Pricing Assignment Report

Team Members: Deep Goon, Chris Lin, Dalton Nycz, Prince Musonerwa Jr., Yue Yu

Data Preprocessing and Feature Engineering:

First, we combined the sales and ticketing data in R and created a combined dataset for training. We did the same with the prediction data. The data preprocessing part also involved aligning the training data (df) and the prediction data (df_pred). The primary goal was to ensure that these datasets were compatible, which was achieved by removing columns from df_pred that were not present in df. An important aspect of the data processing was the transformation of the Date column into a datetime object and further decomposed into year, month, and day features. This process added valuable temporal dimensions to the dataset. Additionally, the code distinguished between categorical and numerical data types, leading to a comprehensive preprocessing pipeline, involving scaling, one-hot encoding, and imputation strategies. These preprocessing steps played a crucial role in maintaining data integrity and significantly improved the dataset's overall quality.

Model Training, Evaluation, and Selection:

During the model stage for the analysis, a variety of regression models were evaluated, including Ridge Regression, Decision Tree, Random Forest, Gradient Boosting, AdaBoost, XGBoost, LightGBM, and SVR. This approach provided a wide spectrum for identifying the most suitable model. Hyperparameter tuning was conducted for each model using GridSearchCV, employing MAE as the evaluation metric. The Ridge Regression model achieved a best cross-validation MAE score of 4199.845 and a test MAE of 4210.896. the standout result was the XGBoost model, which recorded the lowest MAE of 2810.999 with the parameters model_learning_rate: 0.1, model_n_estimators: 100.

Prediction Process:

We first differentiate and process categorical and numerical features in df_pred using a preprocessing pipeline. This pipeline includes imputation of missing values (using the mean for numerical and the most frequent value for categorical data), standardizing numerical features, and applying one-hot encoding to categorical features. An XGBBoost regression model, previously optimized with specific hyperparameters (learning rate and number of estimators) applied. The preprocessed df_pred is fed into this model to forecast the 'Tickets' sales. These predictions are then appended to the df_pred data frame, and the enhanced dataset with the predicted Ticket values is saved as a CSV file.

Conclusion:
Our analysis shows a methodical and detailed procedure for forecasting ticket sales, underscored by thorough data preprocessing, comprehensive evaluation of various models, and judicious selection of the most suitable model. The findings particularly emphasize the effectiveness of the XGBoost model, as evidenced by its considerably lower MAE in comparison to alternative models. This methodical approach ensures that the predictions are as accurate and reliable as possible, making it a valuable tool for data-driven decision-making in ticket sales forecasting.

Our work can be found at the forked repo:
https://github.com/dgoon29/braves_pricing_2024

All of our work lives in the *project* folder

Code for data processing and combining data:

- *project/data_pre_processing.Rmd*

Model building in a python notebook:

- *project/submission/Braves_pricing.ipynb*

Output predictions in CSV and Excel file:

- *project/submission/solution_prediction.csv*
- *project/submission/solution_prediction_excel.xls*

Report describing our approach:

- *project/submission/Pricing Assignment Report.pdf*