishes the reputation of airlines and airports alike. It is important to study airline flight delays, and how airlines can modify their operational practices to predict and handle these events.

Our approach involves meticulously examining three years’ worth of data, spanning temporal, meteorological, airline carrier, and airport information, with an emphasis on refining data quality through techniques like balancing, encoding, and the integration of supplementary datasets. Beyond achieving delay predictions, our objective is to identify influential factors at each stage of a flight's lifecycle, offering valuable insights for enhancing airline operational efficiency. This paper delineates the developmental journey from initial data exploration and feature augmentation to model tuning and interpretation.

We utilize several machine learning models suited for binary classification including Logistic Regression, Gradient Boosting, Random Forest, and Neural Networks, to predict delays of 15 minutes or more. We process and segment data into training and testing sets, with the subsequent evaluation and comparison of machine learning models based on performance metrics. For this project we focus on Area Under the Curve (AUC), F1 Score, and Accuracy as our primary metrics to measure binary classification. The ensuing comparison of these models using these performance metrics reveals the best-performing model recommended for use by the airline industry.

**Data Exploration and Analysis**

Our investigation leverages data from the Bureau of Transportation Statistics' On-Time Performance dataset, covering 2020, 2021, and 2022. This dataset encompasses a comprehensive array of domestic flight performance metrics, spanning all major airline carriers and encompassing over 300 airports. Complementing this primary dataset, we incorporate airliner carrier employment counts, obtained from the Bureau of Transportation, and weather data sourced from the National Oceanic and Atmospheric Administration. The weather data, including precipitation and wind speed, is obtained through a lengthy data scraping process involving weather APIs, resulting in the augmentation of approximately 360,000 records. In contrast, the primary dataset comprises an extensive 16 million records. To address the data's inherent imbalance, where on-time flights outnumbered delays 7 to 1, we balanced our dataset with equal parts of both flight types to ensure an unbiased representative model.

Upon curating our datasets, flights categorized as Canceled or Diverted are filtered out, aligning with the scope of this study. Subsequent to this refinement, we conduct a rigorous data exploration to gain nuanced insights into the dataset. A correlation heatmap, applied to numerical features (refer to Figure 1.1), facilitates the identification of redundancies and potential predictors. Notably, Departure Delay and Arrival Delay exhibit a high correlation coefficient (0.97), suggesting a robust predictive relationship. Further examination through a scatterplot visually corroborates this association, revealing a linear relationship with a regression line slope approaching unity (refer to Figure 1.2). The standard deviation lines, situated at one standard deviation above and below the mean, encapsulate most data points, elucidating the consistency of the delay pattern.

In addressing our dataset, a key challenge we faced was managing the extensive dimensionality presented by the carriers and airports. To tackle this, we experimented with various encoding techniques, including binary encoding, label encoding, custom grouping, and one-hot encoding. Ultimately, we chose a combination of binary encoding for airports and one-hot encoding for carriers.

This decision was guided by the dataset's structure: it comprised approximately 15 carriers, a number manageable enough to employ one-hot encoding without excessively increasing the dataset's dimensionality. Conversely, the dataset featured a substantial number of airports, totaling 374 which were captured as both destination and arrival airports. Following an in-depth exploration of various approaches, we opted to binary encode the airport data. This methodology efficiently condensed the airport information into 9 columns each for both destination and arrival airports, thus addressing the challenge of high dimensionality while preserving essential information.

The subsequent application of the ANOVA test yielded a F-statistic of 158.78, accompanied by an effective p-value of 0. This result firmly establishes the statistical significance of variations in arrival delays across different airline carriers. Furthermore, a T-test comparing weekend and weekday departure delays produced a t-statistic of 8.18 and an extremely low p-value (2.74e-16), reinforcing the notion that the day of the week is an important feature.

After exploring and analyzing potential features within the dataset, we settled on three feature sets each serving distinct business objective, and categorized based on their availability before prediction can occur (Figure 1.3):

**Feature Set 1**: Known over a week before the flight, aiding in the modification of airline

scheduling and other long-term strategies.

