Part 1:

Defining customer value seems like a straightforward concept, but there are challenges within the insurance which make the definition of customer value more challenging. Firstly, when pricing insurance, insurers are charging a premium for a product before actual costs of product are knonn. Data is analyzed from prior years to determine a premium that is appropriate for the risk, but there is still a level of uncertainty present because potential future events are being insured. Secondly, there are limitations on how the premium is charged from year to year. Third, there is not much of a difference when it comes to rate calculations between companies. These, plus other influences and considerations, have led to different ways of determining which customers are more valuable.

To address the issues in defining a valuable customer, there needs to be a definition of this value that considers all of the elements important in determining value, and one that can be consistently applied across an insurance company organization.

I would estimate the remaining value of a customer to be:

$$RV = (P_t - E_{loss} - Ex) * (1 - pc_t)/(1 + d)$$

$where, RV = remaining\ value,$

$P_t = Premium\ until\ time\ t$

$E_{loss} = Expected\ loss$

$Ex = Expected\ loss$

$pc_t = Probability\ of\ cancellation\ during\ time\ t$

$d = discount$

This formula takes into consideration a lot of factors like expected loss, expenses, and probability of cancellation, rather than just the premium and total costs. This can be improved by including the probability of renewal at the end of period t.

$$RV = R_t * (P_t - E_{loss} - Ex) * (1 - pc_t)/(1 + d)$$

$where, R_t = Probability\ of\ renewal\ after\ time\ t$

This measure of value can be extended for the potential customers as well. Once a potential customer receives a premium quote from an insurance company, they have the option of either accepting or rejecting that quote. So, the expected value of business that has been quoted is the adjusted value for the probability that the customer accepts the quote.

Once a consistent measure of customer value has been developed, then a company can begin incorporating that measure of value into its business processes. There are a number of ways that this value can be recognized, and the recognition of this value will vary based on business function. The company can use this information for:

1. Underwriting: Underwriting has focused on identifying customers with better than average expected claim frequencies. This gives a chance to the companies to evaluate risks involved.

2. Actuarial: Ensure that the risks are being priced at a level that allows the company to make a reasonable profit, if not avoid loss. The customer value helps determine short term and long term profits.

3. Marketing: The marketing department can incorporate the expected value of the customer in determination of target market. This will not only focus on customers that are likely to be attracted to the purchase, but also whether the purchased are likely to be profitable customers in the future.

4. Management: The product managers can use this information to evaluate the overall performance of the product. The managers can take courses of action that might impact the expected value.

There are many other applications that could be developed in other areas such as claims, and customer service that could help enhance customer value.

Part 2:

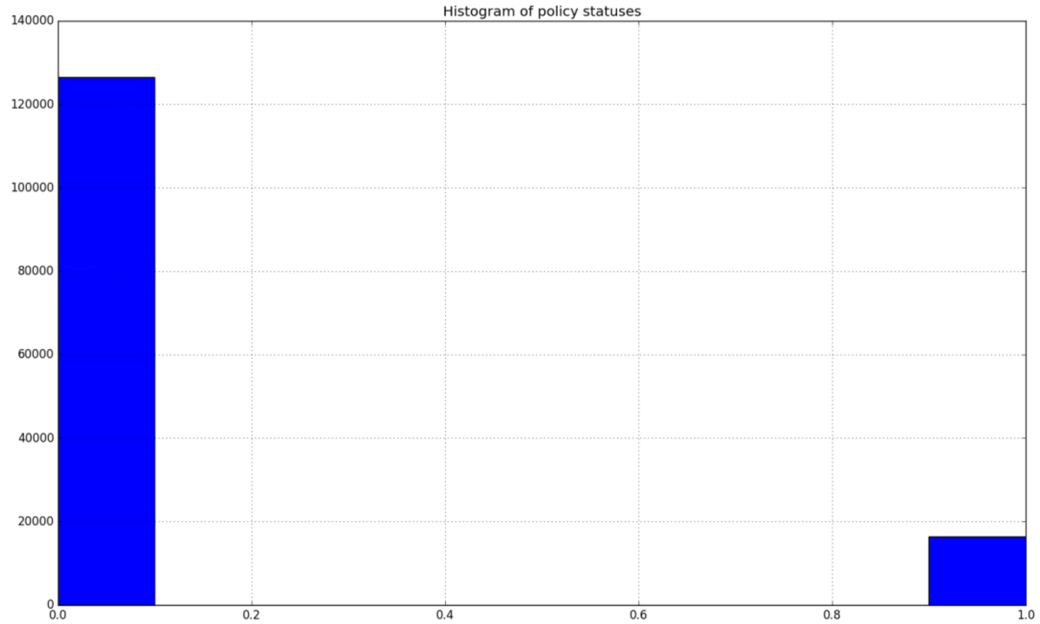
I chose to predict the future cancels; the output should be a probability of cancellation for each policy in the following month. Looking at the data, I felt like there were very few features given. So, I evaluated some features from the given features for better analysis. The approach that I took to come up with the solution that I came up with is as follows:

1. Data preprocessing: I loaded the data in both the given files into one data frame using pandas. Merging both the data sets resulted in a lot of NA values. I replaced all the claimed amount, and paid amount NAs to 0s. Then I replaced all the null cancel dates to a valid date for preprocessing. I converted the datetime format to (year: month) format. This makes it easier to calculate the number of months involved.

