Insurance Cross Sell Prediction

Ruofan Chen

Contents

Abstract

Cross-selling is the action or practice of selling an additional product or service to an existing customer. The objective of cross-selling can be either to increase the income derived from the client or to protect the relationship with the client or clients. This project uses data provided by JantaHack to discover the problem if the customer who already purchased health insurance from one company, would be interested in the same company's vehicle insurance, the logistic regression model with regularization and support vector machine (SVM) are used. After clarifying that the distribution of respond variable is unbalanced, two methods are utilized in these two models, including downsampling technique and weighted loss function. The data preprocessing process includes checking for missing values and selecting variables related to the response, and converting some categorical variables. Data description is carried out, related summary tables and plots are generated. To find the optimal penalized parameters, a 5-fold cross-validation grid search is performed on specific intervals. Logistic regression model and support vector machine model are established by using weighted loss function or downsampling method. Finally, the weighted logistic regression model with the highest area under the ROC curve (AUC) as 0.84476 is selected as the best model.

Key Words: GLM, logistic regression, support vector machine (SVM), regularization, down sampling, weighted loss function, area under ROC curve (AUC)

Section 1. Introduction

Cross-selling identifies products or services that satisfy additional, complementary needs that are unfulfilled by the original product that a customer possesses. Oftentimes, cross-selling points users to products they would have purchased anyway; by showing them at the right time, a store ensures they make the sale. For the insurance industry, an insurance company usually provides coverage and policies for various products, such as health insurance, vehicle insurance, and residence insurance. Establishing a model to predict whether a customer would be interested in Vehicle Insurance is extremely helpful for the company because it can then accordingly plan its communication strategy to reach out to those customers and optimize its business model and revenue.

This project is a Kaggle competition conducted by a company, and some participants released their methods on the Leaderboard. This project uses different methods and models from them, combined with different processing methods for unbalanced data, which is not common in R problem solving (some methods use only one remedy, such as resampling technique).

The project aims to conduct research: predict whether previous health insurance policyholders are interested in purchasing vehicle insurance. This project uses the logistic regression model with regularization and SVM model. In the following sections of the report, the most important characteristics of the dataset will be discussed, how to develop the model, and some final results of the data will be used as conclusions. In section two, data description develops some basic information of the dataset. The third section, Models and Methods, discusses the final chosen model and its performance. The summary and the concluding remarks illustrate the final results and some comments about every model. The Appendix contains the definition of variables, plots, and model-related model supplements.

Section 2. Data Description

The dataset used in the project is a public dataset on Kaggle, see https://www.kaggle.com/shivan118/crosssell-prediction. There are 381,109 observations, including 12 variables and no missing values. For this research, the variable Response is set as the dependent variable. Before determining to set which variables as predictors, examination and description of the data are necessary.

Table 1 shows the first six observations. For the definition of the variables, please see Appendix A.2.

Table 1: First Six Observations

id	Gender	Age	DL	RC	PI	VA	VD	AP	PSC	Vintage	Response
1	Male	44	1	28	0	> 2 Years	Yes	40454	26	217	1
2	Male	76	1	3	0	1-2 Year	No	33536	26	183	0
3	Male	47	1	28	0	> 2 Years	Yes	38294	26	27	1
4	Male	21	1	11	1	< 1 Year	No	28619	152	203	0
5	Female	29	1	41	1	< 1 Year	No	27496	152	39	0
6	Female	24	1	33	0	< 1 Year	Yes	2630	160	176	0

After examining all variables, it shows that the variable 'id' has nothing to do with 'Response'. The variable 'RC' represents the code of the customer's area, with a total of 53 levels. Variable 'PSC' indicates how the company can outreach the customer and has 155 levels. Levels with a few observations will be removed. All observations that do not belong to the highest 6 frequency levels will be placed on a new level, named '200'. The purpose of this dimensionality reduction is to make the design matrix smaller, which can save a lot of time-consuming the algorithm. Another transformation is about the Response variable: set indicator 0 to 'No' and 1 to 'Yes'.

Output 1 is a summary of these variables. The Response variable distributes unbalanced. Some variable descriptive analyses are provided in Appendix A.3.

