Policy Purchasing

A look on predicting options chosen based on a user's history

Lab Group: DS3 Group 10

Team Members:
Chopra Dhruv, Krithika Jayaraman Karthikeyan, Louis Wirja, Neha Ramesh

Dataset at a Glance

Dataset from Kaggle : "Allstate Purchase Prediction Challenge" by Allstate Insurance
Source: https://www.kaggle.com/c/allstate-purchase-prediction-challenge/data (requires login)

	customer_ID	shopping_pt	record_type	day	state	location	group_size	homeowner	car_age	car_value	...	C_previous	duration_previous	A	B	C	D	E	F	G	cost
0	10000000	1	0	0	IN	10001	2	0	2	g	...	1.0	2.0	1	0	2	2	1	2	2	633
1	10000000	2	0	0	IN	10001	2	0	2	g	...	1.0	2.0	1	0	2	2	1	2	1	630
2	10000000	3	0	0	IN	10001	2	0	2	g	...	1.0	2.0	1	0	2	2	1	2	1	630
3	10000000	4	0	0	IN	10001	2	0	2	g	...	1.0	2.0	1	0	2	2	1	2	1	630
4	10000000	5	0	0	IN	10001	2	0	2	g	...	1.0	2.0	1	0	2	2	1	2	1	630

Columns correspond to a customer’s characteristics and the policy coverage options.

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

Objectives

1. Predicting the price a customer has to pay using a **Regression** model.
2. Predicting the policy coverage options purchased by a customer based on their characteristics and history using **Random Forests**.

Exploratory Analysis

Statistics, Observations and Inferences

Data Cleaning

Rows of data with NaN values are removed from the dataset.

```
data.dropna(subset=["car_value","C_previous","duration_previous","risk_factor"],inplace=True)
```

We observe that we have large number of NAN values in risk_factor column. Later on, we'll find that risk_factor is a very important variable in determining what policy the customer will be purchasing and what the cost of that policy will be. Thus, it would be wrong to blindly fill in the missing values with the median as that will dilute the relationship of the risk_factor with other variables.

Encoding

Categorical variables are encoded using OneHotEncoding and LabelEncoding to give them numerical values.

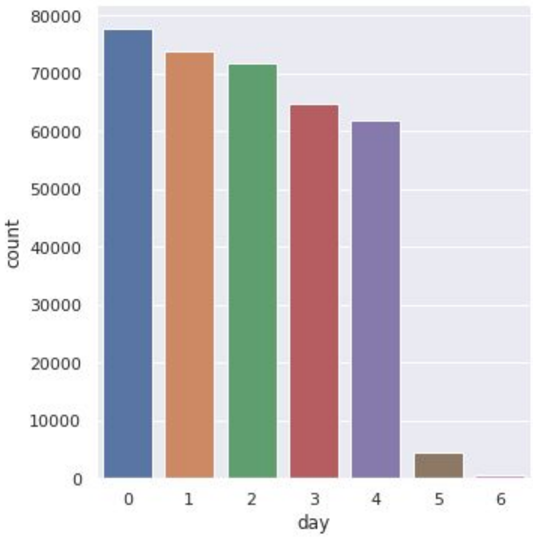
```
hot = OneHotEncoder()
```

```
hot.fit(data[["state"]])
```

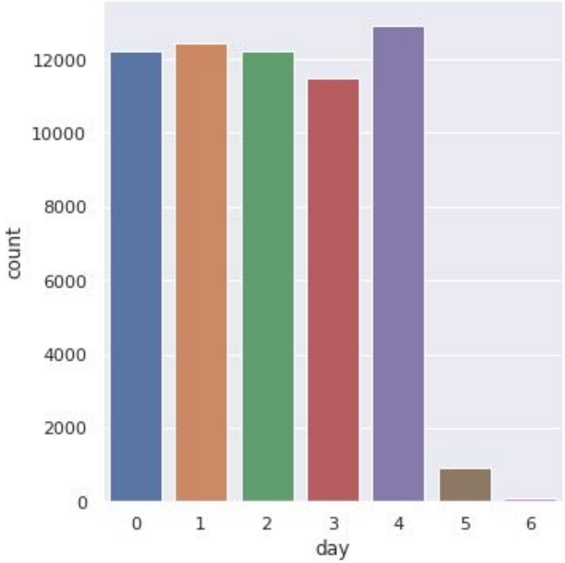
```
OneHotEncoder()
```

```
newstate = pd.DataFrame(hot.transform(data[["state"]]).toarray(), columns=hot.get_feature_names())
```