2. Feature extraction: I evaluated the number of months that each policy existed for. Using the number of months, I evaluated the total premium paid by the customer until the cancellation/present date. I merged these values and sorted them by the policy ID. I also added a feature which tells if the customer's policy is cancelled or not. The months along with the status of policy tells us how many months the customer has been with the company. Then I grouped all the columns by policy ID and Claim date and summed up the claimed amounts, paid amounts, and number of claims by monthly (Number of claims is a new feature that I introduced in here because the policy might get cancelled if there are too many claims). I then evaluated the value remaining of the customer at every month involving the claims made by the customer. I then evaluated the Paid to Claimed ratio (The customer may cancel his policy if the company does not cover a large amount of the claims).

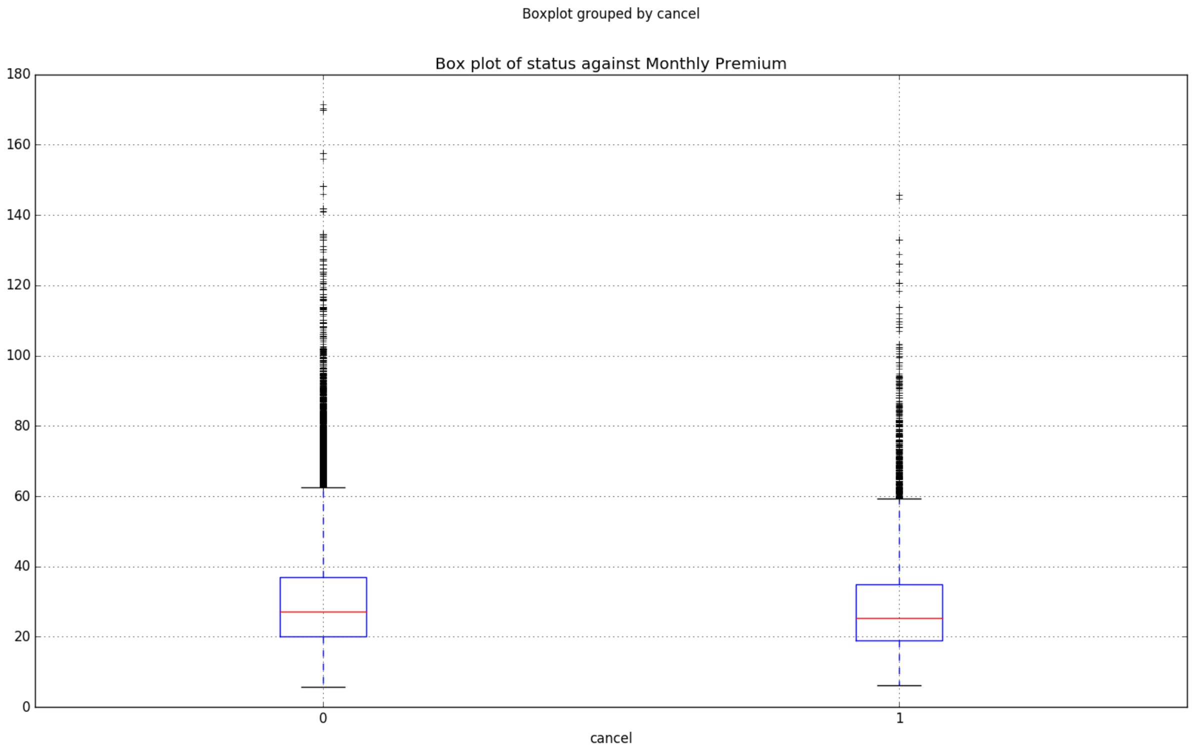
3. EDA: Then I performed Exploratory Data Analysis on the given data. EDA is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.

