9321 ASS3 REPORT

Haoze Zhang z5364634

April 21, 2024

1 Preprocessing & Postprocessing

For the aspect of data preprocessing, I translate the value with OneHotEncoder in the columns of 'area_cluster', 'make', 'segment', 'model', 'fuel_type', 'engine_type', 'rear_brakes_type', 'transmission_type' and 'steering_type', which are discrete categorical features. Then replace True and False boolean value with 1 and 0 after the OneHotEncoder. About the columns of 'max_torque' and 'max_power' which consist of two sets of number and their measurement units, I split them into four columns and remove their measurement units. Besides, in my opinion, columns of length, width and height can be combined together due to their similarity. So I create a new column called 'volume' composed of the product of these three. Finally, delete useless columns, 'max_torque', 'max_power', 'length', 'width' and 'height'.
As for data postprocessing, a method called SelectFromModel from the sklearn library and one function of XGBoost model called feature importances are both used in combination. Firstly, feature importances provides threshes which are how frequently each feature is used to split the data across all the trees in the ensemble, weighted by the improvement to the model's performance brought by each split. Through iterating over threshes, the size of the training dataset becomes smaller and smaller. In each iteration, I can evaluate which size is best for my model to predict or classify targets.

2 Machine learning algorithms and Tune

XGBClassifier and XGBRegressor models from XGBoost are used in the classification and prediction tasks respectively. Because XGBoost uses a technique called "tree pruning" to remove splits that have little impact on improving the model's performance. This helps to build simpler and more interpretable models, while still maintaining high predictive accuracy. And XGBoost provides feature importance scores, which allows me to understand the relative importance of each feature in making predictions. This can be valuable for feature selection and understanding the underlying patterns in the data.
A method called GridSearchCV from sklearn library is used to tune hyperparameters in models I chose before. In the tasks of prediction and classification, A few of the biggest concerns in the model are 'max depth', 'learning rate', 'n_estimators', 'reg_alpha' and 'reg_lambda'. Besides, the extra considered parameter in the classification task is 'scale_pos_weight' due to the imbalanced dataset. Then define a paramter grid composed of them and their ranges of possibly accessible values. Put the grid into the GridSearchCV and obtain the best results of the parameter combinations. Finally, input the combination into the chosen model and predict the target value.

3 Business Value

My model can segment customers based on their risk profiles, preferences, and behaviors. This segmentation enables insurers to tailor their products and services to specific customer segments, leading to higher customer satisfaction and retention. Personalized offerings can also drive cross-selling and upselling opportunities.
And the solution can predict the likelihood and severity of insurance claims based on historical data and current trends. By anticipating claim volumes and costs, insurers can allocate resources more effectively, streamline claims processing, and improve customer satisfaction through faster settlements.