# Vehicle Insurance Cross Sell Prediction Using Logistic Regression

Zehui Yu

09/12/2020

## Abstract

Insurance has evolved as a process of safeguarding the interest of people from loss and uncertainty. Having a car insurance is required by law in most states. And each state's laws set a minimum auto liability coverage limits that drivers are required to buy. From the buyers' perspective, car insurance serves as a financial and life protection. While from the insurance company's side, vehicle insurance helps them to generate profit though the large amount of annual premium since only a few of customers would get accident and use the insurance that year and not everyone suffers from that. Everyone shares the risk of everyone else, where insurance company takes advantage from that. In order to generate substantial profit, company need to identify their target customer who would like to pay large amount of premium. Building a model to predict whether a customer would be interested in updating 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. In this report, we aim to find out the possible attributes that may affect a customer's decision towards the annual premium thy would like to pay. I demonstrate this approach by creating forecasts from a public survey dataset: a series of cross-sell health insurance owner's features. Logistic Regression helps to find out the probability that a person would like to update their vehicle insurance package. It is also interesting to know the gender influence by using a propensity score matching process to discover if there exists a causal influence on vehicle insurance.

## Keywords

Logistic Regression, Propensity Score, Causal Inference, Gender, Age, Vehicle damage, Previously Insured, Annual Prmium

## Introduction

Logistic regression models the probabilities for classification problems with two possible outcomes. It's an extension of the linear regression model for classification problems, which finds an equation that predicts an outcome for a binary variable. It helps to forecasts the effects or impact of some specific changes or it can help to predict the trends and future values. Thus having the ability to use logistic regression to build models spurs flexibility of statistical methods to handle categorical data. However, one problem shows up. How do we choose variables and treatments? Statistical analysis and data values are ubiquitous to research analysis. People are fundamentally primed for making causal attributions based

on correlations. One possible causal links inferred from clinical studies will affect the treatment opinions. This implies that researchers must be careful to present their results in a manner that inhibits unwarranted causal attribution.

Regression models are characterized by a parameterization of the relationship between an outcome variable and a set of regressor variables. In this report, I use a logistic regression to discover whether a customer would be interested in updating their vehicle insurance. There are many reasons that may affect a customer's choice. Customer's age, vehicle's age, damage condition are all attributes that need to be considered for a person's annual premium. I investigate below with a logistic model that may have a causal skew by looking at the coefficients of the variables and the interpretation of statistical results.

One effective approach to make causal inference is through propensity score matching. Propensity score matching (PSM) is a quasi-experimental method in which the researcher uses statistical techniques to construct an artificial control group by matching each treated unit with a non-treated unit of similar characteristics. Using these matches, the researcher can estimate the impact of an intervention. In this report, I will use propensity score matching method to discover if there is a causal link between whether or not customer's gender and whether or not a customer would like to pay high vehicle insurance.

A survey data will be used to investigate if a customer would like to pay in high vehicle insurance by makeing the prediction from the logistic regression, and the dataset will also be used to test if there exists a causal link between a customer's gender and his willingness to pay. In Methodology section (Section 2), I will introduce the data and those variables, and the logistic model that used to make the prediction. It will also includes the model that was used to perform the propensity score analysis. Results of the logistic regerssion analysis and the propensity score analysis are provided in the Results section (Section 3). Besides, disscusions and future steps we can move on are illustrated in Conclusion section (Section 4).