Output 1 Summary of Variables

```
     Age              AP            Vintage       PSC           VD
Min.   :20.00   Min.   :  2630   Min.   : 10.0   26 : 79700   No :188696
1st Qu.:25.00   1st Qu.: 24405   1st Qu.: 82.0   122:  9930   Yes:192413
Median :36.00   Median : 31669   Median :154.0   124: 73995
Mean   :38.82   Mean   : 30564   Mean   :154.3   152:134784
3rd Qu.:49.00   3rd Qu.: 39400   3rd Qu.:227.0   156: 10661
Max.   :85.00   Max.   :540165   Max.   :299.0   160: 21779
                                                 200: 50260
        VA         PI           RC         DL           Gender
< 1 Year :164786   0:206481   8  : 33877   0:   812   Female:175020
> 2 Years: 16007   1:174628   15 : 13308   1:380297   Male  :206089
1-2 Year :200316              28 :106415
                              30 : 12191
                              41 : 18263
                              46 : 19749
                              200:177306
 Response
 No :334399
 Yes: 46710
```

Section 3. Models and Methods

3.1 Downsampling Method

DownSample will randomly sample a data set so that the frequency of the majority class is same with the frequency of the minority class. Due to the large sample size, down-sampling is appropriate when considering computer calculation time.

3.2 Logistic Regression with Regularization

Formula 1 is the logistic regression model. Here, 'Response' is set as the response variable, which is labeled Y, and follows the Bernoulli distribution with parameter p. In addition, each observation is independent. The notation p is a binomial parameter representing the probability of occurrence of the 'Response'.

The hypothesis is probability p follows a logistic distribution. β_0 represents the intercept, β_j (j from 1 to 10) represents the partial coefficient of increasing 1 unit on X_j while holding all other predictors fixed, the change of log odds.

$$logit(p_i) = log(\frac{p_i}{1-p_i}) = x_i'\beta = \beta_0 + \beta_1 X_{Age} + \beta_2 X_{AP} + ... + \beta_{10} X_{DL} \tag{1}$$

The objective function for the penalized logistic regression uses the negative binomial log-likelihood and is shown in Formula 2. The elastic-net penalty is controlled by α, and bridges the gap between lasso ($\alpha = 1$, the default) and ridge ($\alpha = 0$). The tuning parameter λ controls the overall strength of the penalty.

$$min_{(\beta_0,\beta)\in R^{p+1}} -[\frac{1}{N}\Sigma_{i=1}^{N} y_i(\beta_0 + x_i^T\beta) - log(1 + e^{(\beta_0 + x_i^T\beta)})] + \lambda[(1-\alpha)\left\|\beta\right\|_2^2/2 + \alpha\left\|\beta\right\|_1] \tag{2}$$

3.3 Support Vector Machine with Gaussian Kernel

Formula 3 is SVM with kernel. The kernel is a function to transform features from the Euclidean space to Hilbert Space. The Gaussian RBF kernel $k(x, y) = exp(-\frac{1}{2\sigma^2}\left\|x - y\right\|^2)$ is used in this part since Figure 1 shows it is a non-linear SVM.

Dual problem:

$$maxL_D(\alpha_i) = \Sigma_i\alpha_i - \frac{1}{2}\Sigma_{i,j}\alpha_i\alpha_j y_i y_j K(x_i, x_j) s.t. \Sigma_i\alpha_i y_i = 0, 0 \le \alpha_i \le C \tag{3}$$

3.4 K-fold Cross-Validation

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. The k-fold cross-validation method evaluates the model performance on a different subset of the training data and then calculates the average prediction score. Here, this method is used to find the optimized penalized parameters by picking the penalized parameters who has the highest AUC performance.

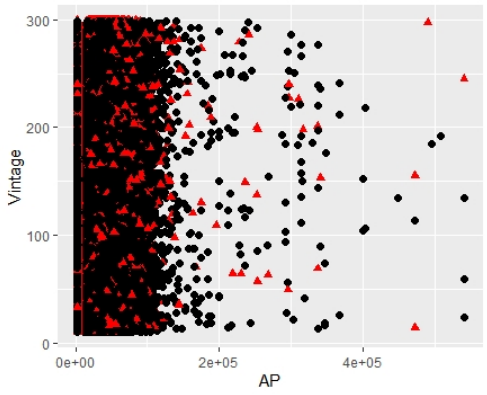

Figure 1: Vintage and AP

Section 4. Method Implement and Model Analysis

4.1 Logistirc Regression with Downsampling Method

After downsampling the dataset, a new dataset with the same number of two different levels is generated. At the same time choose two types of regularization terms L1 norm and L2 norm, 5-fold cross-validation, and grid search from the interval (0, 0.001) to find the optimized penalty parameters. The coefficients of the model are in Table 2. It gives information indicating that age, year, and policy sales channel are negatively correlated with Response. By keeping all other predictors in same level, when Age increases by 1, the odds of Response occurrence increases by $exp(-2.70 * 10^{-2}) = 0.9733612$ times.

