

Stats 101C Final Project

Prediction of U.S. Domestic Flights' Cancellation Status

Lecture 2 Group I:

Miko Farin (mikofarin12@gmail.com)

Jack Krupinski (krupinskij2@gmail.com)

Wonjae Lee (edd6dd@gmail.com)

Matt Turk (mturk96@gmail.com)

Overview

1. **Background**
2. **Cleaning and Preparing the Data**
3. **Modeling the Response**
4. **Conclusion**



The background is a solid pink color. In the top right corner, there is a geometric pattern consisting of several squares and triangles in different shades of pink, creating a stepped or architectural effect.

Background

Goal

The Data:

- The U.S. Department of Transportation collected data about delays and cancellations on 69,225 flights from major airline carriers
- The data has 47 features per observation, such as origin and destination airports, flight duration, date, airline, number of passengers, etc.

Objective: Utilize the U.S. Department of Transportation data on 69,225 flights to develop a model that best predicts the cancellation status of a given flight



Cleaning and Preparing the Data

Data Exploration

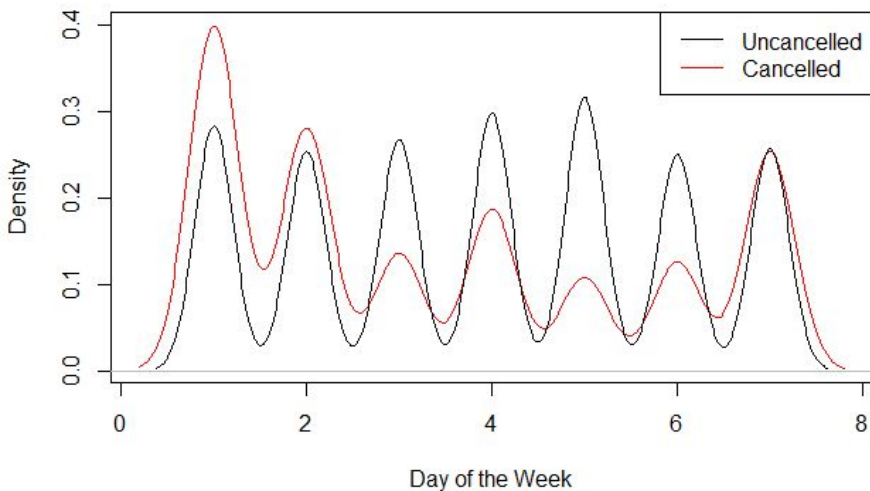
- GGpairs illustrates correlations and relationships between predictors
- We also visualize the distributions of our predictors



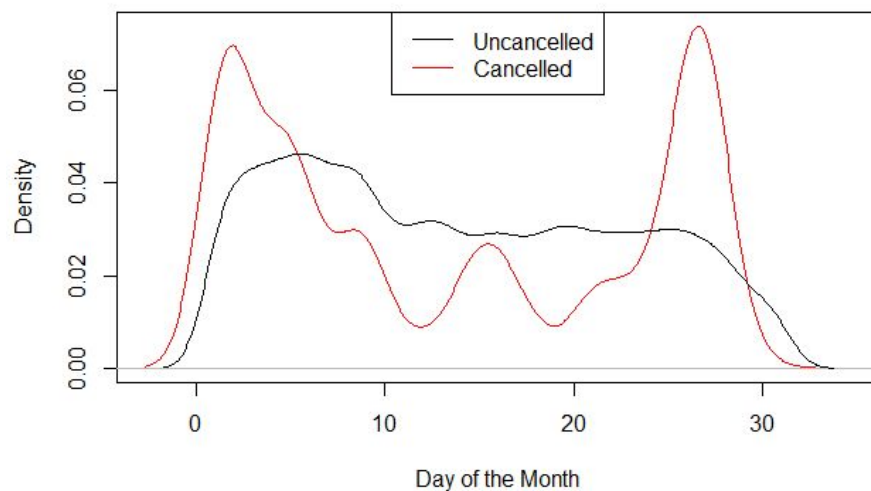
Data Exploration

- Density plots give us a rough idea about how well a predictor can separate cancelled flights from uncanceled flights