Days of Viewing and Purchase

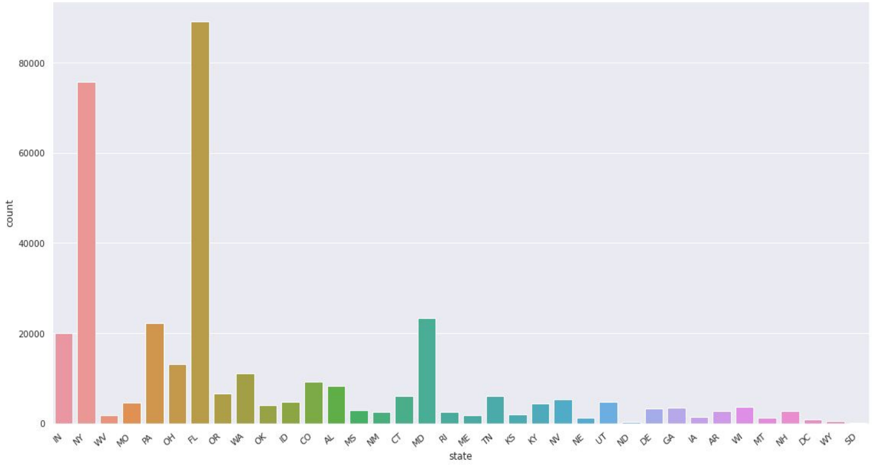


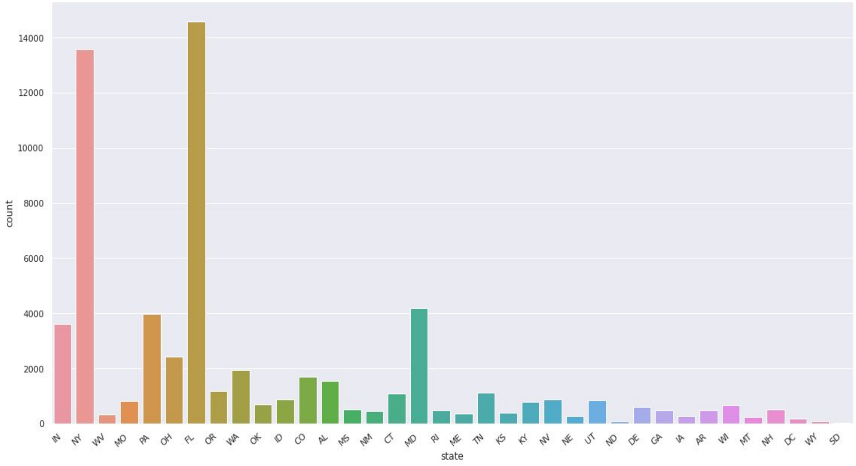
Number of viewings on specific days



Number of purchases on specific days

Location of Viewing and Purchase

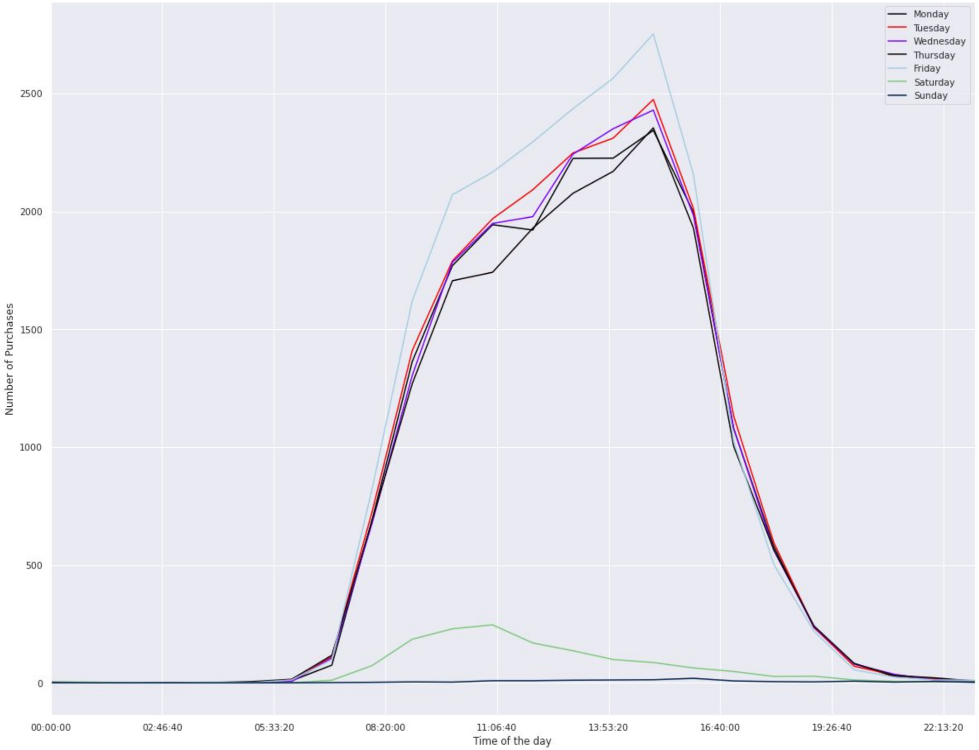




Number of viewings and purchases in different states of the U.S.

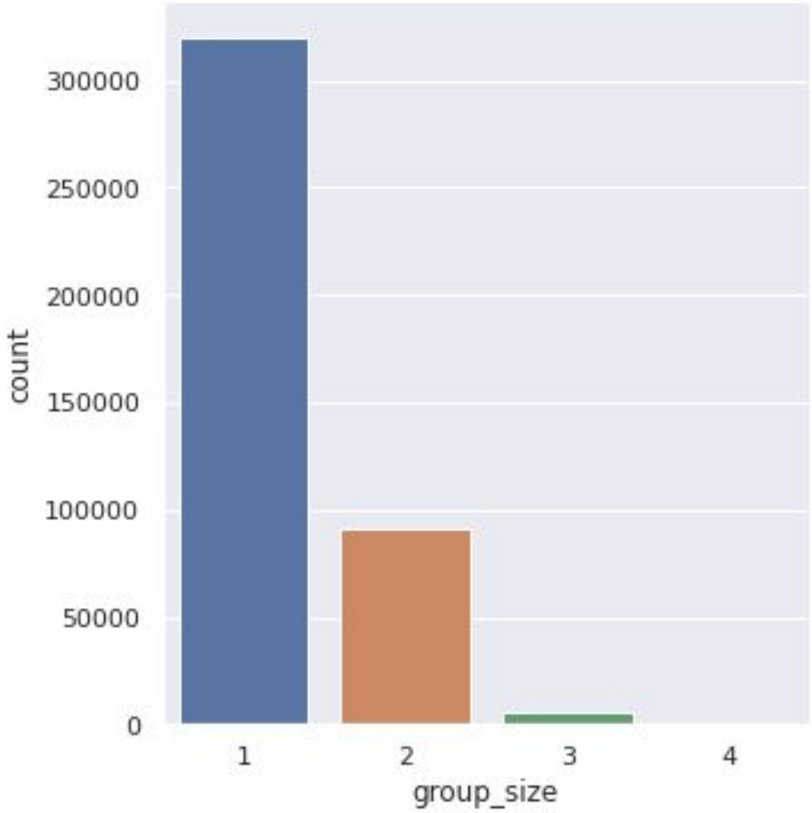
Timeframe of Purchase

A weekly timeframe showing the general trend in purchasing times.