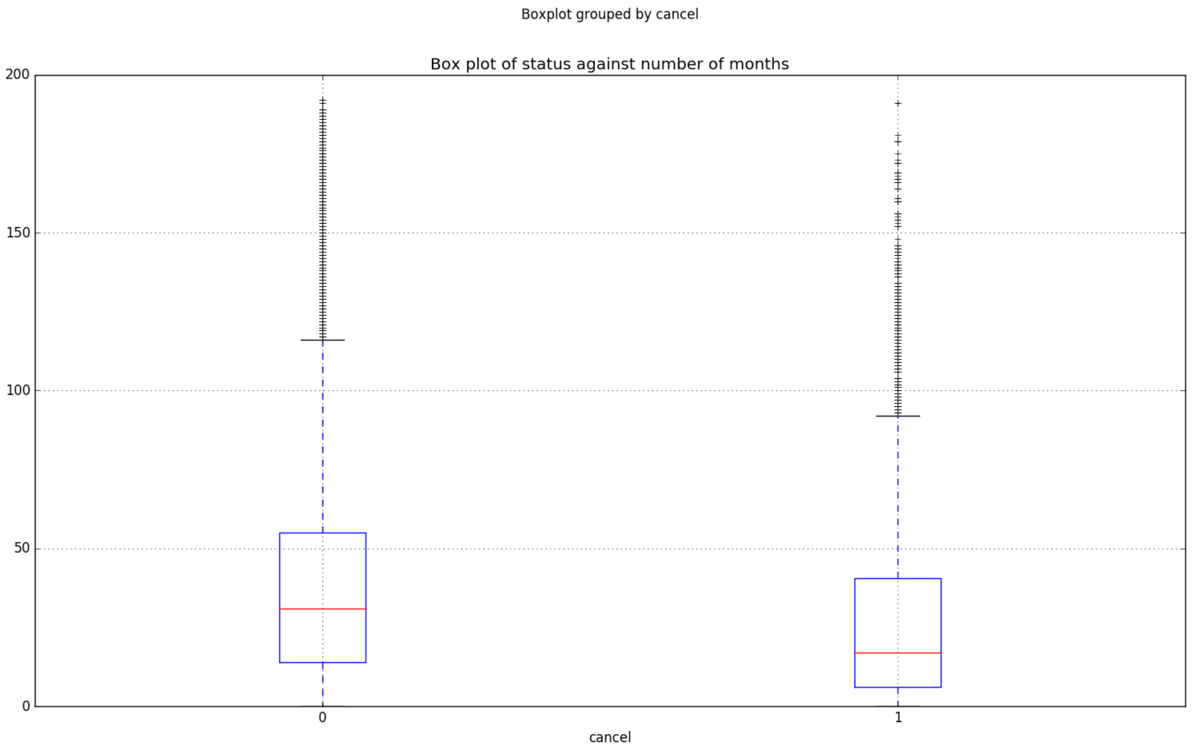
First, I looked at the number of cancelled and active policies to get an idea of what I am working with.



This histogram shows that the number of cancelled policies (1 for cancelled, 0 for active) is low compared to the total number of policies. This means that one need to be careful while predicting probabilities and classification of the policies. There is a possibility that we might get an accuracy of almost 90% if all the claims are predicted to be not cancelled. This is wrong because it predicts all the cancelled policies to be not cancelled which is a huge mistake.

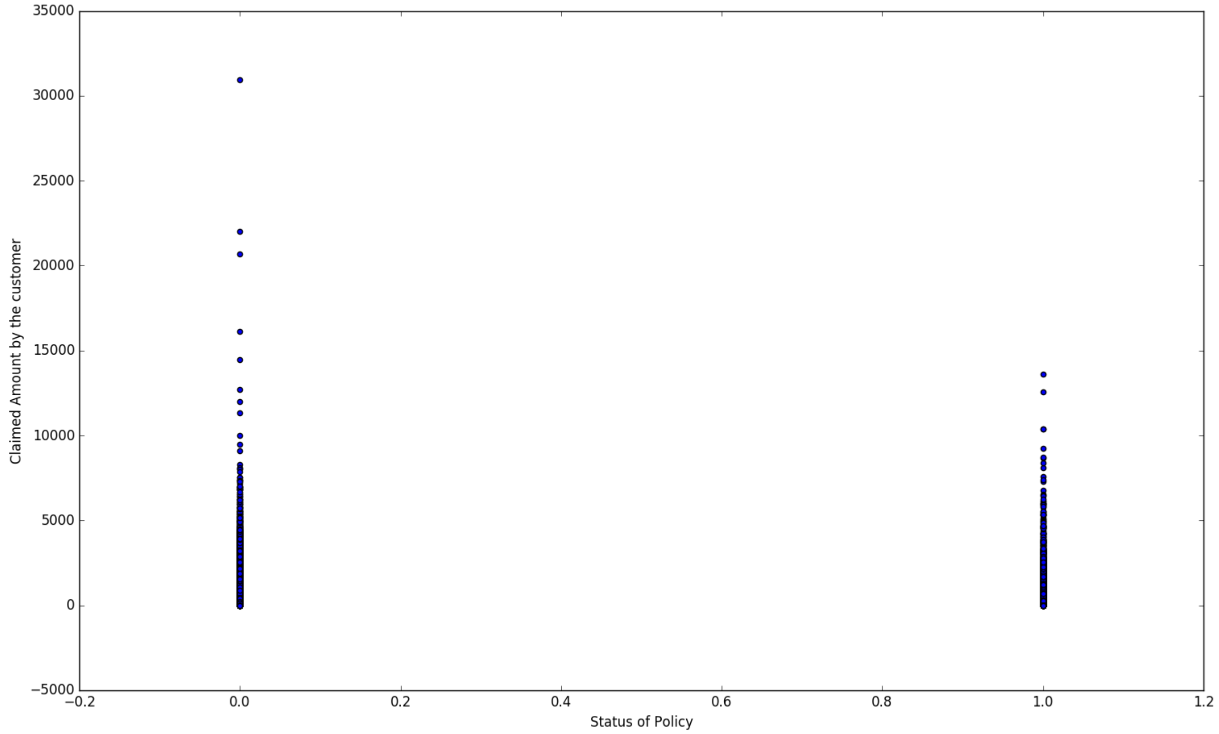
Then, I plotted the box plots of status of the policy against the monthly premium and the number of months of the policy to see if there is an effect on the status if monthly premium changes.