**Density of Day of the Week:
Cancelled and Uncanceled Flights**



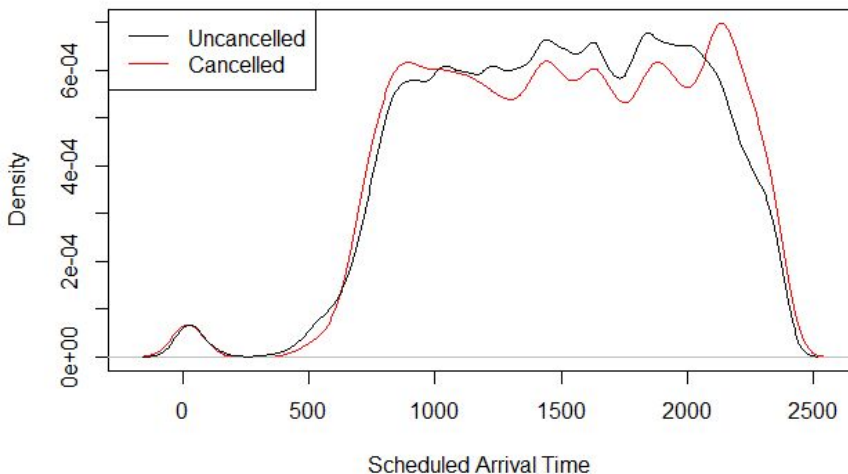
**Density of Flight Day of the Month:
Cancelled and Uncanceled Flights**



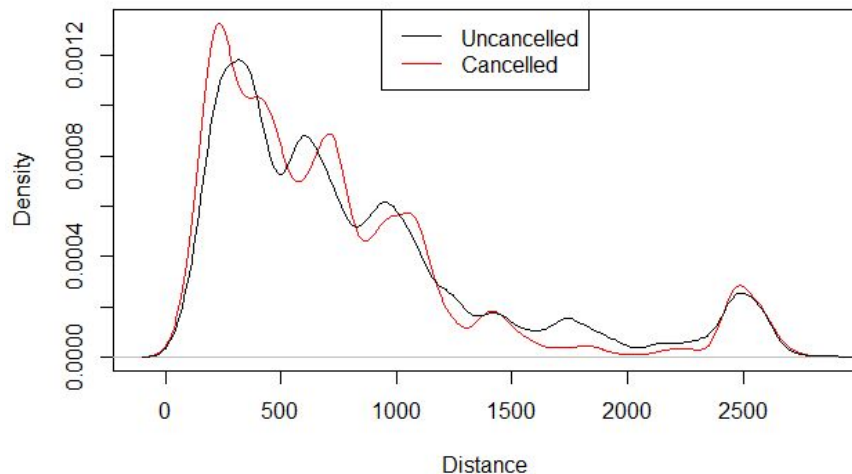
Data Exploration

- Many of the predictors were ineffective in separating the cancelled flights from uncanceled flights

**Density of Scheduled Arrival Time:
Cancelled and Uncanceled Flights**



**Density of Distance:
Cancelled and Uncanceled Flights**



Missing Values

Variable	Missing Value Percentage
share_white	97.46%
share_black	97.46%
share_native_american	97.46%
share_asian	97.46%
share_hispanic	97.46%
Median Income	97.46%
poverty_rate	97.46%
percent_completed_hs	97.46%

-Variables with greater than 80% of their values missing were deleted

-The remaining missing values after deletion were imputed with the “mice” package



Missing Values

Variable	Missing Value Percentage
AIR_SYSTEM_DELAY	85.02%
SECURITY_DELAY	85.02%
AIRLINE_DELAY	85.02%
LATE_AIRCRAFT_DELAY	85.02%
WEATHER_DELAY	85.02%
TAIL_NUMBER	9.40%
Aircraft.Movement	2.06%
Pass.Traffic	0.87%

