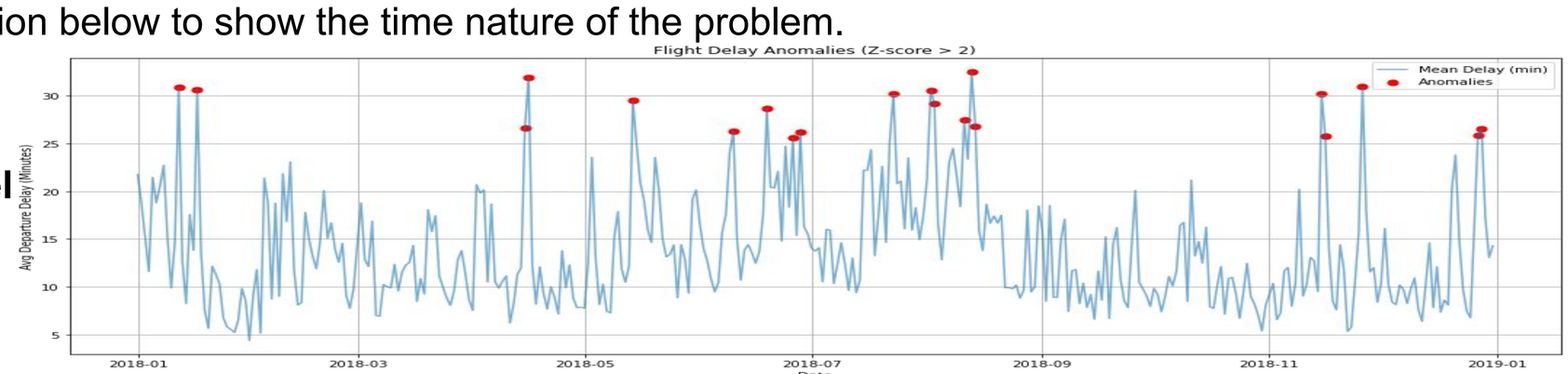
Flight Delay Prediction

Team 068: Katha Korgaonkar, Rebecca Barth, Matthew Rozal, Soma Yamaoka, Joel Westmark

Motivation/Introduction

We aim to empower passengers with better estimates for their itineraries, giving them more autonomy in their travel decisions. Just in one year alone, there were over a million hours lost due to delays. Furthermore, a single minute could save an airline millions of dollars in crew and fuel costs. Through different modeling techniques and interactive visuals, we allow passengers to make more informed decisions about their itineraries and also educate the industry about what is the greatest contributor to delays. See the anomaly detection below to show the time nature of the problem.



Approaches

Random Forest/Markov Chain Hybrid Model

- Random Forest (RF) – Predicts the initial arrival delay for each flight, and also provides a base departure delay prediction
- Delay Propagation Model – Uses the RF's previous flight arrival predictions as inputs and estimates how much of that delay carries over to the current flight's departure(via markov-based flight chaining).
- Final prediction is a linear combination of RF and Markov departure delay predictions
- Bagging these models gives us a prediction that gets the RF model's higher accuracy while still capturing delay propagation dynamics between flights.
  - Innovation is bagging these two approaches, as we often see them as separate models.

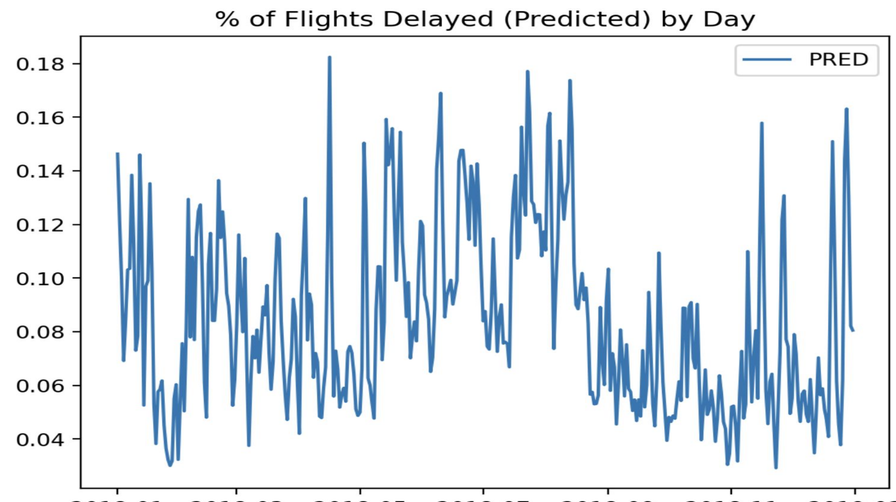
Boosted Classification Model *Classifies flights with > 20 minutes delay*

- Optimized using train/test splitting and gridsearch parameters and feature tuning
- Boosted models handle nonlinear data better than other classifiers for this problem.
- Innovation is adding flight density and airport popularity as core continuous features along with delay propagation. We also use SHAP values to find feature importance.

