Insurance cross sale prediction

Brown University
Dongjun Shin

Github link : https://github.com/josh7197/midterm

1. Introduction

- Why cross-selling is important to companies

In general, insurance companies sell a variety of insurance policies, such as property insurance, health insurance, and car insurance. The companies want to sell additional products to current policyholders since companies think that selling extra products to the current customers is easier due to a pre-established relationship. We call marketing tactics for the current customers who have already purchased their products or services as cross selling.

Cross-selling is so critical to companies, including insurance companies. First, the companies can save marketing or promotion costs through the cross-selling because the companies do not need to spend customer acquisition costs. Second, cross-selling enables the company to fulfill efficient marketing. If the current customers have a positive impression of the company, they are more inclined to purchase an extra product of the company than others. Third, the companies are more interested in cross selling as the proportion of direct sales through the internet has soared compared to indirect sales through brokers or agents.

- The motivation and goal of the study

Nowadays companies use mixed marketing skills between cross-selling and target marketing. If companies spend their budgets intensively for current customers who are more likely to be interested in their diverse products, the possibility of successful cross-selling should climb.

The goal of this project is that the insurance company distinguishes current customers who are interested in a car insurance policy as an additional product in the current customer pool. All customers in the dataset already bought a housing policy of a certain insurance company. The consulting firm whose client is the insurance company did a survey on whether the policyholders are interested in car insurance. This dataset was offered to Kaggle by the consulting firm under an agreement of the insurance company.

- An analytical model and a prior research

The analytical model for the goal is a classification because the target variable is a dummy variable, “Response”, with 1 or 0 if an insurance holder is interested in a vehicle insurance is 1 or 0. The data set has 12 features, including the target variable and ID, and 381,109 rows(data points). However, ID does not explain whether someone is interested in the insurance plan. Therefore, I omit this variable in the model. Therefore, the dataset consists of 10 explanatory variables and a single target variable with 381,109 data points.

Table 1. Features and a target variable

Variable	Classification	Explanation
Gender	Categorical	Customer Gender(M:1, F:0)
Age	Numerical	Customer Age
Driving_License	Categorical	Having a DL has 1 or 0
Region_Code	Categorical	Customer region code
Previously_Insured	Categorical	Already having a car insurance has 1 or 0
Vehicle_Age	Categorical	Vehicle Age
Vehicle_Damage	Categorical	Damaged car has 1 or 0
Annual_Premium	Numerical	The annual insurance premium

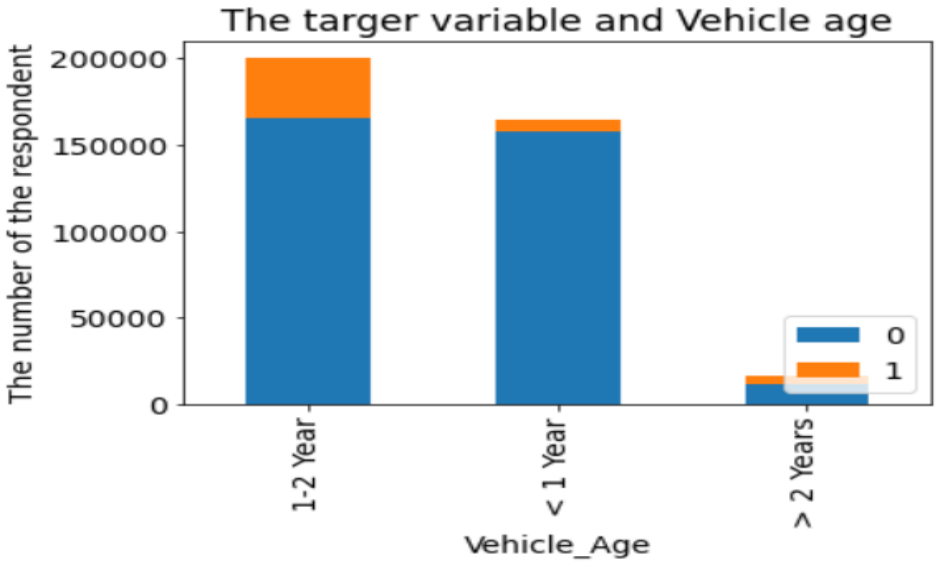
Policy_Sales_Channel	Categorical	Contact channel Code
Vintage	Numerical	Days when customers has been with the company
Response	Categorical	Being interested has 1 or 0

According to prior analysis, the target variable is unbalanced because most responses are negative(=0). The percentage of the negative responses of the dataset is 88%. Two logistic regression models were conducted for this classification problem. An unbalanced regression model is less accurate than a balanced regression model. For instance, the accuracy score of the balanced one is 92% while the figure of the unbalanced model is 84%.

