**Health Insurance anD Vehicle Insurance Cross Sell Prediction**

Group5

Fangzhou Yuan

Xiwei Shen

Zijia Liu

Table of Contents

[1. Abstract 2](#_Toc57583047)

[1.1 Predictions 2](#_Toc57583048)

[1.2 Inferences 2](#_Toc57583049)

[1.3 Brief conclusion summary 3](#_Toc57583050)

[2. Data Exploration and Visualization 3](#_Toc57583051)

[2.1 Dataset description 3](#_Toc57583052)

[2.2 Data Exploration and Visualizations 4](#_Toc57583053)

[3. Methodology 6](#_Toc57583054)

[4. Models 7](#_Toc57583055)

[4.1 PCA Visualization 8](#_Toc57583056)

[4.2 Logistic Regression 9](#_Toc57583057)

[4.3 Decision Tree 10](#_Toc57583058)

[4.4 Random Forest 11](#_Toc57583059)

[4.5 Gradient Boosted Trees 11](#_Toc57583060)

[5. Conclusion 12](#_Toc57583061)

[6. Appendix 15](#_Toc57583062)

[Reference List 15](#_Toc57583063)

[Additional Information 15](#_Toc57583064)

# Abstract

Insurance is a protection that insurance companies provide indemnity guarantees based on customers' losses in unforeseen accidents. In order to get the coverage, 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 in purchasing the newly developed product, vehicle insurance from the company based on customers’ personal information and vehicle-related data. We are also interested in investigating the customer behavioral features, which may lead to a higher probability of a positive response to vehicle insurance. From a higher perspective, this project can benefit the company for a need to pursue potential customers when developing new services and products. In this project, the health insurance and vehicle insurance cross sell prediction will be made based on five models: Decision Tree, Logistic Regression, Random Forest, and Gradient Boosting Machine.

## Predictions

The predictions will be made based on the models mentioned above. The prediction questions we are trying to solve in this 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?

## Inferences

The inferences will be mainly based on the distribution visualization, and variables correlations.

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.

1. 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.
2. 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?

## Brief conclusion summary

All of the proposed prediction and inference questions are solved with our data exploration and models.

For prediction Q1: Vehicle is damaged before but not insured

For prediction Q2: Middle aged and old customers

For inference Q1: Statistically, middle aged customers are more interested in vehicle insurance. There is no obvious distinction of the interest to vehicle insurance between female and male customers.

For inference Q2: “Previously Insured” and “Vehicle Damage” according to the models we built For inference Q3: There is no obvious increase or decrease in interest to vehicle insurance for loyal customers.

For inference Q4: If a customer has his/her vehicle insured before, the customer is not interested in the newly developed insurance at this company. (See Appendix Figure. 2)

# Data Exploration and Visualization

## Dataset description

The training dataset consists of 381109 observations and 12 columns with a mix of categorical and numeric variables. Specifically, there are six numeric variables, which include “Customer ID”, “Age”, “Vintage”, “Annual Premium”, “Region Code” and “Policy Sales Channel”, and six categorical variables, including “Gender”, “Vehicle Age”, “Vehicle Damage”, “Driving License”, “Previously Insured”, and “Response”. The dataset can be divided into two parts: vehicle information, and health insurance information. According to the goal we have set, the “Response” column will be the target variable, and other columns are candidates for predicting variables. The data set is as shown in Figure. 1 below.

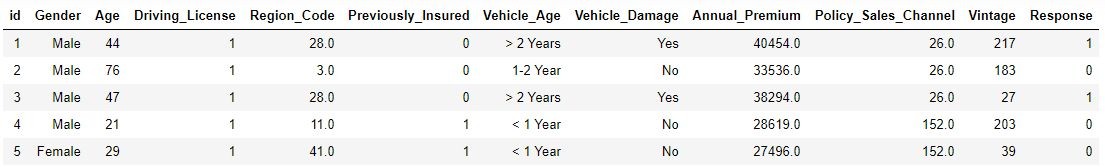


Figure. 1

## Data Exploration and Visualizations

After checking the summary statistics, there’s no NAs and outliers in this dataset. Excluding the However, according to the “Response” column, 334399 customers responded 0 for their interest to the vehicle insurance; while only 46710 customers would like to know more about the vehicle insurance, indicating that the positively labeled data points account for less than 13% of the total. Therefore, if models we build perform poorly on the original dataset, resampling techniques, such as undersampling and oversampling, will be employed to provide an unbiased dataset for the machine learning models to train on.