Boxplot grouped by cancel
Box plot of status against Monthly Premium
180
160
140
120
100
80
60
40
20
0
0
1
cancel

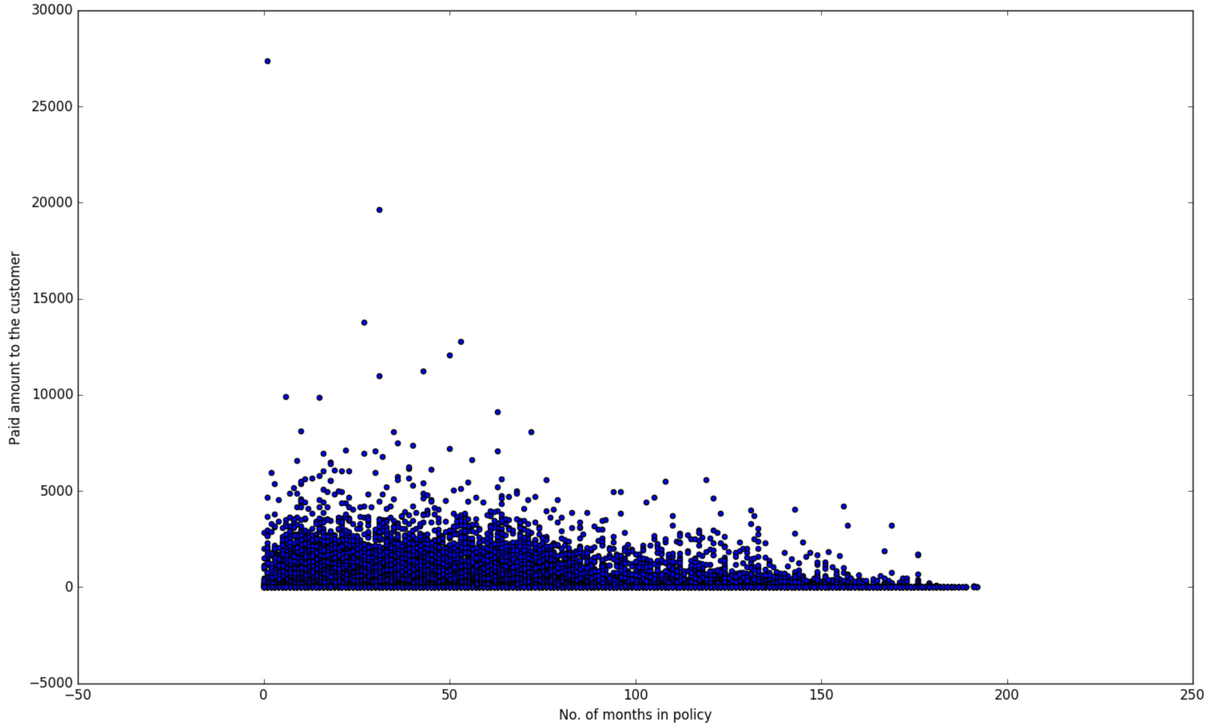

Boxplot grouped by cancel
Box plot of status against number of months
200
150
100
50
0
0
1
cancel