2. EDA

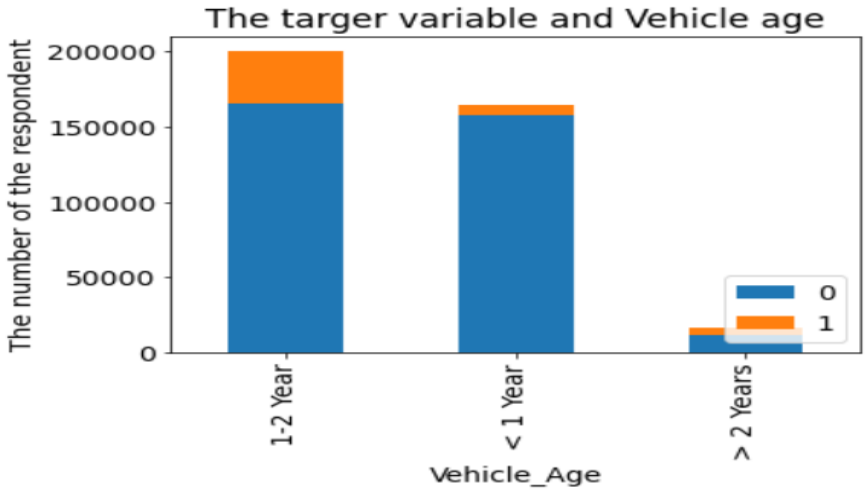
The most interesting finding is that the target variable, “Respondence”, is unbalanced because 88%(=334,399) of customers represent no interest in a car insurance plan. However, the imbalance is serious because the threshold for a process of imbalance is 5%. The second interesting point is that most customers are between 20 and 50 years old. However, the younger generation between 20 and 30 are not likely to be interested in the vehicle plan while the middle-aged generation between 40 and 50 are more likely to be interested in the plan. The third most remarkable result is that customers with vehicles aged 1~2 years are more interested in the plan than comparable groups, customers with vehicles less than 1 year or more than 2 years old.

Figure1. The target variable and vehicle age



In addition, the time that customers have contacted the company does not seem to affect the interest in the car insurance because boxplots between 0 and 1 are very similar when I conducted two variable analyses. Lastly, customers with damaged cars are more likely to be interested in the vehicle plan than ones who had undamaged cars because the majority of customers who responded they were interested in the insurance plan had damaged cars.

Figure2. The target variable and vehicle damage



3. Method

- **Check points before preprocessing**

  - IID : The dataset is IID because each row has its own ID and there is no same ID.

  - Imbalance : This dataset is a little imbalanced, but it is not serious because the ratio is 12%, higher than 5% where stratified splits are needed.

  - Missing value : There is no missing data

- **Splitting Strategy**

I selected a general splitting strategy with a train size of 0.2 and K-fold validation. First, this data is iid, which means that this data is not group data and time series data. Second, the imbalance of the data is not serious. Third, I used a 3-fold validation method although the dataset is large enough for a simple validation. That is because I would like to conduct a robust experiment despite a time-consuming process.

- **Preprocessing**

I used OneHotEncoder for "Gender", "Driving_License", “Region_code”, 'Policy_Sales_Channel', 'Previously_Insured', 'Vehicle_Damage', and 'Vehicle_Age”. First, the type of “Vehicle_age”, “ Gender”, and “Vehicle_Damage’ is an object, so we can regard these variables as non-ordinal categorical data. Second, although the type of “Driving_Licinese” and “Previously_insured” is an integer, it is a dummy variable, which means categorical data. Third, “Region code” and “Policy_Sales_Channel” have numerical figures, but these are used only for classification. There is no ordinal data in this dataset, I used only OneHotEncode for these categorical data.

Secondly, I used StandardScaler for “age”, “annual_Premium” , and “Vintage” because I am not sure that these have reasonable Max and Min and that these features might have a long tail structure. Finally, I have preprocessed 226 features, which excluded the target variable and ID.