# Methodology

## Data

We used a data sample from the company's historical customer background information database covered with 381109 observations and 12 variables. Data was collected via computer assisted telephone interviews. Our response variable is whether or not the customer would like to update their vehicle insurance package saved as "Response" that is a categorical variable with only two outcomes which is also a binary variable. We are interested in the customer's age, gender, their insurance stats(customer already has vehicle insurance or not), vehicle's age, vehicle's past damage status, their current annual premium, the number of days customers have been associated with the company based on previous studies which summarizes characteristics regarding vehicle package ("What Factors Affect Car Insurance Rates?", Sean Jackson), and those variables takes upon the explanatory variables.

| | Overall |
|---|---|
| n | 381109 |
| id (mean (SD)) | 190555.00 (110016.84) |
| Gender = Male (%) | 206089 (54.1) |
| Age (mean (SD)) | 38.82 (15.51) |
| Driving_License (mean (SD)) | 1.00 (0.05) |
| Region_Code (mean (SD)) | 26.39 (13.23) |
| Previously_Insured (mean (SD)) | 0.46 (0.50) |
| Vehicle_Age (%) | |
| < 1 Year | 164786 (43.2) |
| > 2 Years | 16007 ( 4.2) |
| 1-2 Year | 200316 (52.6) |
| Vehicle_Damage = Yes (%) | 192413 (50.5) |
| Annual_Premium (mean (SD)) | 30564.39 (17213.16) |
| Policy_Sales_Channel (mean (SD)) | 112.03 (54.20) |
| Vintage (mean (SD)) | 154.35 (83.67) |
| Response (mean (SD)) | 0.12 (0.33) |

By creating a table one, we could first glance the data and have a general idea of this dataset. Table one summarizes both continuous and categorical variables mixed within one table. Categorical variables such as gender, vehicle damage, response can be summarized as counts. Continuous variables such as age, annual premium, vintage (the number of days customers have been associated with the company) can be summarized with their means and standard deviations.

## Model

Since our response variable is a categorical variable with binary outcomes, we build a logistic regression model to discover whether a customer would like to update their vehicle insurance. Given the fact that women tend to be more likely to have traffic accidents, and with a previous survey mentioned that female would be more likely to choose a high vehicle insurance premium, we want to find out if there is possible causal links between gender and their willingness to pay. One organized way to find out the causal link is using propensity score matching, that match treated and controlled observations on the estimated probability of being treated. In our example, our treatment is gender and the variable of interest is the response which represents the customer's willingness to pay. We focus on customer's age, gender, previous insurance stats, vehicle's age, vehicle's damage status(whether or not the vehicle has a car accident before) and its current annual premium. By using a matching function to help us map those treated variable with untreated together to reduce the data, we are able to proceed followed by a logistic regression to discover the effects of other element that may affect a customer's decision and exam the effect of being treated on gender in the usual way. Our formula is

$$log(P/(1-P)) = \beta_0 + \beta_1 age + \beta_2 ismale + \beta_3 PreviouslyInsured + \beta_4 VehicleAge + \beta_5 VehicleIsDamaged + \beta_6 AnnualPremium + \beta_7 Vintage$$

```
Call:
glm(formula = Response ~ Gender + Age + Previously_Insured +
    Vehicle_Age + Vehicle_Damage + Annual_Premium + Vintage,
    family = binomial, data = raw_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3097  -0.7347  -0.0373  -0.0344   3.8783

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -3.244e+00  4.489e-02 -72.275   <2e-16 ***
GenderMale          1.389e-01  1.096e-02  12.675   <2e-16 ***
Age                 9.526e-04  3.737e-04   2.549   0.0108 *
Previously_Insured -4.092e+00  8.273e-02 -49.469   <2e-16 ***
Vehicle_Age1       -2.121e-01  1.895e-02 -11.190   <2e-16 ***
Vehicle_DamageYes   2.133e+00  3.418e-02  62.407   <2e-16 ***
Annual_Premium      2.676e-06  2.884e-07   9.277   <2e-16 ***
Vintage            -3.699e-06  6.376e-05  -0.058   0.9537
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 283546  on 381108  degrees of freedom
Residual deviance: 214854  on 381101  degrees of freedom
AIC: 214870

Number of Fisher Scoring iterations: 9
```

P is the probability of a customer choosing to update their vehicle insurance package.

$log(P/(1-P))$ represents the log odds of a customer choosing to update their vehicle insurance package.

