Flight Price Analysis Using PySpark

NIDHI DESAI
CS777

INTRODUCTION

- Analyzing flight prices by using pyspark
- Focus on data preprocessing, exploratory data analysis (EDA), and predictive modeling.

DATASET

- Sourced from Expedia for major US airports between April 16, 2022, and October 5, 2022

DATA PREPROCESSING

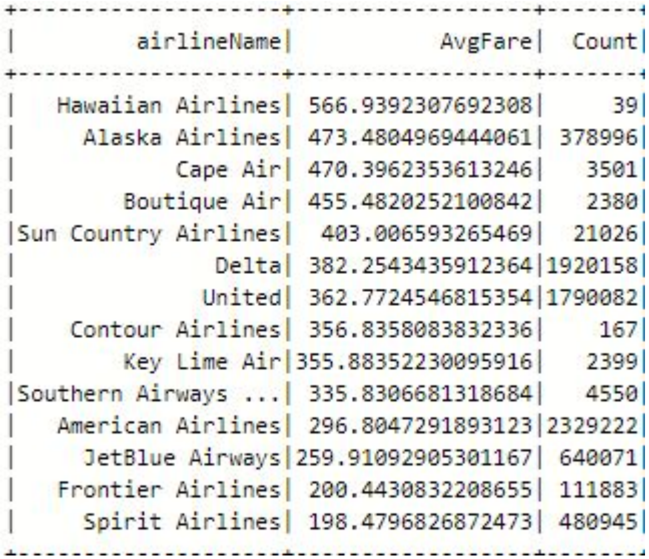
- Column selection and type conversion.
- Handling missing valueS
- Outlier detection and removal.
- Deduplication of data entries.

EXPLORATORY DATA ANALYSIS (EDA)

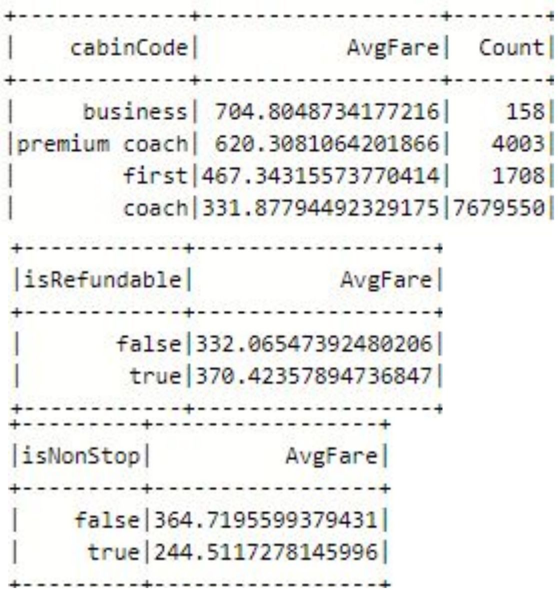
- Descriptive statistics calculation: Summary statistics for numerical columns: mean, median, standard deviation, etc.
- Correlation analysis: Pearson's correlation coefficient between numerical features and totalFare.

EXPLORATORY DATA ANALYSIS (EDA)

- Trend analysis over months and weekdays.
- Average fare analysis by airline, cabin code, refundability, and non-stop status.



```
+--------------------+------------------+-------+
|         airlineName|           AvgFare|  Count|
+--------------------+------------------+-------+
|   Hawaiian Airlines| 566.9392307692308|     39|
|     Alaska Airlines| 473.4804969444061| 378996|
|            Cape Air| 470.3962353613246|   3501|
|        Boutique Air| 455.4820252100842|   2380|
|Sun Country Airlines|  403.006593265469|  21026|
|               Delta| 382.2543435912364|1920158|
|              United| 362.7724546815354|1790082|
|    Contour Airlines| 356.8358083832336|    167|
|        Key Lime Air|355.88352230095916|   2399|
|Southern Airways ...| 335.8306681318684|   4550|
|   American Airlines| 296.8047291893123|2329222|
|     JetBlue Airways|259.91092905301167| 640071|
|   Frontier Airlines| 200.4430832208655| 111883|
|     Spirit Airlines| 198.4796826872473| 480945|
+--------------------+------------------+-------+
```



```
+-------------+------------------+-------+
|    cabinCode|           AvgFare|  Count|
+-------------+------------------+-------+
|     business| 704.8048734177216|    158|
|premium coach| 620.3081064201866|   4003|
|        first|467.34315573770414|   1708|
|        coach|331.87794492329175|7679550|
```

```
+------------+------------------+
|isRefundable|           AvgFare|
+------------+------------------+
|       false|332.06547392480206|
|        true|370.42357894736847|
+------------+------------------+
```

```
+---------+-----------------+
|isNonStop|          AvgFare|
+---------+-----------------+
|    false|364.7195599379431|
|     true|244.5117278145996|
+---------+-----------------+
```

Predictive Modeling Using Random Forest Regressor

Predicting the total fare of flight tickets based on various features related to the flight and booking details.

Why Random Forest Regressor?

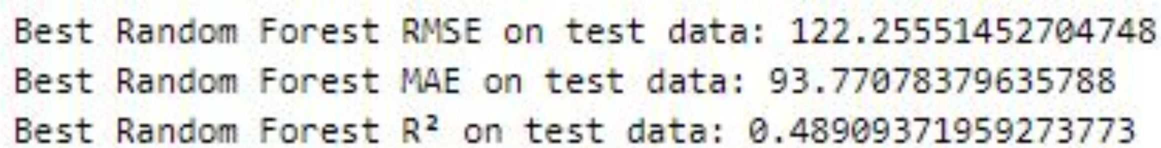
Robustness to Overfitting: Random Forests aggregate multiple decision trees, reducing the risk of overfitting compared to single decision trees. This ensemble method enhances model generalization to unseen data.