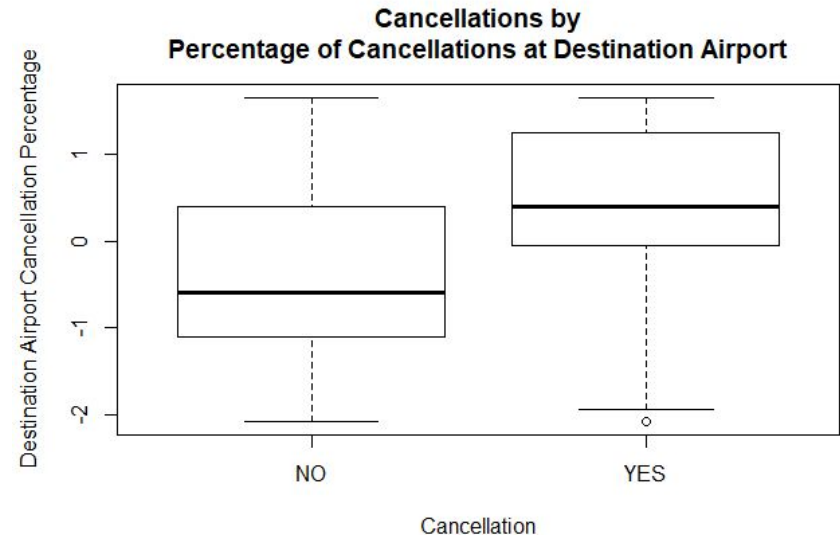
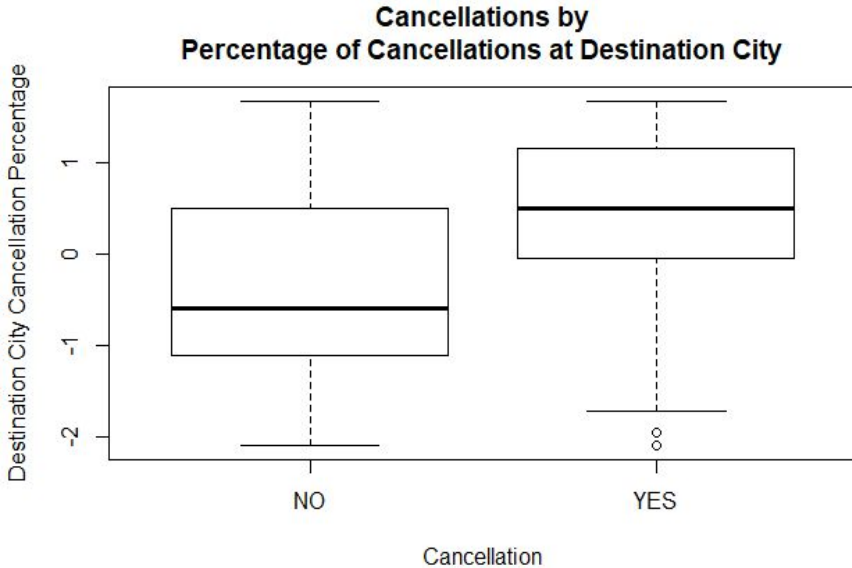
- Note that most predictors with many missing values relate to flight delays or passenger demographics