$\beta_0 = -3.244$ means when a person is a male of age zero that have not payed insurance before who owns a new car without damaged and does not pay any annual premium, log odds of response yes equals to -3.244.

$\beta_1 = 9.526e-04$ means for every additional unit increase in age, we expect the log odds of choosing to update their package increase by 9.526e-04.

$\beta_2 = 0.1389$ means when a person changes from male to female, we expect the log odds of choosing to update their package increase by 0.804.

$\beta_3 = -4.092$ means that when a person changed from not have insurance before to used to have vehicle insurance, we expect the log odds of choosing to update their package increase by 2.133.

$\beta_4 = -2.121$ means for every additional unit increase in vehicle's age, we expect the log odds of choosing to update their package increase by -2.121.

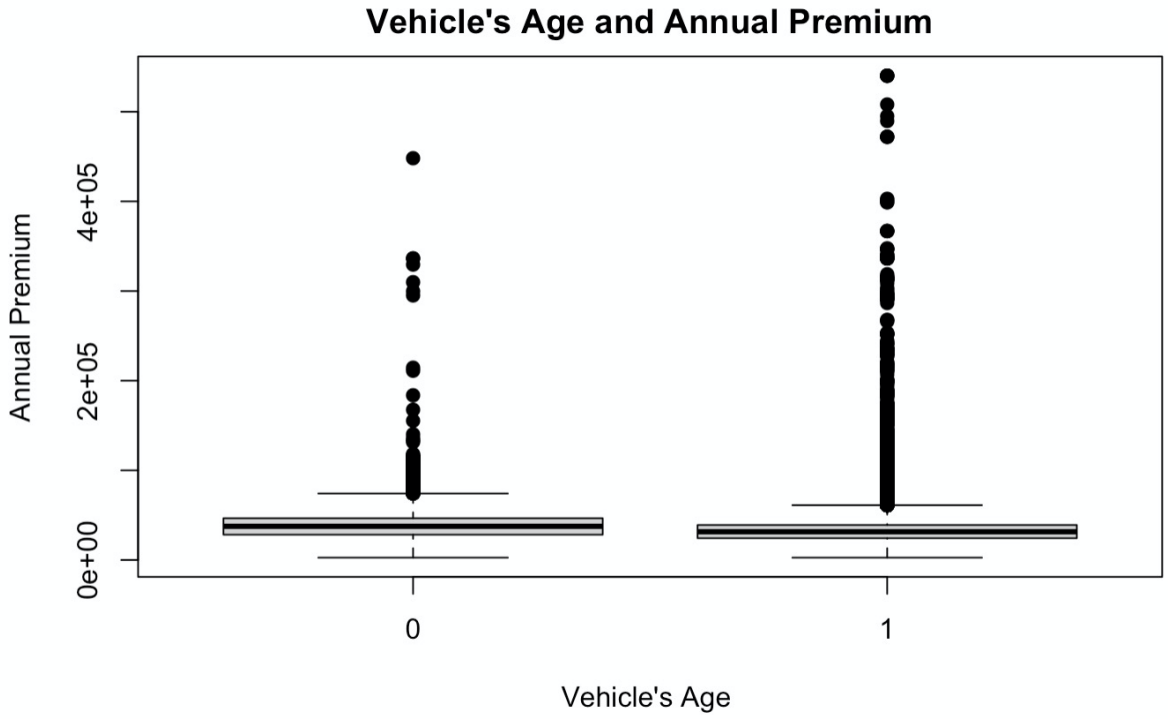
$\beta_5 = 2.133$ means when a person used to have a car accident that damages the car compared with people who does not have car accident before, we expect the log odds of choosing to update their package increase by -4.092.

$\beta_6 = 2.676e-06$ means for every additional unit increase in annual premium, we expect the log odds of choosing to update their package increase by 2.676e-06.

$\beta_7 = -3.699e-06$ means for every additional unit increase in the number of days the customer has been associated with the company, we expect the log odds of choosing to update their package increase by -3.699e-06.

