MSBA7027 Machine Learning

Vehicle Insurance Customer Targeting

22 December 2021

Group 17

3035886269 CHEN Anqi

3035875430 LIU Ruolin

3035881477 WEI Yuhan

3035873626 ZHAN Mingqi

Executive Summary

In this project, we build a model to predict whether a customer from the past year would extend his or her vehicle insurance offered by this company. We figure out the most important features that drive the customer to extend the insurance, which are *Income*, *Renew Offer Type*, *Total Claim Amount* and *Customer Lifetime Value*. We recommend the insurance company to reach out the target customers with higher probability to extend the insurance and give different renew offer type to different customer.

1. Introduction, Problem Description & Relevance

Our dataset is about 9134 customers who have taken vehicle insurance in the past year. We found this dataset in Kaggle: https://www.kaggle.com/ranja7/vehicle-insurance-customer-data . Our objective is to predict whether a customer from the past year would extend his vehicle insurance by this company. Thus, we build a model to find those targeted customers and figure out the most important features that drive the purchase behavior. The company can benefit from reaching out to those targeted customers effectively and thereat increasing the conversion rate of marketing campaigns.

2. Data Description & Preliminary Data Analysis

2.1 Data Description

There are 24 variables and 9134 instances in total, and the categorical variable *Response* is the response variable. Due to the space limitation, only relatively important variables are listed.

Variable Name	Definition	Type
Response	"Yes" if the customers would like to renew their insurance. "No" if the customers would discontinue their insurance.	Character
Customer Lifetime Value	Value of customers insurance, equals to Customer Value * Customer lifespan	Numeric
Education	Background education of customers (High School or Below, Bachelor, College, Master or Doctor)	Character
Effective To Date	The first date when customer would like to activate their car insurance	Character
Employment Status	Customer employment status (Employed, Unemployed, Medical Leave, Disabled, or Retired)	Character
Income	Customers income	Integer
Renew Offer Type	Each sale offers 4 types of new insurances to customers (Offer 1, Offer 2, Offer 3 or Offer 4)	Character
Total Claim Amount	Number of Total Claim Amount customers did based on their coverage and other considerations.	Numeric

2.2 Preliminary Data Analysis

Figure 1 shows that the proportion between "*Response* = 1" and "*Response* = 0" is around 7:1, which indicates a category imbalance problem. Figure 2 demonstrates that customers who have relatively high *Total Claim Amount* tended not to renew their vehicle insurance. One preliminary guessing here is these customers are provided with more strict types of offers, which are not appealing. It we look at Figure 3, we can see that the performance of different offer types is different. It illustrates that Offer2 is more attractive to existing customers, while Offer1 is somehow undesirable.