Our visualizations are created based on the inference questions and assumptions we proposed and the results from the correlation matrix. (see Appendix Figure 1)

**Age Group vs Response.** Customers of different ages may respond differently to auto insurance. Then it is worth exploring the distribution of response among the different age groups. Figure. 2 below shows a histogram of the age distribution. Since the distribution is right-skewed, we decide to divide customers into three groups by age. Figure. 3 demonstrates how customers in different age brackets react to the vehicle insurance. Although young people account for the largest proportion of the total count, only a small portion of young customers are interested in the vehicle insurance. In terms of proportion and counts, the middle age

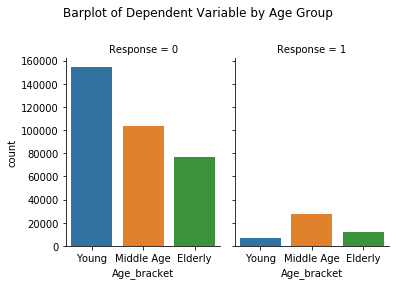
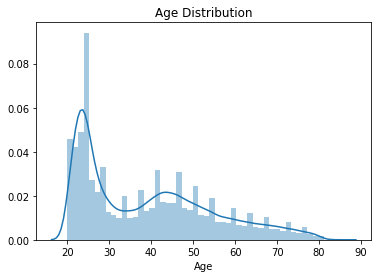


Figure. 2 Figure. 3

group, i.e. from 30 to 50 years old, should be prioritized for marketing the newly developed vehicle insurance.

**Vehicle Age vs Response.** As shown in Figure. 4, most customers’ vehicles are less than two

years old. Comparing the vehicle age distribution in two response groups, we found customers who have their vehicle between 1-2 years old are more interested in vehicle insurance, which can be explained by how vehicle insurance is purchased in real life. When a new car is purchased/leased, the customer must get the vehicle covered before driving away. Whereas loyal customers can usually get a lower rate as a reward for staying longer with an insurance provider. Therefore, 1-2 years could be an appropriate time frame to explore if there are premiums with lower rates or more coverage from other companies.

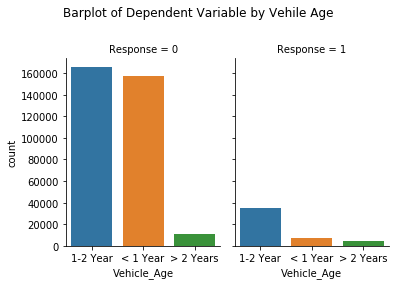
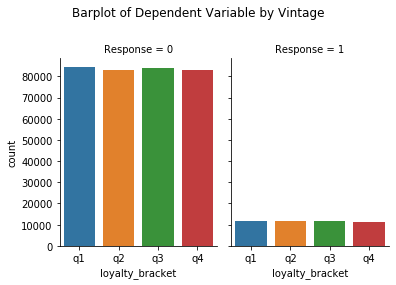


Figure. 4 Figure. 5

A picture containing icon

Description automatically generated**Vintage vs Response.** The “Vintage” variable is a measurement of the number of days a customer has been covered by the insurance provider, which represents the customer loyalty. The median value of the variable in our dataset is 154, and the third quartile is 227. Figure. 5 below shows the distribution of “Response” by “Vintage”. The number of days a customer associates with the company is not observably related to whether the customer is interested in car insurance according to the barplot, which is against our assumption. Hence, we do not expect to find the “Vintage” variable to be a strong predictor in the models we build.

Figure. 6

**Driving License vs Response.** Another assumption we made from common sense is customers without a driving license should not be the target for the vehicle insurance. As shown in Figure. 6, our assumption is consistent with the data. We conclude that the dataset is filtered out with the “Driving License” variable equals to “1” by the data source provider.

**Policy Sales Channel vs Response.** As the metadata provides, **“**the Policy Sales Channel column includes anonymized codes for the channel of outreaching to the customer, i.e. different agents, over mail, over phone, in person, etc.” Although the information stays confidential, we did observe that channel 150 has a dramatic difference in the customers’ response. So customers reached out by this method are less likely going to be interested in the vehicle insurance.