**Feature Set 2**: Available after the aircraft departs from the origin airport, enhancing the

accuracy of same-day delay predictions and providing insights into operational

Inefficiencies.

**Feature Set 3**: Mainly comprising dependent variables, with additions such as ‘Air

Time’, 'WheelsOn' and 'TaxiIn’. This feature set is better for statistical analysis than

modeling, as the data pollutes the model with arrival details.

Together, these feature sets enable a holistic and nuanced approach to machine learning, equipping us to develop models that not only predict delays accurately but also contribute to informed decision-making within the aviation industry.

**Data Modeling Algorithms**

**Logistic Regression**

Due to the strong linear relationship between Departure Delay and Arrival Delay, it was logical to start modeling with a Logistic Regression model. Using only this feature provides an AUC of .908, indicating that there is a critical threshold where if reached, nearly guarantees a delayed arrival. Performing a mean comparison between delayed and on-time flights shows that a delay of 1 hour implies the flight will be delayed with 95% accuracy. Other factors also contribute to delayed likelihood. Layering in the rest of the features across sets 1, 2, and 3 boots the model performance to an AUC of .957. For a simple model, this is a great result that shows promise as simpler models are easier to understand and generalize better than complicated ones.

Removing feature set 3 reduces the mode’s AUC to .926 since it no longer considers Airtime, which is an important predictive feature as pilots may try to make-up time by traveling faster. Further reducing the features by eliminating set 2 has a significant impact on the AUC, bringing it down to .616. Without the strongest indicator Departure Delay, the Logicist Regression model struggles to do better than random chance by using base information about the flight date, projected departure/arrival times, airport, weather, and carrier data.

Attempting to train the model with the full 3-year dataset highlighted the imbalance issue present between Delayed and On-time flights, and the model generated a “Perfect separation detected” error. After balancing the dataset with a proportional number of on-time flights, the model still did poorly with an AUC of .547, barely better than random guessing. The additional reduction of AUC is likely due to missing weather data from our test set.

**Gradient Boosting**

Training the XGBoost model with all three feature sets, and tuning the parameters, provided an impressive AUC of .999, and F1 score of .989. For the hyper parameter tuning we utilize a stratified K-Fold with grid search for the n\_estimators and learning\_rate parameters. The initial run indicates a learning rate .1 or .2 might be best. Since the log loss graph showed both learning rates increasing along with n\_estimates, a second tuning run with a larger range of n\_estimator was used. The learning rate of .1 and n\_estimators of 800 was the clear winner in the second run (see figure 1.4)

Reducing the feature set down to 1 and 2 degrades the model performance slightly with a resultant AUC of .968. These are still good metrics indicating possible model usefulness for airline operations up till the plane departs the airport. Reducing the feature set further down to only set 1 has a large impact on the performance resulting in an AUC of .692. Removing weather reduces the AUC to .675, indicating that weather is not as useful as we originally thought it might be. Again, utilizing only pre-departure data reduces the feasibility of the model.

To test the model for overfitting, we utilize a “Test Data Set” with 7 months of flight data from 2023, a year outside of our initial analysis. Running with the model using Feature Set 1 only results in an AUC of .63 which is slightly lower, indicating a bit of overfitting is present.

**Random Forest**

When training a random forest model, the primary focus was on Feature Set 1. One key aspect of our investigation was to understand the trade-off between a smaller dataset enriched with weather features, comprising 350,000 records, and a larger dataset consisting of 8 million records but with fewer features. Initially, there was a lack a clear intuition regarding the impact of large training sets on model performance. Specifically, it proved intriguing to ascertain whether the smaller dataset, augmented with weather data, would outperform the larger, less feature-rich dataset in terms of accuracy.

The primary insight from this analysis is the importance of a larger dataset. Despite the smaller dataset showing increased accuracy when including additional weather features, increasing our data volume by approximately 20-fold proved more effective. While we couldn’t enrich the larger dataset with weather data, the improvements seen in the smaller dataset suggest that incorporating such features could potentially enhance accuracy by a few percentage points.

**Neural Network**

  In developing a neural network model to predict flight delays, our process focused on the following crucial steps for optimal performance.

