Big Data Management

Project nr 4: Predicting Flight Cancellations

Team members: Anna Maria Tammin, Maria Anett Kaha

Data Preparation

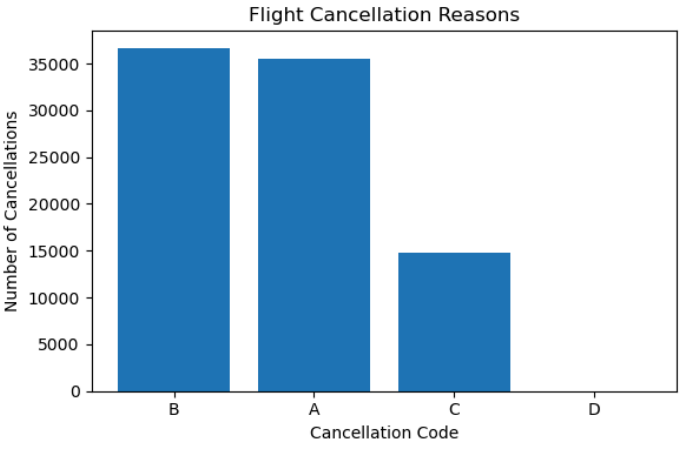
The aim of this project was to train four models in SparkML to predict flight cancellations. For this, we used flight data from two different years - 2009 data for training and 2010 for testing. We cleansed the data by extracting out diverted flights as was instructed for the project. After filtering out diverted flights, we dropped the column altogether. We renamed all of the columns to be more generally understandable compared to the provided column names.

We removed most of the original columns from the dataset as they represented values which can only exist for departed and landed flights. Keeping those features in the model would have given the model an unfair advantage or mislead the model if those values were replaced by either column means or common values. We created 4 additional columns from the date feature: DayofWeek, Month, Season and IsWeeked.

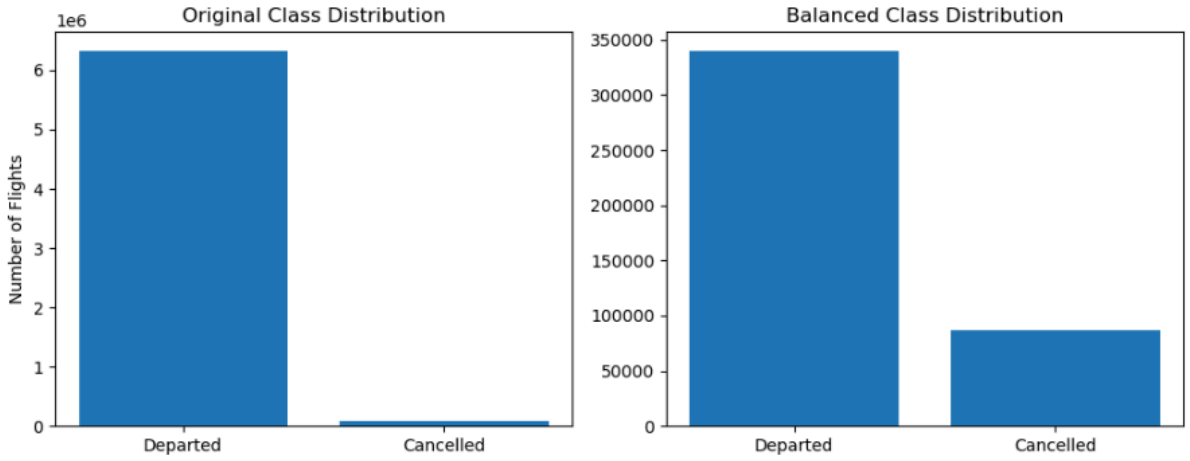
The flight dataset was heavily unbalanced, out of all the flights in the 2009 dataset, only around 1% were cancelled. This meant we had to balance the classes before training the model, otherwise the model would not predict any cancellations. We sampled 5% of all departed flights into the training data so the final training set contained around 80% departed flights and 20% cancelled flights (imbalance ratio of 0.26). We did not balance the dataset to be a 50-50 distribution as it would negatively affect the model making it well overpredict the number of cancelled flights. After all the preprocessing steps, the resulting dataframe contained around 400k entries.

Exploratory Analysis

Out of all the cancellations, the two top reasons for cancelling the flight were weather (B) and carrier (A) issues, and some due to the National Aviation System (C). There were



rarely any cancellations due to security reasons (D).



In the original 2009 dataset, the imbalance ratio was 0.012. After balancing the dataset, the imbalance ratio was 0.26. Above is class distribution visualized for cancellation status of flights before and after balancing.

Out of all airlines in the dataset, Southwest Airlines (WN) had the highest number of flights in 2009 with 69k flights (based on the balanced dataset). This is significantly more than all other airlines, the second and third largest carriers, American Airlines (AA) and SkyWest Airlines (OO), had 38k and 36k flights. Other top carriers had all less than 30k flights, except for Envoy Air (MQ) with 33k flights. Although MQ was the 4th largest carrier, it had the largest number of cancelled flights compared to other airlines.

