Flight Delay Predictions

By: Fucheng Yao, Limei Huang, Eman Nagib, Kwangwoo Kim, Huaiping Wang

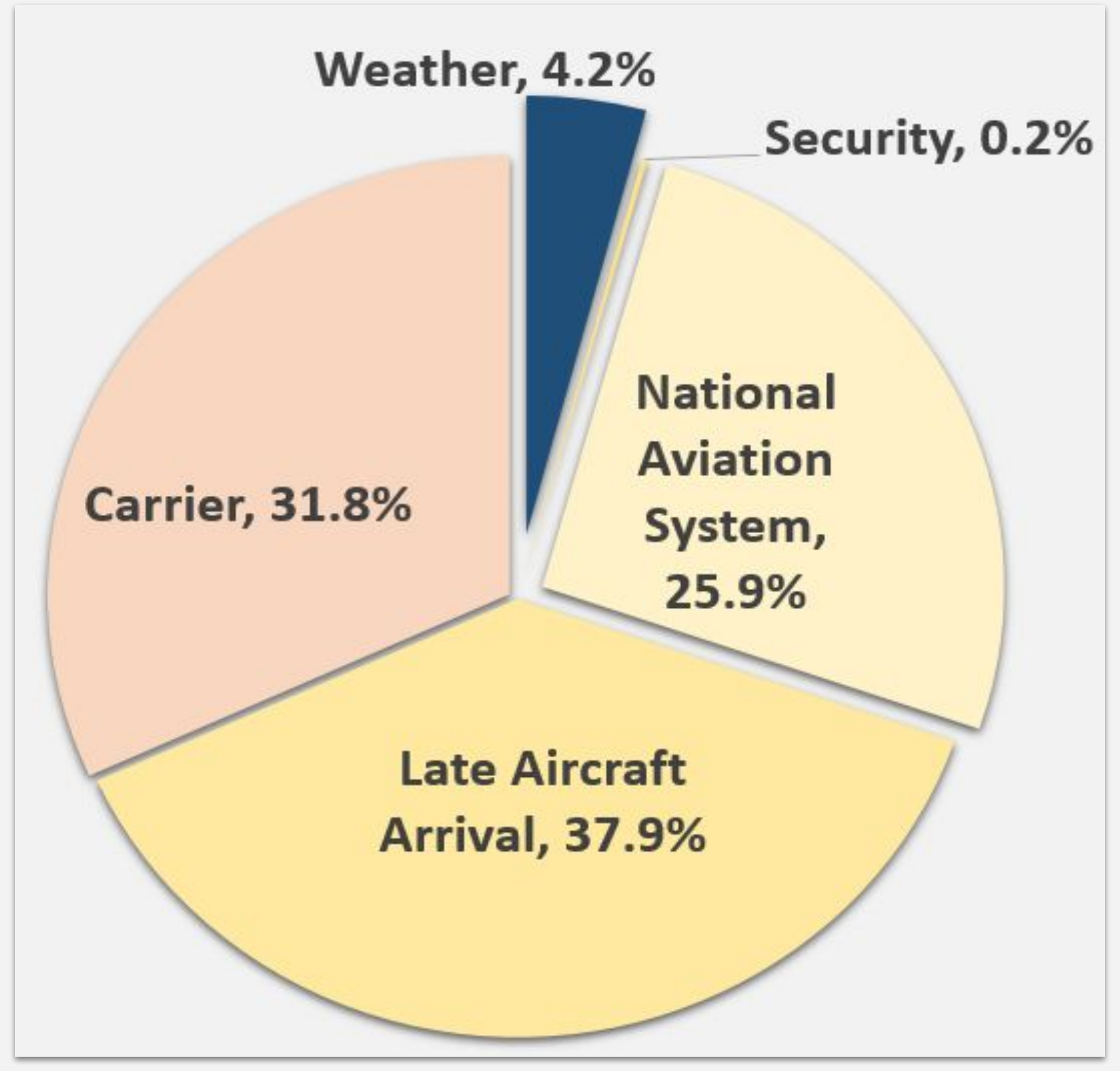
Introduction

With the top U.S. airlines current high congestion rates and ever-increasing passenger traffic, the frustrations that come with unforeseeable delays are undeniable. Flight delays stem from several issues including carrier complications, late aircraft arrival, unfavourable weather, security threats, National Aviation System congestion, and other issues that might arise. Since these issues are not necessarily avoidable, the potential for predicting delays with respect to a specific airline, route, and date remains invaluable.