## Results

By using logistic regression compensated with propensity score matching, we build a model to estimate the probability of customer's willingness to update their vehicle insurance. Our model is based off our logistic analysis of the customer's willingness to pay, which accounted for customer's age, gender, previous insurance status, vehicle's age, vehicle's damage status(whether or not the vehicle has a car accident before) and its current annual premium. We found that an elderly female person with cars that already have insurance and used to have car accident happened with higher annual premium would like to update their package more. With all the other situations in the same condition, female tend to be more likely to pay a higher insurance package than man with a log odds probability of 13.89. Older vehicle will have high insurance package than new vehicles as well since they are more likely to encounter car accident and they need more protect. However, one interesting thing to notice is that the number of days that a customer has been associated with the company does not help to make a difference on their decision making of choosing to update their vehicle insurance package. We can verify this by referring to the p-value. Since the p-value of the vintage is 0.9537 which is greater than 0.05, we have strong evidence to against the alternative hypothesis that beta vintage is not zero. Therefore, we can adjust our communication strategy when promoting the updated package even with new customers given the fact that the customers' decision will not largely depend on the familiarity with insurance's company.




An insurance policy is an arrangement by which a company undertakes to provide a guarantee of compensation for specified loss, damage, illness, or death in return for the payment of a specified premium. A premium is a sum of money that the customer needs to pay regularly to an insurance company for this guarantee. By referring to the plot which draws two boxplots accross the groups with cars less than 2 years and cars above 2 years, we noticed that despite the age of the vehicle, car's owners would pay a roughly same amount of money for their vehicle insurance. This is because some factors that may affect the insurance premiums are your car, demographic factors and the coverages, limits and deductibles you choose. These factors may include things such as your age, anti-theft features in your car and your driving record. While it may be tempting to reduce or eliminate coverages to help lower car insurance premium, such as other factors , for instance your driving habits, your cars status that may also affect the price you pay.

## Discussion

## Summary

Our goal is building a model to predict whether a customer would be interested in updating their package from the insurance company's perspective which helps to accordingly plan its communication strategy to reach out to those customers and optimize its business model and revenue. We used a survey data covered 381109 variables from the company's historical customer information database. The first step is to clean the dataset, we only focus on eight variables, ranged from gender, age, whether a customer is previously

insuranced or not, the vehicle's age, vehicle's damage, its currently annual premium and the number of days it has been stayed with the company saved as "Vintage". All of these cleaning steps are helpful and essential for later comparison and analysis. Since observational data is often more feasible, and arguably more reliable, the causal links inferred from the study may have affected the outcome, thus we used one of the popular way to make causal influence through propensity score matching. We usually think women are more likely to have traffic accidents, which means they are more likely to have updated package to insurance their safety. As a result, we used propensity score matching to discern if there is a causal link between whether or not gender would be a determined element that affect a customer's choice. We construct a logistic regression model that explains whether a female was treated as a function of the variables that we think explain it. Then we are able to add our forecast to our dataset. By using our forecast, we can create matches. For every person who was actually a female, we want a male who was considered as similar to her based on propensity score as much as possible. With the help of a matching function from the arm package, we find the closest of the ones that were male to female and we reduced the dataset to just those that are matched. Finally, we proceed to use a logistic regression tested the log odds of customer's response against gender, age, whether a customer is previously insuranced or not, the vehicle's age, vehicle's damage, its currently annual premium and the number of days it has been stayed with the company to analysis the effect on customer's willingness to update their insurance package. From the logistic regression model, we noticed that except vintage that the customer has been stay with the company will not make a difference on their decision, all the other variables will have an effect on the customer's decision to choose whether to update their package or not.

