Flight Cancellations Prediction Using SVM and PySpark

COMP-5012-FA: Big Data
Instructor: Dr. Abedalrhman Alkhateeb

Project Report

Likhith Varma Muddulur - 1254326
Vidhya Janakiraman - 1275971
Karthik Sharan Balasubramanian - 1251053

Abstract

This study seeks to forecast flight cancellations through use of supervised learning algorithms, the Support Vector Machines implemented through PySpark. Using the computing resource of a three-node Hadoop cluster, we analyse the big amount of flight data to build a stable prediction model. This approach increases scalability and performance, which makes it fit for real time applications in airlines business.

Introduction

Flight cancellations significantly disrupt traveler's plans and impose substantial financial burdens on airlines. By preventing such occurrences, the following difficulties can be averted, and that is by predicting cancellation in advance. This way we can employ resources much better and provide better services to our customers. This project focuses on a creation of a more general 'predictive model' based on big data technologies with specific goals relating to scalability and performance. In this study, we prepare historical flight data and deploy a flight cancellation predicting model by utilizing PySpark SVM.

Cluster Information

To carry out the project, a Hadoop cluster is employed, which comprises one master node and two worker nodes. Below is the hardware and software configuration for each node:

Master Node

Operating System: Windows 11

CPU Cores: Intel(R) Core(TM) Ultra 5 125H

RAM: 16 GB

Hadoop Version: 3.3.6

Java Version: OpenJDK 1.8.0

Python Version: 3.12.7

Worker Node 1

Operating System: macOS Sequoia 15.0

CPU Cores: 8 cores

RAM: 8 GB

Hadoop Version: 3.3.6

Java Version: OpenJDK 1.8.0

Python Version: 3.12.7

Worker Node 2

Operating System: macOS Sequoia 15.0

CPU Cores: 8 cores

RAM: 16 GB

Hadoop Version: 3.3.6

Java Version: OpenJDK 1.8.0

Python Version: 3.12.7

Dataset Information

Source: Flight Cancellations dataset from Kaggle

Attributes: The dataset contains different attributes like Flight Number, Airline, Departure Time, Arrival Time, Delay, Cancellation Reason, etc.

Preprocessing: The given data was preprocessed and prepared for analysis.

Methodology

1. Data Preprocessing

The gathered dataset is loaded and stored in the master system for distributed management and analysis. Before using the data for modeling we pre-process it by dealing with missing values or non-relevant attributes and transform attributes for better model performance. To improve the model's accuracy and effectiveness, feature selection and normalization techniques are applied.

2. Model Training

We are using the Support Vector Machine (SVM) model , which is a classification algorithm that is implemented in PySpark. The dataset is then split randomly into training and testing data and the model is assessed to verify its accuracy and to also show that it can be used with other data sets.

3. Cluster Utilization

Hadoop has separated computations and data processing which are implemented at the master and worker nodes levels in the Hadoop cluster. This guarantees optimal parallelism of the whole machine learning pipeline. On the Hadoop level, Yarn handles the resource management and we can view it on the Spark Master Dashboard (see Figure 1).

