Flight Delay Prediction and Visualization for US Travelers

Motivation

Currently, consumer rely on certain internet based real time flight tracking services through certain government websites such as FAA's National Airspace System to provide the statistical data on delay, airport events and real time monitoring of commercial flights. Considerable non-commercial work has been done on the back end to build ML models for flight delay prediction. However, none of these models are accessible to general public as they lack an interactive and easy to use interface.

Commercial

- ✓ Simple Flight Stats
- ✗ ML based prediction
- ✓ Consumers have access

Private websites, National Airspace System Status

Academia

- ✓ Simple Flight Stats
- ✓ ML based prediction
- ✗ Consumers have access

Various ML models built on historical flight data

Impact

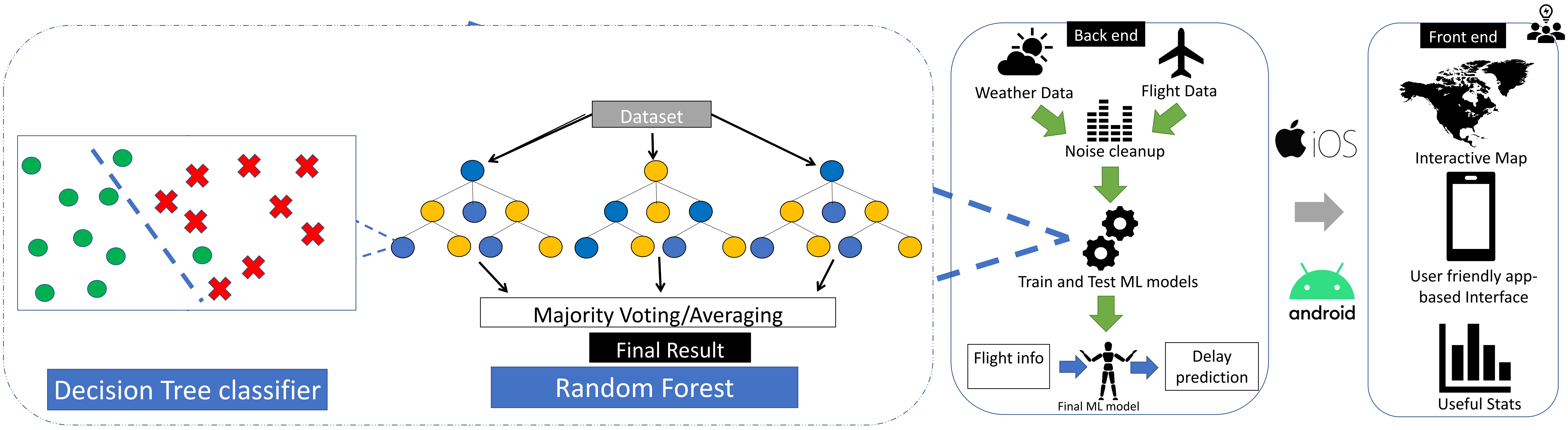
Our objective is to provide highly transparent delay prediction information to the air travelers in the US. The application primarily targets air travelers in the US with airlines being the secondary stakeholder as they can directly implement the app and enable passengers to view expected delays while booking their flights. For travelers, the visualization tool can help make more informed flight booking decisions by providing information on flight delays and cancellations, allowing them to avoid flights that are prone to delays, thereby saving time and improving their overall quality of life. Additionally, airlines can benefit from these tools as well, as passengers having a better flying experience commonly leads to improved customer ratings and reviews.

Impact

- Travelers avoid delays; time savings
- More productivity for business travelers
- Avoid hotel/rental car cancellations; cost savings
- Increased passenger satisfaction with airlines

Solution approach

Our high-level solution approach is to use flight and weather datasets to train a ML model to predict flight delay. This model is made accessible to a consumer through an iOS based interactive front end. We used python for tasks pertaining to feature selection, feature analysis and building models. Flight and weather-related datasets were used for data exploration and to identify essential features. The data was then split into training and test sets to fit and evaluate multiple models such as decision trees and linear regression. Evaluation metrics such as accuracy, recall, f1 score, and precision were used to evaluate the performance of the models. Our final modeling approach was to divide flight delays into multiple categories and apply Random Forest Classifier to a multicategory classification, enabling us to predict delay duration. The most innovative aspect of our approach is the creation of an iOS based front end for the model to present delay prediction information to travelers in a transparent and comprehensible manner. The GUI enables travelers to input their flight information and receive output information regarding their flight. This provides a level of accessibility to the consumer not available before.



Data

The flight delay and cancellation data was collected and published by the Department of transportation's Bureau of Transportation Statistics and downloaded from Kaggle. The flight data contains all or most domestic flights in the US for the year 2015. The size on disk for the flight's dataset is 76.3 MB. The weather data was extracted through an API from the National Weather Service website and it contains weather data for every hour for US airports throughout 2015 with total size of 0.47 GB. Datasets for both flights and weather are temporal datasets.

