Airline Data Analysis Using SparkML

Project Goal

The project aims to analyze the US domestic flight dataset using PySpark Dataframes and predict which flight/flight carrier is most likely to be canceled or delayed.

Dataset

The dataset is taken from Kaggle and consists of 7 CSV files containing information about airlines, delay information, location details (origins and destinations), and cancellation (reasons labeled as cancellation codes) over the period between 2009 to 2015 [1].

You need to do the following steps:

1. Load data

- Load the 2009 CSV file and combine the dataset (2009 to 2015 data).

2. Data Preprocessing

- Remove the unnamed columns and remove records containing null values.

3. Data Analysis

- Find the top 10 airlines with the most flight operations from 2009 to 2015.
- Visualize the proportion for the total flight cancellation reasons across 2009 to 2015 is shown below

4. Model Prediction

To predict whether the flight would be canceled or not, the dataset is modeled as a binary classification problem with categorical variables.

- Prepare the data for machine learning: use StringIndexer, OneHotEncoder, and VectorAssembler to transform our features
- Split the data into a 70/30 test and train ratio
- Apply models: logistic regression, decision tree classifier, random forest, and gradient boosted trees
- Compare all the models' accuracy for predicting cancellations

References

[1] Mu, Y. (2019, August). Airline Delay and Cancellation Data, 2009–2018. Retrieved April 2020 from https://www.kaggle.com/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018/data