-The 31 features not listed had no missing values



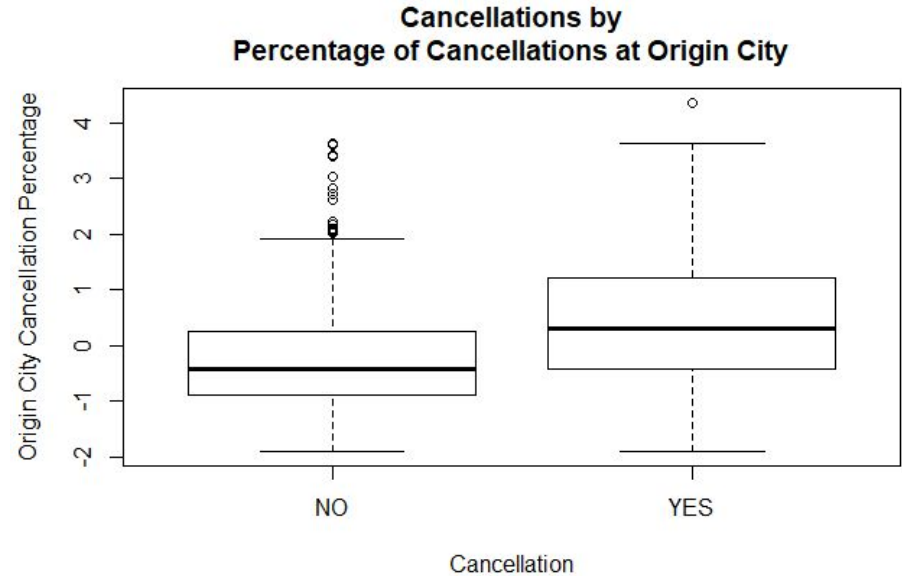
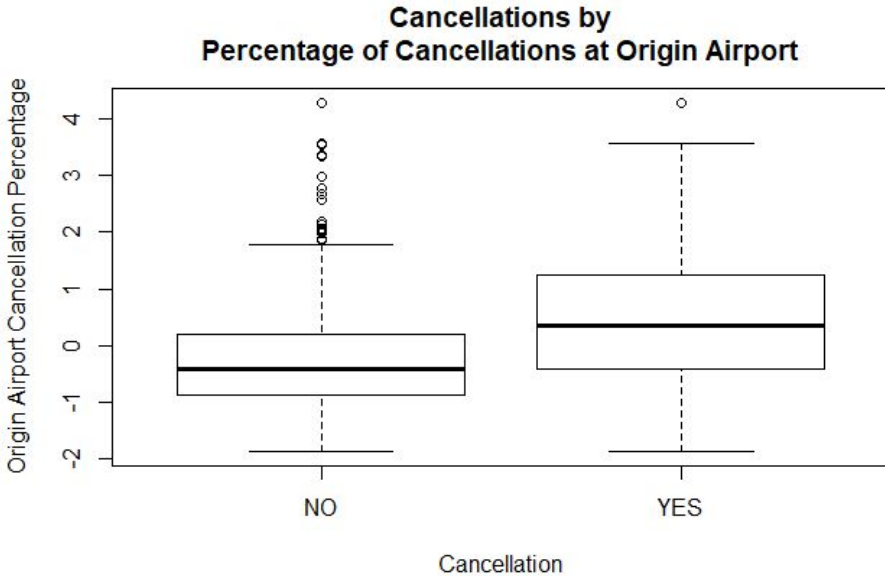
Variable Creation

-We created new variables that expressed cancellations as a percentage of total flights by destination airport, origin airport, airline, destination city, and origin city



