Predicting Flight Delay at U.S Airports using Regression and Clustering

Neel Khakhar [1] Abhinav Padmanabha Mukthavaram [2] Srishti Kachhara [3]

Introduction

Our project looks at two different things about flights. First, we study data from many years to figure out which routes are popular and predict how many people might be on those flights. Then, we focus on the year 2015 and look at why some flights were delayed. Our models, including Logistic Regression and Decision Tree, predicted flight delay status and time with high accuracy. The goal is to give useful tips for making flights go smoothly, and managing delays better and grouping airports for regional flights.

Methods

The project commenced with the collection of two primary datasets, US Domestic Passenger/Departure data 1990-2009 and US Domestic flights delay data for 2015. We hypothesize that clustering airports (start-end airport route as one) according to passengers travelling, can help in prediction of flight delay. A pivotal step involved merging the KMeansCluster drawn from historical 20 years of data to 2015's delay dataset. We create our 'Status' target variable, based on how long the flight is delayed <15 min Status 0; >15 and <60 min Status 1; >60 min Status 2; Diverted route Status 3; Cancelled Status 4. One-hot encoding was applied to categorical variables laid the groundwork for subsequent machine learning tasks as a preparation of dataset. We train models for Decision Tree and Logistic Regression classifiers on 2 datasets, one containing historical derived passenger clusters and one without. We finally create 4 models using The models were trained, evaluated, and saved for future use, with assessments conducted using confusion matrix and accuracy to gauge their effectiveness in predicting flight status.

