Machine Learning Project by Maria

1. Project Overview

Machine Learning Problem

- Continuous target variable ‘flight price’
- Regression

Topic Relevance

- Predicting flight prices is relevant for consumers
- Ticket buying behaviour could be adjusted based on insights

2. Data Selection & Preparation

Dataset

- 300.000 rows with information on flights between India's top 6 metro cities
- Source: Kaggle, original data scraped from Easemytrip (11.02-31.03.2022)

Features (11)

- Categorical: airline, flight, source city, destination city, departure time, arrival time (morning, early morning, afternoon, evening, night, late night), class
- Continuous: duration, days left, price

Data Cleaning & Wrangling

- Creating ‘euro_price’ column by dividing rupee price through exchange rate
- Dropping unnecessary columns: unnamed column, flight code, rupee price

3. Feature Engineering & Selection

Transformation of Categorical to Continuous Variables

- One Hot Encoding using pd.get_dummies for ‘airline’, 'source_city', 'destination_city', 'departure_time', 'arrival_time’
- Binary categorical variable class (business/ economy) to 0/1 using lambda
- Number of stops (zero, one, two or more) to 0/1/2 using lambda function

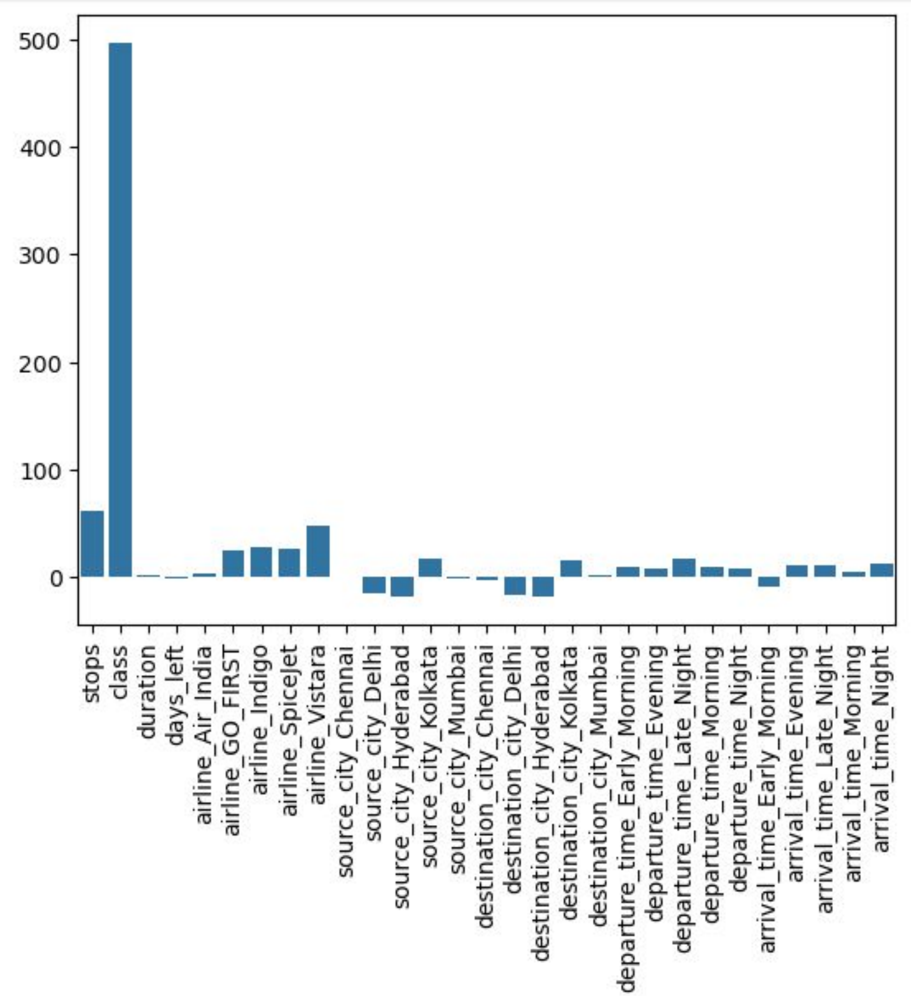
Feature Selection

- Correlation between features very low
- Correlation between features and target sufficient
- No features removed