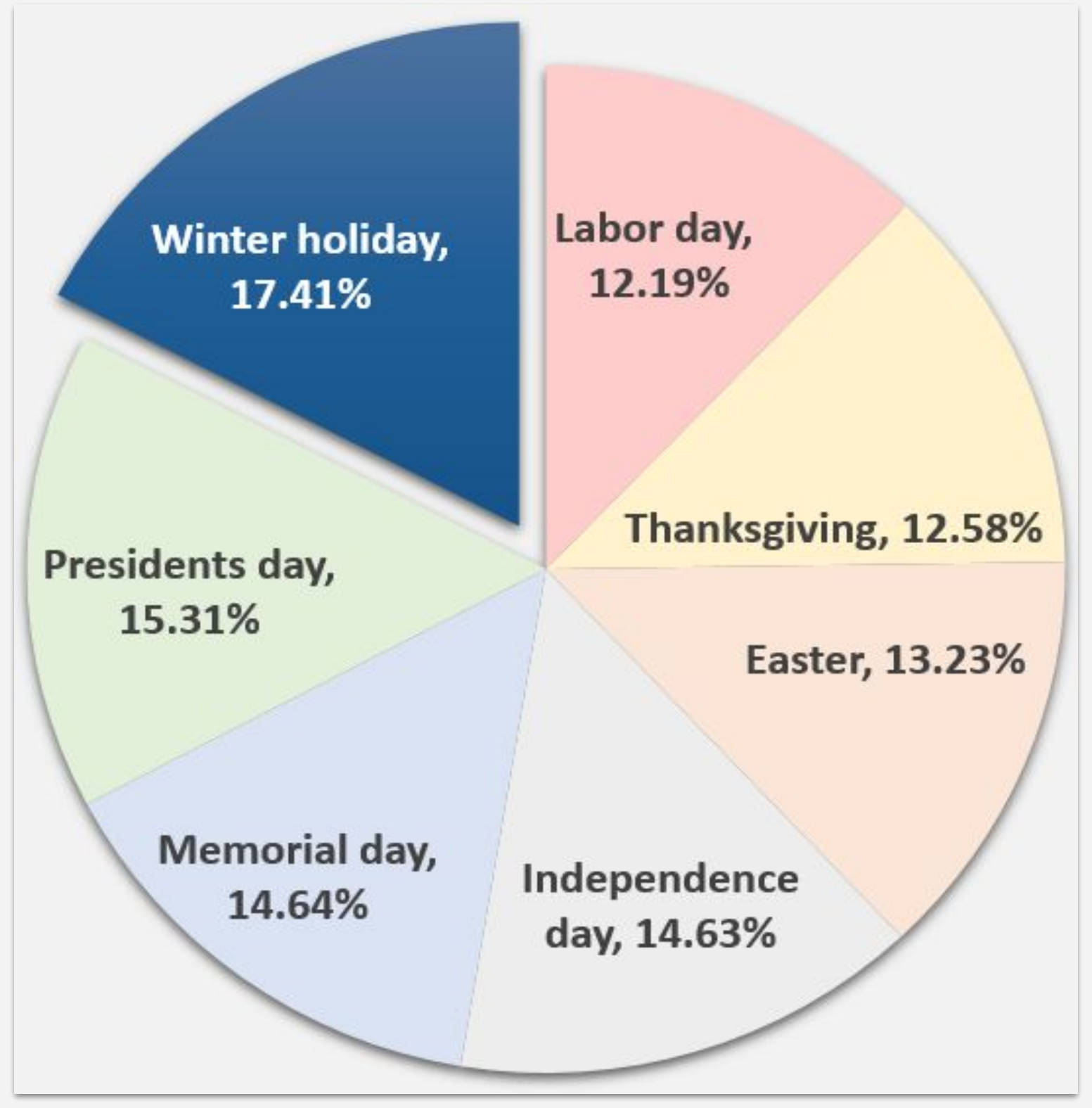
A few websites exist, such as Google Flights, that have recently started to provide some insight concerning potential flight delays. Google Flights displays these insights during the ticket purchase process, by showing 30-minute delay estimations as a precaution for their customers. The following project aims to take delay predictions a step further and leverage insight from historical data and machine learning algorithms to allow customers to see a refined delay prediction in respect to their future flight.