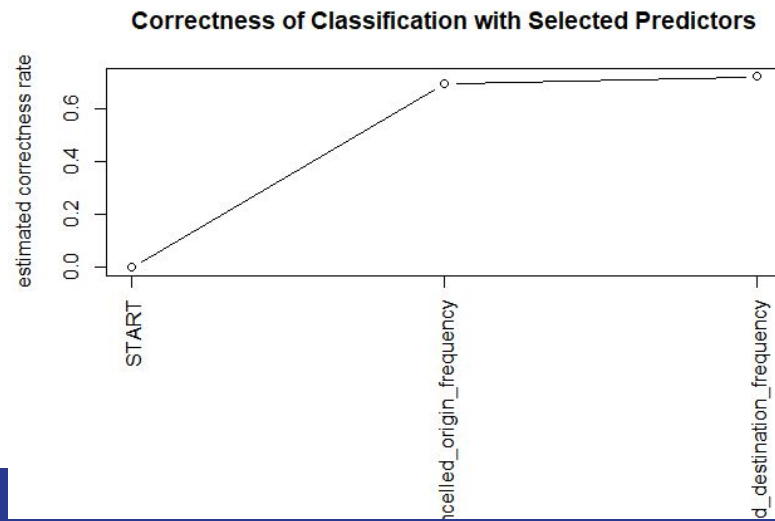
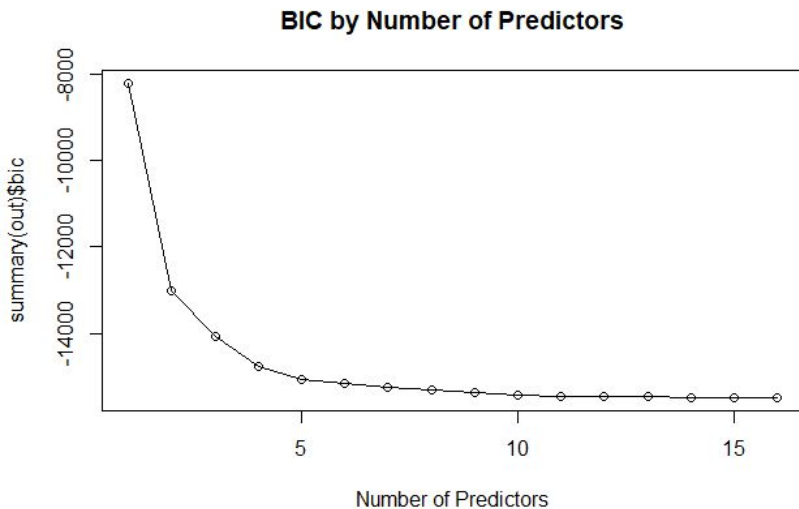
Variable Creation

-As expected, more cancelled flights occur at cities/airports with a higher cancellation **rate**



Determining the Best Predictors

- First, we used the “stepclass” function from the “klaR” package to find the best predictors
- We then used stepwise feature selection with BIC as our selection criterion



Scaling and Splitting Data

Scaling

-Although scaled variables are not necessary for every model we created, we did scale the numerical variables to make the data more “versatile”

Splitting

-For the purposes of testing potential models, we split the given training data into a testing subset (20%) and a training subset (80%)



Modeling the Response

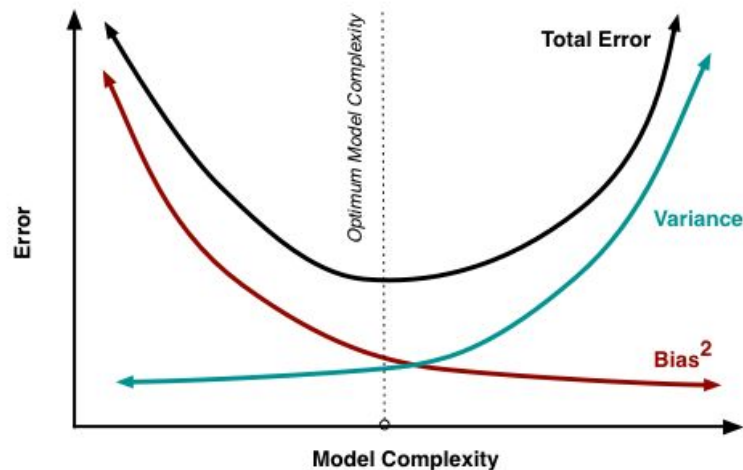
Choosing an Appropriate Model

We needed a model that would:

- Be effective with a large data set (many predictors)
- Yield a high accuracy on both the training and testing sets
- Minimize bias and variance

Candidate Models: LDA, Random Forest, Boosting,

Neural Net



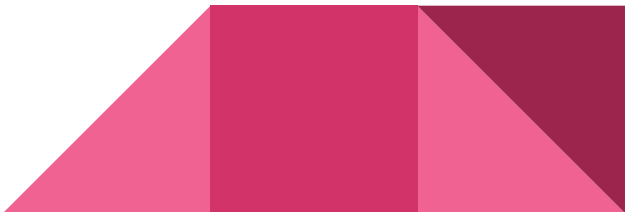
LDA

-The LDA model served as a relatively simple baseline; given its poor performance, we looked to build a model that could accommodate a more complex decision boundary