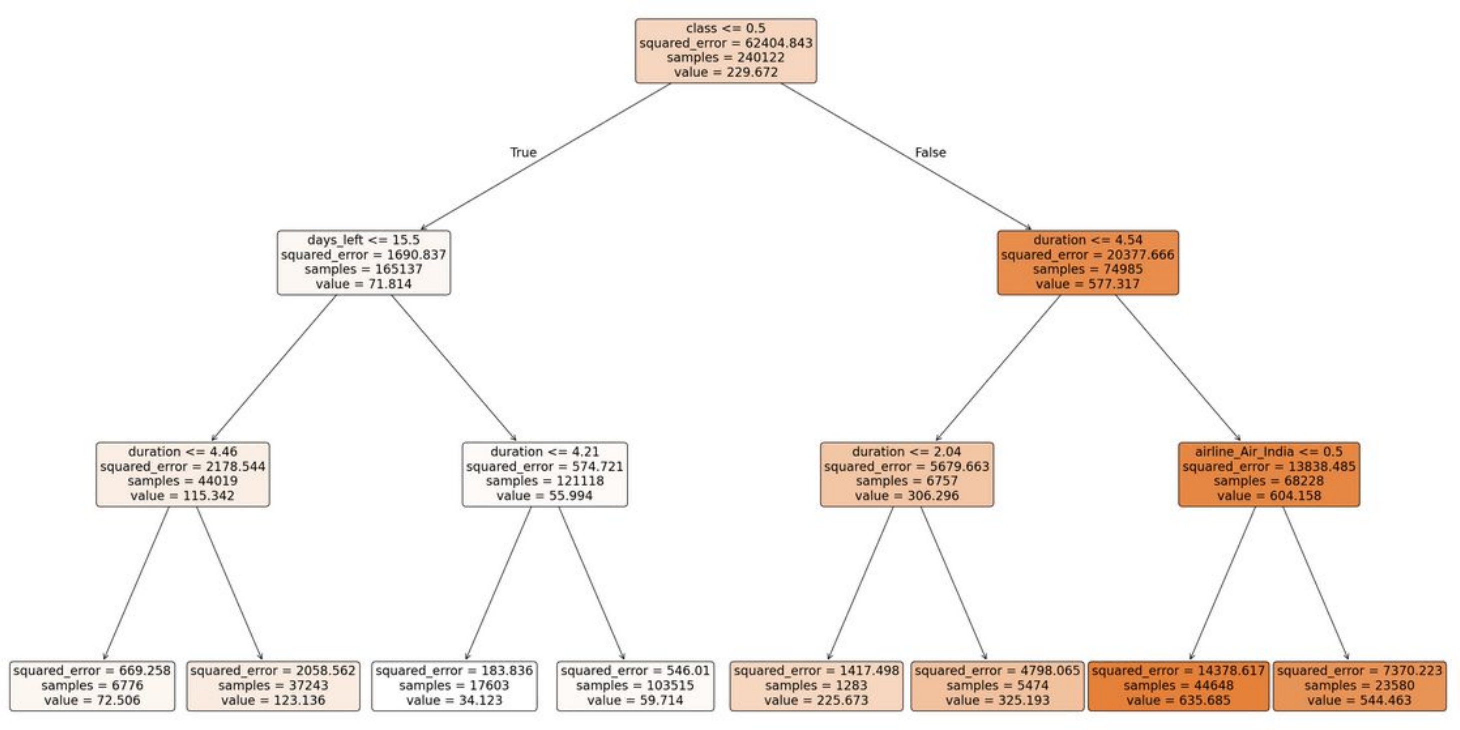
4. Model Building & Evaluation

MODEL NAME	EVALUATION METRICS
KNN Basic	R^2 0.76
KNN Normalized	**R^2 0.97**
KNN Standardized	R^2 0.96
Linear Regression	R^2 0.91 | MAE 49.39 | RMSE 74.35
Linear Regression Normalized	R^2 0.91 | MAE 49.39 | RMSE 74.35
Decision Tree	R^2 0.93 | MAE 38.08 | RMSE 62.89
Random Forest	R^2 0.96 | MAE 26.35 | RMSE 47.27
AdaBoost	**R^2 0.9800 | MAE 20.16 | RMSE 35.14**
Gradient Boost	**R^2 0.9827 | MAE 12.22 | RMSE 32.62**

5. Key Findings & Insights



5. Key Findings & Insights



6. Challenges & Learnings

Understanding the Scope of the Project

- Progress in content bit by bit
- Difficult to choose dataset that is fit for all applications

Time Constraints

- Finished presentation only late Thursday evening
- E.g. realised only then that duration column could have been transformed to integer/full hours for more intuitive interpretation but no time to run all models again
- Did not have time to make graphs pretty etc.

7. Future Work & Improvements

Limitation

- Class (business/economy) had highest impact in predicting flight price

Improvement in Future Work

- Class could be excluded from the model
- Or economy/ business analysed separately to investigate the impact of other factors

Conclusion

+ High explanatory / prediction power for flight prices in all presented models

+ Best models: AdaBoost and Gradient Boost (R^2 0.98) and normalized KNN (R^2 0.97)

- However, importance of class (business/economy) takes away explanatory power from other, more nuanced factors such as number of stops, time of departure and arrival etc.

Thank you!

Flight Price Prediction

Machine Learning Project by Maria