* Feature Selection and Encoding: We identified essential features like temporal variables and categorized top airports, leading to a dataset with 16 million records and 16 features. One-hot encoding was applied to carriers, enhancing our model's ability to interpret this data.
* Data Partitioning and Oversampling: The dataset was split into training, validation, and test sets, with an equal 70/15/15 distribution. We undersampled the majority class of on-time flights, ensuring a fair representation of classes.
* Initial Model Construction and Evaluation: The initial model included dense layers with ReLU activation, batch normalization, and dropout to mitigate overfitting. Compiled with an Adam optimizer, it achieved an accuracy of 66.56% on feature set 1 and  91.84% on feature sets 1 & 2.
* Complex Model Exploration: To enhance performance, we built a more complex model with varied dropout rates and additional layers. However, it also reached a similar accuracy level, indicating a possible limit in predictive capability with the current data.

     Both models achieved comparable accuracy, suggesting the chosen features effectively captured delay influences. The similar performance of both models indicates potential predictive limits with the specific features chosen.  In summary, the models established a solid foundation for predicting flight delays, emphasizing the importance of feature selection, data balancing, and model architecture.

**Summary**

Our research into factors contributing to delayed arrivals was enlightening, with several key concepts surfacing. Looking at the data features, it is no surprise that “Departure Delay” was the most impactful. Using means testing to identify the 1-hour mark as a general indicator for arrival delay is valuable for further investigation (potentially looking at flights of different lengths do find other boundaries). Weather variables, and information about the carrier employee counts, were both low drivers in arrival delays making up only ~2-5% of model improvements when included.

With a single feature the Logistic Regression model achieved an initial AUC of 0.908. Incorporating additional features increased this score to 0.957, highlighting the effectiveness of a simple model. Removing specific feature sets significantly impacted performance with the model’s AUC dropping to 0.616 without ‘Departure Delay’. Training with a larger dataset initially led to a “perfect separation detected” error. After balancing, the model performed marginally better than random guessing (AUC of 0.547).

The Random Forest model demonstrates larger datasets significantly enhance model performance, as evidenced by the superior results of an 8 million record dataset without weather features compared to a smaller, enriched dataset of 360,000 records with weather data. This underlines the crucial role of data volume in predictive modeling.

The Gradient Boosting model, specifically XGBoost, demonstrated high performance with all three feature sets, achieving an AUC of 0.999 and an F1 of 0.989. There was a notable decrease in performance when the feature sets were reduced from 3 to 1 (AUC of 0.968 - 0.692). A test for overfitting by using a dataset from a different year (2023) indicated a slight overfitting in the model.

In the development of the Neural Network models, two different architectures were constructed, both achieving an accuracy of approximately 91.84%. This is indicative of a potential ceiling in the predictive capabilities of the models given the selected features. This highlights the importance of feature selection, data balancing, and model architecture in predicting flight delay.

One area that warrants additional research would be identifying overfitting in the models. When we tested a couple with data from 2023, we saw lower results that may imply some data memorization by the models. Due to time constraints and issues using matching the categorical encoding across the 2023 and historical data, we were unable to explore this topic in detail.

These models demonstrate varied success in predicting U.S. flight delays, with the volume and quality of data, along with feature selection, playing pivotal roles in determining model effectiveness. The Random Forest and Neural Network models highlight the importance of data size, while the Gradient Boosting and Logistic Regression models underscore the significance of feature selection and the challenge of data imbalance.

When looking at the final accuracy results for the models across feature sets, we see Feature Set 1, comprising pre-departure flight details, yields the highest accuracy in the XGBoost Classifier (0.6931), closely followed by Random Forest (0.6916) and Neural Net (0.6656).  
When incorporating Feature Set 2, there's a notable accuracy boost across models. Random Forest leads (0.9208), with Neural Net (0.9184) and XGBoost (0.9135) closely behind. This indicates that post-departure data significantly enhances model accuracy, highlighting its importance in predicting flight-related outcomes more effectively (see Figure 1.7 and Figure 1.8).