Table 3 shows the optimized parameter and model performance. Logistic regression model with regularization uses downsampling technique has an AUC of 0.8434906. The importance of the variables in Table 4 shows that the top three important variables in the model are: previous insured with the answer 'Yes', policy sale channel with the level of 160, and vehicle damage with the answer 'Yes'.

Table 2: Coefficient of DownSampled Logistic Regression Model

Variables	Coefficient	Variables.1	Coefficient.1
(Intercept)	-1.40e+00	VA>2Years	0.6060
Age	-2.70e-02	VA1-2Year	0.4530
AP	7.00e-07	PI1	-3.8400
Vintage	-9.33e-05	RC15	-0.0988
PSC122	-2.89e-01	RC28	0.2030
PSC124	-1.48e-01	RC30	0.3640
PSC152	-1.17e+00	RC41	0.3670
PSC156	-2.55e-01	RC46	0.0656
PSC160	-2.16e+00	RC200	0.0729
PSC200	-2.18e-01	DL1	1.1800
VDYes	2.00e+00	GenderMale	0.1010

Table 3: Performance of DownSampled Logistic Regression Model

Alpha	Lambda	ROC	Sens	Spec	ROCSD	SensSD	SpecSD
1	0.000303	0.8434906	0.6326054	0.9508671	0.0017126	0.0035667	0.0014579

Table 4: Variable Importance of DownSampled Logistic Regression Model

Variables	Total	Variables.1	Total.1
PI1	100.00000	VA>2Years	15.797085
PSC160	56.33730	VA1-2Year	11.798251
VDYes	52.12801	RC41	9.573333
DL1	30.66595	RC30	9.501387
PSC152	30.51902	PSC122	7.527867

4.2 Logistirc Regression with Weighted Loss Function

Formula 4 is to calculate the new weights for the minority class and the majority class of 'Yes' and 'No'. With default weights, the classifier here will assume that both kinds of label errors have the same cost. But for this unbalanced dataset, the wrong prediction of the minority is worse than the wrong prediction of the majority class. Use the entire dataset to construct a logistic regression model with regularization, along with weighted loss function, a grid search is performed on the interval (0,0.001), and it is found that the optimized penalized parameter lambda is 0.0003232323 when the L1 norm regularization is selected. The coefficients of this model are in Table 5. Table 6 shows the optimized parameter and model performance. It has an AUC as 0.8447601. The importance of variables in Table 7 shows the important variables in the model are basically the same as the down-sampled logistic regression model.

$$weights = \begin{cases} \frac{1}{\text{number of majority class}} * 0.5 \text{ for majority class} \\ \frac{1}{\text{number of minority class}} * 0.5 \text{ for minority class} \end{cases} \tag{4}$$

Table 5: Coefficient of Weighted Loss Logistic Regression Model

Variables	Coefficient	Variables.1	Coefficient.1
(Intercept)	-1.29e+00	VD>2Year	0.6180
Age	-2.72e-02	VD1-2Year	0.4200
Annual_Premium	9.00e-07	PI1	-3.8400
Vintage	-2.13e-05	RC15	-0.1010
PSC122	-2.62e-01	RC28	0.2240
PSCl124	-1.59e-01	RC30	0.3430
PSC152	-1.22e+00	RC41	0.4240
PSC156	-2.42e-01	RC46	0.0880
PSC160	-2.17e+00	RC200	0.0785
PSC200	-2.08e-01	DL1	1.1100
VDYes	1.98e+00	GenderMale	0.0851

Table 6: Performance of Weighted Loss Logistic Regression Model

Alpha	Lambda	ROC	Sens	Spec	ROCSD	SensSD	SpecSD
1	0.0003232	0.8447601	0.636826	0.9479769	0.0005215	0.0040551	0.0032068