Descriptive Analysis

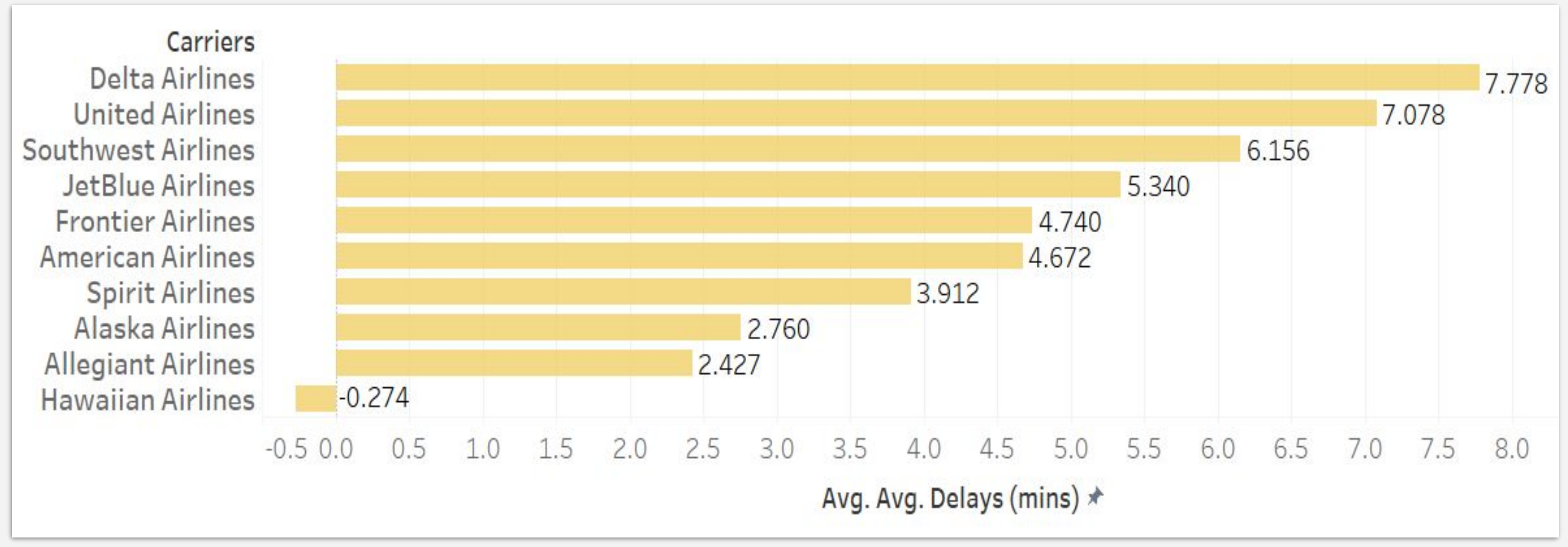
Major Flight Delay Reasons:



Holiday Delay Statistics:



Airlines' Compensation of Departure Delays:



- Airlines can make up delay time by reducing airborne time
- Highest departure delays: Allegiant, Jet Blue, and Frontier Airlines.
- Worst airlines at compensating for delays: Delta Airlines, Allegiant, Alaska

Project Objective:
"Using historical data, identify which model best predicts future flight delays and in turn identify top indicators of departure and arrival delays."