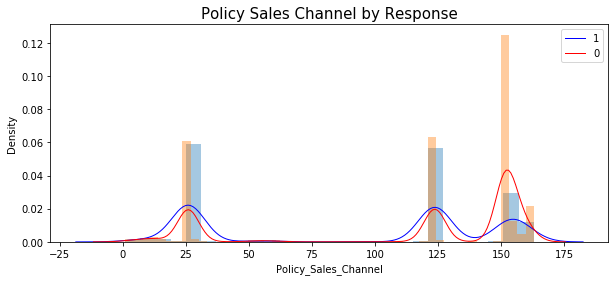
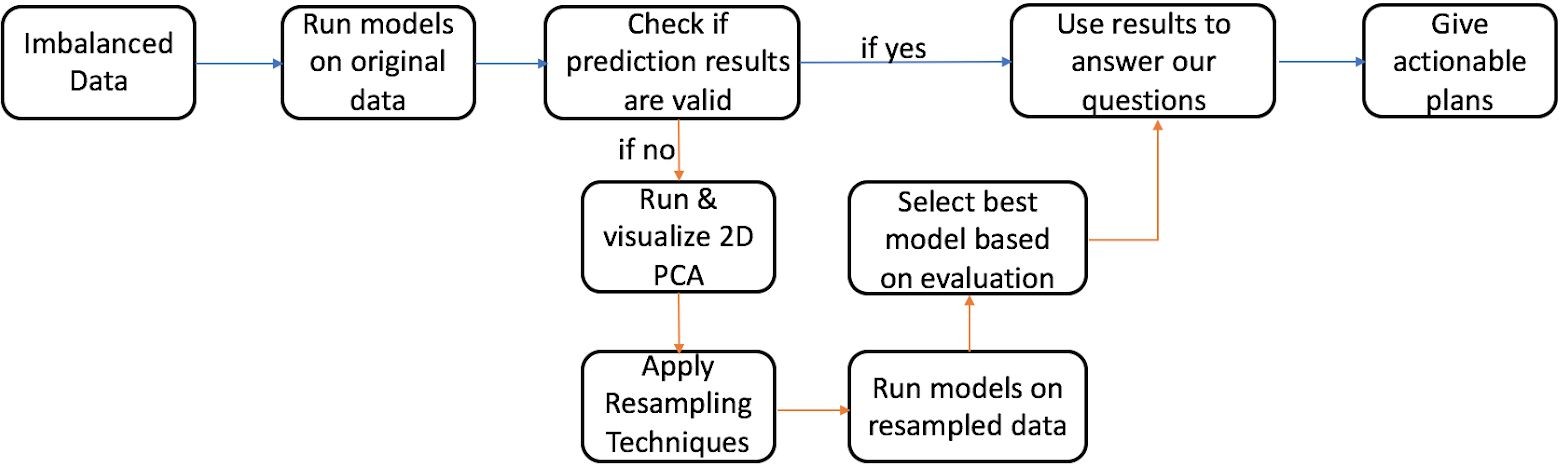


Figure. 7

# Methodology

As explained in the data visualizations section above, the class label “Response=1” only constitutes a very small proportion of the training dataset, which skewed classification may challenge the models’ performances. We would like to check whether the original dataset's performance is valid enough to make the predictions first. If the models perform poorly on the biased dataset, data resampling techniques will be applied to overcome the biasness. We plan to build four models to predict and solve our inference questions. Predicting features that lead to Response=1 and 0 will be given from the Decision Tree model. For Random Forest, Logistic Regression, and Gradient Boosting Machine, we will fine tune the hyperparameters to achieve optimum AUC score. Then we can extract feature importance to answer the second question in our inference questions list. Combining all the results from the best models, we can achieve the business goal of the project: identify suitable customers for promoting the vehicle insurance and giving actionable plans to the insurance provider.

# Models

As described in our Methodology section, we would start building our models on the original dataset to see if the imbalance property of the dataset affects model performance. Given by the distribution of our target variable, “Response”, the majority vote methods of predicting all

the cases as 0: not interested can yield an accuracy of approximately 88%. In Table. 1 shown below, none of the models we built exceed the accuracy of majority vote, which drove us to use resampling techniques to prepare a balanced dataset for training.

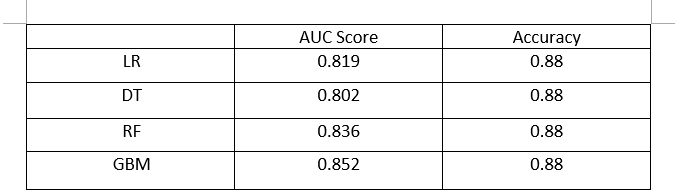


Table. 1

## 4.1 PCA Visualization

Before implementing resamples, we thought visualizing the scatter plot of PC1 and PC2 would help us better understand how the data points are distributed in two dimensions. As shown in Figure. 8 below, data points with positive response are colored in yellow, and negative ones are colored in purple. The two classes are not linearly separable in this scatter plot.