**References**

Anupkumar A., (2023), Investigating the Costs and Economic Impact of Flight Delays in the Aviation Industry and the Potential Strategies for Reduction, California State University San Bernardino, <https://scholarworks.lib.csusb.edu/cgi/viewcontent.cgi?article=2885&context=etd>

Bureau of Transportation Statistics. (2023). Carrier Employment Data (2005-present). Retrieved from <https://www.transtats.bts.gov/Employment/>

Bureau of Transportation Statistics. (2023). On-Time: Reporting Carrier On-Time Performance (1987-present). Retrieved from<https://www.transtats.bts.gov/Fields.asp?gnoyr_VQ=FGJ>

Dissanayaka, D., Adikariwattage, V., Pasindu, H. (2018). Evaluation of Emissions from Delayed Departure Flights at Bandaranaike International Airport (BIA). In Proceedings of the 11th Asia Pacific Transportation and the Environment Conference (APTE 2018), Malang, Indonesia, Atlantis Press: Paris, France, 2019; pp. 143–146.

Kost N., Morris B., & Youngers, A. (2023). MS-AAI-501-Team-Project. GitHub. <https://github.com/NedKost/MS-AAI-501-Team-Project>

National Oceanic and Atmospheric Administration (accessed in November 2023): Global Hourly–Integrated Surface Database (ISD). Available online: <https://www.ncei.noaa.gov/products/land-based-station/integrated-surface-database>

**Appendix**

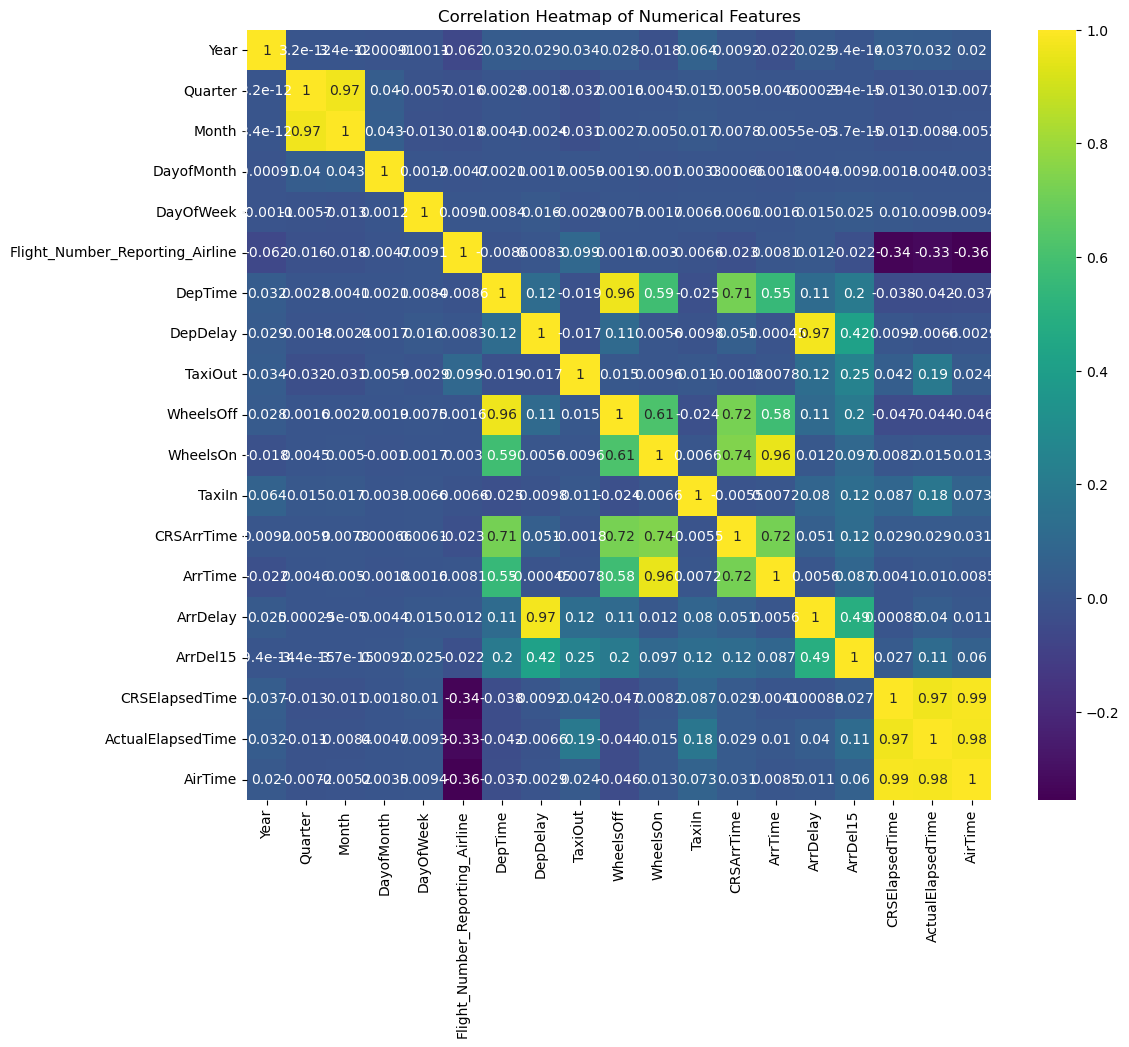
Figure 1.1 Correlation Heat map.

