Allstate Purchase Prediction Challenge

Submitted by: Parvez Rafi (rafiparvez@tamu.edu)

Variable Descriptions

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

Product Options

Each product has 7 customizable options selected by customers, each with 2, 3, or 4 ordinal values possible:

Option	Possible Values
A	0,1,2
B	0,1
C	1,2,3,4
D	1,2,3
E	0,1
F	0,1,2,3

G	1,2,3,4

A product is simply a vector with length 7 whose values are chosen from each of the options listed above. The cost of a product is a function of both the product options and customer characteristics.

Exploratory Data Analysis and Data Cleaning

- **Analyzing and imputing missing training and test values**

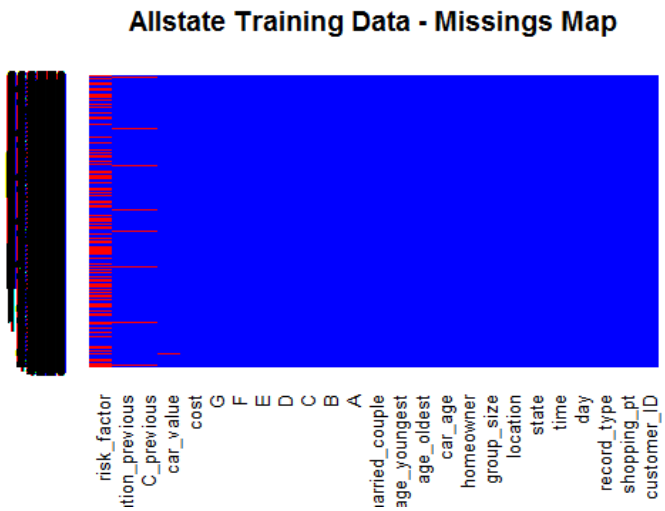


Figure: Missing values in Training data

	risk_factor	C_previous	duration_previous	car_value
% of missing value	36.1395508	2.8126311	2.8126311	0.2301394

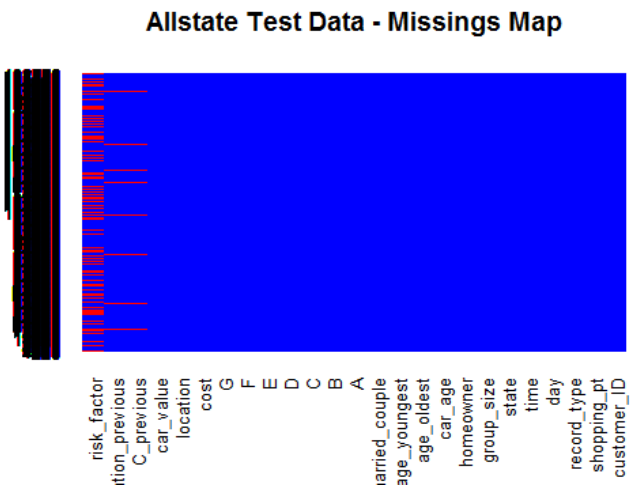
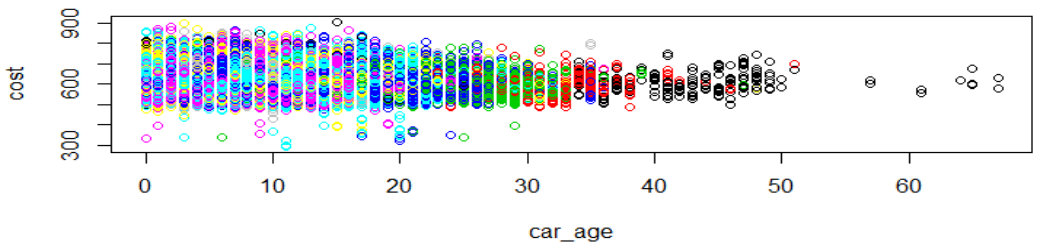


Figure: Missing values in Test data

	risk_factor	C_previous	duration_previous	car_value	location
% of missing value	37.9606348	4.9126001	4.9126001	0.3716257	0.3409502

- **car_value:** It is the value of the car at the time it was bought and should remain same for a car for a specific customer. So, I imputed the missing car_value by picking non-missing value in another row for same customer. I could still see car_value missing for 369 records in testdata. These were the car_value for which there no data was available for a specific customer. Assuming car_value to be function of cost and car_age, I generated following scatter plot.



I imputed remaining missing car_value values by generating a multinomial classification model with attributes cost and car_age.

- **C_previous:** It is what the customer formerly had or currently has for product option. For missing values for a specific customer, if customer has a C_Previous value in other row, I picked that value. If a customer has C_Previous missing for all the entries, I imputed the value to 0 assuming it's a new customer.

