Airfare Prediction and Optimization with PySpark

IST 718: Big Data Analytics

Final Project Report

Team Members: Raghuveera Narasimha, Arunava Das, Tejas Mistry

1. Project Overview

The airline industry operates within a complex market influenced by a multitude of dynamic factors. Airfares are not static, they fluctuate based on real-time demand, seasonal trends, competitive pricing, fuel costs, and even day-to-day variations. This project, "**Airfare Prediction and Optimization with PySpark**" addresses the challenge of accurately predicting airfares using a substantial dataset of historical flight data. Our objective is twofold: first, to develop a robust predictive model that empowers travelers to make informed booking decisions by anticipating fare changes; and second, to explore potential optimization strategies for airlines to maximize revenue by dynamically adjusting prices based on predicted demand. The sheer volume of data necessitates the use of big data technologies, and we chose Apache Spark with PySpark as our primary tools for data processing, feature engineering, and model training. This choice allows us to efficiently handle the dataset's scale and complexity.

2. Prediction, Inference, and Other Goals

Our project had several interconnected goals:

- **Primary Prediction Goal:** To construct a high-accuracy predictive model for airfares. We aimed for a low RMSE (indicating accurate predictions in the original price units) and a high R-squared (indicating a strong fit of the model to the data). We set a target of achieving an RMSE within a reasonable range (within 10-15% of the average fare) and an R-squared above 0.9.
- **Key Inference Goals:**
    - **Feature Importance:** Determine the most influential factors driving airfare fluctuations. For instance, we investigated the relative importance of time-related features (days until departure, day of the week), route characteristics (origin, distance), airline reputation, and cabin class.
    - **Variable Relationships:** Analyze the nature of the relationships between features and airfare. For example, we explored whether the relationship between days until departure and fare is linear, or if there's a "sweet spot" for booking. We investigated interaction effects between features (does the impact of days until departure vary depending on the route?).
    - **Seasonality and Temporal Effects:** Quantify the impact of seasonality (peak travel seasons, holidays) and temporal factors (day of the week, time of day) on airfares.
- **Optimization Exploration:** Investigate how airlines could leverage our predictive model to dynamically adjust pricing. We considered scenarios like:
    - **Demand-based pricing:** Increasing prices when demand is predicted to be high and lowering prices when demand is low.

3. Data Exploration

Our dataset comprised millions of flight itineraries, providing a rich source of information. Key variables included:

- **Temporal Data:** searchDate, flightDate, elapsedDays, travelDuration
- **Flight Details:** originAirportCode, destinationAirportCode, segmentsAirlineCode, segmentsEquipmentDescription, totalTravelDistance, isNonStop
- **Pricing Information:** baseFare, totalFare
- **Other Attributes:** seatsRemaining, isBasicEconomy, isRefundable, cabinCode

Our exploratory analysis revealed several interesting patterns:

- **Fare Distribution:** The distribution of totalDuration and totalFare exhibited a strong right skew, with a long tail of expensive flights. This indicated the need for transformations (square root) to improve model performance.

