684 Midterm Project Proposal

Allstate Purchase Prediction

Mengyun Li

November 13, 2017

1. Introduction

As a customer shops an insurance policy, he/she will receive a number of quotes with different coverage options before purchasing a plan. This is represented in this challenge as a series of rows that include a customer ID, information about the customer, information about the quoted policy, and the cost. The task is to predict the purchased coverage options using a limited subset of the total interaction history.

2. Data and Method

2.1 Data Source

```
#Data clean
data_test <- read.csv("test_v2.csv")
#data_test$car_value <-gsub(pattern=NULL, replacement=NA, x= data_test$car_value)
#levels(data_test$car_value)[levels(data_test$car_value)=="/s"] <- "NA"
data_test <- na.omit(data_test)
data_new <- data_test[1:12000, ]
head(data_new, n=5)
```

```
##    customer_ID shopping_pt record_type day  time state location group_size
## 5     10000003           1           0   3 17:12    AR    10004          1
## 6     10000003           2           0   3 17:12    AR    10004          1
## 7     10000003           3           0   3 17:13    AR    10004          1
## 12    10000004           5           0   1 12:53    OK    10005          1
## 13    10000004           6           0   1 12:54    OK    10005          1
##    homeowner car_age car_value risk_factor age_oldest age_youngest
## 5          0       4         d           4         26           26
## 6          0       4         d           4         26           26
## 7          0       4         d           4         26           26
## 12         0      13         f           3         22           22
## 13         0      13         f           3         22           22
##    married_couple C_previous duration_previous A B C D E F G cost
## 5               0          3                 1 1 0 1 1 0 2 2  628
## 6               0          3                 1 1 0 2 1 0 2 2  625
## 7               0          3                 1 1 0 2 1 0 2 2  628
## 12              0          1                 3 2 0 1 1 0 3 2  673
## 13              0          1                 3 2 0 1 1 0 2 2  683
```

2.2 Variable Discriptions

customer_ID - A unique identifier for the customer
shopping_pt - Unique identifier for the shopping point of a given customer

record_type - 0=shopping point, 1=purchase point
day - Day of the week (0-6, 0=Monday)
time - Time of day (HH:MM)
state - State where shopping point occurred
location - Location ID where shopping point occurred
group_size - How many people will be covered under the policy (1, 2, 3 or 4)
homeowner - Whether the customer owns a home or not (0=no, 1=yes)
car_age - Age of the customer's car
car_value - How valuable was the customer's car when new
risk_factor - An ordinal assessment of how risky the customer is (1, 2, 3, 4)
age_oldest - Age of the oldest person in customer's group
age_youngest - Age of the youngest person in customer's group
married_couple - Does the customer group contain a married couple (0=no, 1=yes)
C_previous - What the customer formerly had or currently has for product option C (0=nothing, 1, 2, 3,4)
duration_previous - how long (in years) the customer was covered by their previous issuer A,B,C,D,E,F,G - the coverage options
cost - cost of the quoted coverage options

2.3 Goal of analysis

Use multilevel linear model to find the relationship between different kinds of customer with the quote price of car insurance.

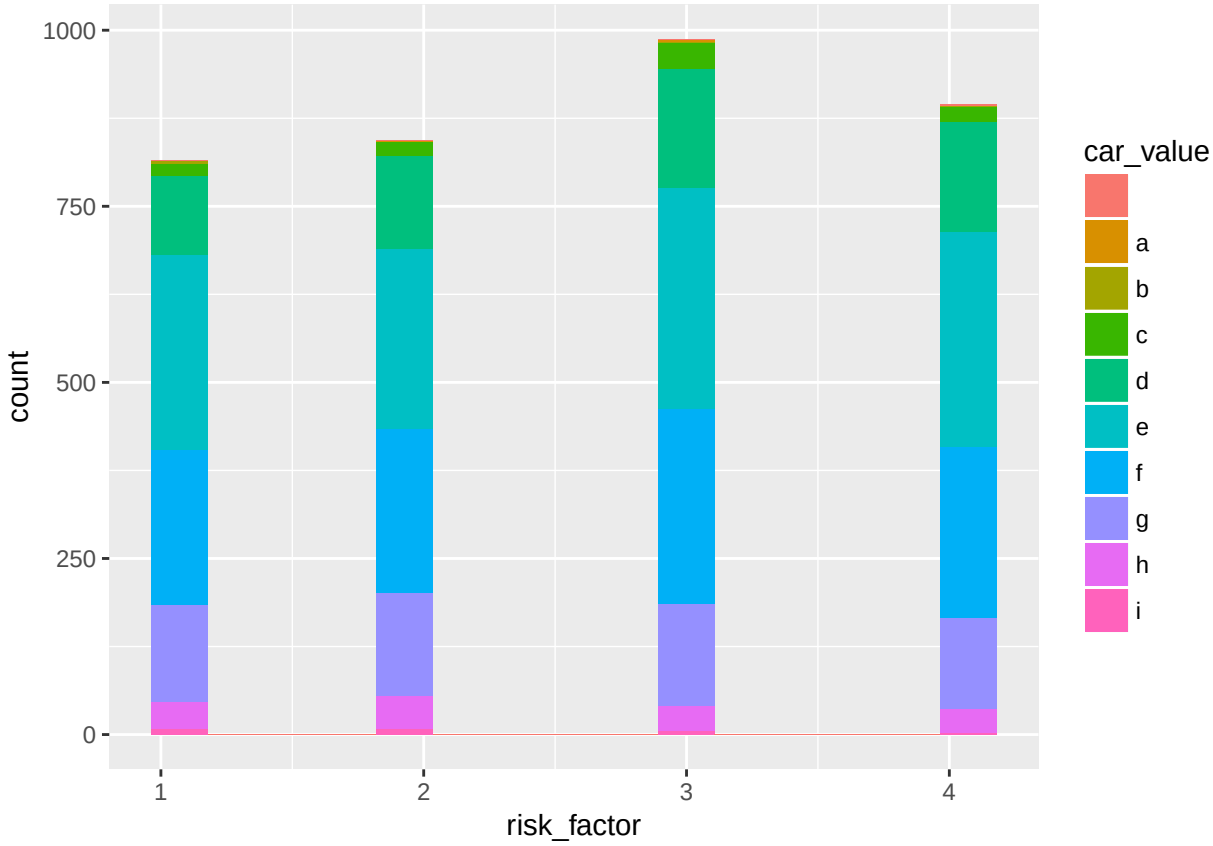
3. Exploratory Data Analysis

3.1 Data Visuallization

Numbers of customers in each risk factor (fill by different car value).

```
data.count <- data_new%>%
  select(customer_ID,risk_factor,car_value)%>%
  count(customer_ID,risk_factor, car_value)%>%
  group_by(customer_ID)

ggplot(data=data.count)+
  geom_histogram(mapping=aes(x=risk_factor, fill = car_value), bins=15)
```


1000
750
500
250
0
count
1
2
3
4
risk_factor
car_value
a
b
c
d
e
f
g
h
i