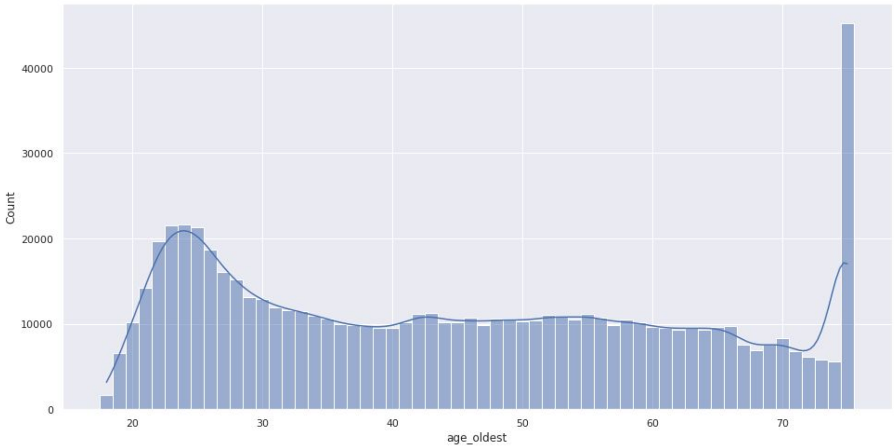


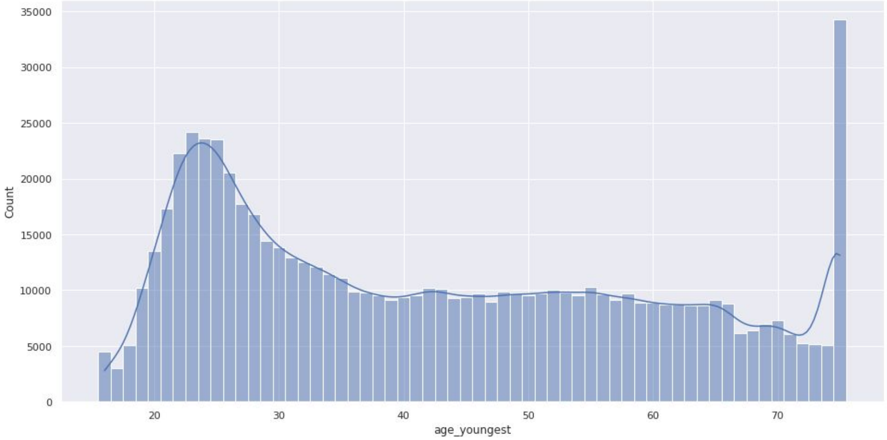
Group Size

Number of people
covered under the policy



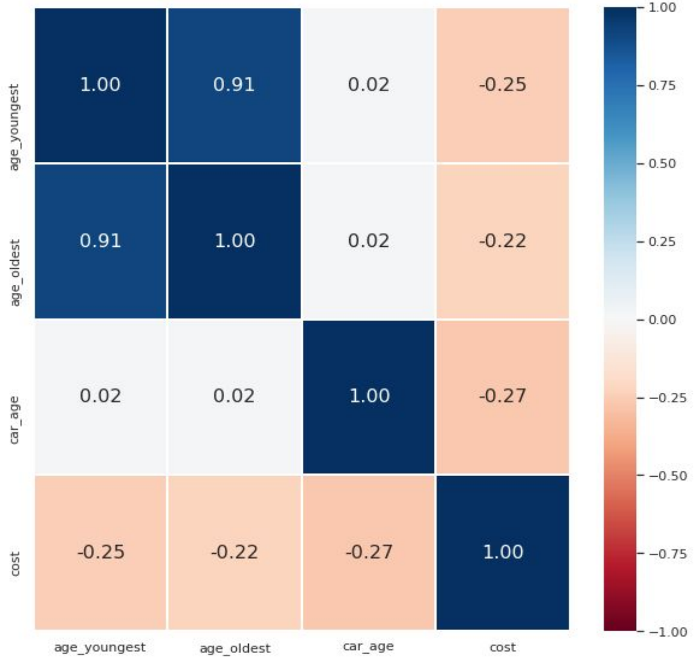
Customer's Age



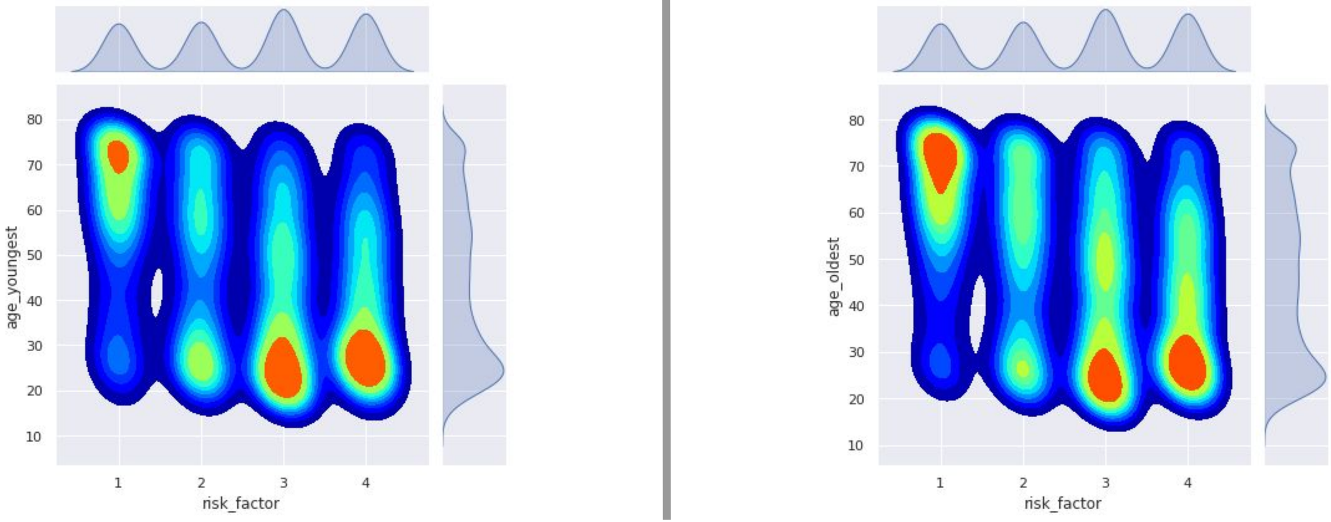


Correlation Between Variables

A heatmap is plotted between the numerical variables to analyse the correlation between each variable.

