Project Report of Vehicle Recall Analysis

In fulfillment of Take-Home-Test as part of interview

Ailing Liu, alingliu88@hotmail.com

Project background:

The manufacturer is deciding if a specific vehicle should be recalled based on the historical data of models released in last summer. The provided dataset has 100 instances, 8 attributes, and 1 recall result. My goal is to build not only a machine learning model to recommend recall, but also analyze the data to draw insights that support stakeholders to make recall decision.

In the first half of the project, I used statistical analysis, visualization tools, parametric and non-parametric models that successfully identified multiple predictors driving recall result. In the second half, I built a supervised learning model to recommend whether a model should be recalled. A few metrics were also proposed to address business interest and evaluate model performance. In below, I would discuss the challenges I faced in this project and the solutions and execution to accommodate the problems.

Challenges:

The main challenge of this project was the small data set. While data is small, over-fitting becomes much harder to avoid, and the model is prone to lose generosity in predicting unseen data. In our dataset, not only the number of cases was limited, the size of the features was also inadequate. Therefore, finding the key driver from these features became more difficult. Last but not least, the subject of the problem focused on whether to recall a specific vehicle, but the provided data was aggregated summary by model specifics, including quantity of vehicles sold. The mismatch of data and desired outcome required careful selection of features. For example, quantity sold is not a valid feature for a specific vehicle, but model year, variant, etc. is.

Solutions and Execution:

<u>Visualize data and Conduct statistical analysis</u>

Machine learning model is not the only solution. A good visualization gives immediate insights to viewers as clear conclusion. In this project, model year played a key role in predicting recall. Just eyeballing the distribution, we can see there was obvious recall differences. Moreover, from the plots of each given feature by recall result group, we could spot some differences in feature level, some of which were more likely to be recalled than others. In addition, we also found out the quantity sold in recall group have recurring pattern in time series manner. (Suggestion) Provided with panel data, we can build a time series model to predict vehicle sold in next model.

Statistical tests, parametric models, non-parametric models, bootstrapping, and other useful mathematical tools are the domain of classical statistics. In this project, using Z test, we compared the recall rate in two wheel types and concluded they did not make a difference in recall result. On the other hand, variant type played a key role in recall prediction. In particular, vehicle with Base variant was twice more likely to be recalled than vehicle with GT variant. Apart from statistical test, logistic regression was used to interpret feature effects in relation to recall. As a result, four prominent features out of eight were selected to be involved in the later supervised learning model. Finally, we created decision tree graph that

gave if-else condition to chain multiple variables that led to recall result. This served as guidance for manufacturers to make judgement call on recalling a vehicle.

Cross validation to maximize usage of data

In traditional machine learning models, we tend to split data into train, validation, and test set. The first set is used to train model; the second set is used to select model; the last set is used to simulate the real-work situation to evaluate how well the model will perform.

In our project, the split of data would result in inadequate training data. So instead of splitting in three sets, I used 80% of the data to train, 20% of data for testing. During training, the 80% of the training data was split into three folds for both training and validation. In this case, we reduced overfitting while maintained as much training ability as we could.

(Suggestion) Another way I want to explore with stakeholders but beyond the scope of this project, is to identify the possibility to collect more data, and build infrastructure to support this creation. If the model is deployed, this infrastructure will push live data to the model, so we do not need to manually download the data and/or retrain the model.

Define evaluation metrics

Metrics is an essential part in building successful machine learning model. A good metric should address stakeholders' interest and maximize model performance. For example, if the cost of fixing vehicle is more than implementing massive recall, then high precision is in favor, since it tries to reduce the false recalled vehicle. Conversely, if implementing massive recall is more expensive than fixing vehicle, we prefer high recall, where we correctly predict as many problematic vehicles as possible. In addition, similar to Kaggle competition, models are evaluated by metrics, so we know if it performs well equally on unseen data.

In this project, I selected ROC AUC, F1 Score, and Accuracy as my model metrics. These three metrics capture the popular classification performance standard. Specifically,

- Area Under ROC Curve

ROC curve plots the *true positive rate* against the *false positive rate*. AUC provides an aggregate measure of the performance across all classification thresholds under ROC curve.

- F1 Score

The F1 score can be interpreted as a weighted average of the *precision* and *recall*.

- Accuracy

The accuracy measures the *total number of samples correctly predicted* by our model in both recall and not recall group.

Use ensemble model

In this project, I started with 6 models, then narrowed down to 3 models after retrained and tuning. Each of these models performed well in one metric, but none of them performed equally well in all metrics. Therefore, I decided to assign weights into each model then averaged their probability prediction to get the final outcome. In this way, the final model was optimized and performed well in all metrics defined.

The End of The Report

Thank you for your time!