Table 7: Variable Importance of Weighted Loss Logistic Regression Model

Variables	Total	Variables.1	Total.1
PI1	100.00000	VA>2Years	16.083772
PSC160	56.40790	RC41	11.043422
VDYes	51.40938	VA1-2Year	10.924762
PSC152	31.65903	RC30	8.917569
DL1	28.96431	PSC122	6.824376

4.3 Support Vector Machine with Downsampling Method

For a large datasets, SVM runs slower than the logistic regression. To deal with the time complexity, 1% down-sampled data is selected to find the penalized parameters C and sigma. From the interval C = 2 ^ (-5:10), sigma = 2 ^ (-10: 3), a grid search is performed to find the optimized penalty parameter that makes the AUC the highest. As a result C=1 and sigma=0.001953125, the AUC is the highest, which is 0.7739307. After a train-test split with 0.7:0.3 in the downsampled dataset, the AUC of the test set is equal to 0.632.

4.4 Support Vector Machine with Weighted Loss Function

A train-test split at a ratio of 0.7:0.3 in 20% of the entire dataset, and use the previous results C=1 and sigma=0.001953125 to construct a model with a weighted loss function. Prediction on the test set gives an AUC equaling to 0.620. Figure 2 is the ROC curve.

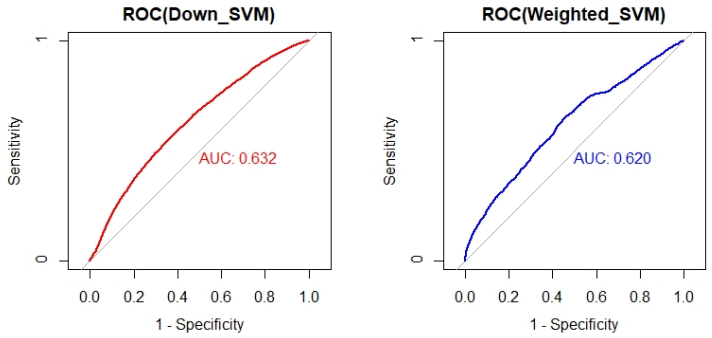

Figure 2: SVM ROC curve

Section 5. Model Evaulation and Concluding Remarks

Two models are used, namely logistic regression with regularization and SVM, along with two different methods (downsampling and weighted loss function) for processing imbalanced dataset. To evaluate the model properly, AUC is adopted. The implicit goal of AUC is to deal with the highly skewed distribution of the dataset, and not to overfit a single class.

Table 8 summarizes these four combinations. Logistic regression with weighted loss function has the highest AUC of 0.8448. Moreover, the performances of the two SVMs are not as good as the logistic regression model. Considering the time required to construct the model, it is concluded that SVM is not an optimal model for large sample size data.

Table 8: AUC Summary

Model	AUC
Downsample Logistic regression	0.8435
Weighted Loss Logistic Regression	0.8448
Downsample Support Vector Machine	0.6320
Weighted Loss Support Vector Machine	0.6200

In the future data preprocessing process, one option is to eliminate outliers. In a logistic regression model with regularization, it is difficult to perform a goodness-of-fit test using the current package. However, some papers are discussing these tests. Interaction terms and quadratic terms or higher-order terms may be considered in the model. In addition, CART model may be a better choice for large sample size data.

Appendix

A.1. References

1. Chao-Ying Joanne Peng, Kuk Lida Lee Gary M. Ingersoll (2002). An Introduction to Logistic Regression Analysis and Reporting Article. The Journal of Educational Research. September 2002 (3-14)
2. Roger Koenker, Jungmo Yoon. Parametric Links for Binary Choice Models: A Fisherian-Bayesian Colloquy. http://www.econ.uiuc.edu/~roger/research/links/links.pdf
3. Yihui Xie, Christophe Dervieux, Emily Riederer. R Markdown Cookbook
4. R. Berwick. An Idiot's guide to Support vector machines (SVMs). https://web.mit.edu/6.034/wwwbob/svm.pdf
5. Trevor Hastie, Junyang Qian. An Introduction to glmnet. https://cloud.r-project.org/web/packages/glmnet/vignettes/glmnet.pdf