Confusion Matrix	NO	YES
NO	7923	2384
YES	1290	2248

Misclassification Rate = 26.54%

Accuracy = 73.46%



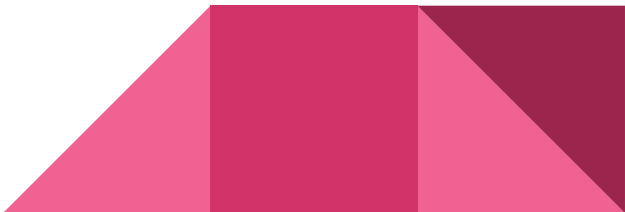
Boosting

- Boosting is a **slow learning method** that is designed to reduce bias and variance

Confusion Matrix	NO	YES
NO	8597	1047
YES	616	3585

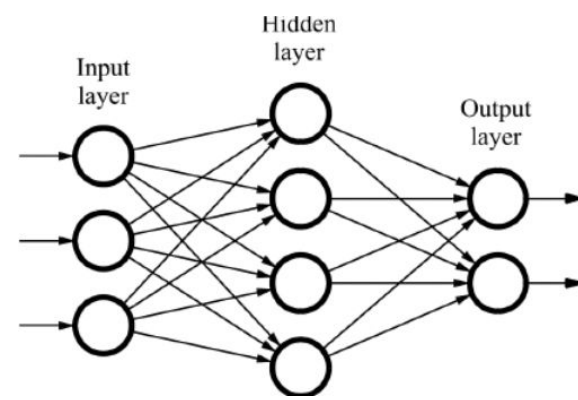
Misclassification Rate = 12.01%

Accuracy = 87.99%



Neural Network

-A powerful classification tool that works best with large training sets and numerical features



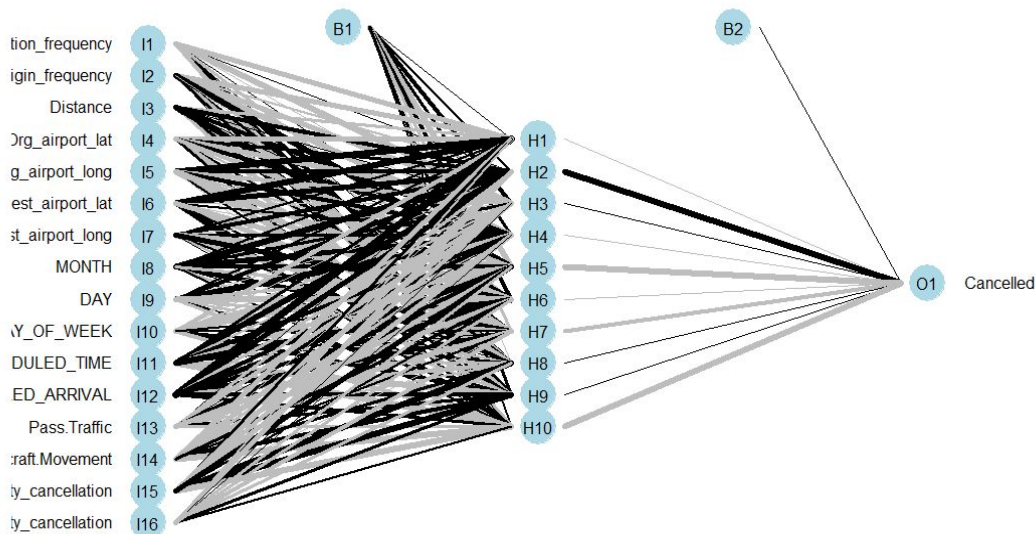
Confusion Matrix	NO	YES
NO	8367	966
YES	783	3729

