Final Project

-Sutianyi Wen (Writer)

11/18/2020

Summary

What types of customers are more interested in buying vehicle insurance? Is the number of days a customer has been associated with the company statistically significant to his/her interest in car insurance? Does customers' interest in car insurance differ by region and policy sales channel? The multi-level logistic regression is ideal to answer the three questions above because the response variable is binary and there are naturally grouped variables. It turns out customers who didn't have car insurance and customers who had a car damaged before are the two most scientific significant groups. And the number of days a customer has been associated with the company is not statistically significant to understand his/her interest in car insurance. Lastly, Customers' interest in car insurance differs by regions and sales channels.

Introduction

Cross-selling refers to sell one product with another that customers already purchased. For example, if a customer bought a new TV at Target, Target will also want to sell a TV stand to the customer because he/she needs to place the new TV at home. An insurance company that plans to sell new vehicle insurance in addition to health insurance would like to know what types of customers are more interested in buying its vehicle insurance. Moreover, the company also wants to know whether the number of days a customer has been associated with the company is significant to the customer's interest in buying car insurance(i.e.the time a customer enrolled in the company's health insurance policy). Finally, the company wants to know if the interest in buying vehicle insurance would vary by sales channels and regions.

Exploratory Data Analysis

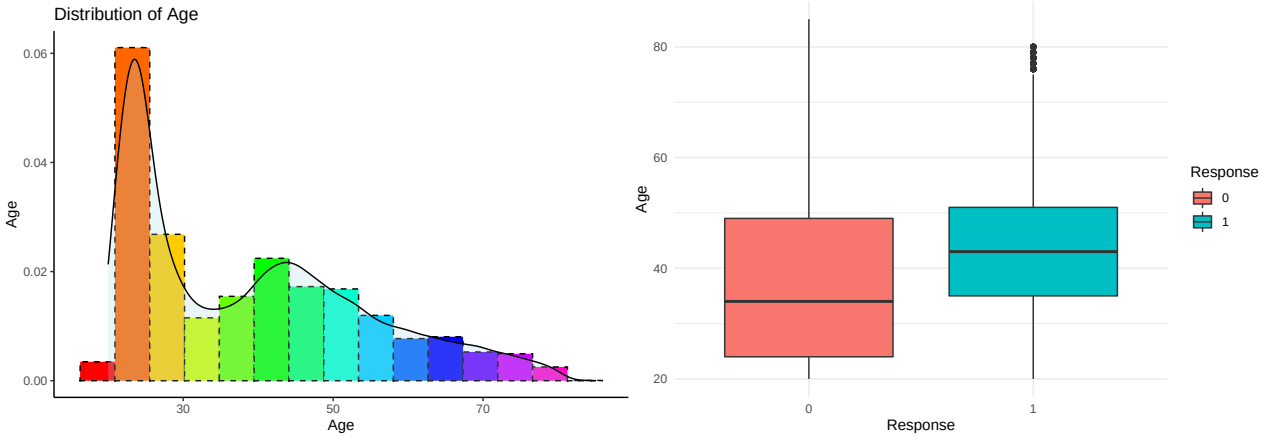
The original dataset has 381,109 observations and 11 predictors. Because the given dataset is so large that significantly slowing down the fitting process, I decided to take a random sample with 130,000 entries which are roughly 34% of the original dataset. "Response" is a binary response variable indicating whether a customer is interested in buying the vehicle insurance (1:interested, 0:not interested). Gender is a factor predictor with 100,000 more male observations than female. Age is a numeric variable ranging from 20 to 85. Driving_License indicates whether a customer has a license and it appears 99.7% of customers have a license. Region_Code is a categorical variable with 53 categories and each code doesn't contain geographical meaning in the real world because the company didn't provide information. Besides Region_Code, Policy_Sales_Channel contains 145 categories in number, such as sale via phone and manager, but the company also didn't provide information about what each number stands for. Previously_Insured is a binary variable suggesting if a customer already had vehicle insurance. Vehicle_Damage is a binary predictor indicating whether a customer got his/her vehicle damaged before. Vehicle_Age is a categorical variable with three levels, where only 4% of customers' vehicles are older than 2 years. Annual_Premium and Vintage(number of days associates with the company) are continuous variables.