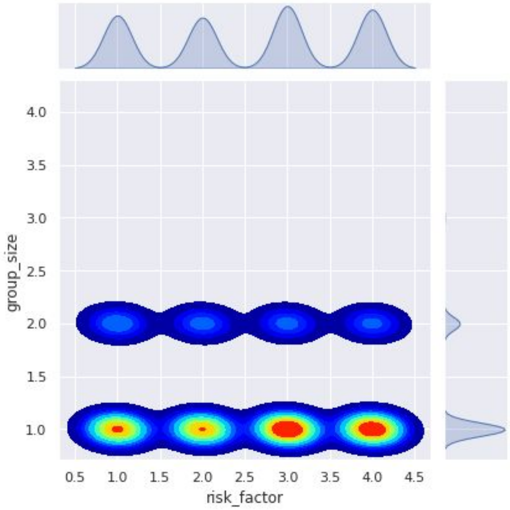


Riskiness of Customers

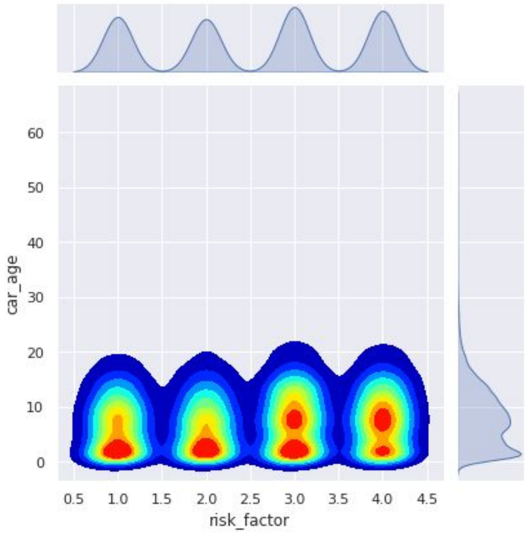


Density plots of the ages of customers and their values of risk

Risk Factor

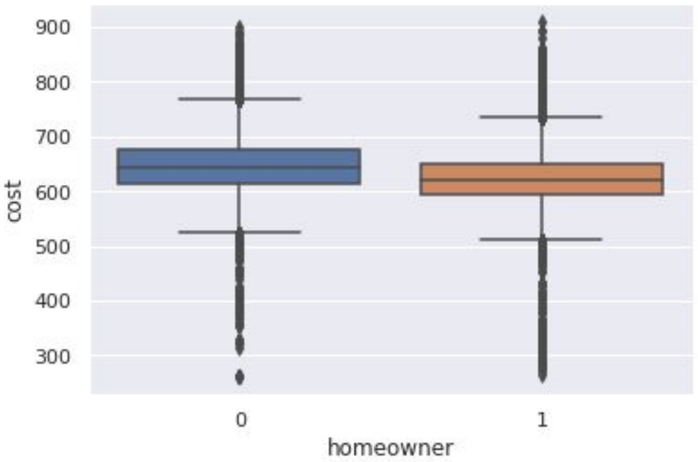


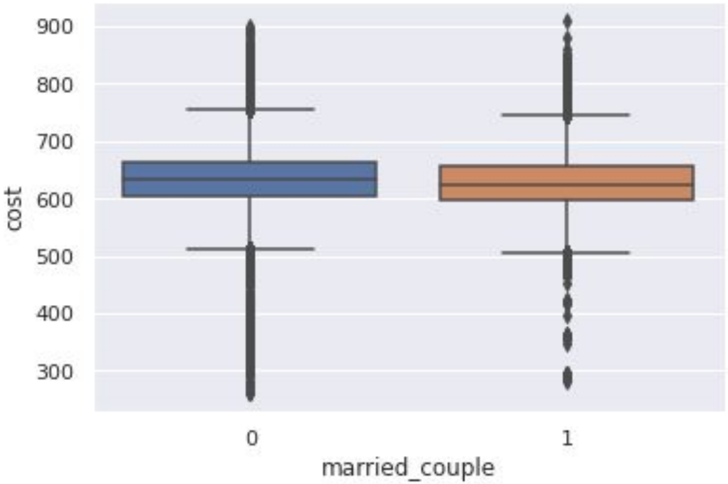
Risk factor among different group size of customers.



Car age compared with the customers’ risk factor.