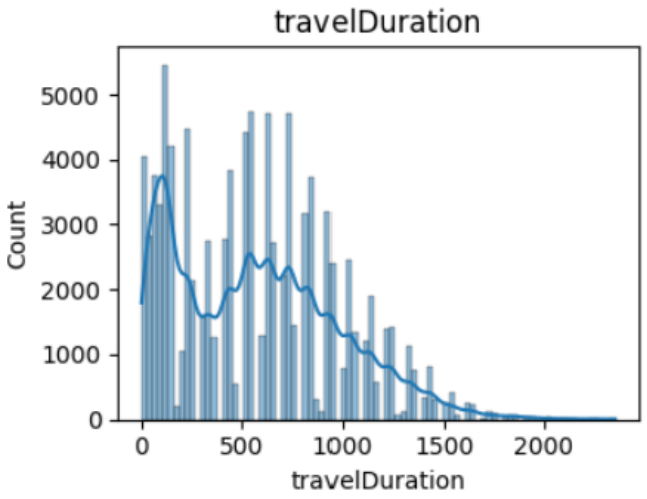

Fig. 1

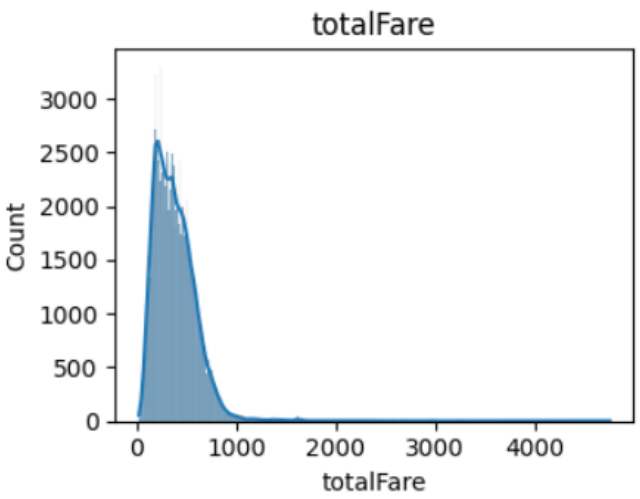

Fig. 2

- **Outliers:** Box plots revealed outliers in travelDuration and totalFare, likely representing unusual flight itineraries or data errors.

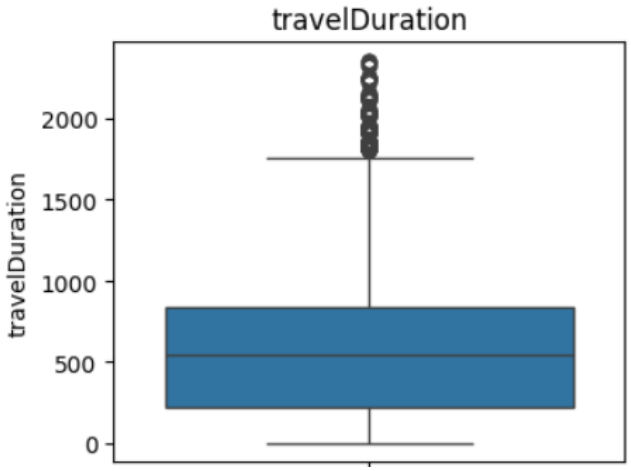

Fig. 3

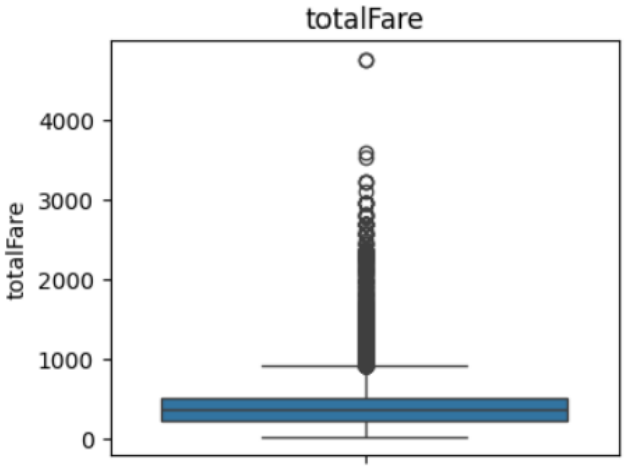

Fig. 4

- **Seasonality:** We found clear seasonal patterns in airfares, with higher prices during peak travel periods (summer, holidays)