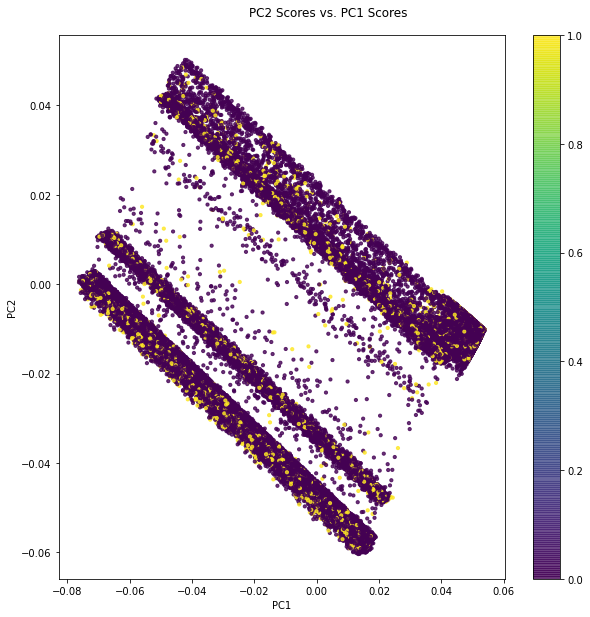


Figure. 8

As the noisiness represented in the scatter plot, we could not give a fair justification of which resampling technique would work the best on this dataset, so we employed three methods from the “Imbalanced-Learn” package: Over-sampling, Under-sampling, and Combination of Over & Under-sampling (SMOTEomek). We obtain a balanced dataset from each method, then run and fine tune our models to see which sampling method works the best by looking at AUC, Recall, and Accuracy.

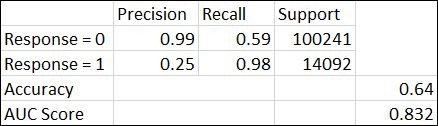
After building models and evaluating the performance, we decided to rely on Under-sampling for model training. We infer that the noise of the data affects the quality of Over-sampling.

Because we are creating new data points for the minority class, which do not have clear

distinctions from the majority class.

## 4.2 Logistic Regression

To make a binary classification, Logistic Regression is our first go-to option. As we have found in the PCA visualization, the data is not linearly separable. Thus, as shown in Table. 2, logistic regression does not give a very high accuracy and AUC score, but yields a recall of 0.98, which fits our goal of capturing customers who are interested in vehicle insurance. Top 3 features that the Logistic Regression uses for making predictions are “Previously Insured”, “Vehicle Damage”, and “Vehicle Age”. The importance is ranked by the coefficient of each predictor.



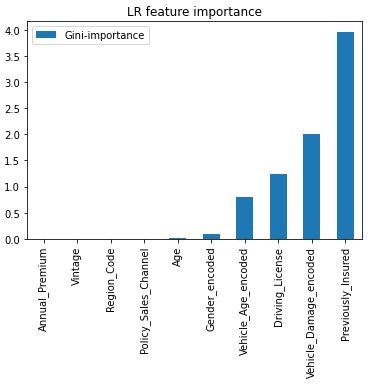
Table. 2

Figure. 9

## 4.3 Decision Tree

In order to get inferences and insights of the predicting variables, we have built a Decision Tree model. Our best model is generated with “maxDepth” and “maxBins” set to 10, which is too complicated for visualization. So Figure. 10 below is a post-pruned tree with less depth and width for explaining the inferences. The root node tells us that if the customer’s vehicle is damaged before, he/she is more likely going to be interested in the newly developed vehicle insurance, as the customer’s current premium can be high due to the negative driving history. As we found in the Data Visualization section, the “Policy Sales Channel” can be another important factor to look at. In addition, the tree can grow with “Customer Age” and “Vehicle Age”, older customers and vehicle models can also lead to an increase in the interest to our new service.



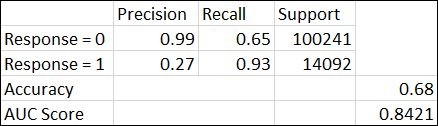
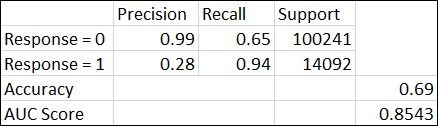
 Figure. 10

Table. 3

## 4.4 Random Forest

We expect to build ensemble learning algorithms to achieve optimal AUC score and accuracy for predicting the response column. The accuracy and AUC score of the Random Forest model is slightly higher than Decision Tree and Logistic Regression. But as for achieving maximum recall for “Response = 1”, Logistic Regression outperforms Random Forest.