A.2. Variable Definitions

Table 9: Variable Definition

Variables	Definition
Id	Unique ID for the customer
Gender	Gender of the customer
Age	Age of the customer
DL	0 (Customer does not have DL) 1 (Customer already has DL)
RC	Unique code for the region of the customer
PI	1 (Customer already has Vehicle Insurance) ,0 (Customer doesn't have Vehicle Insurance)
VA	Age of the Vehicle
VD	1 (Customer got his/her vehicle damaged in the past), 0 (Customer didn't get his/her vehicle damaged in the past)
AP	The amount customer needs to pay as premium in the year
PSC	Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc.
Vintage	Number of Days, Customer has been associated with the company
Response	1 (Customer is interested), 0 (Customer is not interested)

A.3. Basic Summary Statistics

In addition to summary tables, graphical methods can also be done very well, which can display data more intuitively. Select some variables for analysis. Figure 3 summarizes these variables. Due to limitation of picture size, variables 'PSC' and 'RC' are not included. Most importantly, the response variable is highly unbalanced. Therefore, for an unbalanced datasets, necessary data processing should be considered. The variable 'Age' is skewed to the right. The 'Vintage' boxplot does not show any outliers. Annual Premium seems to has some outliers.

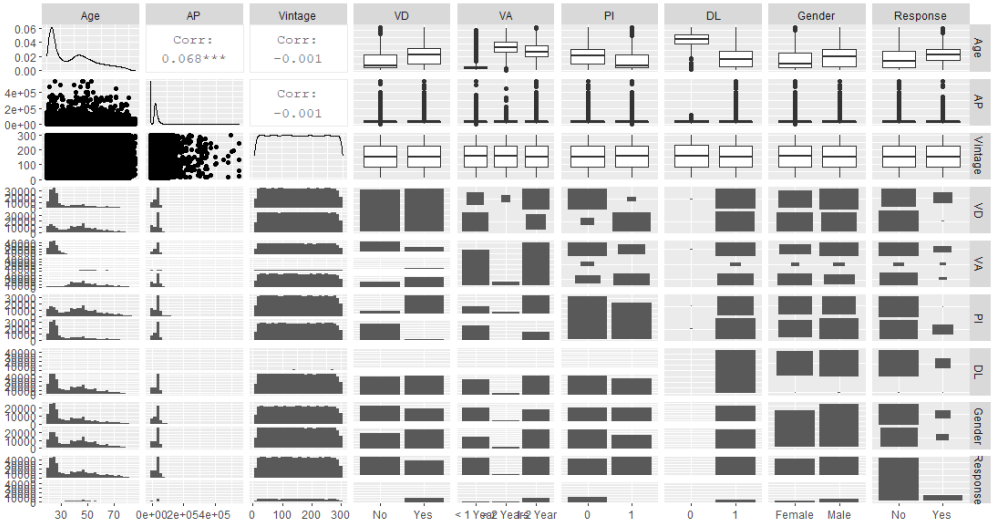

Figure 3: Summary of Variables

A.4 Different Parameter Combination of SVM

For the SVM model, there are two parameters, C (Cost) and sigma. For the combination of two parameters and intervals separetely, a grid search is performed. The two optimized hyperplane parameters with the highest AUC are selected. Figure 4 shows the hyperplane parameters combination and its performance.

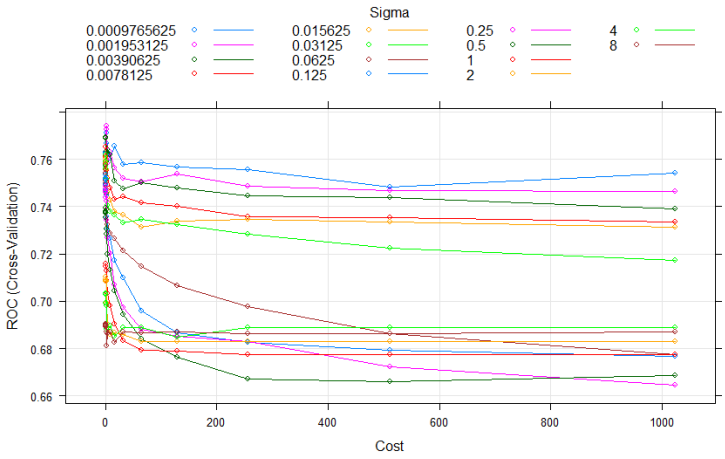

Figure 4: Different Parameter Combination SVM