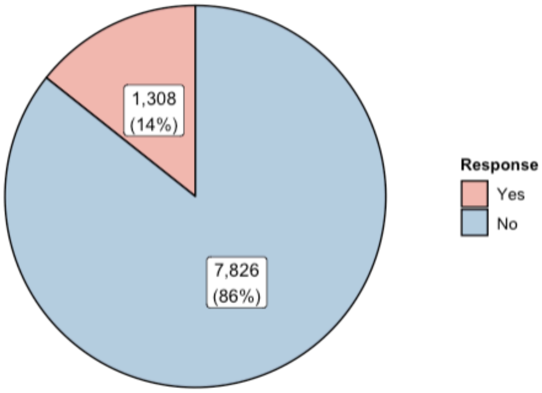


Figure 1 The Constitution of Response

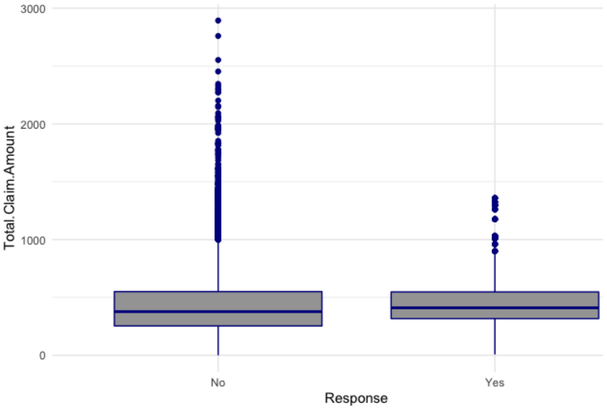


Figure 2 The Box Plot of Total Claim Amount

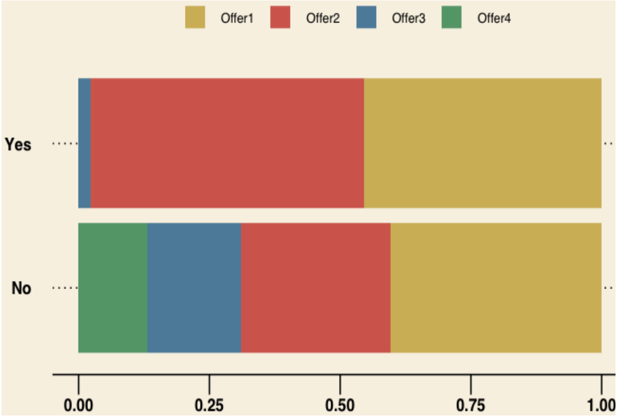


Figure 3 The performance of Offers

3. ML Methods, Analysis & Comparison

3.1 Data Preprocessing

1) **Deal with missingness:** Check to make sure there is no missing value.
2) **Feature filtering:** Delete non-informative variable Customer ID, and check to make sure there is no near-zero variance variable.
3) **Categorical variable engineering:**

- Considering computation efficiency, lump all the dates in a month together, i.e., transform *Effective.To.Date* to *Effective.To.Month*. Then, set threshold 0.01 to check if there are other categorical variables needed to lump.
- Divide the variable into nominal and ordinal variables, and then do dummy encoding and label encoding.

4) **Data splitting:** Divide the dataset into 70% training data and 30% testing data by using stratified sampling.
5) **Normalizing and Standardization**

3.2 Machine Learning Methods

We select the following nine models to compare since our case is a supervised learning and classification problem. Our metrics are accuracy, run time and interpretability. The models are listed according to descending accuracy order.

	Hyperparameter	Accuracy	Run time	Interpretability
Random Forest	num.trees = 220, mtry = 7, min.node.size = 1,replace = FALSE, sample.fraction = 0.8	0.9949	short	medium
Basic GBM	n.trees = 3000, shrinkage = 0.3, interaction.depth = 3, n.minobsinnode = 5	0.9898	long	medium
SVM - radial	cost = 5, gamma = 0.5	0.9376	long	low
KNN	K = 5	0.9143	short	medium
SVM - linear	cost = 1	0.8723	medium	low
LDA	NA	0.8719	short	medium
Decision Tree	NA	0.8709	short	high
Logistic Regression	NA	0.8709	short	high
Basic GBM	n.trees = 3000, shrinkage = 0.3, interaction.depth = 3, n.minobsinnode = 5	0.9898	long	medium
Logistic Regression	NA	0.8709	short	high
SVM - polynomial	cost = 10, degree = 3	0.8639	medium	low

We found that the most important variables of the four best models are different.

	4 most permutation-based important variables			
Random Forest	Income	Renew.Offer.Type	Total.Claim.Amount	Customer.Lifetime.Value
Basic GBM	Customer.Lifetime.Value	Income	Effective.To.Month	EmploymentStatus
SVM - radial	Effective.To.Month	EmploymentStatus	Renew.Offer.Type	Education

We find this strange because we get very similar predictions using different features. It’s a Rashomon Effect. We think it is caused by the different mechanisms of the machine learning methods and maybe high consistency is not desirable here because the models do not rely on similar relationships.

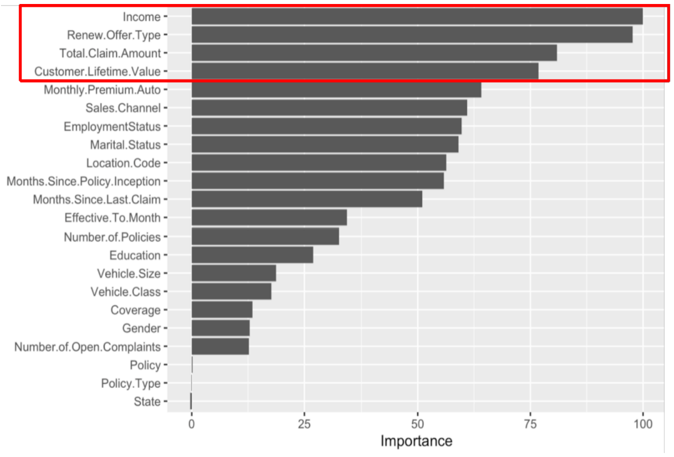
We also quantify the strength of two-way interaction effects between some features using vint().