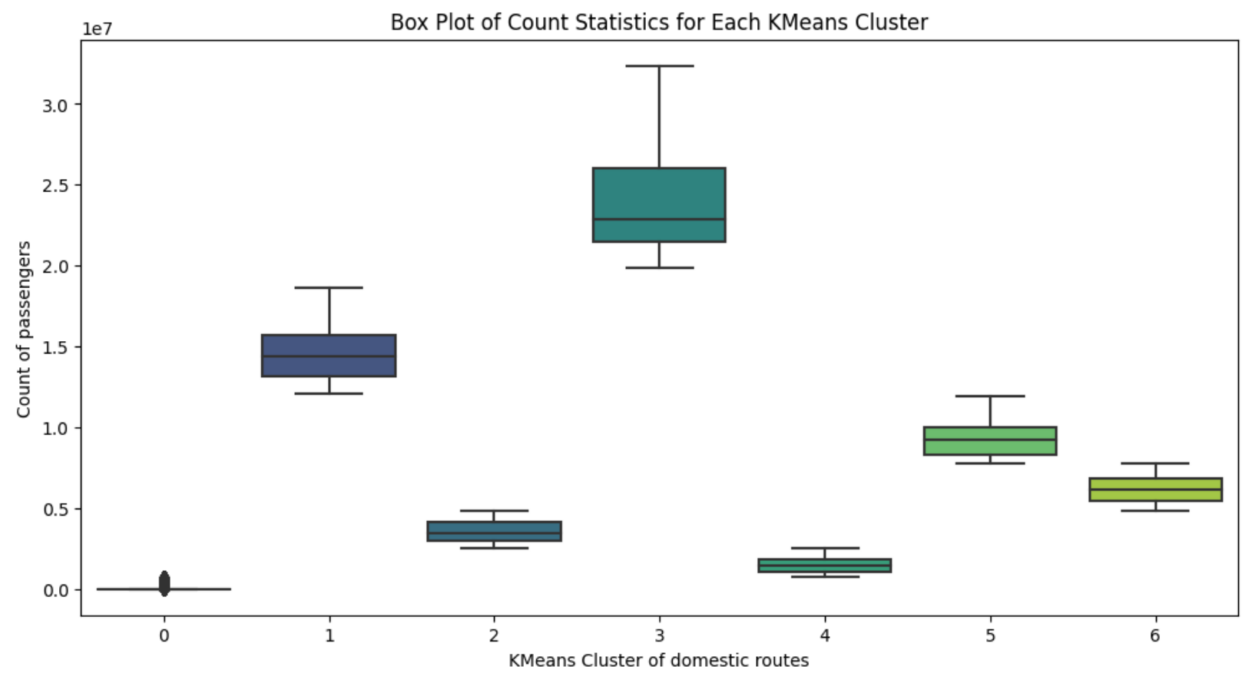

Figure 1. Count of Travelling Passengers for each KMeansCluster

Data Description and Preprocessing

Dataset Overview: The US Flight delay dataset includes over 5.8 million entries, detailing flight dates, airlines, flight numbers, origin and destination airports, scheduled and actual departure/arrival times, delay durations across various categories, cancellation details, wheels on/off time and reasons for delays. Preprocessing Steps: Data preprocessing involved several key steps: handling missing and null values, particularly in delay and cancellation columns; encoding categorical variables using one-hot encoding and label encoding methods; normalizing time-related features after converting HHMM format to normal datetime; We also customize wheels on and wheels off feature as a difference from Arrival and Departure times.

Models Used

- Model Selection: We utilized Logistic Regression for its efficiency in classification and Decision Trees for their interpretability and capability in handling non-linear relationships. These models were chosen while one uses random features to split the dataset, the other depends on statistical values of variances contributed by each feature to net split. We use Gini scoring and random splitting of features for Decision Tree with max depth set to 20.
- Model Training and Validation: The dataset was split into 80% training and 20% testing segments. We create 18 Models: 2 Architectures * (7 KMeans Cluster + 1 Using All data with Cluster + 1 Using All data without Cluster)

Mathematical Formulae Used

- Logistic Function (Sigmoid): $\sigma(z) = \frac{1}{1+e^{-z}}$

- Logistic Regression Hypothesis: $h_\theta(x) = \sigma(\theta^T x)$

- K-means Update Rules: $\mu_j = \frac{\sum_{i=1}^{m} x^{(i)} \nVdash \{c^{(i)}=j\}}{\sum_{i=1}^{m} \nVdash \{c^{(i)}=j\}}$