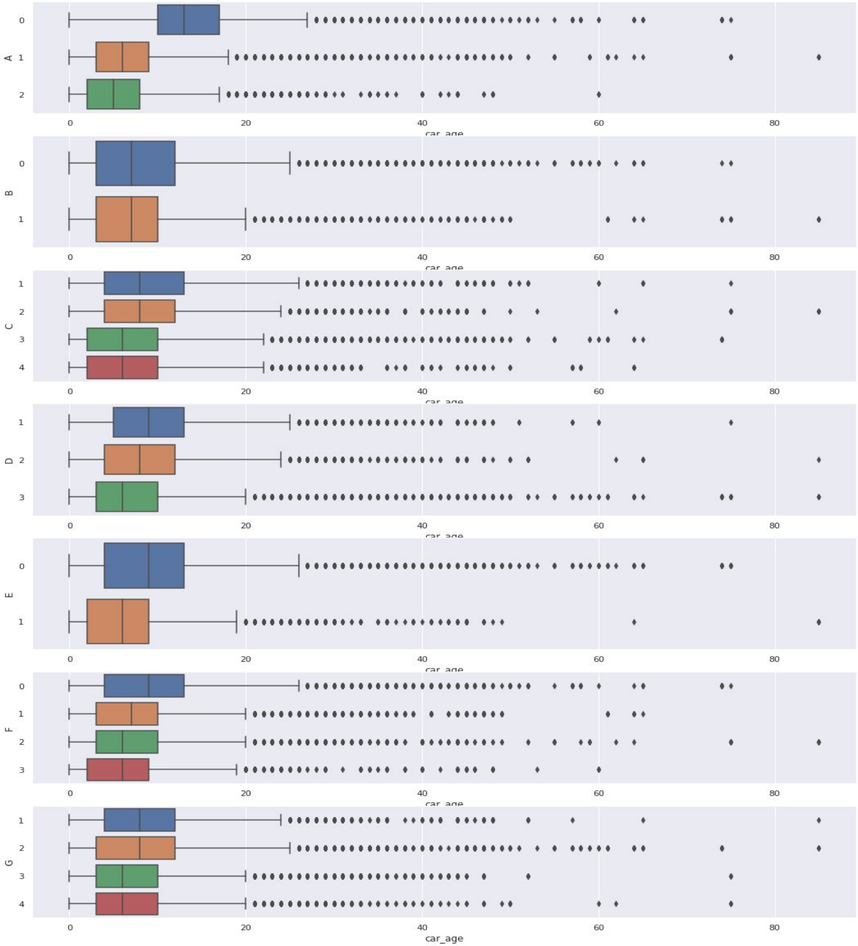
Cost of Policies Purchased





Policy Coverage Options

Different policy coverage options purchased in relation with the customer’s car age.



Modelling

Regression Analysis

Creating a Model for Cost

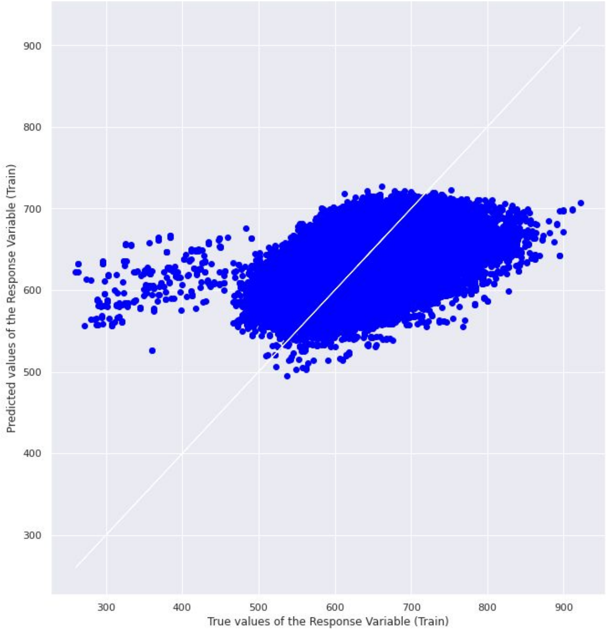
Using a Regression model, we want to predict how much a customer has to pay based on their purchased options and their characteristics.

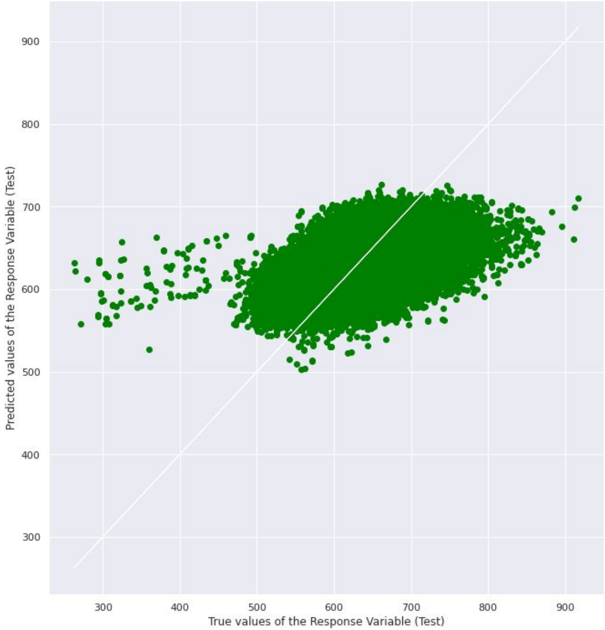
customer_ID	cost
10000000	633
10000005	630
10000007	630
10000013	630
10000014	630

Initial Linear Model

```
Goodness of Fit of Model        Train Dataset
Explained Variance (R^2)        : 0.4095468292336204
Mean Squared Error (MSE)        : 1249.4500978975996

Goodness of Fit of Model        Test Dataset
Explained Variance (R^2)        : 0.40925251749311664
Mean Squared Error (MSE)        : 1248.5005277652797
```