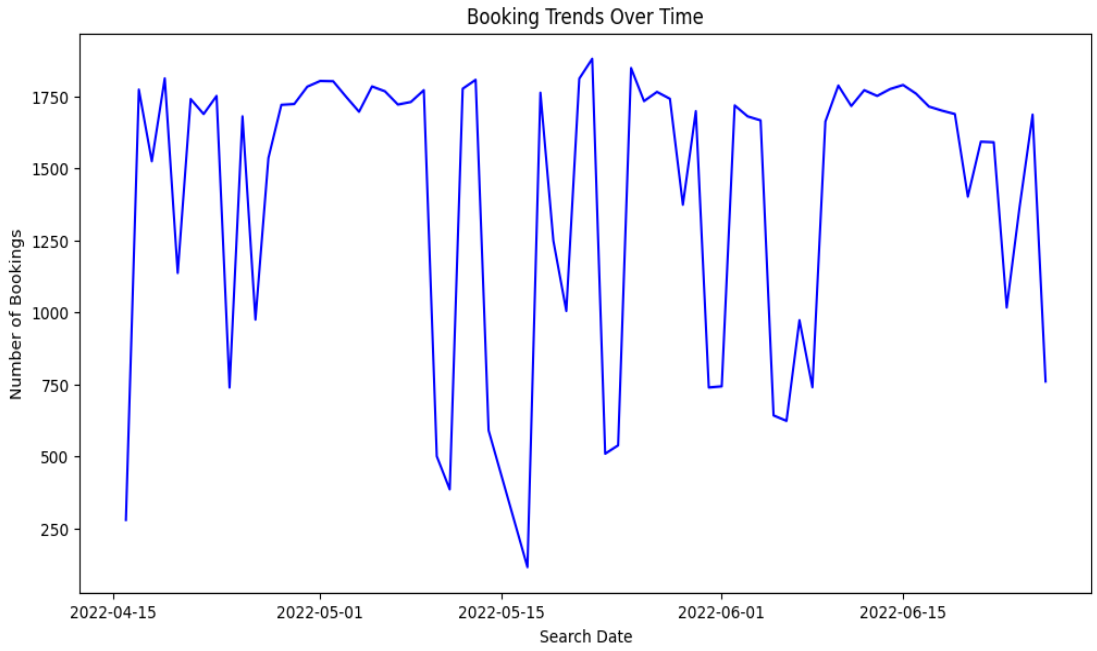

Fig.5

- **Missing Values:** We identified missing values in certain columns, particularly in segment-related information, which we addressed during data cleaning.

- **Other Insights From the data:**

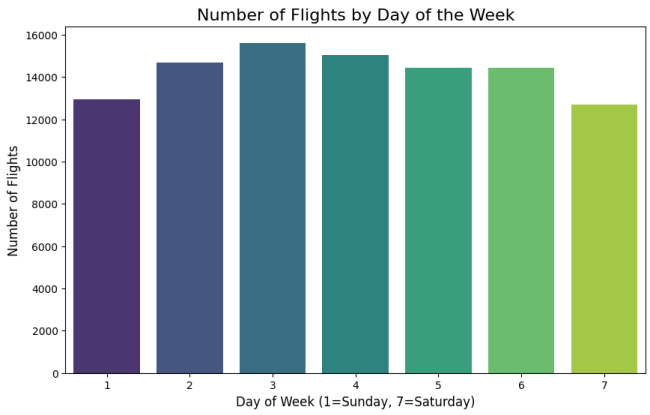

Fig. 6

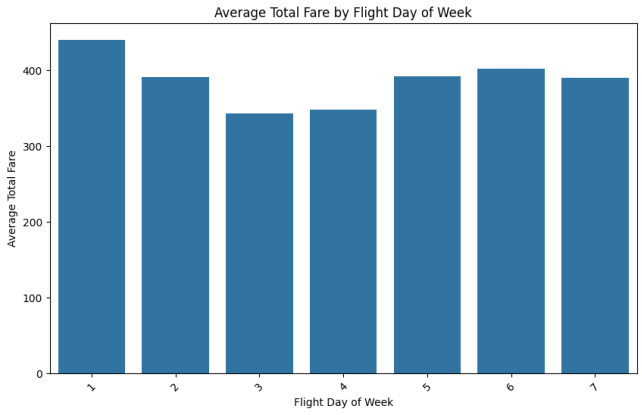

Fig. 7

- Figure 06 reveals a distinct weekday peak in flight activity, with a significant drop in flight numbers during the weekend. This pattern aligns with typical travel trends, where weekdays see higher demand due to business and leisure travel.

- The analysis of average total fares across different days of the week (figure 07) reveals minimal variation. While there are slight fluctuations, there is no strong evidence of a significant impact of the day of the week on average flight prices.