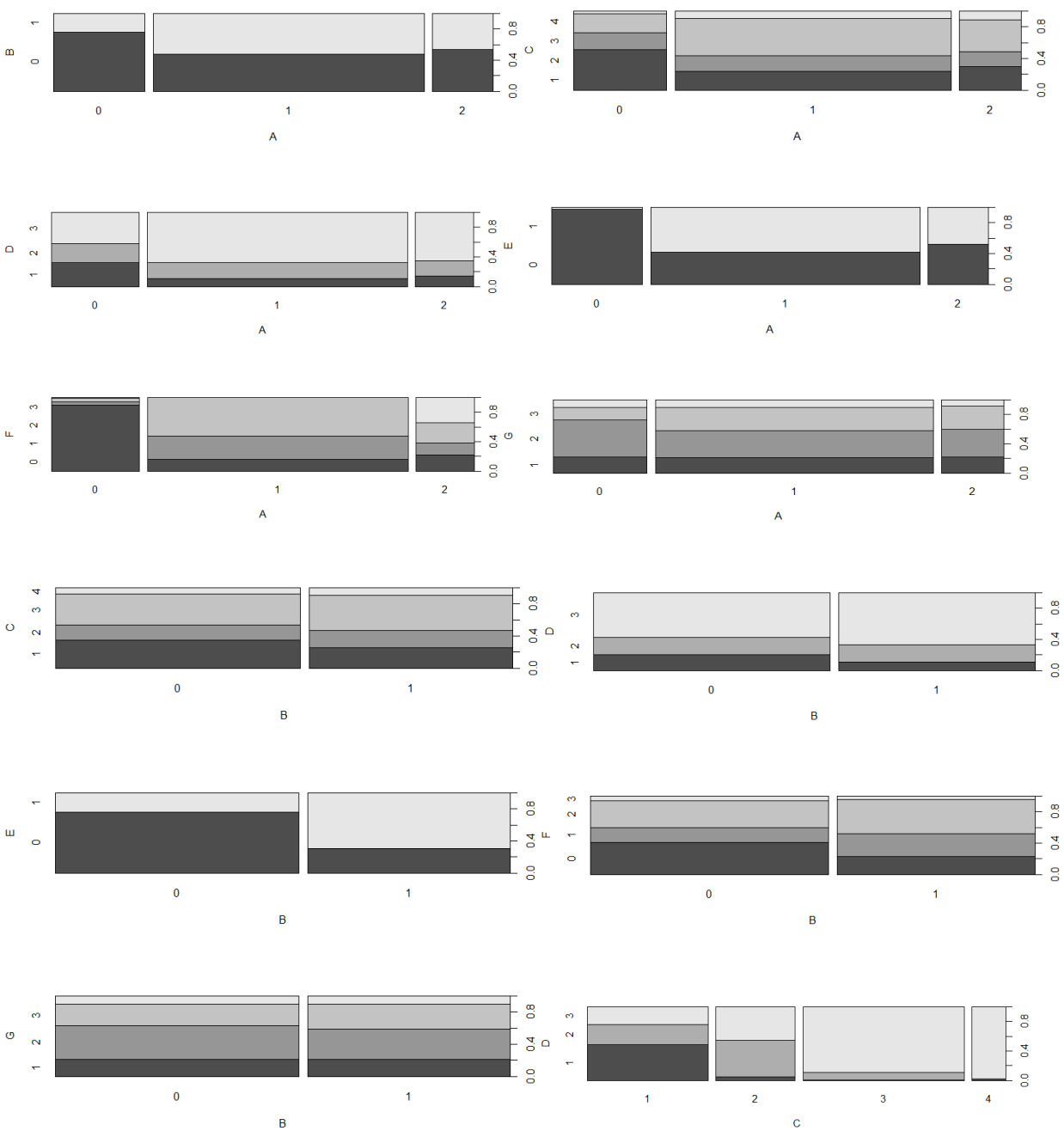
- **Duration_previous:** It is how long (in years) the customer was covered by their previous issuer. For missing values for a specific customer, if customer has a Duration_previous value in other row, I picked that value. If a customer has Duration_previous missing for all the entries, I imputed the value to 0 assuming it's a new customer.