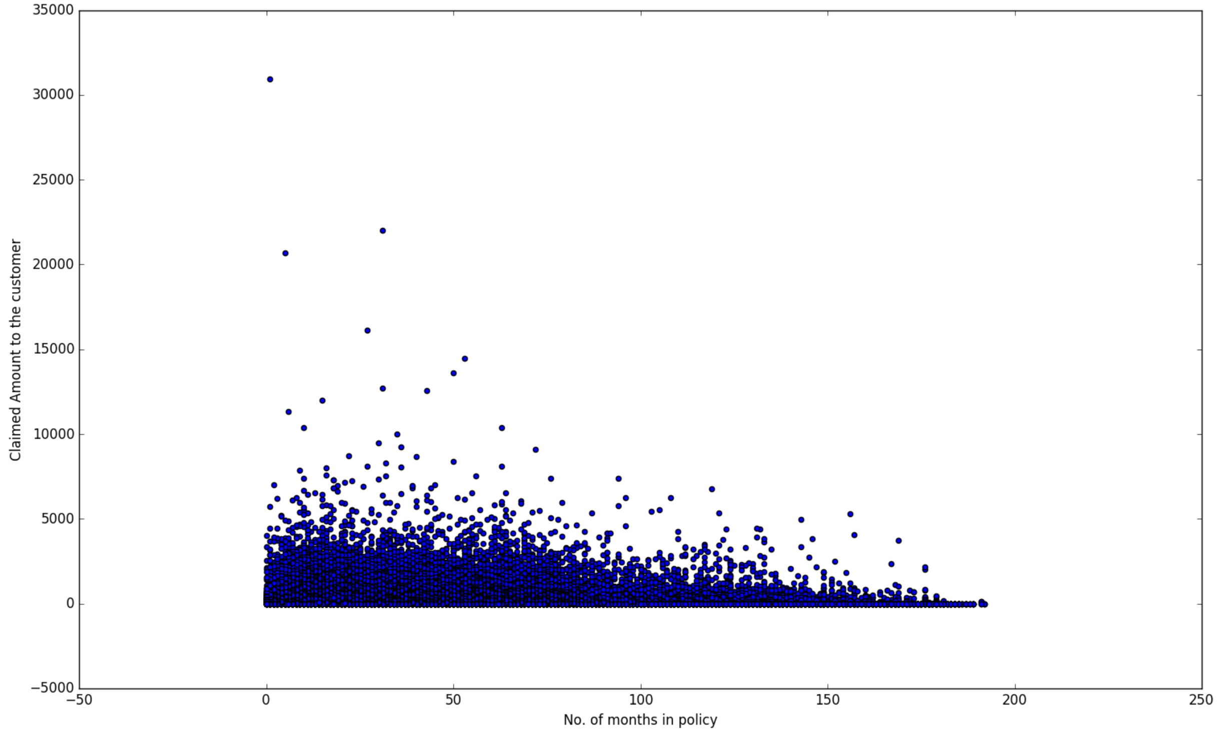
I found some very interesting results here. The mean of total number of months of policies which were cancelled are less than that of non-cancelled policies. This means that the policies being cancelled are during the early stages of the policies. It would be helpful if the reason of cancellation can be found out for further analysis.

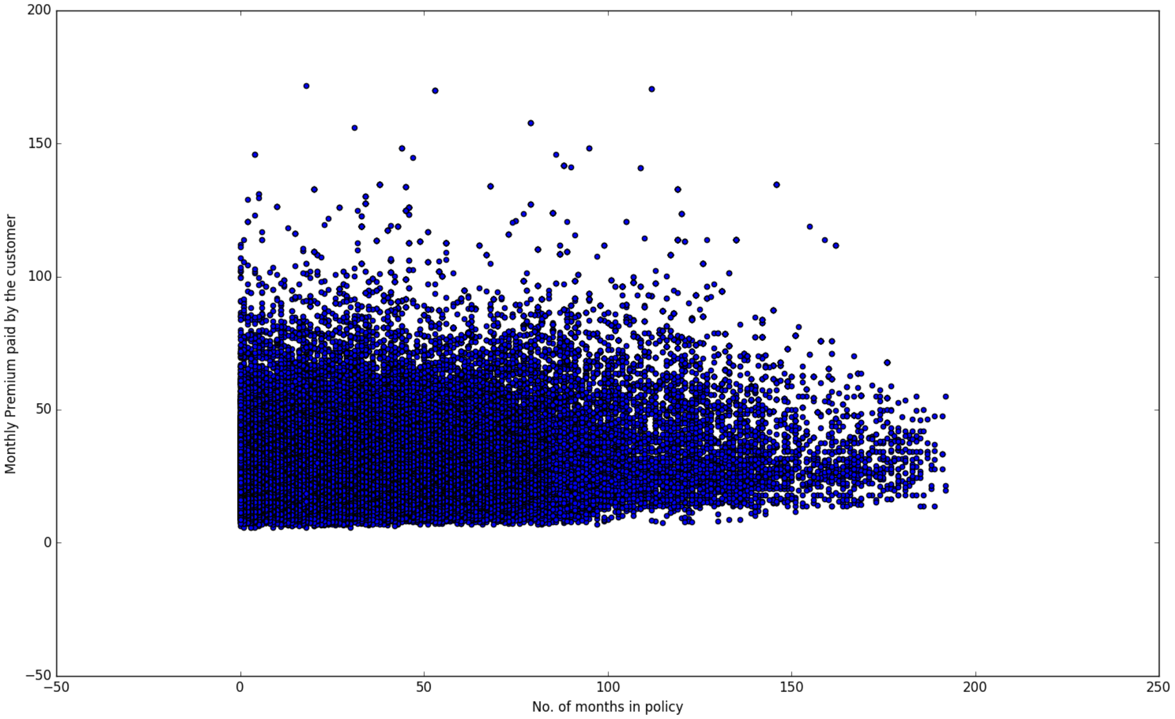
I also found out that the less the claimed amount, more is the probability that the policy is cancelled.