Misclassification Rate = 12.63%

Accuracy = 87.37%

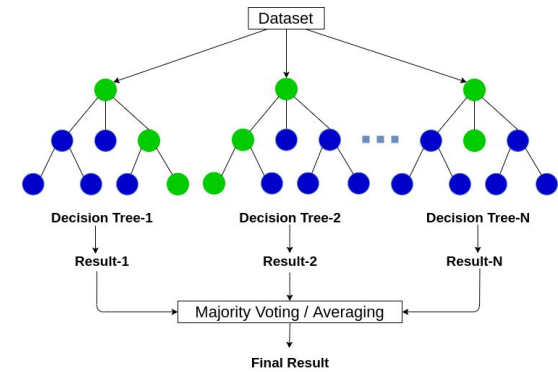
Neural Network

- Minimizes the value of a loss function, similar to SLR
- Tends to be computationally expensive (depending on the complexity of the data, can take days to run)



Random Forest

-Although the random forest with 16 predictors had the greatest accuracy, we gain simplicity and lose little predictive power with just 8 predictors



Confusion Matrix (16 predictors)	NO	YES
NO	9182	11
YES	31	4621

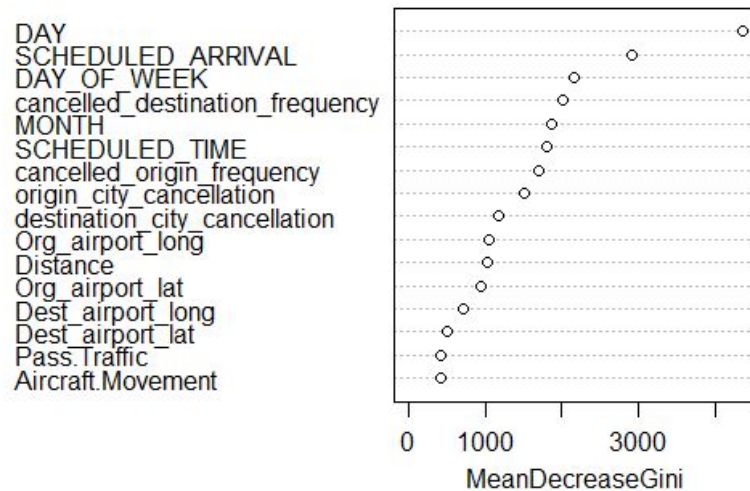
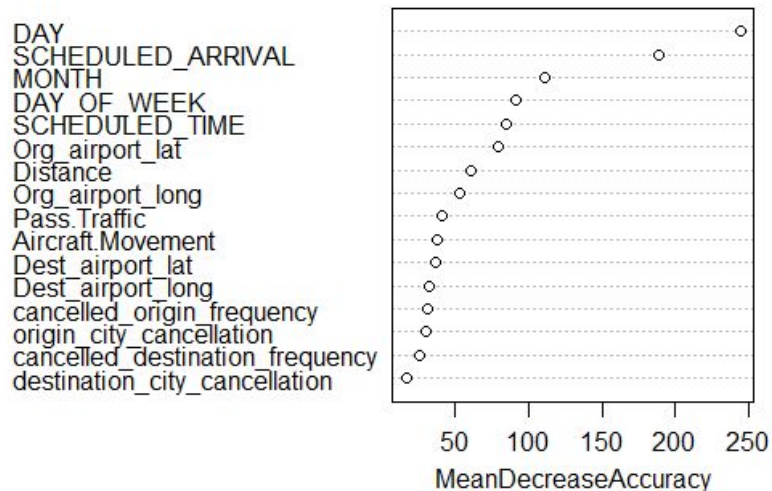
Misclassification Rate = 0.30%

Accuracy = 99.70%

Random Forest

Variable Importance Plot:

random_forest_flight



Conclusion

Conclusion

- How We Rank on Kaggle?

99.764 % accuracy on Public Leaderboard

- Advantage of Using Random Forest:

Recorded highest accuracy

Lower Risk of Overfitting



Conclusion

- Shortcoming Of Neural Network:

 - Blackbox nature

- Shortcoming of Random Forest:

 - Less Interpretable



Conclusion

Why was the random forest the best classification model? What were the most significant predictors?

How do we know the model is valid?

What, if anything, could we do to improve the model?

In what other situations would a model like ours be useful?