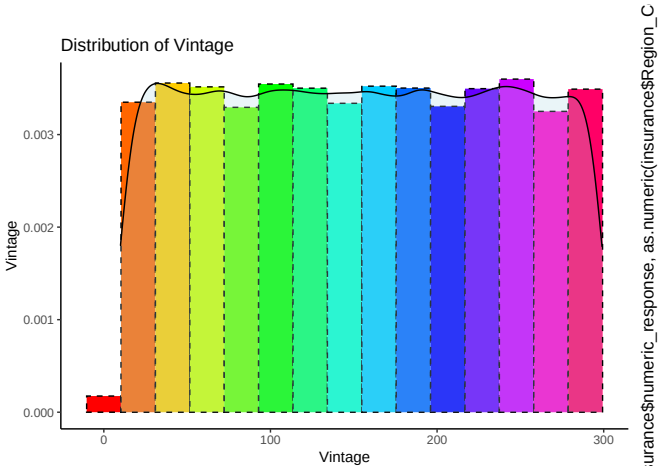
Gender	Age	Driving_License	Region_Code	Previously_Insured

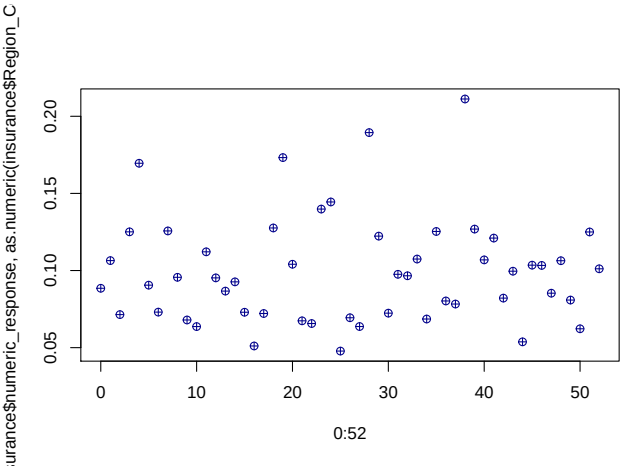
Gender	Age	Driving_License	Region_Code	Previously_Insured
Female:59767	Min. :20.00	0: 279	28 :36389	0:70297
Male :70233	Median :36.00	1:129721	8 :11589	1:59703
	Mean :38.83		46 : 6727	
	Max. :85.00		(Other):75295	

Vehicle_Damage	Annual_Premium	Sales_Channel	Vintage	Response	Vehicle_Age
No :64557	Min. : 2630	152 :46108	Min. : 10.0	0:114015	< 1 Year :56326
Yes:65443	Median : 31771	26 :27467	Median :154.0	1: 15985	1-2 Year :68332
	Mean : 30661	124 :25157	Mean :154.2		> 2 Years: 5342
			Max. :299.0		

The distribution of Age is not normal and there are a significant number of customers with age less than 30. From the Age Vs Response boxplot, it appears the medium age of customers who are interested in buying vehicle insurance is higher than the medium age of customers who are not interested. The distribution of variable Vintage is uniform, which suggests the dataset contains approximately the same number of observations for each numeric value. Because we would like to use a hierarchical model to fit the data, it's better to bin up both Age and Vintage to avoid the "Can not converge" error. Lastly, the probability of interest in vehicle insurance VS region plot suggests customers in different regions have a different probability of interest in buying the new insurance. The highest-interest region is 38 and the lowest is region 17 so it makes sense to fit the "glmer" model to see the variance between regions. I also tried with different interactions but none of them seems to be interesting. Finally, I calculated conditional probability for factor variables and all of them bring interesting findings such as customers' interest in vehicle insurance appears to differ by gender, whether they already had vehicle insurance and whether their vehicles were damaged before.







	0	1		0	1
Female	0.90	0.10	Insured before	0.77	0.22
Male	0.86	0.14	Not-Insured before	1	0

	0	1		0	1
Vehicle age $<$ 1 Year	0.96	0.04	Vehicle not damaged before	0.99	0.01
Vehicle age 1-2 Year	0.83	0.17	Vehicle damaged before	0.76	0.24
Vehicle age $>$ 2 Year	0.70	0.30			

Model Building and Assessment

First, it's necessary to bin up two continuous variables: Age and Vintage, and save them as Age_fac and Vintage_fac. Age_fac has three levels: less-than-30, 30-to-50, and above-50. Vintage_fac has three levels: less-than-100, 100-to-200, and above-200(thresholds are set based a domain knowledge). Also, I transfer Annual_Premium to a factor variable since the distribution contains several gaps between values.