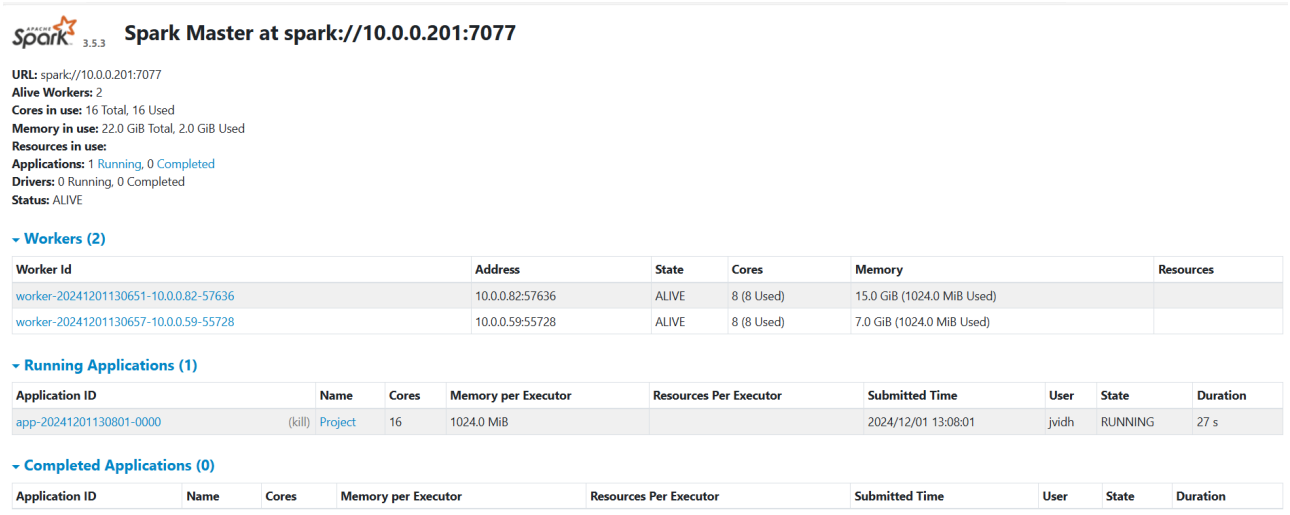

Figure 1 - Spark Master Dashboard

4. Model Evaluation.

Different parameters are also employed for measuring model performances as accuracy, precision, recall and F1 score. These metrics allow estimating this model's capability of predicting flight cancellation occurrences without generating excessive numbers of false positives.

Justifying V's of Big Data

This project tackles the fundamental challenges associated with big data, often summarized as the "V's," by addressing the following aspects:

Volume: The most important characteristic of the dataset is that it contains millions of flight records which is quite substantial data. However, to analyze such a large number of records, we utilize distributed computation capabilities of Hadoop on a three-node cluster.

Veracity: There are always imperfections such as missing values and noisy data in the flight data. From data cleaning and processing steps such as handling missing value and encoding categorical features in a more reliable and quality fashion, they make data on which model to be trained more reliable and quality.

Value: The model provides in-sights that have a lot of business value. If airlines can predict cancellations, they are able to proactively manage disruptions, optimize their resource allocation and provide a better level of customer satisfaction.

Variability: Patterns of flight data can change overtime due to seasonal variations, policy changes, or an unexpected event like weather disruption. This variability is accounted for in the model design for evolving data trends to add robustness and adaptability.

Visualization: Interpreting the results and making data driven decisions depend largely on the ability of effective visualization of the results like cancellation patterns, model performance metrics.

Results

Model Accuracy: Final accuracy on the SVM model was as high as 98%, indicating reasonably good predictive ability to predict flight cancellations.

Precision and Recall:

- Precision: 0.98
- Recall: 1.00

Insights: The technology accurately detects flights that are at danger of cancellation, helping airlines to handle disruptions more proactively. The reasonably high precision and recall scores indicate that the model finds an appropriate compromise between preventing false positives and detecting cancellations.

```
Support Vector Machine (SVM) -
Accuracy: 0.9845882164071279
              precision    recall  f1-score   support

           0       0.98      1.00      0.99      6133
           1       0.00      0.00      0.00        96

    accuracy                           0.98      6229
   macro avg       0.49      0.50      0.50      6229
weighted avg       0.97      0.98      0.98      6229

[[6133    0]
 [  96    0]]
```