Figure 1.2 Arrival Delay vs Departure Delay Scatter Plot.

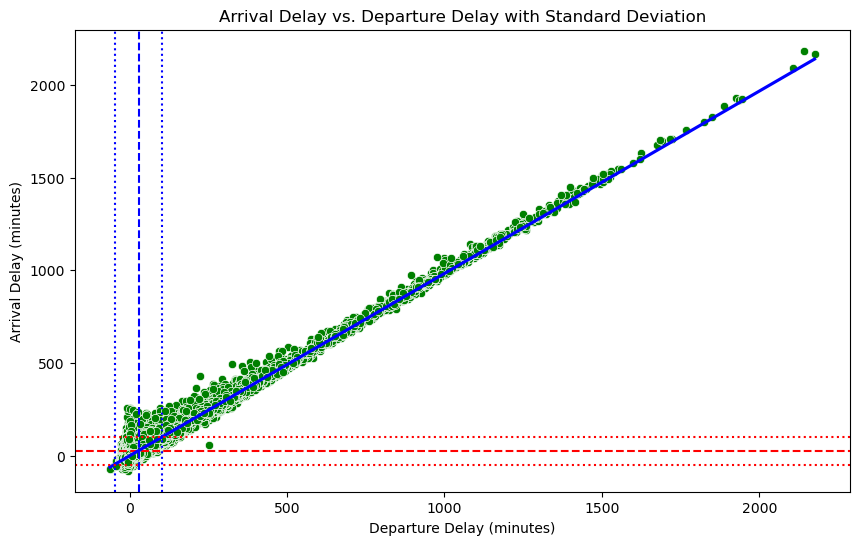


Figure 1.3 Feature Sets

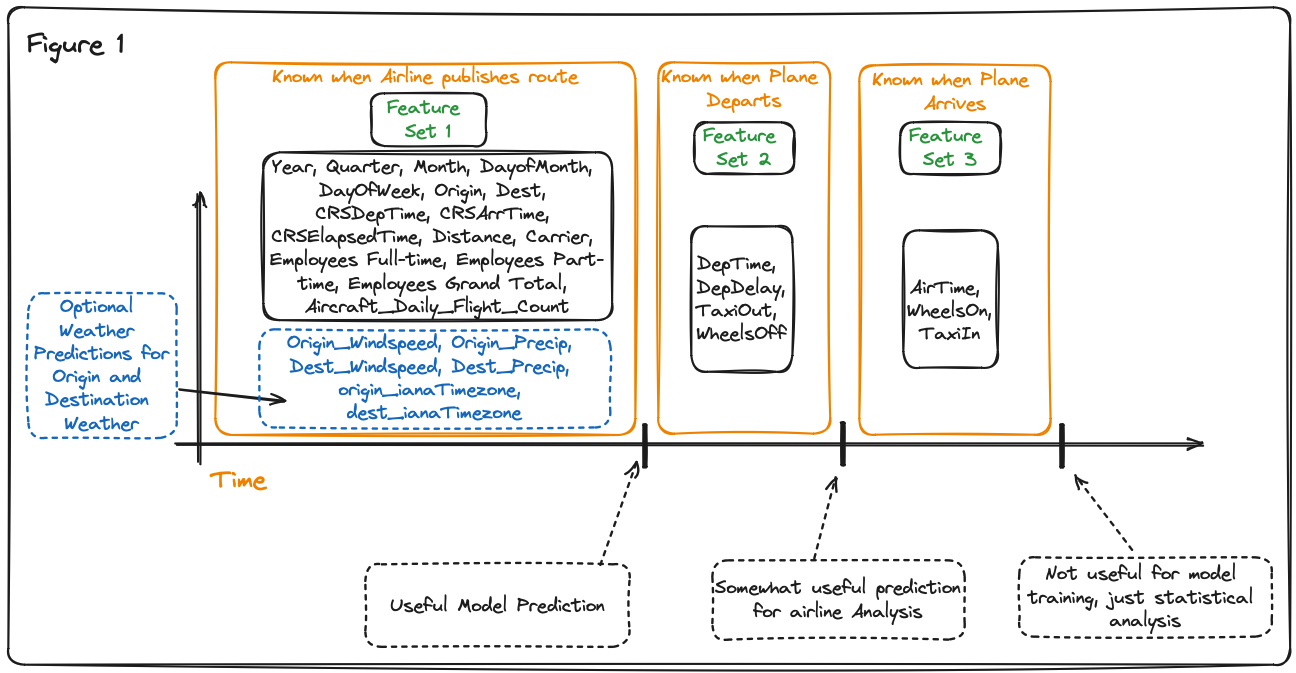


Figure 1.4 Gradient Boosting Loss Graph

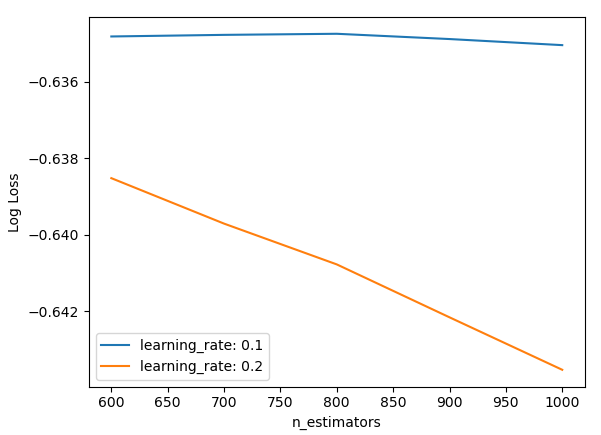


Figure 1.5 Average Arrival Delay by Distance Category

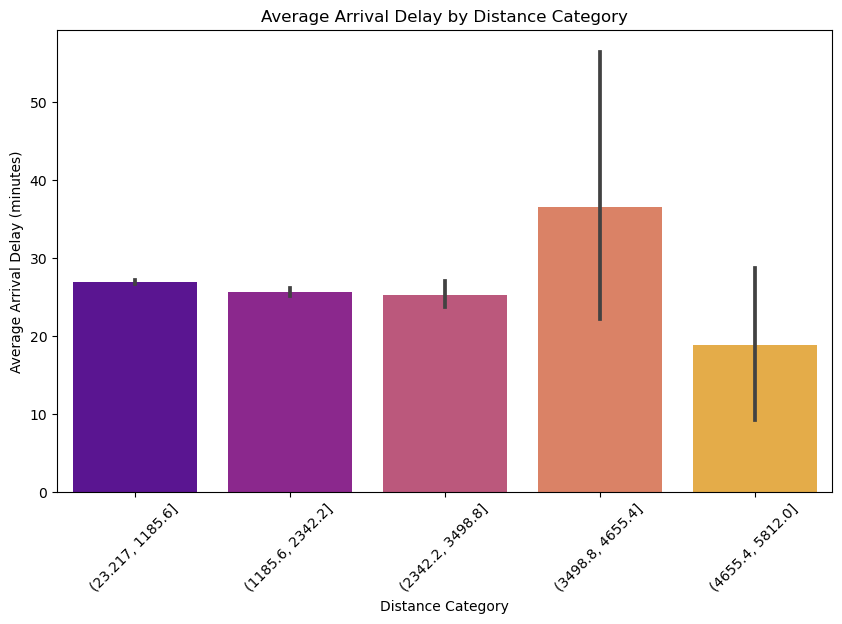


Figure 1.6 Right Skew of Departure Delay

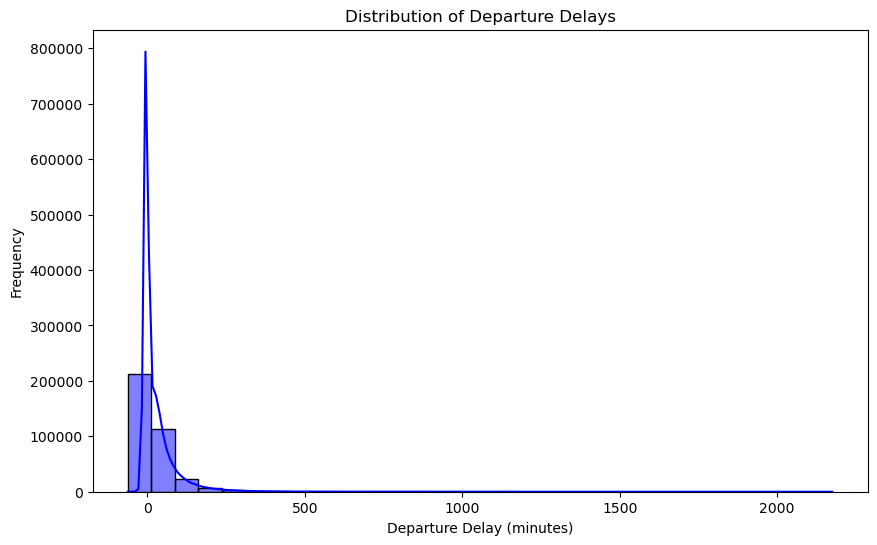


Figure 1.7 Model Performance Scores

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model Name** | **Feature Set** | **Dataset** | **AUC** | **F1** | **Accuracy** |