For model development, I start with a model with fixed predictors: Gender, License, Previously_Insured, Vehicle_Damage, Vehicle_Age, Annual_Premium_fac, Age_factor, Vintage_fac, and two grouping terms to capture varying intercept effects by region and sales channel. Then I run ANOVA tests to see whether I need to drop any fixed predictors and it turns out $License$ and $Vintage\_fac$ are not significant. To if test random effects are significant, the AIC score increases 200 if taking out the random effects of Region_Code and the AIC score increases 800 if taking out the random effects of Sales_Channel, which convinces me to keep both grouping term. Therefore the final model is below

$$Response = \beta_0 + \gamma_{0k[i]}^{Region\_Code} + \gamma_{1k[i]}^{Sales\_Channel} + \beta_{Gender} Gender_i + \beta_{Age\_fac} Age\_fac_i +$$

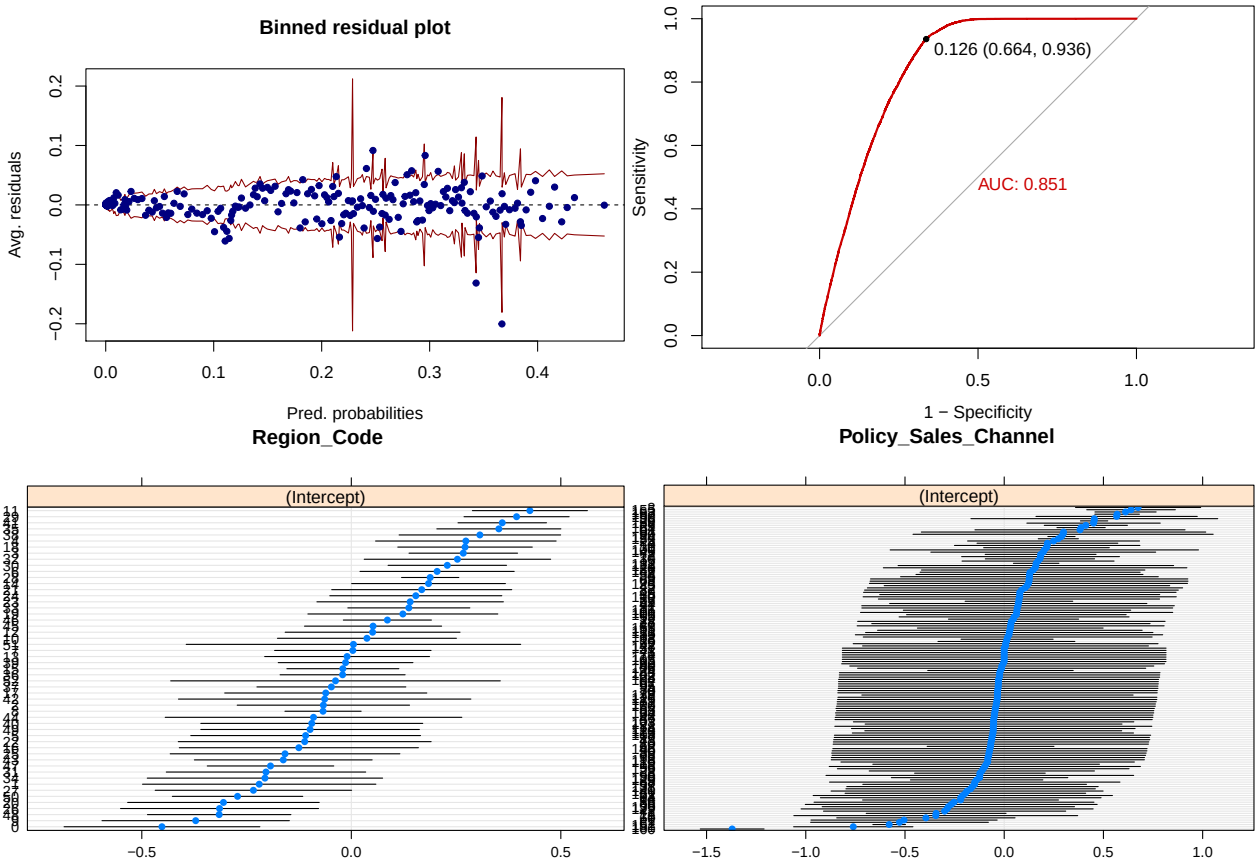
$$\beta_{Vehicle\_Damage} Vehicle\_Damage_i + \beta_{Previously\_Insured} Previously\_Insured_i +$$

$$\beta_{Vehicle\_Age} Vehicle\_Age_i + \beta_{Annual\_Premium\_fac} Annual\_Premium\_fac_i$$

$$\gamma_{0k} \sim N(0, \sigma^2_{Region\_Code}), \quad \gamma_{1k} \sim N(0, \sigma^2_{Sales\_Channel})$$