## Conclusion

The logistic regression model predicts the the binary outcome of whether a customer would like to update their vehicle insurance package or not based on a set of independent variables ranged from gender, age, whether a customer is previously insuranced or not, the vehicle's age, vehicle's damage, its currently annual premium and the number of days it has been stayed with the company saved as "Vintage". It assumed that the logit transformation of the outcome variable has a linear relationship with the predictor variables, thus it yields a log odds which is the probability of a customer would like to update their package taken the logarithm of the odds. From the model, we noticed that elderly female person would increase the lod odds by -3.244. With age of the cars increase and the driver's age increase, customer tend to more likely to pay for higher insurance package.

The propensity score analysis showed that people who are female were 13.89 (p-value < 0.002) times more likely to choose to update their insurance package than male. Based off this result it appears as though that women are less likely to be in an accident than males and that males are more likely to drive without their seatbelt, it may surprise drivers to learn that, on average, females pay more for car insurance than their male counterparts. According to The Zebra, women pay $740 for their six-month premium versus men who can except to pay an averge of $735 for six months. ("Which Gender Pays More for Car Insurance?", David Robinson)

## Weakness and Next Steps

By using a logistic regression model, we are assuming that each observations are independent from each other. However, this may not be true in real life thus we ignore some random variability. Besides, the data set is not representative enough. There are other factors that should be take account of this analysis as well. For instance, the driver's credit level, the marriage status may also have an impact on the customer's willingness to pay. In order to be more accurate and specific, we could also use a more up-to-date database and proceed our regression models.

We can increase the sample size to get more data, getting more people involved and collect more information, thus make a more reliable and representative conclusion based a relatively larger sample size. Also, we will collect more detailed factors such as driver's credit level, the marriage status to better explain what factors will influence the customer's willingness to pay and how these factors influence, since different credit level requires different minimum premium settings that may affect a person's decision. From the economic perspective, the willingness to pay will be determined by the demand and supply of the current package, therefore we may also apply the basic demand and supply model to improve this analysis. The last but not the least, adding Bayesian inference is also helpful for our study.

## References


1. Yoshida, Kazuki. Introduction to Tableone, 25 July 2020, cran.r-project.org/web/packages/tableone/vignettes/introduction.html.
2. Austin, Peter C. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies." Multivariate Behavioral Research, Taylor & Francis, May 2011, www.ncbi.nlm.nih.gov/pmc/articles/PMC3144483/.
3. Samantha-Jo Caetano(2020), data cleaning code (01-data_cleaning-post-strat1.R).https://q.utoronto.ca/courses/184060/files/
4. Samantha-Jo Caetano(2020), Matching-PropensityScore-Amazon (02-data_propensity_score.R).https://q.utoronto.ca/courses/184890/files/
5. R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
6. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
7. Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

8. Hadley Wickham and Evan Miller (2020). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files. R package version 2.3.1. https://CRAN.R-project.org/package=haven

9. David Robinson and Alex Hayes (2020). broom: Convert Statistical Objects into Tidy Tibbles(Version 0.5.3) [R Package]. https://CRAN.R-project.org/package=broom

10. RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/

11. Kumar, Anmol. "Health Insurance Cross Sell Prediction   ." Kaggle, 11 Sept. 2020, www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction.

12. /, Sean Jackson, et al. "What Factors Affect Car Insurance Rates?" Coverage.com, 8 Nov. 2020, www.coverage.com/insurance/auto/factors-affecting-your-auto-insurance-rates/.

13. Research, Hearst Autos. "Which Gender Pays More for Car Insurance?" Car and Driver, Car and Driver, 13 Nov. 2020, www.caranddriver.com/research/a31268333/which-gender-pays-more-for-car-insurance/.

14. says:, Jack King, et al. "What? Women Pay More Than Men for Auto Insurance? Yup." Insurance Journal, 13 Feb. 2019, www.insurancejournal.com/news/national/2019/02/12/517466.htm.

15. "6 Factors That Affect Car Insurance Premiums." Policybazaar, www.policybazaar.com/motor-insurance/car-insurance/articles/factors-that-affect-car-insurance-premiums/.

## Github

https://github.com/Zehui-Yu/Sta304-Final

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.