(Update cluster centers) $c^{(i)} = \arg\min_j ||x^{(i)} - \mu_j||^2$

$- Gini(D)$: Gini(D) $= 1 - \sum_{i=1}^{c}(p_i)^2$

- Accuracy: Accuracy $= \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$

- Precision: Precision $= \frac{\text{TP}}{\text{TP+FP}}$

Data Analysis and Visualisations

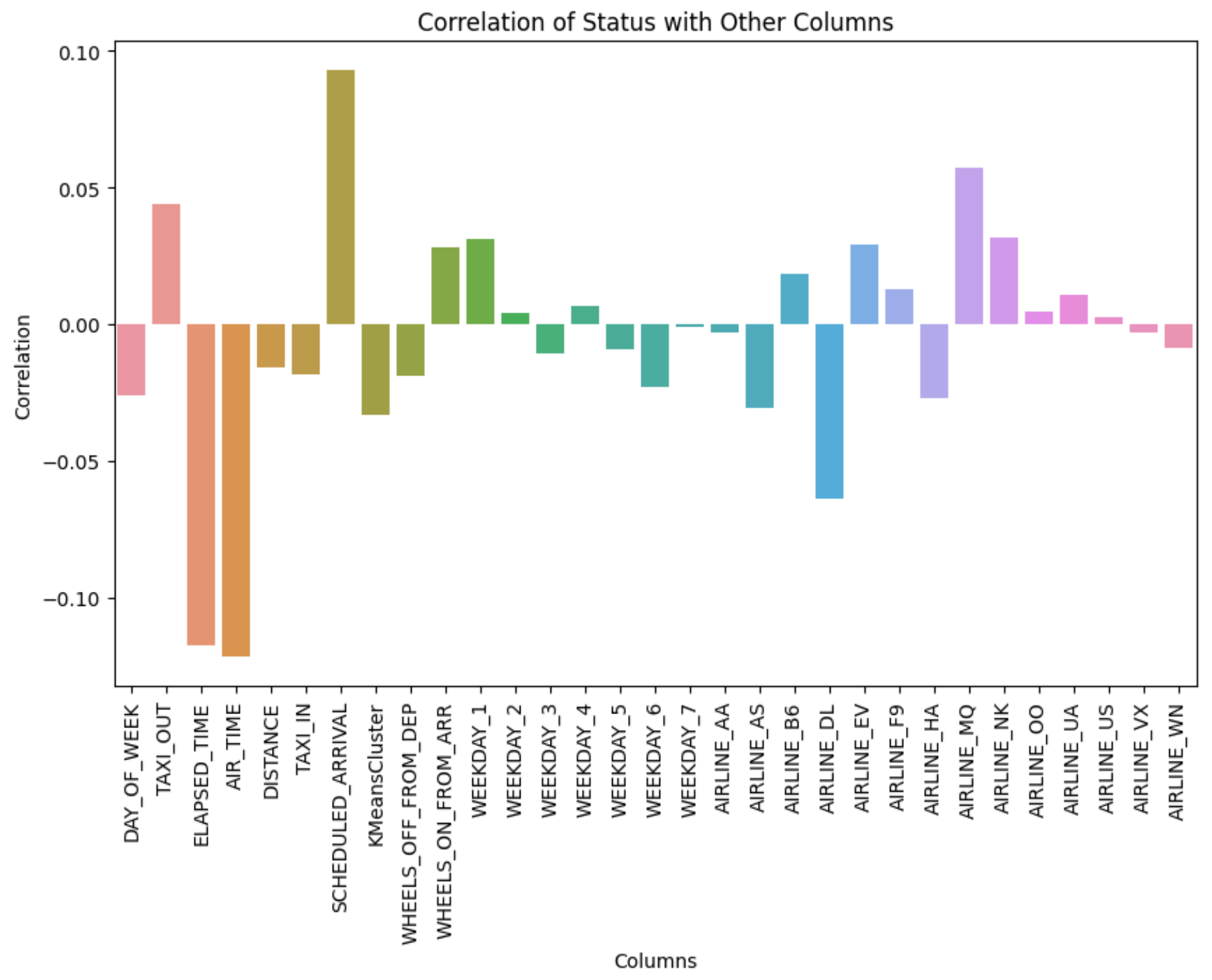

Figure 2. Correlation with Status - We can notice significant correlation with cluster