- **Correlation Analysis:**

Strong Positive Correlations:

totalFare and travelDuration (0.95): A strong positive correlation indicates that flights with longer travel durations tend to have higher fares. This is logical, as longer flights generally have higher operating costs for airlines.

totalFare and totalTravelDistance (0.76): A positive correlation suggests that longer flights are typically associated with higher fares. This aligns with the expectation that longer distances often translate to higher costs.

Strong Negative Correlations:

days_until_departure and fare_difference (-0.48): A moderate negative correlation suggests that as the departure date approaches, the difference between the fare and some reference fare tends to decrease. This could be due to last-minute price adjustments or a general trend of decreasing fares closer to the departure date.

Other Notable Correlations:

Intra-feature Correlations: Several features with their square root transformations exhibit perfect correlations (e.g., search_is_weekend and search_is_weekend_sqrt). This is expected as the square root transformation of a binary variable (0 or 1) maintains its original relationship.

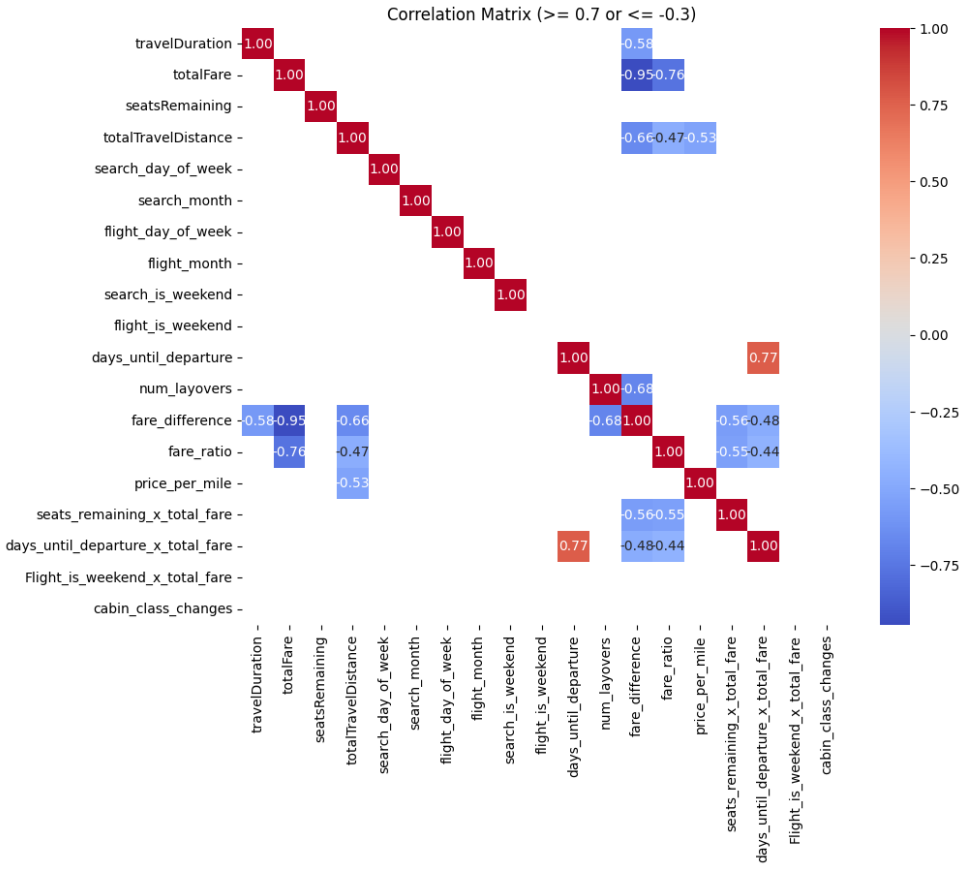

Fig. 8

4.Interesting/Surprising Results

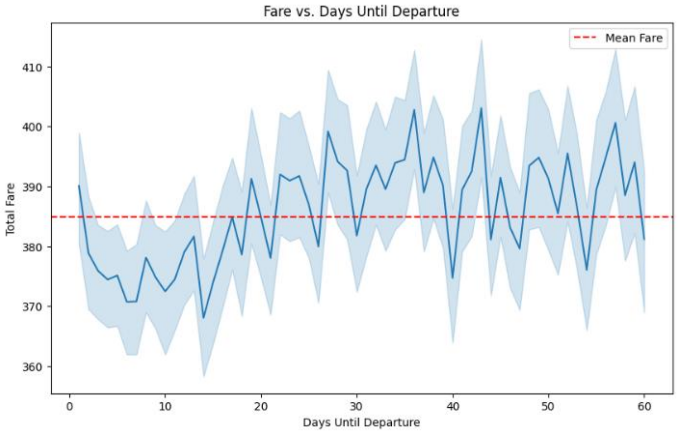

Fig. 9

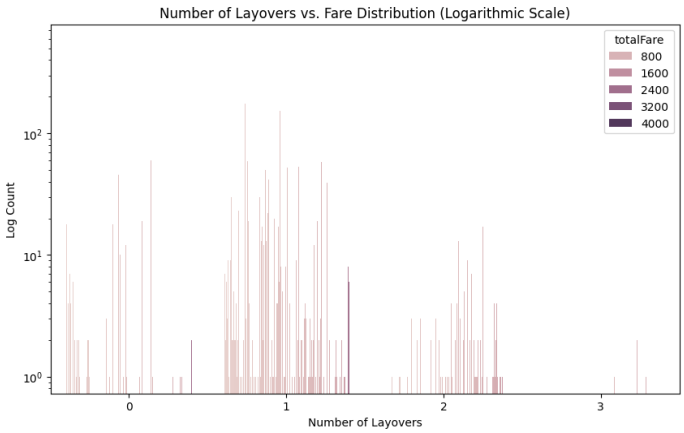

Fig. 10

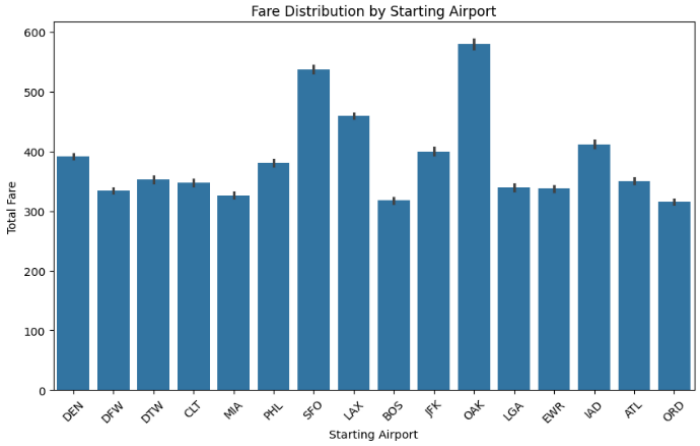

Fig. 11

- **Non-Linearity of Price vs. Time:** We initially expected a simple linear relationship between days until departure and fare. However, our analysis showed a more complex pattern, with fares often decreasing closer to the departure date (but sometimes increasing sharply very close to departure).
- **Impact of Layovers:** We hypothesized that more layovers would lead to lower fares. While generally true, we found cases where itineraries with a single layover were more expensive than direct flights, likely due to route popularity or airline pricing strategies.