- **Model selection**

I selected and conducted four different machine learning models : KNN classification, logistic regression with l2, Random forest, and XGboost classification. I tuned hyperparameters through grid research to find the optimal parameter set for each model and each random state. I repeated this process with 5 different random states. The parameters of each model are below.

Table2. model parameter

Model	Parameters
KNN	n_neighbors': 3,5,7,20,30 weights: uniform, distance
Logistic Regression	C : 0.01, 0.1, 10, 50, 100
Random Forest	Min_samples_split : 16, 32, 64, 128 Max_depth : 10,30,100,300
XGboost	Learning_rate : [0.03,0.05] model__max_depth : 5,10,50,30,50

I chose an accuracy score that evaluates a model’s performance because we need to evaluate true negatives and true positives at the same time. The goal of this model is to spend more budget on true positives and not spend any budget on people who are not interested in the car insurance policy. In addition, I calculated the means of scores for each model to compare the model's performances. I also yielded the standard deviation of scores in order to measure uncertainties due to splitting.

- **Reducing the number of data points**

Due to a large number of data points, more than 380,000, I conducted stratified random sampling with the same ratio of the target variables as the full data set and then implemented each model with these reduced data points. In fact, I tried to conduct the model with full data points, but my laptop did not work. In addition, I did this random sampling with different random states and compared outputs. The reason why I chose this method is the dataset would be iid.

4. Results

- **Comparison of test scores**

Table3. Comparison of test scores

Model	**Means**	**std**	**std from the baseline**
KNN	87.79%	0.009019476	0.045895229
Logistic Regression	87.76%	0.008146931	0.018636186
Random Forest	87.73%	0.007802294	-0.014136288
XG boost	87.76%	0.008146931	0.018636186

The baseline score is 87.74% because the rate of negative responses is 87.74%. If we predict all data to the negative, the score should be 87.74%. Among four models, KNN records the highest score, but the standard deviation is the highest score which means variation through each random state is larger than other three models. However, Ithink that KNN is the most predictive model because the standard deviation is so small despite the highest score.

- **Global feature importance**

I calculated the global feature importance through coefficients of logistic regression, MDI, permutations, and SHAP. I used an un-preprocessed test set for permutations while I used a preprocessed test set for the rest of three calculations. For the comparison, I introduced these three calculations. We can see that the global importances are very similar to each other.

Figure3. Global feature Importance through coefficients of logistic regression.