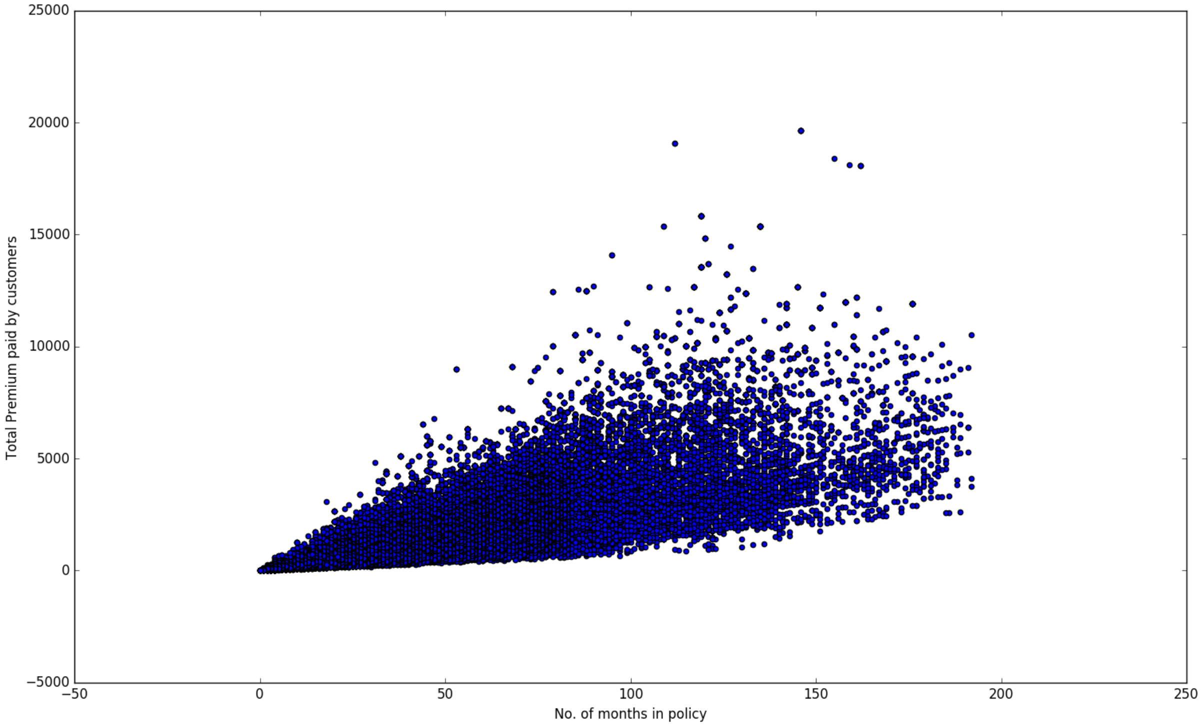


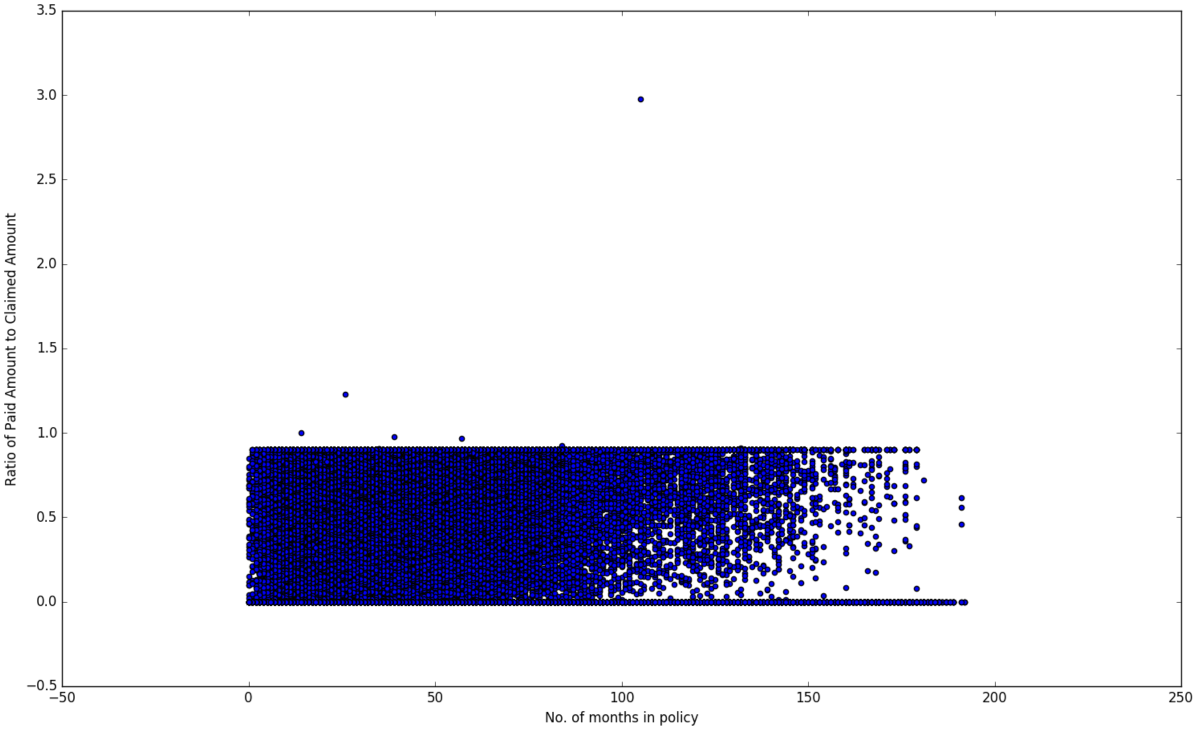
Then, I looked at the total number of months a policy existed against the monthly premium, total premium paid, paid amount to claimed amount ratio, paid amount, and claimed amount. The findings here are interesting as well.


30000
25000
20000
15000
10000
5000
0
−5000
−50
0
50
100
150
200
250
Paid amount to the customer
