Predictive Analysis of Vehicle Pricing: Exploring Influential Factors

Data Cleaning

In this predictive analysis competition, I delved into the various factors influencing vehicle prices, considering both features and conditions. The initial step involved a thorough data cleaning process. Approximately 11% of the data contained null values across several independent variables. To address this, I merged the analysis and scoring datasets using the rbind() function, ensuring consistency in data cleaning.

Subsequently, I modified data types to better reflect their nature. For instance, I converted the 'is\_new' variable into a logical type using as.logical(). For numeric variables like 'city\_fuel\_economy', 'engine\_displacement', and 'horsepower', which correlate with 'body\_type', I applied a median imputation strategy. Grouping these variables by 'body\_type' and using their median values to fill in missing data proved more effective than deletion or zero substitution, as evidenced by a reduced Root Mean Square Error (RMSE).

Seller ratings with missing values were assigned the median rating based on their wheel system. For 'mileage', new cars were assigned a zero value, while used cars had their mileage estimated based on the car's age and average yearly mileage (14,263 miles). 'Owner\_Count' was similarly treated, with median values based on the car's year.

A challenge arose with variables like 'frame\_damaged', 'has\_accidents', 'salvage', and 'fleet'. Initial attempts to introduce an 'Undisclosed' category for these boolean variables led to errors and uninterpretable coefficients in the model, prompting me to abandon this approach. Predictive models, including random forests and logistic regression, were tested to impute these variables. However, they significantly increased the final model's RMSE, indicating a negative impact on predictive accuracy.

Feature Selection

Feature selection involved a careful examination of correlations between variables. For example, a strong linear relationship (over 0.9 correlation) between 'city\_fuel\_economy' and 'highway\_fuel\_economy' suggested that choosing either variable would suffice. After data cleaning, the analysis and scoring data were separated again. The feature selection process, guided by a hybrid stepwise method, identified key predictors: 'horsepower', 'mileage', 'engine\_type', 'is\_new', 'wheel\_system\_display', 'body\_type', 'engine\_displacement', 'year', 'seller\_rating', 'height\_inches', 'isCab', 'franchise\_dealer', 'owner\_count', 'daysonmarket', 'maximum\_seating', 'city\_fuel\_economy', and 'is\_cpo', plus 'wheelbase\_inches'. These variables were instrumental in constructing effective predictive models.

Model Selection

With the relevant features identified, various models were considered for predicting vehicle prices. The models tested included linear regression, generalized additive models, regression trees, and a random forest model using ranger. The linear model's RMSE on test data was 8000.6, while the generalized additive model, which applied smoothing to variables like 'mileage' and 'seller\_rating', achieved a slightly lower RMSE of 7863. The regression tree model underperformed, with an RMSE of 8599.6, suggesting limitations in its tree structure and the need for a different feature selection approach. The effective model was the random forest using ranger, which yielded the lower RMSE of 4570, significantly outperforming the other models. However, using xgboost yield an average RMSE of 100 round resulting 3784. In this case, Xgboost is the most effective prediction on used car prices.

Conclusion

In summary, the random forest model with ranger and xgboost emerged as the most effective approach for predicting vehicle prices in this dataset. This model's superior accuracy demonstrates its robustness in handling the complex relationships between vehicle features and pricing.

Possible Improvement

There are few possible improvement methods on handling NA values, feature selection and models evaluations. To begin with, there are more complex models can be applied in predicting missing values, such as KNN. Also, for feature selection, there are some hidden correlations in my independent variables, such as year and mileage are somehow correlated. What’s more, there are more models that I can choose from. Also, I can improve my model by finding the best tune and other parameters.

Conclusion

As a result, this analysis was able to pinpoint important variables affecting car costs and identify the best possible prediction model from a pool of contenders. The random forest model with ranger, which stands out for having the lowest RMSE, shown to be especially skilled at managing the dataset's complexity. However, there is room for development. More sophisticated imputation methods might enable more precise management of missing data, while feature engineering and model optimization could enhance the forecast precision even further.