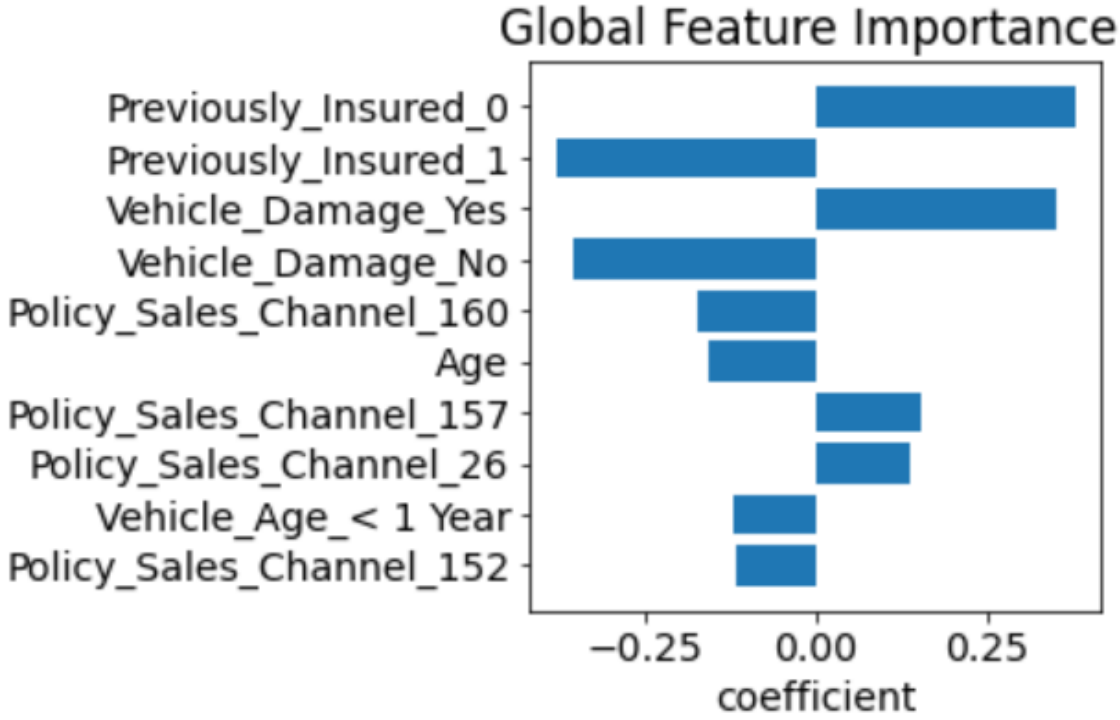


Figure4. Global feature Importance through coefficients of MDI

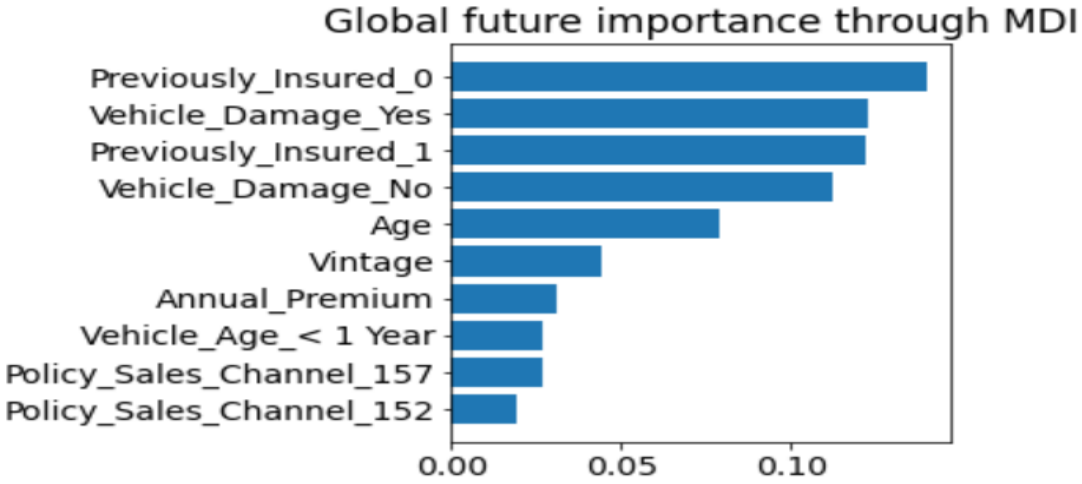
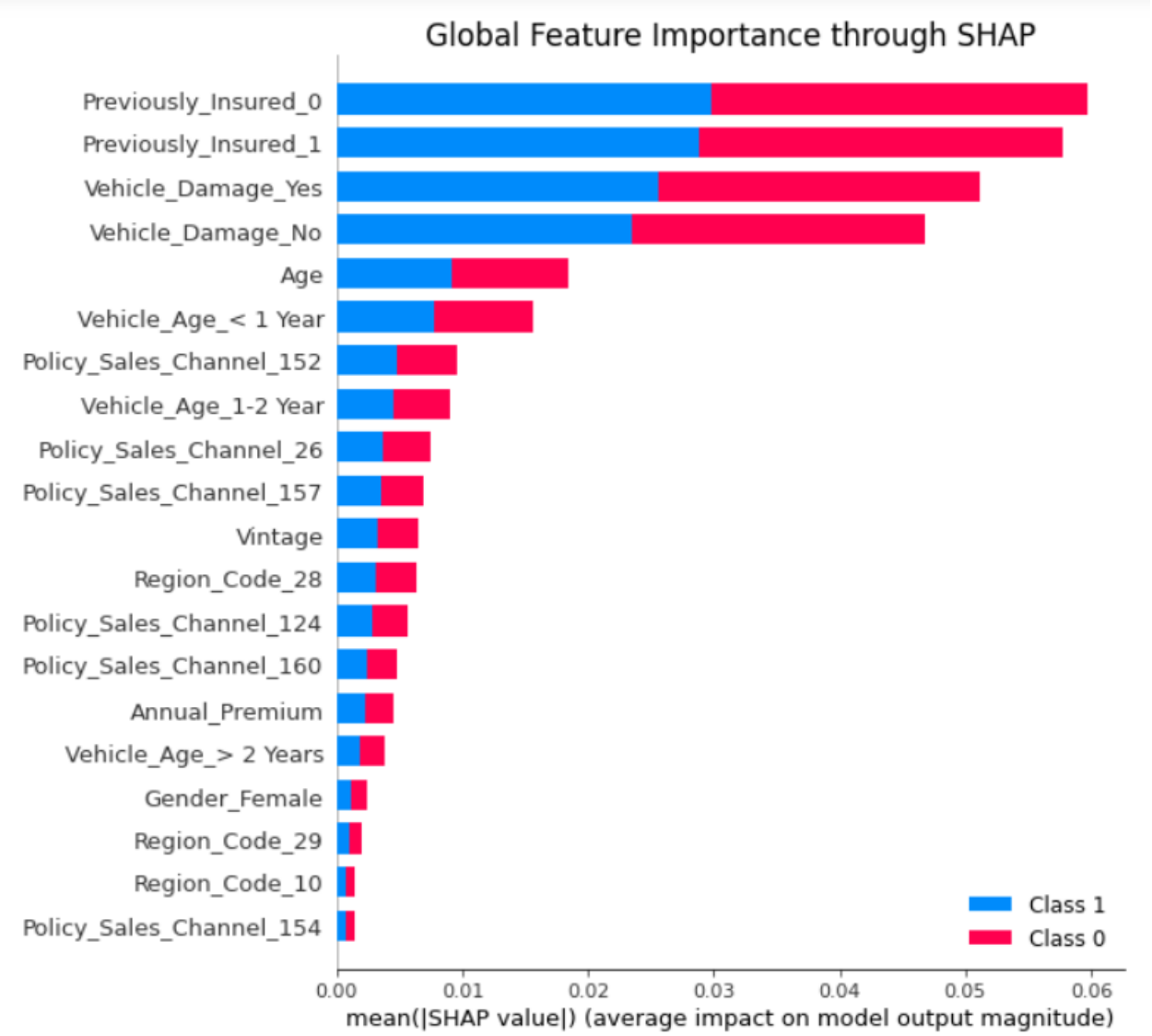