In my binned residual plot, most of the points lie between the bounds of the 95% confidence interval, and they are randomly distributed. But the model seems to struggle for fitted values close to 0.0. The ROC curve suggests using a threshold equals to 0.126. the AUC is 0.851. In the dot plot for Region_Code, 21 regions are outliers. Among all outliers, region 0 is most negatively different and region 11 is most positively different from other regions, which suggests customers at region 0 are least likely to be interested in buying

car insurance and customers at region 11 are most likely to be interested compared to customers in other regions. For the dot plot of sales channels, sales channels 160 is most different from others.



Model Results & Interpretation

If a customer is female with an age less than 30, her car, not damaged and insured before, is less than 1 year and the policy premium is low, the probability of the customer interested in buying vehicle insurance from the company is 0.03. Given every other predictor unchanged, a male customer is 6% more likely to be interested in car insurance than a female customer. And holding everything else the same, a customer with age between 30 and 50 is 60% more likely to be interested in buying car insurance than a customer with age less than 30 and a customer with age above 50 is 10% less likely to be interested in car insurance compared to a customer with age less than 30.

Holding everything unchanged, a customer who had car damages before is 612% more likely to be interested in buying car insurance compared to a customer who never had vehicle damages. Giving every other variable staying the same, a customer who already insured his/her car is 98% less likely to be interested in car insurance. Given everything staying the same, a customer who has a vehicle with an age between 1 and 2 years is 20% less likely to be interested in buying car insurance compared to a customer who has a vehicle with an age less than 1 year. Lastly, holding everything else the same, if the annual premium is between 8103 and 36315 instead of below 8103, a customer is 11% more likely to be interested in buy car insurance; if the annual premium is above 36315 instead of below 8103, a customer is 12% more likely to be interested in buy car insurance.

The random effect variance of Region_Code is 0.06 and the random effect variance of Policy_Sales_Channels is 0.17

Predictors	Odds Ratios	CI	p
(Intercept)	0.03	0.02 – 0.04	<0.001
Gender [Male]	1.06	1.02 – 1.10	0.002
Age_fac [30-to-50]	1.60	1.47 – 1.74	<0.001
Age_fac [above-50]	0.90	0.82 – 0.99	0.023
Vehicle_Damage [Yes]	7.12	6.34 – 8.00	<0.001
Previously_Insured [1]	0.02	0.01 – 0.02	<0.001
Vehicle_Age [> 2 Years]	0.98	0.88 – 1.09	0.655
Vehicle_Age [1-2 Year]	0.80	0.74 – 0.88	<0.001
Annual_Premium_fac [high]	1.11	1.04 – 1.18	0.001
Annual_Premium_fac [medium]	1.12	1.06 – 1.19	<0.001
Random Effects			
σ^2	3.29		
τ_{00} Policy_Sales_Channel	0.17		
τ_{00} Region_Code	0.06		
ICC	0.07		
N Region_Code	53		
N Policy_Sales_Channel	145		
Observations	130000		
Marginal R2 / Conditional R2	0.699 / 0.719		

Conclusion & Limitation

According to the result of the final model, male customers are more interested in new car insurance than female customers. Customers age between 30 and 50 are more interested in buying car insurance compared to other age groups. The insurance sales team should reach out to customers who had a car damaged before or customers who are not insured since the two predictors are most scientifically significant. Moreover, it appears customers are less likely to purchase car insurance as vehicles get older and customers to pay for high premium car insurance.

The final model doesn't include the Vintage variable, which provides information about how many days a customer has been associated with the company because the ANOVA test suggests it's not significant. Lastly, customers' interests in buying car insurance vary by region and sales channel according to model results. And sales channels have a larger variance than regions.

It can't be ignored that the dataset has limitations. First, variables Region_Codes and Policy_Sales_Channels don't carry real-world meaning in this analysis. It's impossible to draw meaningful conclusions in real-world settings, like customers in California are more interested in car insurance than customers in Washington. Besides that, Policy_Sales_Channels has 161 levels and the boundaries between each level are ill-defined. For instance, level 1 and level 2 could be "Sale Agent A" and "Sale Agent B" then it makes sense to combine the two levels. However, I can't bin up Policy_Sales_Channels because the codebook is unclear about it. Second, my analysis is based on a random sample instead of the whole dataset, which waste some data entries.