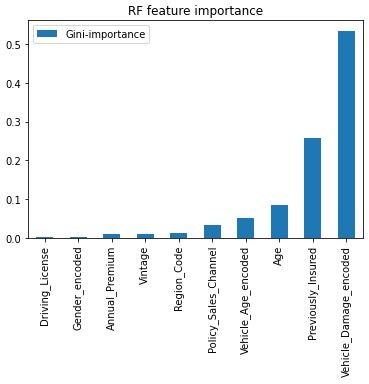
Table. 4

Figure.11

The most important three features for the Random Forest model are “Vehicle Damage”, “Previously Insured”, and “Age”.

## 4.5 Gradient Boosted Trees

As shown in the evaluation matrix, the GBT model predicts the validation dataset with a higher accuracy and AUC score than the Random Forest model. But similar to the Random Forest, the

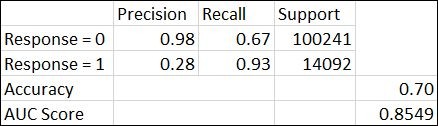
GBT model gives a recall of 0.93 for the positive label class. Speaking of feature importance, the top three variables are the same with Random Forest, except for the 1st and 2nd rank places are reversed.

Table. 5

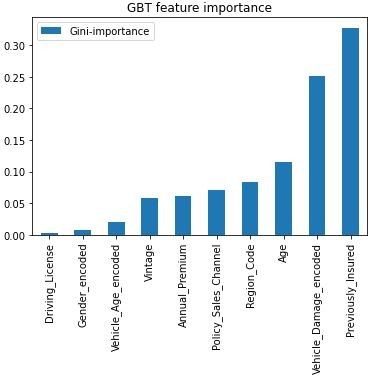


Figure. 12

# Conclusion

The prediction and inference questions that we listed in our proposal were all properly solved by our analysis work. We also want to deploy the XgBoost classifier on our dataset, but we found that PySpark does not support this model yet. I would say this is one of the limitations of our project. Another limitation would be that our dataset doesn't contain variables of health

insurance history, which is possibly because the data is confidential. It might not be a good option to apply our output to a real-world problem directly.

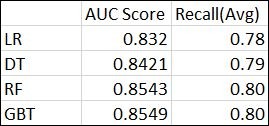


Table. 6

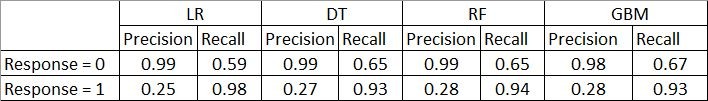
Table. 6 is a summary table of all of our models. After resampling our original dataset, there is a small increment in the AUC score for some of our models, but the difference is not significant in general. For each model, the average recall rises from 0.5 to about 0.8.

Table. 7

Based on Table. 7, we can see that there is a large increment in the recall of the positive label class. For data with the response variable equals to 1, all models now have a recall greater than

0.9. If we look at recall value in general, the Decision Tree, Random Forest, and GBT Classifier

have a very similar result. Logistic Regression has the highest recall for data with the response variable equals 1, and the lowest recall for data with the response variable equals 0. Our models are able to identify the large proportion of actual positive label class correctly, which is exactly our desired result. If the insurance company wants to locate the target group of people to send out advertisements based on the model prediction, then Logistic Regression is the best option since it has the highest recall for the positive response class. For other cases, Random Forest would be a better option, since it has a higher AUC score and high enough recall for both classes. For future works, we will try to increase the precision of data with the response variable equals to 1, and keep the recall at the same level. Right now, the precision scores we have for the positive label class are lower than 0.3. It means that only a small proportion of positive identifications was actually correct. Since we are using Under-sampling data for our optimal models, the number of data points we have is less than the original dataset. In this case, Cross-Validation with multiple folds might be a good option to increase model stability. We would like to do some experiments on Cross-validation and check if it will improve the precision for the positive label class.

# Appendix

## Reference List

Package “dtreeviz” for Decision Tree Visualization, <https://github.com/parrt/dtreeviz> Package “Imbalanced-Learn” for resampling balanced dataset,

<https://imbalanced-learn.org/stable/user_guide.html> Handling imbalanced datasets in machine learning, B. Rocca,

[https://towardsdatascience.com/handling-imbalanced-datasets-in-machine-learning-7a0e8422](https://towardsdatascience.com/handling-imbalanced-datasets-in-machine-learning-7a0e84220f28) [0f28](https://towardsdatascience.com/handling-imbalanced-datasets-in-machine-learning-7a0e84220f28), Last Accessed on Nov.29.2020.

## Additional Information