No. of months in policy


35000
30000
25000
20000
15000
10000
5000
0
−5000
−50
0
50
100
150
200
250
Claimed Amount to the customer
No. of months in policy

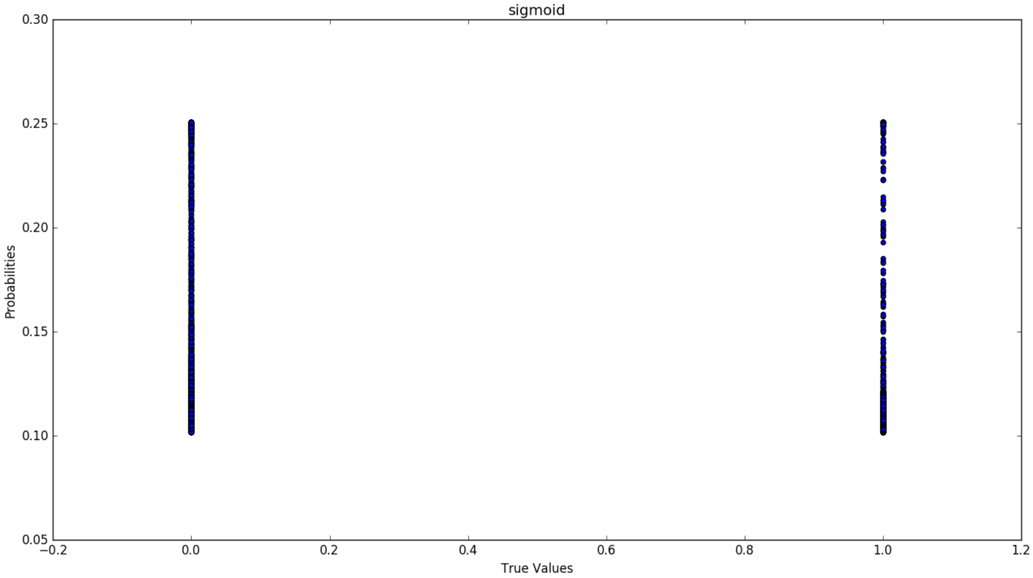

200
150
100
50
0
−50
Monthly Premium paid by the customer
−50
0
50
100
150
200
250
No. of months in policy


25000
20000
15000
10000
5000
0
−5000
Total Premium paid by customers
−50
0
50
100
150
200
250
No. of months in policy

