Departure Delay Prediction for Flights

Akash Kashyap, Anant Gupta, John Kalathil, Matthew A. Lanham

Purdue University Krannert School of Management

kashyap5@purdue.edu; gupta499@purdue.edu; jkalathi@purdue.edu; lanhamm@purdue.edu

Abstract

The aim of this project is to predict the flight departure delay for a given set of flight attributes such as airport origin, destination airport, departure time, etc. The motivation for this study is that flight delays lead to economic losses of over $30 billion to both aviation industry and passengers. The delays are caused by a confluence of controllable and uncontrollable factors. Using R, we performed extensive data preparation using the **caret** package, and built several classification models including neural networks, naïve Bayes, logistic regression, and a C5.0 decision tree. These models were assessed and evaluated, and the best tuned neural net was picked.

Introduction

Flight delays are a common global phenomenon. According to the Bureau of Transportation Statistics, nearly one in four airline flights arrived at its destination over 15 minutes late.

Formulation of a delay management system can have multiple benefits. For the passengers, it can save them time. For the airline companies, it can save costs by streamlining processes and resources. For airports, it can assist in air traffic management. As a core part of a modern transportation system and economic system, the air transportation will influence the other downstream industries which rely on flights for their commercial operations. This fact motivates the need for accurate and practical prediction of flight delay.

The primary research questions that we are trying to address are:

- Can a flight delay be predicted? If yes, which is the best model?
- How much money can be saved if we are able to accurately predict the delay of flights?

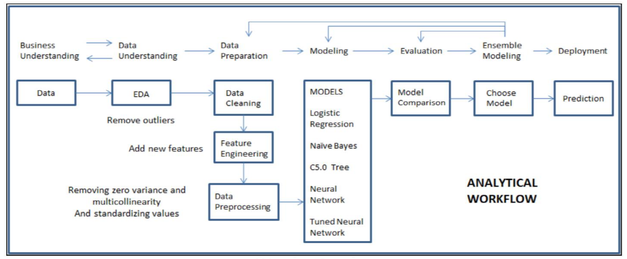
Literature Review

Study	Motivation	Algorithms used	Results
Predicting flight delay based on multiple linear regression	Prediction of arriving flights and prediction of delay	Naive-Bayes and C4.5	Experiment result showed that the model built performed better than the Naive-Bayes and C4.5.
CS229 Final Report: Modeling Flight Delays	Use of flight information and weather data to predict flight delay	Random Forest, Logistic Regression, Neural Networks	Machine learning algorithms predicted flight delay with an F1-score of 82%
Characterization and Prediction of Air Traffic Delay	Prediction of the delay experienced on certain routes	Classification, logistic regression, linear regression, neural nets and Random Forests	The results obtained for the 100 most-delayed OD pairs showed an average test error of 19%

Summary of literature review

In order to do a comprehensive analysis that goes over and above other studies, we captured multiple extra features like population of origin city, latitude, longitude etc. and incorporated feature engineering like binning the airports based on average departure delay of all flights at that airport and classified the airports into high, medium and low quality.

Methodology



Study Design

Data

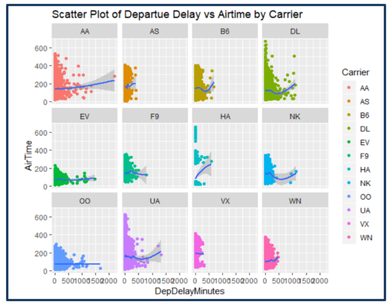
2017 US Domestic Flights US Bureau of Transportation Statistics database, US Census Report and Airports datasets.

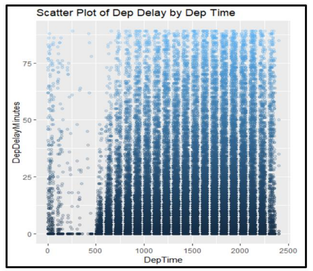
Feature Engineering

Airport Traffic, Airport Quality, Carrier Quality, City Bins based on population and Holidays which shows the days since the closest federal holiday.

Data Cleaning & Pre-Processing

EDA used to remove anomalies and outliers. The **caret** package was used extensively. Categorical variables were one-hot encoded, highly correlated and perfect linear combinations of features were removed, and the numeric features were pre-processed using a Z-Score standardization.





Model Design and Methodology Selection

We partitioned the data into a 60:40 train:set set to reduce the bias introduced during initial random sampling of the overall dataset. Also, 3-fold cross-validated was performed considering the large data size.
Methods: Logistic Regression, Naïve Bayes, Decision Tree, Neural Networks

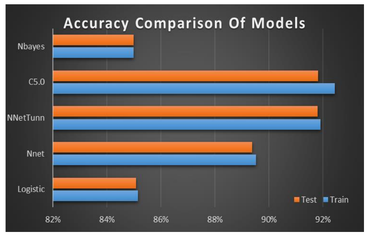
The models investigated were chosen based on various factors such as variance, bias, input features independence and model selection which takes care of linear and complex relationships between input and output variables.

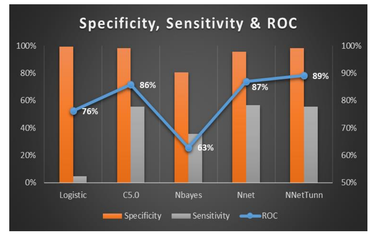
Model Evaluation / Statistical & Business Performance Measures

Models were evaluated on Accuracy, Sensitivity, Specificity and ROC. We used Sensitivity and Specificity because the dataset used had an uneven balance of observations for delayed and non-delayed flights.

Results

We used Sensitivity, Specificity and ROC value to determine the best model.

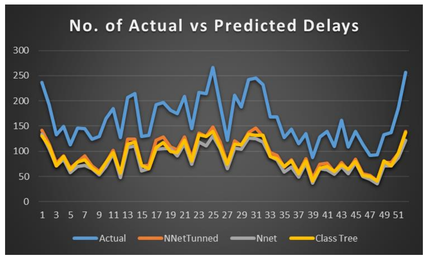


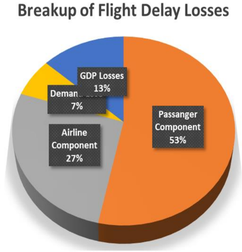


From the above charts, it is observed that NeuralNetTuned and Classification Tree have quite similar prediction accuracy.

The NeuralNetTuned was selected because it led to a higher sensitivity value that corresponded to predicting actual delays. ROC values also reveals that it performed better that the alternatives having a value of 89.17%.

Business Application





Using our best model, we could reduce the financial losses of the airline component completely and the other components by half. This results in an annual savings of almost $20.32 billions to the US.

Models	Money Saved
Logistic	9.1
C5.0	19.66
Nbayes	15.92
Nnet	19.88
NNetTunn	20.32

Potential Money Saved (Billions) by Model

Conclusions

Our business problem revolves around the economic losses of flight delays which are borne by the passengers, airlines, economy and society in general. With prediction accuracy of 92%, the model is an ideal candidate for understanding the critical factors behind flight delays and hence used in decision support systems to revamp the processes. Also, the predictions will empower stakeholders to make informed decisions relating aviation sector.

Model Prototype

R Shiny App: https://anantgupta.shinyapps.io/FlightPlanner/
Video demonstration: https://www.youtube.com/watch?v=aUXQas_iv4M