FLIGHT PRICE PREDICTION REPORT

❖ Introduction

The purpose of this report is to present the findings and evaluation of a Flight Price Prediction project. The project aimed to develop a machine learning model capable of predicting flight prices based on various features. In this report, we will discuss the performance and effectiveness of different regression models used for this prediction task.

❖ Methodology

For this project, we employed several regression models, namely DecisionTreeRegressor, SVR, KNeighborsRegressor, LinearRegression, RandomForestRegressor, AdaBoostRegressor, and GradientBoostingRegressor. Each model was trained on a labeled dataset containing relevant features such as flight duration, departure time, airline, and other factors that influence flight prices. The dataset was split into training and testing sets to evaluate the models' performance.

❖ Results and Evaluation

Let's examine the performance metrics for each model:

I. DecisionTreeRegressor

R2 score: 0.8202

R2 score for train data: 0.9966

Mean Absolute Error: 731.17

Mean Squared Error: 3,869,196.26

Root Mean Squared Error: 1,967.03

The DecisionTreeRegressor model achieved a decent R2 score of 0.82, indicating that it can explain 82% of the variance in the flight prices. The model also demonstrated excellent performance on the training data, with an R2 score close to 1. However, the Mean Squared Error and Root Mean Squared Error suggest that there is room for improvement.

II. SVR (Support Vector Regressor)

R2 score: -0.0215

R2 score for train data: -0.0251

Mean Absolute Error: 3,649.18

Mean Squared Error: 21,980,883.69

Root Mean Squared Error: 4,688.38

The SVR model performed poorly, with negative R2 scores indicating that it failed to capture the relationships between the input features and flight prices. The model's high errors, as indicated by the Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error, further emphasize its ineffectiveness in predicting flight prices accurately.

III. KNeighborsRegressor

R2 score: 0.6570

R2 score for train data: 0.7848

Mean Absolute Error: 1,621.54

Mean Squared Error: 7,380,454.28

Root Mean Squared Error: 2,716.70

The KNeighborsRegressor model achieved a moderate R2 score of 0.6570, suggesting that it can explain 65.70% of the variance in the flight prices. While the R2 score for the training data is higher, indicating a potential overfitting issue, the model's errors are within a reasonable range. However, there is still room for improvement to enhance its predictive performance.

IV. LinearRegression

R2 score: 0.4978

R2 score for train data: 0.4937

Mean Absolute Error: 2,309.73

Mean Squared Error: 10,806,679.96

Root Mean Squared Error: 3,287.35

The LinearRegression model achieved a modest R2 score of 0.4978, indicating that it can explain 49.78% of the variance in flight prices. The R2 score for the training data is similar, indicating a consistent performance. However, the model's errors suggest that it may not be capturing the complex relationships present in the dataset effectively.

V. RandomForestRegressor

R2 score: 0.9058

R2 score for train data: 0.9795

Mean Absolute Error: 597.99

Mean Squared Error: 1,900,286.29

Root Mean Squared Error: 1,378.51

The RandomForestRegressor model demonstrated excellent performance with an impressive R2 score of 0.9058, indicating that it can explain 90.58% of the variance in flight prices. The model also achieved a significantly lower Mean Squared Error and Root Mean Squared Error compared to other models, suggesting its superior predictive capability.

VI. AdaBoostRegressor

R2 score: 0.4734

R2 score for train data: 0.5399

Mean Absolute Error: 2,557.73

Mean Squared Error: 10,626,441.45

Root Mean Squared Error: 3,259.82

The AdaBoostRegressor model achieved a moderate R2 score of 0.4734, indicating that it can explain 47.34% of the variance in flight prices. The R2 score for the training data is slightly higher, indicating a reasonably good fit. However, the model's errors suggest that it may not capture the underlying patterns effectively, leading to less accurate predictions.

VII. GradientBoostingRegressor

R2 score: 0.8323

R2 score for train data: 0.8478

Mean Absolute Error: 1,205.40

Mean Squared Error: 3,383,257.76

Root Mean Squared Error: 1,839.36

The GradientBoostingRegressor model achieved a high R2 score of 0.8323, indicating that it can explain 83.23% of the variance in flight prices. The model also demonstrated good performance on the training data, further validating its effectiveness. With relatively low errors, this model can provide reliable flight price predictions.

❖ Cross Validation and Hypertuning

To mitigate the potential issue of overfitting, we employed cross-validation, a widely used technique in machine learning, on our two best-performing models: Random Forest and Gradient Boosting. Cross-validation helps assess a model's generalization performance by splitting the training data into multiple subsets and training the model on different combinations of these subsets.

After applying cross-validation, we proceeded to fine-tune the hyperparameters of both models using GridSearchCV. It systematically searches through a specified parameter grid to identify the best combination of hyperparameters that optimizes the model's performance.

The resulting scores obtained after cross-validation and hyperparameter tuning for Random Forest and Gradient Boosting are as follows:

Random Forest: 0.8191

Gradient Boosting: 0.8454

These scores represent the mean cross validated R2 scores, indicating the models' overall performance in predicting flight prices. Both models achieved relatively high scores, indicating their ability to capture the underlying patterns in the data and generalize well to unseen instances.

Based on the cross-validation and hyperparameter tuning results, we conclude that both Random Forest and Gradient Boosting models exhibit strong predictive capabilities for flight price prediction.

❖ Conclusion

Finally, to evaluate the models' performance on unseen data, we used the best-performing model, which in this case is the Gradient Boosting model, to predict flight prices on our test dataset. The test dataset contains instances that the model has not seen during training or cross-validation, providing a fair assessment of its generalization ability.

By employing a rigorous evaluation process, including cross-validation, hyperparameter tuning, and testing on unseen data, we ensure that the selected model is robust and reliable for flight price prediction.