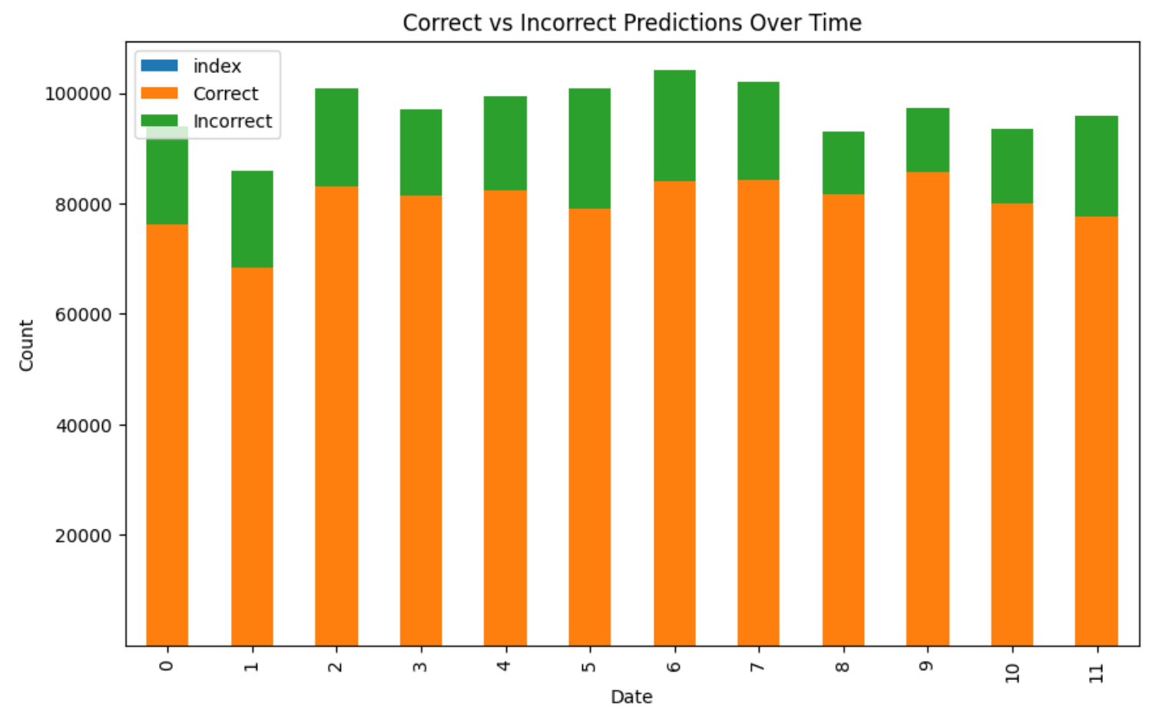

Figure 3. Monthly trend in prediction by Best Model

- Model Performance: The Logistic Regression model demonstrated an accuracy of 81.54%, an AUC of 81.78%, reflecting moderate predictive power. The Decision Tree model exhibited slightly higher accuracy at 82.72% and slightly varied in other metrics, indicating differing strengths and weaknesses in model performance. We also observe increase in accuracy with our new KMeans Cluster as a feature.

- Visualization Highlights: We see significant improvement in performance when clustering is added as a feature. It would be worthwhile to explore creating an ensemble model for each cluster delay prediction to evaluate the impact. The results were evaluated with confusion matrices for each model, ROC curves depicting model sensitivity and specificity, and monthly trends in correct and incorrect predictions.

Results and Discussions

Model	Accuracy
DT With Cluster	86.29
DT Without Cluster	82.72
LR With Cluster	82.75
LR Without Cluster	81.54

- We study the confusion matrix and realize that for the Decision Tree model shows a high number of correct predictions for all classes except Status 1 which suggests effectiveness in these areas. Similarly, Logistic Regression performs better in classifying Status 1. Due to bias of data and unstratified sampling between classes, we see that Decision Tree performs better inspite of biased data in prediction of a multi-class problem, while Logistic Regression struggles handling the bias in data. The macro and weighted averages show less discrepancy than the Logistic Regression model, yet there's a clear indication of class imbalance affecting performance metrics. This model also shows better predictive strength for more prevalent classes and may benefit from strategies to improve its handling of less frequent classes.
- The bar chart illustrates a model's monthly prediction accuracy, showcasing a consistently favorable trend of more accurate predictions than errors throughout the year. This implies that the model generally outperforms random chance. Minor month-to-month variations in correct and incorrect prediction counts suggest potential seasonal effects or time-dependent factors influencing model performance. However, an extraneous 'index' category in blue, likely a plotting artifact, should be removed for enhanced clarity. The x-axis, labeled as 'Date' with numbers 0 to 11, could benefit from using actual month names for improved informativeness. Fluctuations in monthly prediction counts may indicate changes in data volume, posing an important consideration for assessing model consistency and robustness.

Conclusions

In our analysis of 2015 flight data, the Decision Tree model outperformed the Logistic Regression model with an overall accuracy of 86.29% with cluster and 82.72% without cluster, compared to 82.75% for LR with cluster and 81.54% without. Our inclusion of KMeansCluster as a feature improves the performance of both models The Decision Tree model showed exceptional predictive accuracy for certain classes and variability across months, indicating a sensitivity to external factors such as seasonal trends. While both models demonstrated robustness in predicting flight delays, the variability in precision and recall across classes suggests opportunities for improvement, potentially through advanced feature engineering, algorithmic enhancements, and addressing class imbalance to achieve more equitable performance across all classes.

References

[1] Dhendra Marutho, Sunarna Handaka, Ekaprana Wijaya, and Muljono. “The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News”. In: Sept. 2018, pp. 533–538. doi: 10.1109/ISEMANTIC.2018.8549751.
[2] ICEBI '21: Proceedings of the 2021 5th International Conference on E-Business and Internet. Singapore, Singapore: Association for Computing Machinery, 2021. isbn: 9781450385657.
[3] Kerim Kili ̧c and Jose M. Sallan. “Study of Delay Prediction in the US Airport Network”. In: Aerospace 10.4 (2023). issn: 2226-4310. doi: 10.3390/aerospace10040342. url: https://www.mdpi.com/2226-4310/10/4/342.
[4] Ntani, G., Inskip, H., Osmond, C. et al. Consequences of ignoring clustering in linear regression. BMC Med Res Methodol 21, 139 (2021). https://doi.org/10.1186/s12874-021-01333-7