Feature Importance: RandomForestRegressor provides a measure of feature importance, helping in understanding which features contribute most to the prediction. This insight aids in feature selection and model interpretability.

Robust to Missing Values and Outliers: Can handle missing values and outliers more effectively than many other regression algorithms. This robustness makes it suitable for real-world datasets that often have imperfect data.

Model Evaluation

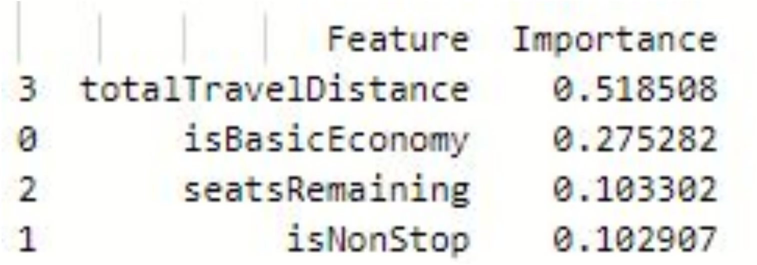
- Root Mean Squared Error (RMSE): Measures prediction error.].
- Mean Absolute Error (MAE): Average absolute error.
- R^2 Score: Proportion of variance explained by the model.



```
Best Random Forest RMSE on test data: 122.25551452704748
Best Random Forest MAE on test data: 93.77078379635788
Best Random Forest R² on test data: 0.48909371959273773
```

Feature Importance

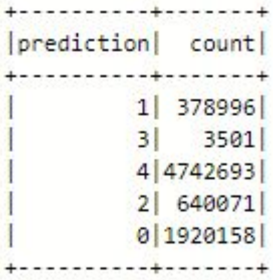
Identification of key features contributing to totalFare predictions.



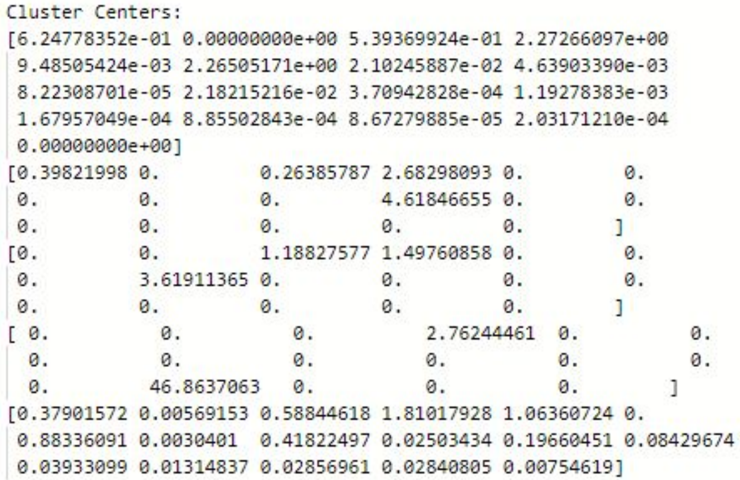
```
               Feature  Importance
3  totalTravelDistance    0.518508
0       isBasicEconomy    0.275282
2       seatsRemaining    0.103302
1            isNonStop    0.102907
```

Clustering

- Categorical to numerical feature conversion.
- Feature normalization with StandardScaler.
- K-Means model training with k=5.
- Evaluation using Silhouette score.
- Cluster center analysis.



```
+----------+-------+
|prediction|  count|
+----------+-------+
|         1| 378996|
|         3|   3501|
|         4|4742693|
|         2| 640071|
|         0|1920158|
+----------+-------+
```



```
Cluster Centers:
[6.24778352e-01 0.00000000e+00 5.39369924e-01 2.27266097e+00
 9.48505424e-03 2.26505171e+00 2.10245887e-02 4.63903390e-03
 8.22308701e-05 2.18215216e-02 3.70942828e-04 1.19278383e-03
 1.67957049e-04 8.85502843e-04 8.67279885e-05 2.03171210e-04
 0.00000000e+00]
[0.39821998 0.         0.26385787 2.68298093 0.         0.
 0.         0.         0.         4.61846655 0.         0.
 0.         0.         0.         0.         0.        ]
[0.         0.         1.18827577 1.49760858 0.         0.
 0.         3.61911365 0.         0.         0.         0.
 0.         0.         0.         0.         0.        ]
[ 0.          0.          0.          2.76244461  0.          0.
  0.          0.          0.          0.          0.          0.
  0.         46.8637063   0.          0.          0.        ]
[0.37901572 0.00569153 0.58844618 1.81017928 1.06360724 0.
 0.88336091 0.0030401  0.41822497 0.02503434 0.19660451 0.08429674
 0.03933099 0.01314837 0.02856961 0.02840805 0.00754619]
```

Conclusion

The results show a comprehensive analysis of flight data. The dataset has been preprocessed, including categorical encoding and feature scaling. The K-Means clustering yielded five clusters with a silhouette score of -0.365, indicating poor cluster cohesion and separation. The cluster centers reflect varying patterns in the scaled features. In the Random Forest regression model, the best RMSE is 122.26, MAE is 93.77, and R^2 is 0.49, suggesting moderate prediction accuracy. Feature importance highlights totalTravelDistance and isBasicEconomy as significant. Clustering results are diverse, with imbalances in cluster sizes, indicating challenges in clustering quality.

Future Work

Advanced Feature Engineering: Explore additional features like flight duration.

Model Selection: Experiment with alternative machine learning models.

Temporal Analysis: Investigate seasonal and time-based trends.

THANK YOU!!!