PREDICTIVE ANALYTICS COMPETITION REPORT

Data Analysis Process and Insights

By – Ushmi Kuvadia

Introduction:

This report presents a comprehensive summary of the data analysis process conducted to predict car prices. The analysis involved exploring the dataset, preparing the data, and experimenting with various predictive models.

Exploring the Data:

Upon loading the dataset, the first step was to conduct exploratory data analysis. This involved:

- Checking the structure of the dataset to understand the variables and their types.
- Investigating missing values to determine their impact on the analysis.
- Exploring summary statistics and visualizations to gain insights into the distribution and relationships between variables.
- Addressing multicollinearity issues by examining correlation matrices and removing highly correlated variables.

Data Preparation:

1. **Handling Missing Values:**

- Rows with missing values in the 'price' column were removed, as it was the target variable.
- Missing values in categorical variables were imputed with the mode to retain valuable information.
- Missing values in numerical variables were imputed with the median to maintain the distribution of the data.

2. Feature Engineering:

- Created a new feature 'age' from the 'year' variable, providing additional information for model training.

Analysis Techniques Explored:

1. **Linear Regression:**

- A simple linear regression model was initially built with only quantitative variables.
- The model was extended to include both quantitative and qualitative variables, aiming to improve predictive performance.
- Different methods for handling missing values and multicollinearity were explored.

2. **Random Forest:**

- Employed the Random Forest algorithm to build a predictive model.
- Various parameter settings were experimented with to optimize the model's performance.

3. **Gradient Boosting Machine (GBM):**

- Utilized GBM to construct another predictive model.
- Hyperparameters were tuned to enhance the model's predictive accuracy.

4. **Decision Trees:**

- Decision Trees were also employed to develop a predictive model.
- The model's performance was evaluated and compared with other techniques.

5. **Bagging:**

- Implemented bagging with Random Forest to further enhance predictive performance.

Model Performance and Insights:

1. **Linear Regression:**

RMSE: 10506.669

2. **Random Forest:**

RMSE: 9632.619

3. **Gradient Boosting Machine (GBM):**

RMSE: 12108.461

4. Decision Trees:

RMSE: 10239.172

Conclusion:

In conclusion, our analysis aimed to develop a predictive model for car prices based on various features. While the linear regression model performed reasonably well, several other models, including Random Forest, Gradient Boosting Machine (GBM), Decision Trees, and Bagging, outperformed it in terms of predictive accuracy. Among the ensemble methods, the GBM model achieved the lowest RMSE, indicating its superior performance in predicting car prices. However, our analysis encountered several challenges. These challenges included handling missing values, addressing multicollinearity, and selecting the appropriate features for the models. Additionally, despite our efforts to improve the model's performance, including feature engineering and parameter tuning, further optimization may be required. In future analyses, we plan to explore additional feature engineering techniques and further optimize the selected models to improve predictive performance. Furthermore, we aim to investigate other machine learning algorithms and ensemble methods to identify the best-performing model for predicting car prices accurately.