Interactive Tableau Viz

- Allows passengers to explore their flight and airline options for specific days/routes by integrating the output from our hybrid random forest/markov chain model
- In comparison to other attempts, ours is much more interactive and actually focuses on delays to help find the best choice for an on-time arrival
- Users can explore airport statistics such as average delays and most delayed airlines and see their top 3 most reliable routes

Predictions from Classification Model



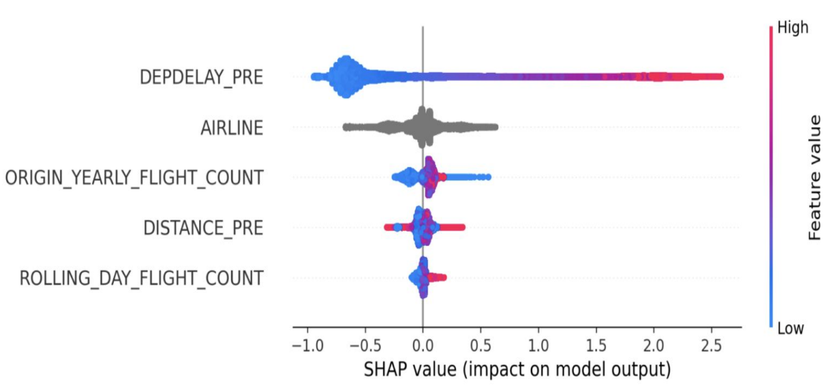
Data

We downloaded the Flight Status Prediction dataset from kaggle (~ 10 GB). Since this was an extremely large dataset, we subsetted the data to just the year 2018 (2 GB) for visuals. There are over 5.6 million flights with information about their date, flight time, delay time, scheduled arrival and departure airline, arrival and departure airport and more.

Experiments and Results

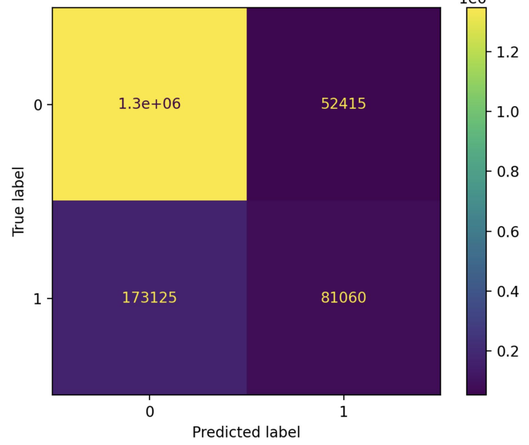
For our boosted classification model, we find that the departure delay for the previous flight was by far the most important indicator, with the specific airlines following it up. We used a self-joined time series with window-function deduplication to assemble our training set. This shows delay propagation (preceding departure delay) is a key factor, and about 5x more important than the next feature we considered. Accuracy and f1 of 0.86 and 0.42, respectively. Results are not strong historically, though our implementation is not best-in-class. Our random forest and markov model results show that the RF model predicts departure delays more accurately than the Markov model. However, the Markov model captures delay propagation that is not captured by the RF model. As a result, we believe the hybrid approach has the most potential.

Classification Feature Importance

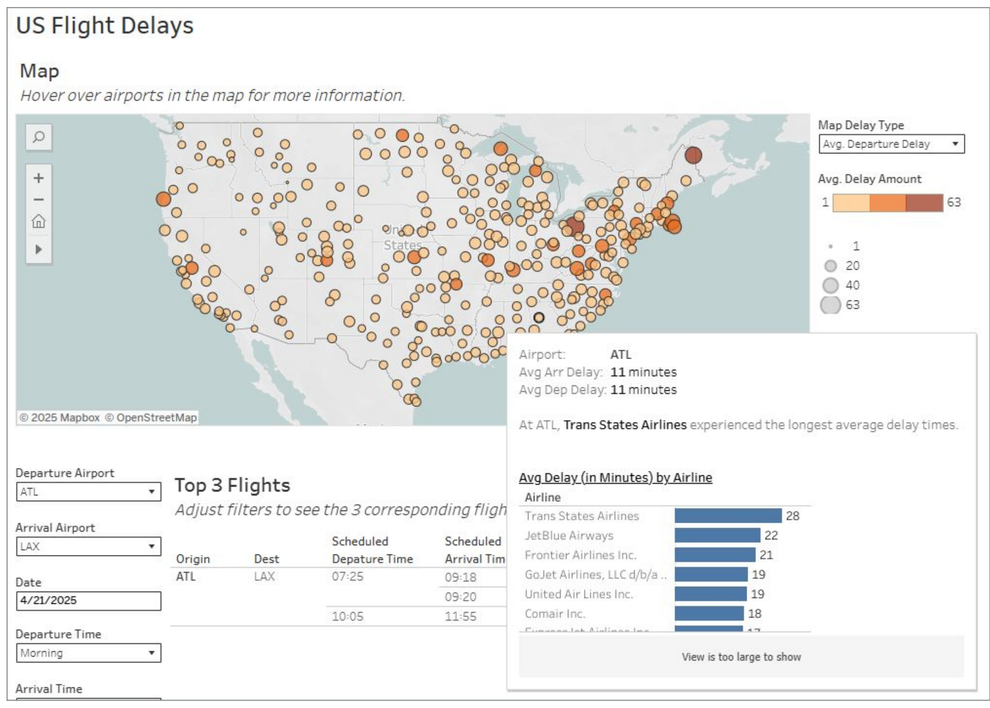


Classification Model Eval Results
Hybrid RF/MC Model Eval Results

RF RMSE	RF MAE	Markov RMSE	Markov MAE	Hybrid RMSE	Hybrid MAE
25.2158	6.7414	46.9454	18.6562	31.0774	10.2227



Interactive Viz with filters



The [interactive visualization] can be seen on the right. We have a map of all US airports with an interactive tooltip showing airport delay statistics **from our Hybrid RF/MC model**. We also have a flight finder that takes in flight inputs to show passengers the top 3 least delayed flights to choose from.

We evaluated our visual by asking people to test it out. We asked them for any suggestions and implemented them to improve their experience. As a result we included a barplot in the tooltip to show top most delayed flights out of airpots. We also included more dropdown options for passengers to choose from.