Airline Data Analysis Using SparkML

Introduction

This report presents an analysis of US domestic flight data using PySpark Dataframes, aiming to predict flight cancellations based on available data. The project focuses on utilizing machine learning models to forecast whether a particular flight or carrier is more likely to experience cancellations or delays.

Data

The data used in this analysis was retrieved from Kaggle and is comprised of seven separate CSV files containing details about airlines, delay durations, location information (origins and destinations), and cancellations (reasons for cancellations are encoded) over the period between 2009 and 2015[1].

Methodology

Data Loading and Preprocessing

The data was loaded using PySpark with the first CSV file from 2009 being the initial dataset, which was then appended with data from 2010 to 2015. Following this, unnamed columns and records with null values in the selected columns were removed to clean the data and prepare it for analysis.

Data Analysis

With the data ready for exploration, the top 10 airlines with the most flight operations from 2009 to 2015 were identified. Subsequently, the proportions of different flight cancellation reasons across this period were visualized, providing insight into the most common causes of cancellations.

Model Prediction

To predict whether a flight would be canceled, the dataset was formulated as a binary classification problem, with 'CANCELLED' as the target variable. The data was prepared for machine learning through the application of StringIndexer and OneHotEncoder to convert categorical variables to numerical ones, and VectorAssembler to assemble these features into a single vector column. The data was then split into training and test sets in a 70/30 ratio.

Four models were trained and evaluated: logistic regression, decision tree classifier, random forest, and gradient-boosted trees. Each model's accuracy in predicting flight cancellations was then compared.

Results

The top 10 airlines with the most flight operations from 2009 to 2015 were identified:

OP_CARRIER	**Count**
EV	109,517
MQ	103,038
OO	84,931

OP_CARRIER	Count
WN	81,400
AA	70,605
DL	46,362
UA	42,303
US	36,711
XE	27,595
B6	25,401

The proportions of different flight cancellation reasons across this period were also evaluated:

CANCELLATION_CODE	Count
B	331,529
D	319
C	129,128
A	247,074

Models accuracies:

The accuracy of the machine learning models used in this analysis varied. The GBTClassifier and DecisionTree models both achieved perfect accuracy scores of 1.0, indicating that they were able to predict flight cancellations flawlessly in the test data. On the other hand, the RandomForest model had an accuracy of 0.98, which, while slightly lower than the GBTClassifier and DecisionTree models, is still a high score that demonstrates a strong ability to predict cancellations accurately. Lastly, the LogisticRegression model had an accuracy of 0.97, which is also quite high and indicates a strong predictive performance. These results suggest that all four models are highly capable of accurately predicting flight cancellations, with the GBTClassifier and DecisionTree models performing exceptionally well.

Conclusion

This project showcased the power of PySpark in handling large datasets and the utility of machine learning models in making predictions based on this data. The insights gained about flight cancellations and the performance of various models in predicting these cancellations could be useful for airlines in mitigating the impact of these events and improving their services.