Synergy Variables

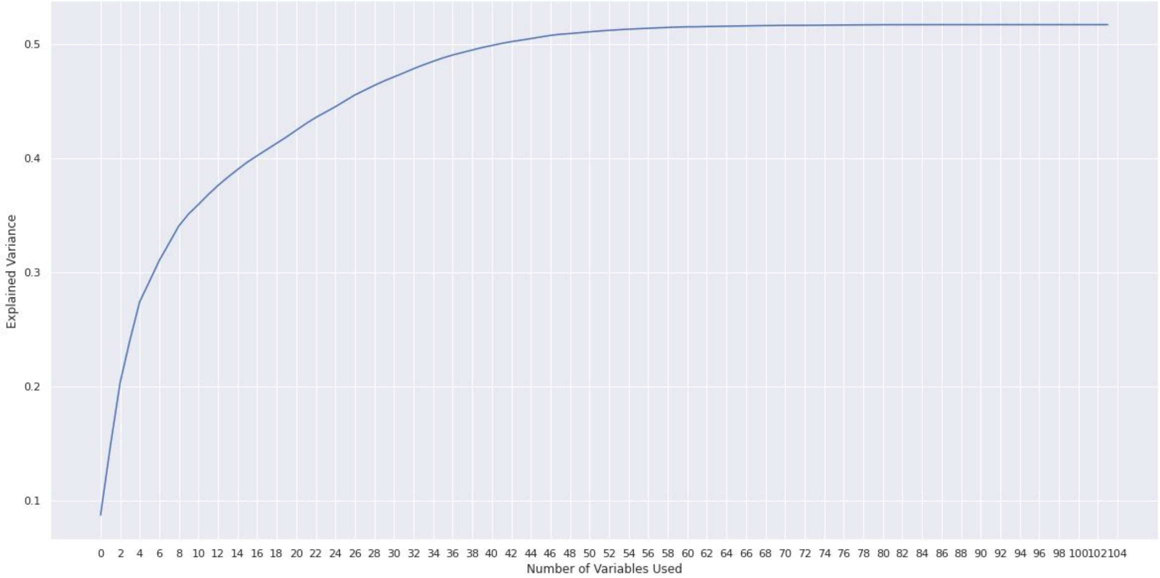
Before creating the final model, extra variables were made in order to increase the accuracy of the model.

The synergy variables consist of the combination, squares, and square roots of the original variables.

10	shopping_pt-group_size	10000	non-null	int64
11	shopping_pt-age_oldest	10000	non-null	int64
12	shopping_pt-age_youngest	10000	non-null	int64
13	shopping_pt-C_previous	10000	non-null	float64
14	group_size-age_oldest	10000	non-null	int64
15	group_size-age_youngest	10000	non-null	int64
16	group_size-duration_previous	10000	non-null	float64
17	car_age-car_age	10000	non-null	int64
18	car_age-risk_factor	10000	non-null	float64
19	car_age-age_oldest	10000	non-null	int64
20	risk_factor-risk_factor	10000	non-null	float64
21	age_oldest-age_oldest	10000	non-null	int64
22	age_oldest-age_youngest	10000	non-null	int64
23	C_previous-duration_previous	10000	non-null	float64
24	duration_previous-duration_previous	10000	non-null	float64
25	sqrt-car_age	10000	non-null	float64
26	sqrt-risk_factor	10000	non-null	float64
27	sqrt-age_oldest	10000	non-null	float64
28	sqrt-age_youngest	10000	non-null	float64
29	sqrt-C_previous	10000	non-null	float64
30	sqrt-duration_previous	10000	non-null	float64

Removal of Predictors

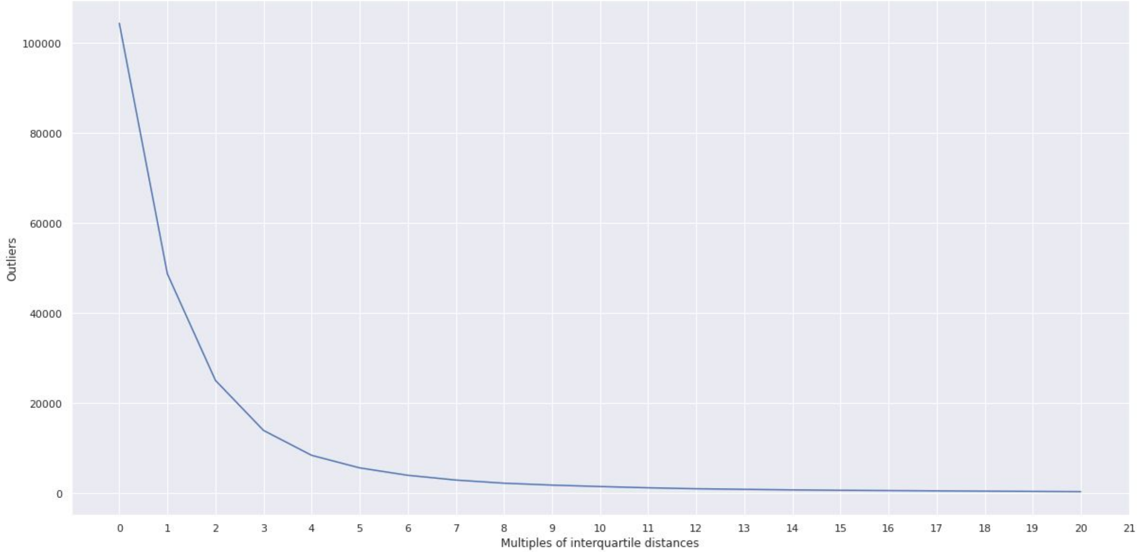
Predictors that did not play a significant role in the prediction of the response variable were removed from the linear regression model.



Removal of Outliers

Outliers of the dataset were removed.