Figure5. Global feature Importance through coefficients of SHAP



- **Local feature importance**

Figure6. Local feature importance through SHAP

I chose the third row for the example. This row should be predicted to 1. “Vintage” variables make the possibility more close to 1 while “previously_insured_0” and

"vehicle_damage_Yes" push the probability to 0. But the value is still larger than 0.5, predicting this data point to 1.

- **Model interpretations**

  Overall, car damages, previously_insured, and age are the most important variables. Through EDA, those three variables are predicted to be more important than others because there are distinct differences between two binary variables. For instance, It is clear that people who have not bought car insurance are more interested in car insurance through EDA. In addition, "driver licence" is the least impactful variable except for "sales-channel" and "region-code". These two variables have a large number of categories. This result was also expected before the modeling because people who do not have feww driver license are interested in a car insurance in EDA.

  However, I do not expect that "gender" is not an influential feature through the result. Through EDA, there is a clear difference between female and male. But the result is different from my expectation.

5. Outlook

Due to the laptop, I had no choice but to reduce the data points, but I tried to sample the dataset points with different random states. Compared to my effort to improve the model accuracy, there are more weaknesses in my experiment. First, I can not use the whole data set, so I tried to use a clouding computer with better capability, but I failed to do so. Second, I want to make interaction variables because there seems to be an interaction between other variables. Third, using a Bayesian muti-level model could improve the accuracy. "sales-channel" and "region-code" have a large number of categories, but data points in certain categories are too small while the number of data points in certain categories are too large. If I control this characteristic, the prediction would be improved.

References

Anmol Kumar, “Predict Health Insurance Owners' who will be interested in Vehicle Insurance”, 2020, kaggle.com

https://www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction

PBRILLIANCE, “Imbalanced Insurance Data Classification“, 2020, kaggle.com
https://www.kaggle.com/pbrilliance/imbalanced-insurance-data-classification

shivan kumar, “JantaHack : Health Insurance Cross sell Prediction”, 2020, kaggle.com
https://www.kaggle.com/shivan118/crosssell-prediction