R Appendix

```
knitr::opts_chunk$set(echo = FALSE)

library(ggplot2)
library(arm)
library(pROC)
library(e1071)
library(caret)
library(lme4)
library(sjPlot)
library(lattice)

setwd("~/Desktop/DukeFA20/IDS702/Final Project")
insurance = read.csv('train.csv')
set.seed(1234)
index <- sample(1:nrow(insurance), 130000)
insurance = insurance[index,]
insurance$numeric_response = insurance$Response
### Data Pre-processing ###
insurance = insurance[ , -which(names(insurance) %in% c("id"))]
insurance$Gender = factor(insurance$Gender)
insurance$Driving_License = factor(insurance$Driving_License)
insurance$Region_Code = factor(insurance$Region_Code)
insurance$Previously_Insured = factor(insurance$Previously_Insured)
insurance$Vehicle_Age = factor(insurance$Vehicle_Age)
insurance$Vehicle_Damage = factor(insurance$Vehicle_Damage)
insurance$Policy_Sales_Channel = factor(insurance$Policy_Sales_Channel)
insurance$Response = factor(insurance$Response)

index_1 = insurance$Vintage < 100
index_2 = insurance$Vintage >= 100 & insurance$Vintage < 200
index_3 = insurance$Vintage >= 200
insurance$Vintage_fac[index_1] = '100-day-less'
insurance$Vintage_fac[index_2] = '200-day-less'
insurance$Vintage_fac[index_3] = '300-day-less'
insurance$Vintage_fac = factor(insurance$Vintage_fac)

index_1 = log(insurance$Annual_Premium) < 9
index_2 = log(insurance$Annual_Premium) >= 9 & log(insurance$Annual_Premium) < 10.5
index_3 = log(insurance$Annual_Premium) >= 10.5
insurance$Annual_Premium_fac[index_1] = 'low'
insurance$Annual_Premium_fac[index_2] = 'medium'
insurance$Annual_Premium_fac[index_3] = 'high'
insurance$Annual_Premium_fac = factor(insurance$Annual_Premium_fac)

index_1 = insurance$Age < 30
index_2 = insurance$Age >= 30 & insurance$Age < 50
index_3 = insurance$Age >= 50
insurance$Age_fac[index_1] = 'below-30'
insurance$Age_fac[index_2] = '30-50'
insurance$Age_fac[index_3] = 'above-50'
insurance$Age_fac = factor(insurance$Age_fac)
```

```r
# Age histogram
ggplot(insurance,aes(Age)) +
  geom_histogram(aes(y=..density..),color="black",linetype="dashed",
                 fill=rainbow(15),bins=15) + theme(legend.position="none") +
  geom_density(alpha=.25, fill="lightblue") + scale_fill_brewer(palette="Blues") +
  labs(title="Distribution of Age",y="Age") + theme_classic()
# Age Vs Response
ggplot(insurance,aes(x=Age,y=Response,fill=Response))+
  geom_boxplot()+coord_flip()+ theme_minimal()
# Vintage histogram
ggplot(insurance,aes(Vintage)) +
  geom_histogram(aes(y=..density..),color="black",linetype="dashed",
                 fill=rainbow(15),bins=15) + theme(legend.position="none") +
  geom_density(alpha=.25, fill="lightblue") + scale_fill_brewer(palette="Blues") +
  labs(title="Distribution of Vintage",y="Vintage") + theme_classic()
# Probability of interesting buying insurance by region code
plot(0:52,tapply(insurance$numeric_response, as.numeric(insurance$Region_Code), mean),col='blue4',pch=1
model1 <- glmer(Response ~ Gender+Age_fac+Vehicle_Damage+Previously_Insured+
                  Vehicle_Age+Annual_Premium_fac+(1|Region_Code)+(1|Policy_Sales_Channel),
                family=binomial(link="logit"),
                data=insurance)
resid = residuals(model1,"resp")
binnedplot(x=fitted(model1),y=resid,xlab="Pred. probabilities",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
roc(insurance$numeric_response,fitted(model1),plot=T,print.thres="best",legacy.axes=T,
    print.auc =T,col="red3")
dotplot(ranef(model1, condVar=TRUE))$Region_Code
dotplot(ranef(model1, condVar=TRUE))$Policy_Sales_Channel
```