	Interaction
Income * Education	0.044
Income * EmploymentStatus	0.013

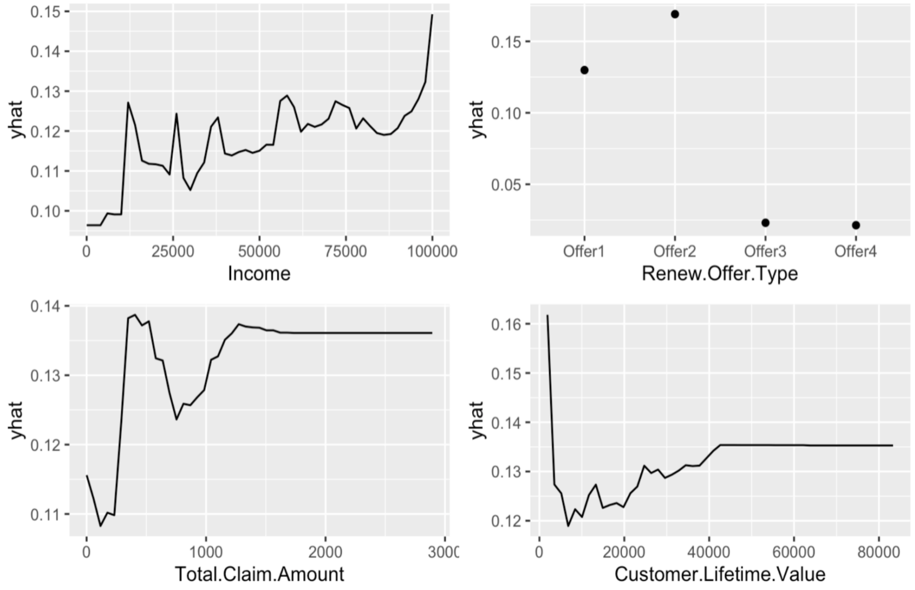
Finally, we choose the random forest model because it has both the high accuracy and high interpretability. What's more, we can also solve the class imbalance problem by adjusting the threshold of random forest in further exploration.

4. Conclusion

By analyzing the results of the Random Forest model, we select the four most important features, i.e., *Income*, *Renew.Offer.Type*, *Total.Claim.Amount* and *Customer.Lifetime.Value*.



We make the partial dependence plots and try to give some reasonable explanations and suggestions to the results. In these partial dependence plots, the horizontal axis are the features, and the vertical axis is the probability of renewal of contracts.



The first feature is *Income*. With the increase of income, renewal rate shows an upward trend, rising from 0.1 to 0.15. This is reasonable in real life, people with lower income are more likely to give up continuing the contract due to some price related reasons, such as new-customer-discount from other insurance companies. In order to prevent this, the company should give some discount according to the customer's income.

The second feature is *Renew.offer.Type*. Which means the type of offer a customer received this year, and it is according to the compensation or other situations in the past. Similar to what we find in EDA, the renewal rate of Offer1 and Offer2 is normal, but almost all the customers who got offer 3 or offer 4 gave up. By further analysis of data, we find the company didn't provide Offer3 and Offer4 to some particular group of customers. So, these two offers are not attractive to customers, they need to be redesigned or give up.

The third feature is *Total.Claim.Amount*, which is the claim amount of the customer in the past year. When the claim amount is lower than some threshold, the renewal rate increases, as people receiving higher claims tend to think the offer is useful. But when the claim amount exceeds the threshold, insurance will increase the premium, this could make the renewal rate no longer increase.

The fourth feature is *Customer.Lifetime.Value(CLV)*. For those new customers who have low CLV, insurance companies tend to give some discount, so renewal rate is high. However, after few years, at some certain stage the renewal rate drops rapidly and reaches the minimum value. After that, the renewal rate increases gradually. Therefore, it is necessary to provide appropriate continuity policies for customers to avoid customer defection.

By using different models to do classification, we find the Random Forest Model not only gives the most accurate prediction, but also very efficiently. We can judge the renewal intention of the customer with an accuracy of 99.5%. In addition, we also give some suggestions where the insurance company needs to improve.

Reference

Sarkar, R. (2019, May 31). *Vehicle_insurance_customer_data*. Kaggle. Retrieved December 22, 2021, from https://www.kaggle.com/ranja7/vehicle-insurance-customer-data

How to calculate customer lifetime value in 2021: Qualtrics. Qualtrics AU. (2021, October 12). Retrieved December 22, 2021, from https://www.qualtrics.com/au/experience-management/customer/how-to-calculate-customer-lifetime-value/

What is customer lifetime value (CLV) and how do you measure it?: Qualtrics. Qualtrics AU. (2021, October 12). Retrieved December 22, 2021, from https://www.qualtrics.com/au/experience-management/customer/customer-lifetime-value/

Molnar, C. (2021, December 19). *Interpretable machine learning*. Christoph Molnar. Retrieved December 22, 2021, from https://christophm.github.io/interpretable-ml-book/