Feature Selection

We used the following features to train the models:

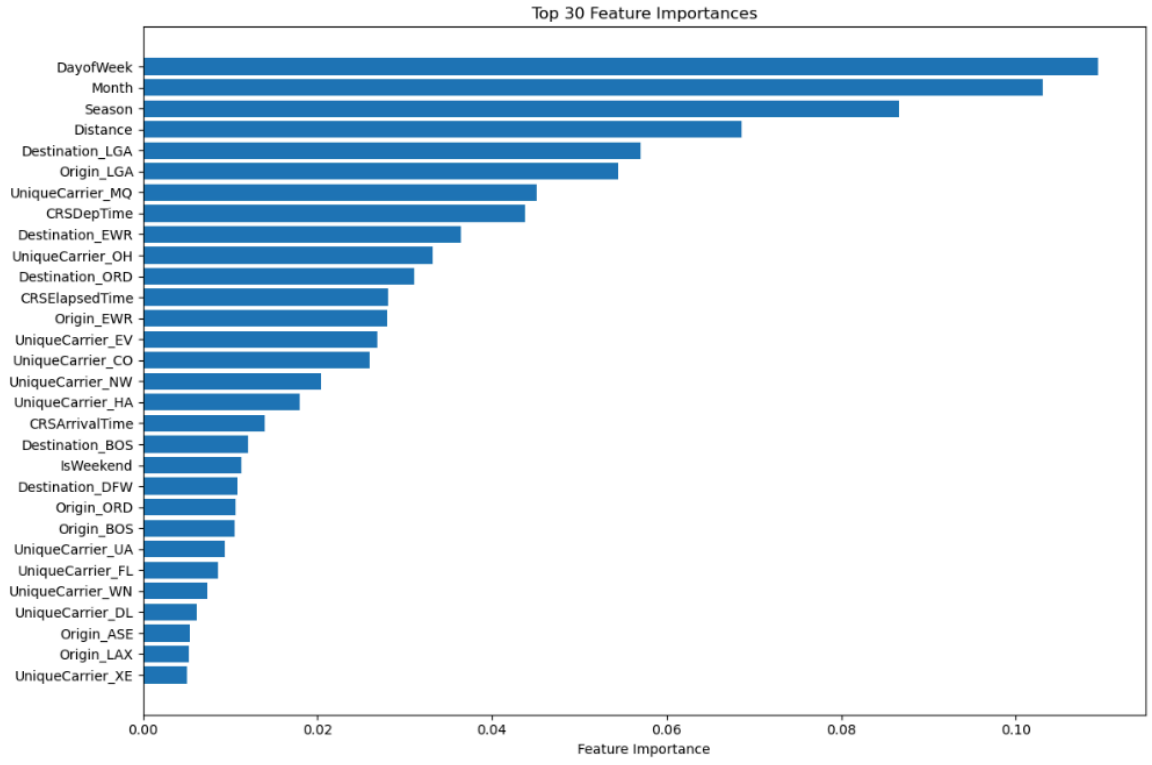
1. origin - the airport code where the flight departs;
2. destination - the airport code where the flight arrives;
3. unique carrier - the airline operating the flight;
4. CRS departing time - the scheduled departure time of the flight;
5. CRS arrival time - the scheduled arrival time of the flight;
6. CRS elapsed time - the scheduled total duration of the flight;
7. day of week - the day of the week the flight occurs (1 - Monday, 7 - Sunday);
8. month - the month in which the flight occurs (1 to 12);

9. is weekend - a binary indicator showing whether the flight is on a weekend (1 - yes, 0 - no)
10. distance - the distance between the origin and destination airports
11. season - the season based on the month of the flight (0 - Winter, 1 - Spring, 2- Summer, 3 - Autumn).

We selected these columns as these values do not depend on whether the flight has departed or was cancelled. We one-hot-encoded categorical features (origin, destination, carrier) and combined the features using a VectorAssembler.

Models

We trained four models on the 2009 flight data: Logistic Regression (LR), Random Forest, Decision Tree and Gradient Boosted Trees (GBT). We used 3-fold cross validation to tune hyperparameters of all four models. We evaluated these models based on accuracy and AUC score. Gradient Boosted Trees had overall the best performance out of all the models: 80% accuracy and an AUC score of 0.75. Logistic Regression AUC score was 0.0005 higher, but as the difference in accuracy was larger, we considered Gradient Boosted Trees to be the best out of two. Random Forest was also a strong competitor and reached near 80% accuracy, but the AUC score was lower (0.7) than for GBT and LR. Decision Tree was the worst performing model with 80% accuracy, but 0.59 AUC score.



We tested the GBT model on 2010 data which we first preprocessed similarly to the 2009 data. The GBT model was 77% accurate on the new data and had an AUC score of 0.69. This meant that the model was not guessing randomly. We also extracted feature importances from the GBT model and found DayofWeek, Month, Season and Distance to be the most influential features in the model. LaGuardia Airport (LGA) as a destination and origin was ranked higher than most other features. The top 30 most important features can be seen in the figure above.