- **Airport-Specific Pricing:** We observed significant variations in pricing strategies across different airports, even for similar routes and flight times. This highlighted the importance of including airline information in our model.

5.Summary of Methods Used

Data Cleaning and Preparation

The initial phase of our project focused on rigorous data cleaning and preparation. We addressed missing values in numerical features using KNN imputation, leveraging the relationships between variables to accurately estimate missing values. Outlier detection and removal were crucial, and we employed the IQR method to identify and handle outliers effectively. To ensure compatibility with Spark operations, we meticulously verified and converted all data types to their appropriate formats.

Feature Engineering

To enhance model performance, we implemented a comprehensive feature engineering strategy. We extracted key temporal features such as days_until_departure, search_day_of_week, flight_day_of_week, search_month, and flight_month to capture the temporal dynamics of flight prices. Furthermore, we incorporated route-specific features like originAirport, destinationAirport, and leg_routes (concatenated airport codes) to capture location-based influences. Price-related features, such as fare_difference, fare_ratio, and price_per_mile, were calculated to provide valuable insights into pricing trends. Additionally, we engineered interaction features, such as fare * days_until_departure, to capture potential non-linear relationships between variables. Finally, we effectively handled categorical features using StringIndexer and OneHotEncoder, transforming them into numerical representations suitable for machine learning algorithms.