Flight Dataset Statistics

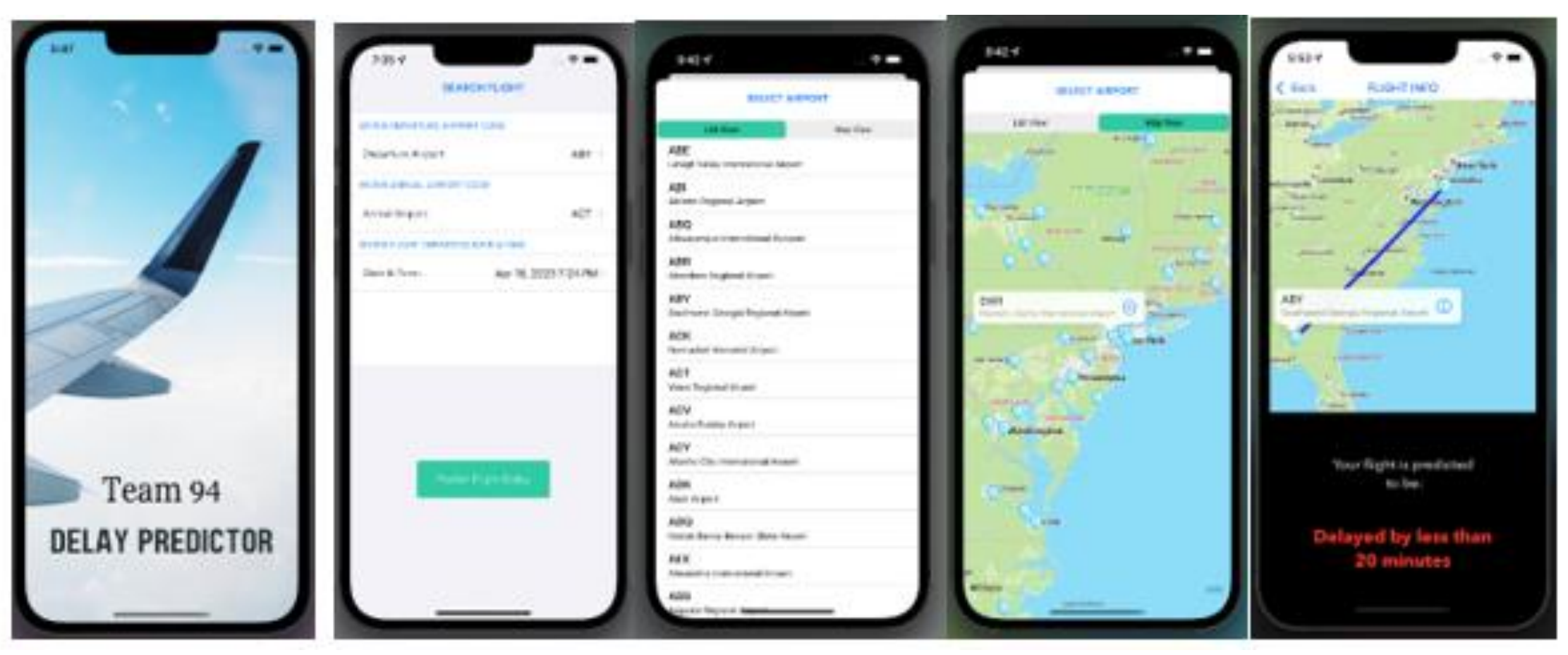
Number of Variables	31
Number of Rows	165843
Missing Cells	783826
Missing Cells (%)	15.20%
Duplicate Rows	0
Total Size in Memory	76.3 MB
Average Row Size in Memory	482.6 B
Variable Types	Categorical: 10 Numerical: 21

Experiments and Results

In evaluating our approaches, we used a combination of statistical metrics such as accuracy, precision, recall and F1-score to assess model performance. We also used cross-validation to ensure the robustness of our models and avoid overfitting. The results of our modeling efforts showed that the Random Forest Classifier was the most effective approach, achieving >70% accuracy in predicting flight delays. We were also able to predict the duration of delays with 60-70% accuracy. Comparing our models to others, we found that our approach outperformed linear and logistic regression in predicting flight delays. Although our modeling approach successfully addressed the class imbalance issue in the data and enabled us to accurately predict delays, we acknowledge that there may be other models or techniques that we did not explore that may yield better results.

For front end visualization, we looked at several options including d3 and python-based toolkits (PyQT, wxPython etc) but finally decided to go with a core ML based iOS app front interface. Our primary evaluation criteria was the ease of use and level of accessibility. Given the widespread use of smart phones among masses, this approach gives us the best level of accessibility to an average traveler compared to the other options. The result is an interactive, user friendly, and easy to use app that can be accessed and used by anyone.

Final product



An interactive flight delay predictor app backed up by Core ML models