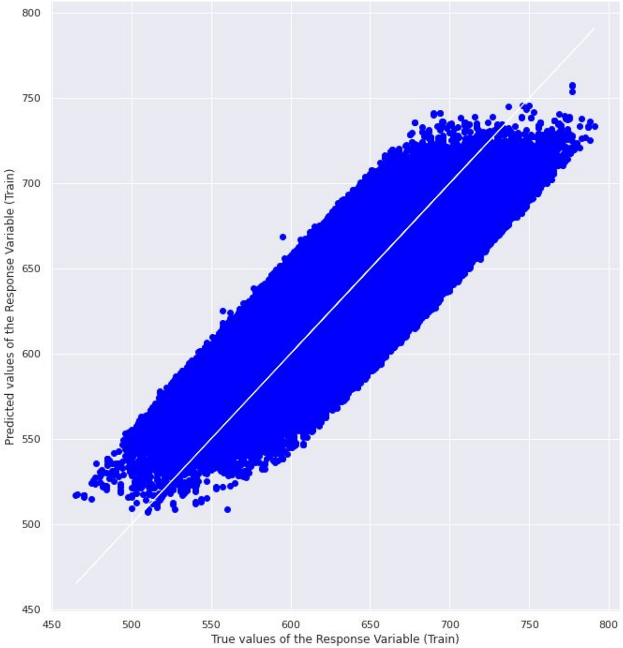
Approximately 6.82% of the data was discarded.

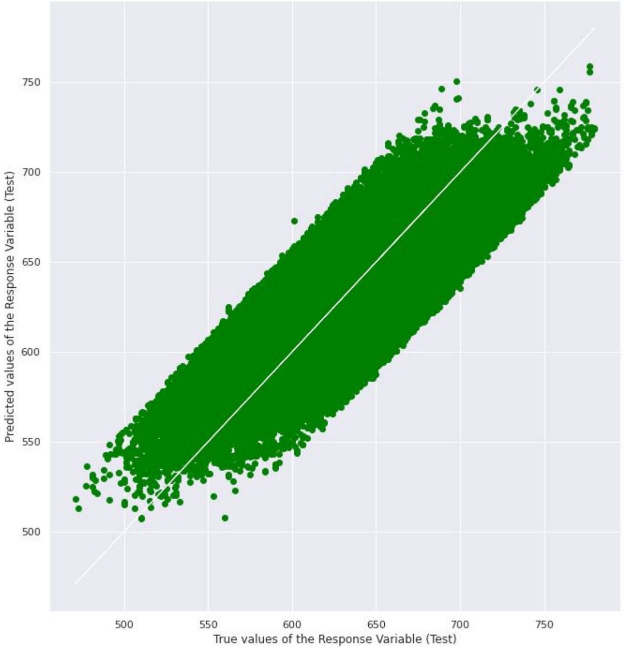


Final Model

```
Goodness of Fit of Model        Train Dataset
Explained Variance (R^2)        : 0.6205012683587205
Mean Squared Error (MSE)        : 664.415946121879

Goodness of Fit of Model        Test Dataset
Explained Variance (R^2)        : 0.6187188823271186
Mean Squared Error (MSE)        : 665.0758580586325
```





Classification

Random Forests
Chi-squared Tests

Coverage Options

A customer can choose a policy made up of components A,B,... G with each of the 7 different policy coverage components to purchase

Objective: Predict how much of each a customer would buy.

	customer_ID	A	B	C	D	E	F	G
90854	10033689	1	0	3	3	1	1	3
166648	10061285	1	1	3	3	1	2	2
360592	10132248	0	0	2	2	0	0	1
236591	10086558	1	1	1	3	1	1	3
376648	10137915	1	0	1	3	0	1	3

Prediction on concatenated strings?

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

=

Concat
1033113
1133122
0022001
1113113
1013013

?

Too many classes!

Dependent on Each Other?

The coverage options, A to G, may have some dependency on each other, e.g. A customer buying option A will also buy option G.

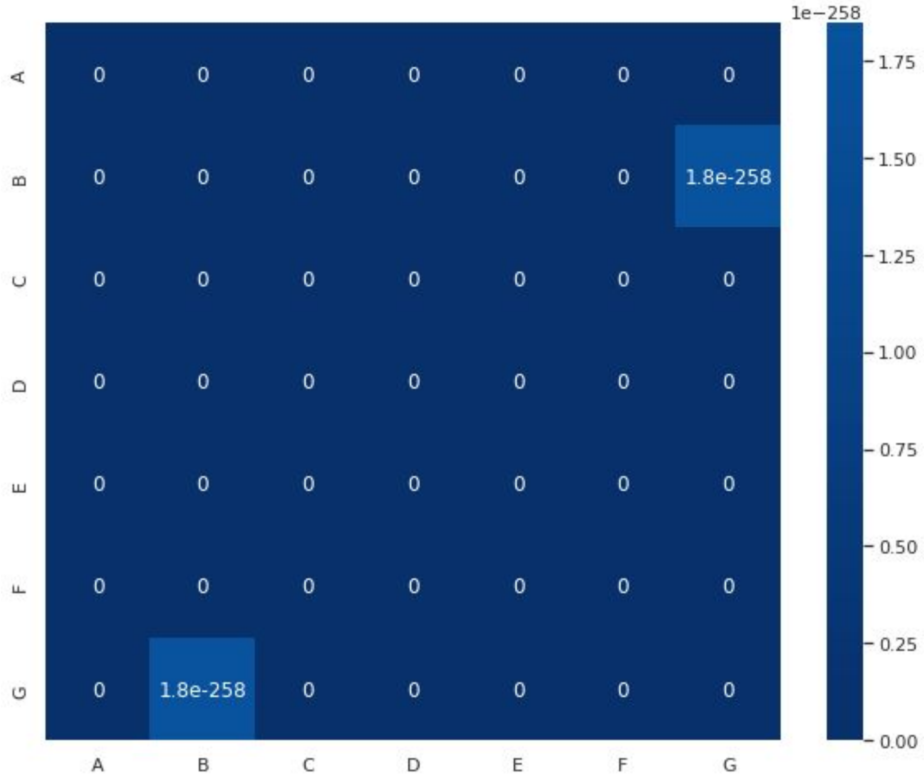
To test whether these variables are indeed dependent on each other, chi-squared tests are conducted on these variables.

Chi-squared Test

A statistical test that determines whether there is an association between two variables.