| Neural Net | 1 | Subset\_no\_weather | 0.6225 | 0.5704 | 0.5848 |
| Neural Net | 1 | Subset\_weather | 0.6401 | 0.5861 | 0.5980 |
| Neural Net | 1 | All\_balanced | 0.6849 | 0.4936 | 0.6656 |
| Neural Net | 1 & 2 | Subset\_no\_weather | 0.9578 | 0.8991 | 0.9026 |
| Neural Net | 1 & 2 | Subset\_weather | 0.9579 | 0.8987 | 0.9021 |
| Neural Net | 1 & 2 | All\_balanced | 0.9644 | 0.8926 | 0.9184 |
| Random Forest | 1 | Subset | 0.6889 | 0.6303 | 0.6368 |
| Random Forest | 1 | Subset | 0.6534 | 0.6022 | 0.6108 |
| Random Forest | 1 | All\_balanced | 0.7372 | 0.5831 | 0.6916 |
| Random Forest | 1 & 2 | Subset\_no\_weather | 0.9610 | 0.9029 | 0.9057 |
| Random Forest | 1 & 2 | Subset\_weather | 0.9633 | 0.9027 | 0.9057 |
| Random Forest | 1 & 2 | All\_balanced | 0.9693 | 0.8979 | 0.9208 |
| XGBoost Classifier | 1, 2 & 3 | Subset | 0.9683 | 0.9118 | 0.9135 |
| XGBoost Classifier | 1 | Subset | 0.6919 | 0.6398 | 0.6334 |
| XGBoost Classifier | 1 | Full Dataset | 0.7380 | 0.5471 | 0.6931 |
|  |  |  |  |  |  |

Figure 1.8 Model Accuracies ChartA diagram of a graph

Description automatically generated