Methodology

Using the Root Mean Square Error (RMSE) as the measure of accuracy, the team evaluated model performance on predicting future flight delays. The methodology included the following steps:

1. The team ran several flexible models including Random Forest, Boosting, and XGBoost.
2. Cross validation was used to adjust parameters that would increase prediction accuracy. These models are slower to run but can inform the "performance ceiling" for the dataset.
3. Next, the team investigated a suite of simple, or interpretative, models: Linear Regression, Lasso, and Forward Selection. Models such as Lasso and Linear Regression are helpful in identifying the most influential variables.
4. Finally, the team identified the most influential variables by examining the most interpretable models and selected the best predictive model based on the RMSE.

Results

Model	RMSE Test (departure delay)	RMSE Test (arrival delay)
Linear Regression	26.27	14.60
Lasso Regression	25.64	14.45
Forward Selection	25.73	14.63
Boosting Trees	25.96	15.42
Random Forest	26.12	13.78
XGBoost	25.38	13.86

Most influential variables to predict departure and arrival delays

- **Weather:** (-) Visibility, (+) Snow_ice_pallette, (+) Thunder, (+) Wind_speed
- **Months:** (+) May, (+) June, (+) July, (+) August
- **Day of week:** (+) Monday, (+) Thursday, (+) Friday
- **Origin/Destination:** (+) JFK, (+) SFO, (+) EWR, (+) LGA
- **Airlines:** (+) Frontier, (+) Jet Blue, (+) South West
- **Holidays:** (+) Winter Holiday, (+) Thanksgiving

Discussion and Conclusion

Ultimately, the most flexible model XGBoost provided the most accurate departure and arrival delays. However, to get a better understanding of the variables that are the best predictors for delays we need to examine a more interpretable model albeit less flexible and less accurate. By examining models such as Random Forest and Regression we can see that weather attributes followed by specific months and days as well as airlines can be the best predictors of delay.

Predicting flight delays is no easy feat. Although historical data is able to shed some insight related to the trends and seasonality in delays it definitely doesn't provide a suitable ground to build upon accurate predictions for future delays on its own. While utilizing real-time data is sure to improve the reliability and accuracy of the delay predictions including additional variables to expand the breadth of the data is just as crucial.

Limitations and Further Research

The ability to predict the impending delays by studying historical data is a huge advantage. Vast amount of historical data can help spot trends and patterns that are likely to influence future course of actions. However, the danger comes when this predictive analysis is mistaken for prescriptive analysis especially in contexts were historical data are not necessarily accurate predictors of an ever-changing world. This current model is essentially a probabilistic model, so it will not detect delays with 100% accuracy.

Consequently, working on enhancing the accuracy of predicted delays based on additional factors, including social abnormalities, such as COVID-19, and real-time data can prove invaluable for further research. This research will address unnecessary airport congestion and as a result, enhance the customer experience.

Bibliography

- https://console.cloud.google.com/marketplace/details/noaa-public/gsod?project=team5-888
- https://stat-or.unc.edu/files/2018/09/Paper3_MSOM_2012_AirlineFlightDelays.pdf
- https://www.bts.gov/topics/airlines-and-airports-0
- Data sets source: https://www.bts.gov/topics/airlines-and-airports-0
- Weather data source: https://www.ncdc.noaa.gov/cdo-web/datatools/records;
- https://console.cloud.google.com/marketplace/details/noaa-public/gsod?project=team5-888