Feature Selection

To optimize model performance and prevent overfitting, a rigorous feature selection process was employed. Correlation analysis was utilized to identify and remove highly correlated features, reducing redundancy and improving model interpretability. Furthermore, we leveraged feature importance scores derived from tree-based models to select the most influential predictors. This focused the model's attention on the most relevant information, enhancing its predictive power and reducing the risk of overfitting.

Model Training and Evaluation

We trained and evaluated several prominent regression models, including GBTRegressor, RandomForestRegressor, LinearRegression, and DecisionTreeRegressor, using PySpark's MLlib library. To ensure robust and reliable model performance, we employed k-fold cross-validation with a dynamically adjusted value of 'k' based on the dataset size. This rigorous cross-validation approach effectively prevented overfitting and provided a more accurate assessment of each model's optimization capabilities. Model performance was rigorously evaluated using key metrics such as Root Mean Squared Error (RMSE) and R-squared, providing a comprehensive understanding of each model's strengths and weaknesses.

This multi-faceted approach, encompassing thorough data preparation, effective feature engineering, rigorous feature selection, and robust model training and evaluation, led to the development of a high-performing airfare prediction and optimization model.

6.Results Summary

Model	RMSE	R-squared
GBTRegressor	12.60	0.99
RandomForestRegressor	16.86	0.98
Linear Regression	18.47	0.98
Decision Tree Regressor	15.25	0.98

The GBTRegressor model consistently outperformed the other models, achieving the lowest RMSE and highest R-squared. This suggests that the non-linear relationships captured by the GBT model are crucial for accurate airfare prediction. The Random Forest also performed reasonably well, indicating the importance of ensemble methods. Linear Regression, assuming linear relationships, had lower predictive power.

7.Problems Encountered

Processing and analyzing the large dataset presented several computational challenges. Running KNN imputation on the entire dataset proved computationally expensive. We addressed this by optimizing Spark configurations, such as increasing memory allocation for executors and drivers.

High multicollinearity among features was initially observed in the data. While Principal Component Analysis (PCA) was explored to address this, it oversimplified the feature space, leading to a significant loss of information and ultimately degrading model performance.

Furthermore, creating effective interaction features and handling high-cardinality categorical variables (e.g. airport codes) required careful consideration and experimentation.

We also had presence of multiple segments within the data, separated by the pipe operator (||). Exploding these segments significantly increased the feature space and model complexity.

8.Summary of How Well Goals Were Achieved

Our initial approach involved developing separate predictive models for various aspects of the airline industry, such as seat availability and customer behavior, with the aim of optimizing pricing strategies. However, computational limitations necessitated a shift in focus.

To address these challenges, we successfully developed a robust airfare prediction model and optimization using GBTRegressor, achieving high accuracy and providing valuable insights into factors influencing prices. This model now serves as the foundation for our subsequent efforts in price optimization. A key component of this optimization strategy involves estimating price elasticity of demand using historical data and the insights gained from our predictive model. This will enable us to dynamically adjust prices based on demand fluctuations, ultimately maximizing revenue while considering factors like such as customer behavior and aircraft seating capacity.

9.Citations

[1] *Flight Prices*. (n.d.). Retrieved December 13, 2024, from https://www.kaggle.com/datasets/dilwong/flightprices

[2] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

[3] *Apache Spark™—Unified Engine for large-scale data analytics*. (n.d.). Retrieved December 14, 2024, from https://spark.apache.org/

[4] *eventplot(D)—Matplotlib 3.9.3 documentation*. (n.d.). Retrieved December 14, 2024, from https://matplotlib.org/stable/plot_types/stats/eventplot.html

[5] *pandas—Python Data Analysis Library*. (n.d.). Retrieved December 14, 2024, from https://pandas.pydata.org/

[6] *Scikit-learn: Machine learning in Python—Scikit-learn 1.6.0 documentation*. (n.d.). Retrieved December 14, 2024, from https://scikit-learn.org/stable/

[7] Waskom, M. (2021). seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021