Based on the chi-squared tests, **all** of the coverage options are **dependent** on each other.

```
X^2 =  48435.68909064619
p =  0.0
Degrees of freedom =  9
Significance level = 0.010, p = 0.000000
Dependent (reject H0)
```



Heatmap of p-values of chi-squared tests

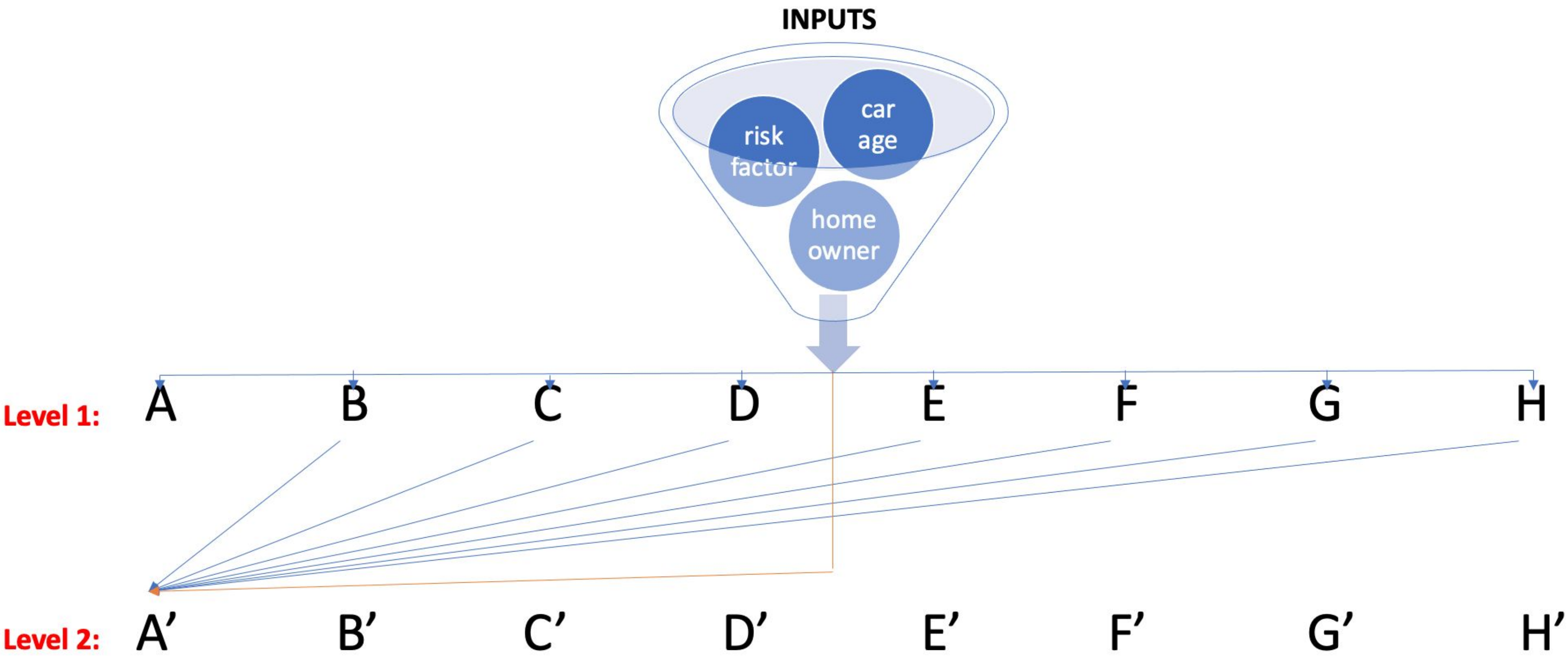
Approach

The coverage options A to G are first assumed to be independent of each other, i.e. they are not predictors of each other.

A random forest is created for each of the coverage options, a total of 7 forests.

The predicted values from 6 of the forests are then used to create another forest for each option, to implement the dependency of the variables, e.g. The predicted values of B to G are used as predictors in the prediction of A.

This iterative process can be repeated as many times as desired.


INPUTS
risk factor
car age
home owner
Level 1:
A
B
C
D
E
F
G
H
Level 2:
A’
B’
C’
D’
E’
F’
G’
H’

Random Forests

Unnamed: 0	A	B	C	D	E	F	G	car_value	shopping_pt	group_size	...	x0_4	x0_5	x0_6	A_new	B_new	C_new	D_new	E_new	F_new	G_new
25008	0	0	2	1	0	0	1	4	4	2	...	0.0	0.0	0.0	0	0	2	1	0	0	1
33231	2	1	3	3	1	0	2	4	9	1	...	0.0	0.0	0.0	2	1	3	3	1	0	1
159604	2	1	2	2	1	2	2	5	4	2	...	0.0	0.0	0.0	1	1	2	2	1	2	2
325205	0	1	2	2	0	0	2	4	6	1	...	0.0	0.0	0.0	0	0	2	2	0	0	2
409773	2	0	3	3	0	0	3	5	7	1	...	0.0	0.0	0.0	2	0	3	3	0	0	3

5 rows × 69 columns

```
A: RandomForestClassifier(max_depth=13, n_estimators=600) 0.7973487157591026
B: RandomForestClassifier(max_depth=14, n_estimators=500) 0.7150863584612732
C: RandomForestClassifier(max_depth=16, n_estimators=800) 0.8689082095947471
D: RandomForestClassifier(max_depth=16, n_estimators=1300) 0.8470955414380933
E: RandomForestClassifier(max_depth=17, n_estimators=900) 0.8249978671132023
F: RandomForestClassifier(max_depth=17, n_estimators=800) 0.8782283043165731
G: RandomForestClassifier(max_depth=17, n_estimators=800) 0.8484296711690641
```

Work Distribution

Exploratory Data Analysis: Krithika, Neha

Dataset Cleaning: Dhruv

Modelling: Dhruv

Classification: Dhruv, Louis

Presentation Slides: Dhruv, Louis, Krithika, Neha

Thank You!