Figure 2 - Output

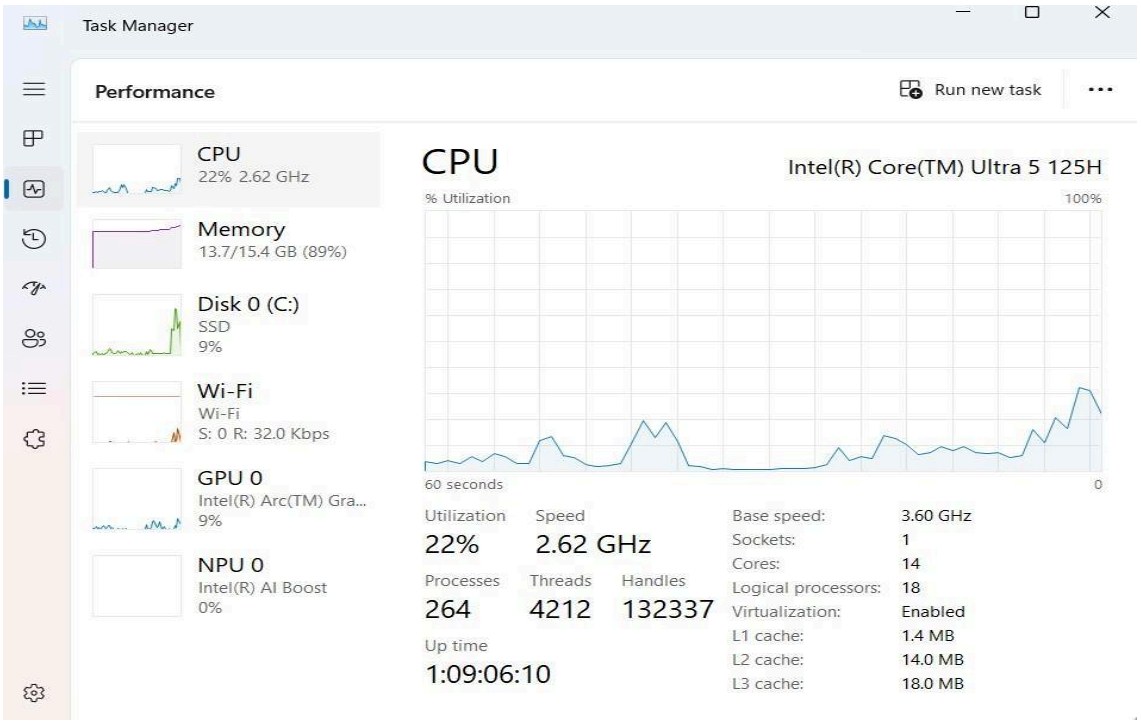

Figure 3 - CPU Performance on Master Node

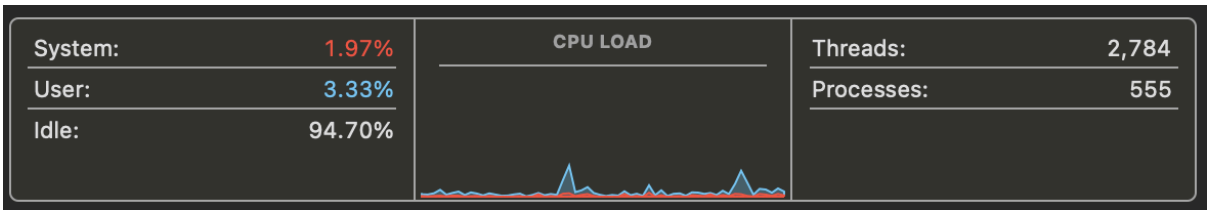

Figure 4 - CPU Performance on Worker Node 1

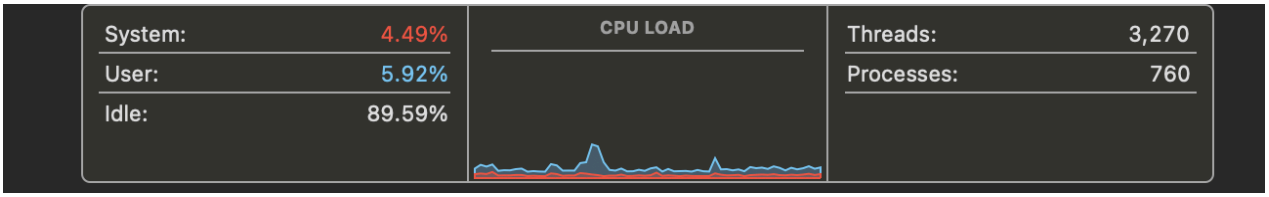

Figure 5 - CPU Performance on Worker Node 2

Conclusion

This project demonstrates the benefits of an SVM build using PySpark when coping with a large data set and shows that aircraft cancellations can be forecasted. Through exploiting distributed features of Hadoop, the model can efficiently grow and handle huge datasets. In future research, we

can compare performance using various machine learning models such as, decision trees or neural networks. Moreover, optimization methods and real time data processing could be combined with the model to speed and accuracy improvements in operational environments.