**Grade: 100%**

**Health Insurance and Vehicle Insurance Cross Sell Prediction**

**Group 5**

Fangzhou Yuan

Xiwei Shen

Zijia Liu

**Objective**

Insurance is a protection that insurance companies provide indemnity guarantees based on customers' losses in unforeseen accidents. In order to get the service, customers need to pay a calculated premium periodically. The dataset we are working on is from a health insurance provider and they would like to expand their business further into vehicle insurance.

Therefore, the purpose of this project is to predict if the customers are interested to purchase for the newly developed product, a vehicle insurance, from the company based on customers’ personal information and vehicle related data. In addition, we are interested in investigating the customer behavioral features which may lead to a higher probability of positive response to the vehicle insurance. From a higher perspective, this project can be beneficial to the company for a need of pursuing potential customers when developing new services and products.

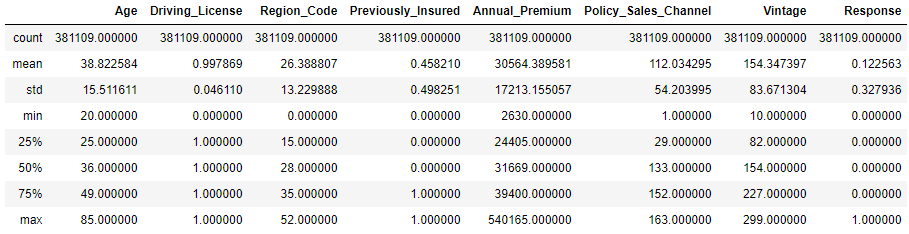
**Data Set Description**

The training dataset contains 381109 observations and 12 columns. There are six columns (“Customer ID”, “Age”, “Vintage”, “Annual Premium”, “Region Code” and “Policy Sales Channel”) are numerical variables, and the rest of the variables are categorical. The dataset shows basic vehicle information (“Vehicle Age”, and “Vehicle Damage”) and health insurance information, e.g. “Age”, “Region Code”, “Previously Insured”, etc., for each customer. According to the goal we have set, the “Response” column will be the target variable, and other columns are candidates for predicting variables. We may reduce the number of predictors depending on the result of feature selection. We discovered some interesting patterns from the correlation matrix. For example, vehicle age is positively related to customers’ degree of interest to vehicle insurance. More systematic predictions will be made after various models are implemented. The dataset can be retrieved from the link below:

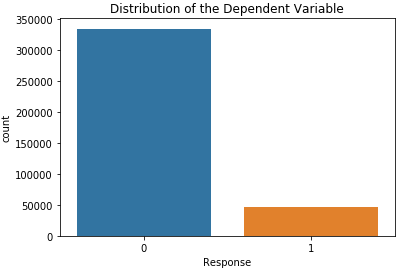
<https://www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction?select=train.csv> .

**Preliminary Data Exploration**

Looking at the summary statistics of the numeric columns in our training data, the first thing we noticed was the mean value of the target variable, “Response”. The mean value and the median of this binary distribution represents the “Response” variable is imbalanced. With the same logic, we can tell there are more data points as “not previously insured” compared to “insured” and most of the customers in the dataset have a driver's license. Besides, the range of the “Age” column is 20 to 85, which is in line with the age requirement of getting a driver’s license.

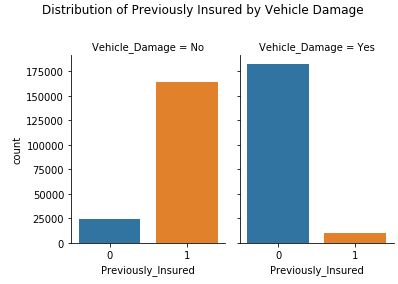


(figure.1 Summary Statistics)



The distribution of the dependent variable, “Response” is consistent with what we found in the summary statistics. The label column is imbalanced and most of the data points belong to the 0 class of “Not Interested”. Due to the biases of the data, we may employ oversampling techniques before building and tuning our models.

(figure.2 Barplot of Dependent Variable)



The distribution of the “Previously Insured” column by “Vehicle Damage” shows a negative correlation between these two variables. In other words, most customers who had previously insured did not have any damage to their vehicle. Extracting the collinearity among independent variables sets us a reminder when building regression models. Thus, we may perform PCA to reduce multicollinearity.

(figure.3 Barplot of Previously Insured by Vehicle Damage)

**Predictions**

The predictions we are trying to make for our project include:

1. What kind of vehicle information will lead the customer to be interested in vehicle insurance?

2. What kind of health insurance information will lead the customer to be interested in vehicle insurance?

Our group is planning to perform predictions through multiple models, such as SVM, Decision Tree, ANN, Logistic Regression, GBM, etc., so each member of the group will be able to take a sufficient workload.

**Inferences**

1. For customers of different gender or age groups (young / middle-aged / elder), is there any significant difference in their interest in vehicle insurance?

2. Which predictors in our dataset are the most important predictors for determining whether the customer is interested in vehicle insurance.

3. The median value of the “Annual\_Premium” variable in our dataset is 31669 and the first quartile is 24405. For customers who have a health insurance policy with an annual premium lower than 24405, how many of them would be interested in vehicle insurance?

4. Comparing the level of interest in vehicle insurance between customers who previously had a vehicle Insurance and customers who didn't have vehicle insurance before.

5. The median value of the “Vintage” variable in our dataset is 154 and the third quartile is 227. For customers who have been associated with the company longer than 227 days, are they more likely to be interested in the company’s vehicle insurance?

By exploring these inferences, we can get a general guideline for our modeling process. We would be able to know whether there are predictors that are strongly correlated with our target variable and whether there are predictors that are irrelevant to our target variable. For the irrelevant predictors, we may consider removing it during the feature selection process.

**NonSpark Packages**

1. Pandas: We are planning to use Pandas to transform the spark output to Pandas data frame, then we can properly annotate the plot.

2. Seaborn & Matplotlib: We are planning to use these two packages to generate visualization outputs during the process of exploratory data analysis, feature engineering, and model evaluation.

3. Imblearn: Since our dataset is unbalanced, we are planning to use RandomOverSampler() from the imblearn package to resample our data. The function will randomly select examples from the minority class, with replacement, and add them to the training dataset. This is one of the main approaches to deal with an imbalanced classification problem.