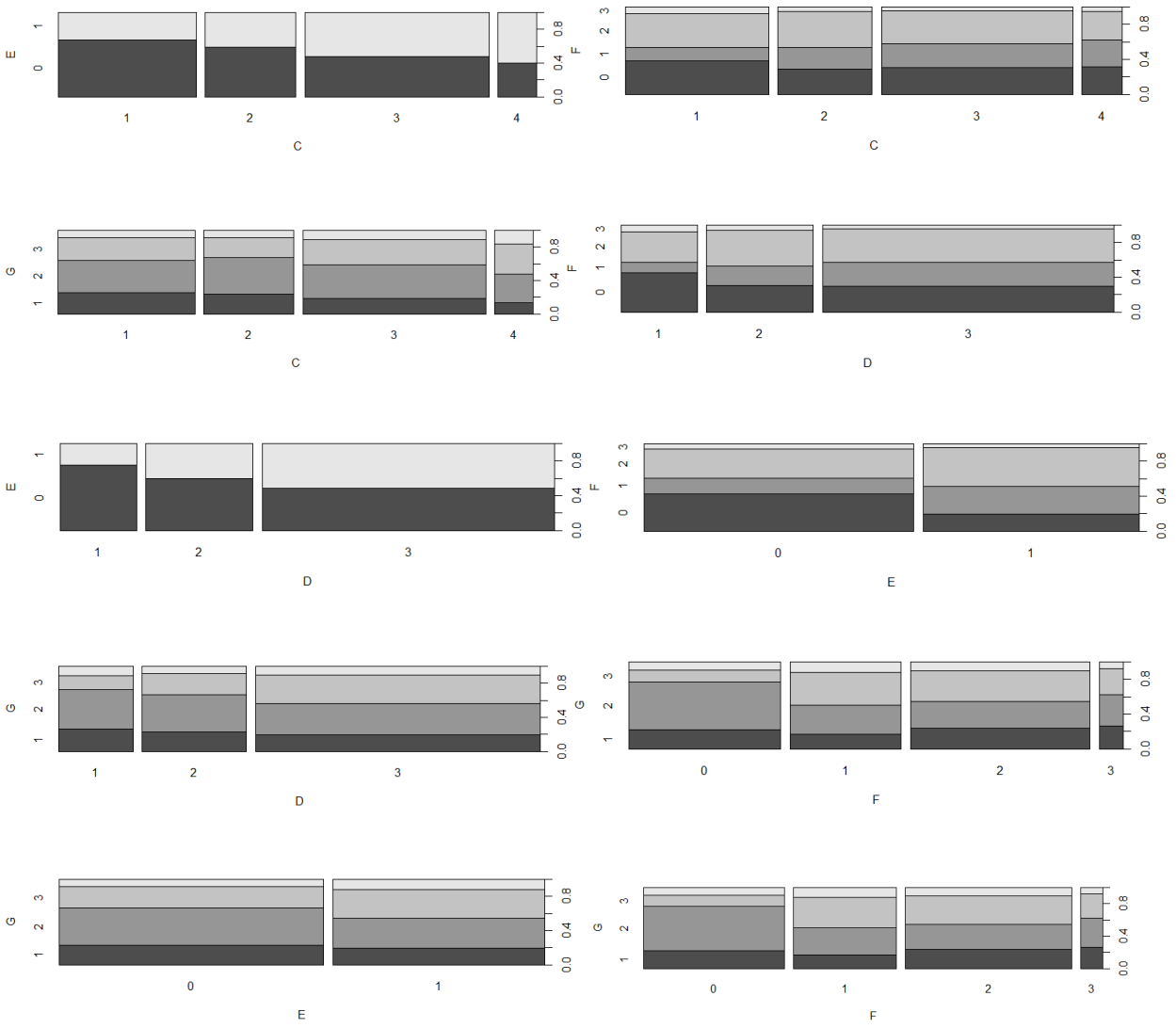
- **risk_factor**: About 36% data was missing from training set and about 38% data missing from test set. This parameter cannot be ignored as it can significantly impact insurance option a customer goes for. Assuming that a customer with specific set of attributes (homeowner + car_age + car_value + age_youngest + married_couple+age_oldest) has a specific risk factor, I predicted risk factor for missing values using a linear regression model.
- **location**: All the location values lie between 10001 and 16581. I decided to replace missing location values with mean of these values.

Data Exploration

- **Analyzing dependency in between plans**

I analyzed the relation in between different plans from A through G to see if there is any strong correlation between them. Some of the graphs are as follows:


B
C
A
0
1
2
D
E
F
G
C
D
B
E
F
G
D
C
3
4



We can see that

(a) when A=0, there is strong possibility of F=0, E=0, and B=0
(b) when A=1, there is strong possibility of D=3
(c) when B=1, there is strong possibility of D=3 and E=1
(d) when C=3 or C=4, there is strong possibility of D=3

- **Analyzing change in plan considered by the customer**

On analyzing the training set, I realized that for about 50% cases, the customer's final purchase did not change from the previous code. So, the next thing I analyzed were the cases in which customer's quote changed. Then, I began analyzing the parameters which were most significant in influencing customer to change the quote.

Approach Towards Creating Predictive Model

(Note: due to lack of time I could write R code till data exploration point, which will be submitted with this paper. However, I am presenting the approach that I would have employed to create my predictive model)

Step 1. Since, for about half of the training data the customer finally went for the quote previously considered, the hypothesis will simply replicate the last option considered by the customer for prediction. Let's call it prediction1.

Step 2. I analyzed some strong correlation among plans A through G. So, next step will be to sequentially change the values of C, B and A in prediction1 based on other options. This will generate prediction2.

Step 3. The last step will be to create to use support vector machine(SVM) ML algorithm to predict values A Though G. The predictors used to derive the hypothesis will be the significant one that I identified during data exploration. I decided to go for SVM because the size of dataset is very large and the relationship of output with predictor was found to be non-linear. The SVM algorithm works very efficiently in these scenarios.