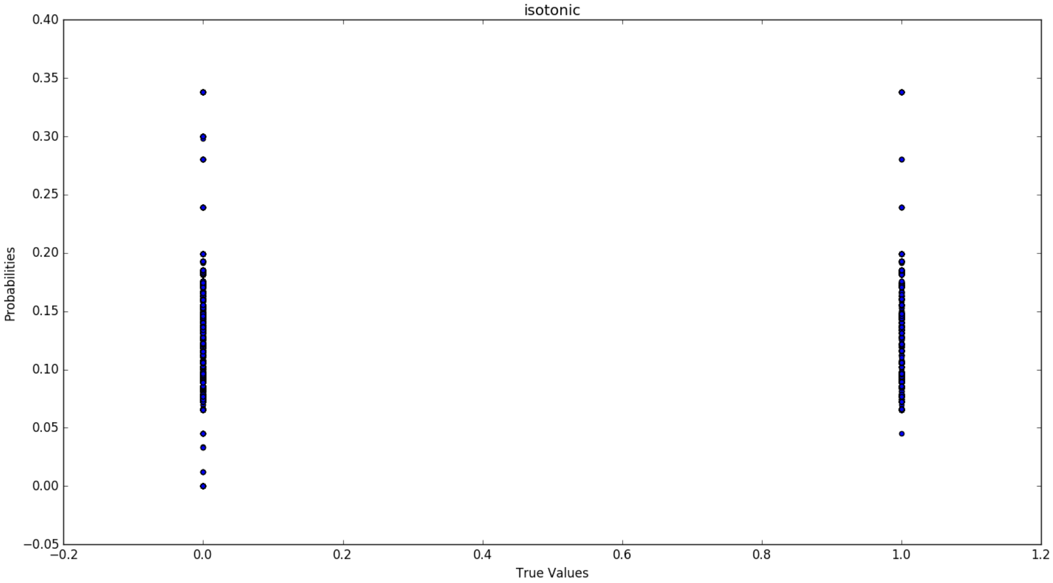


The amount of time a policy existed is inversely proportional to the monthly premium, claimed and paid amount. It is directly proportional to paid amount to claimed amount ratio and the total premium. This means that the probability that the policy getting cancelled is high if the claimed, paid amounts and the monthly premiums are high.

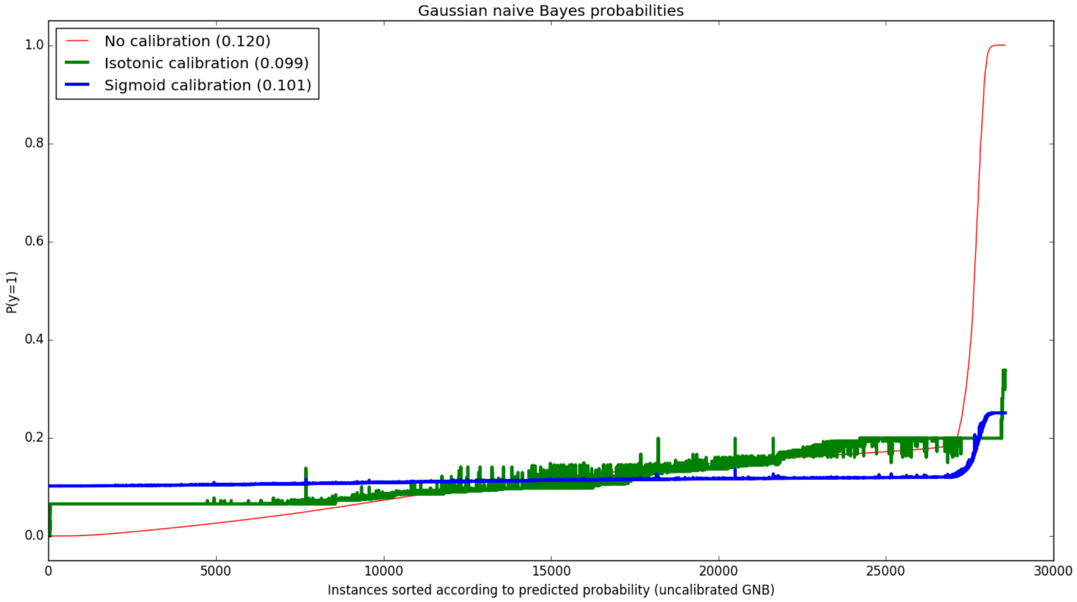
4. Train and test the data: I divided the data set into 80% training set and 20% testing set. Initially I used a regression model to predict probabilities. I expected the outputs to be between 0 and 1 but there were some negative values. I normalized the values and the output came to be around 48 percent. I then implemented a Gaussian Naive Bayes method to predict the probabilities of the test set. The accuracy improved to 67%.



Sigmoid Gaussian Probabilities

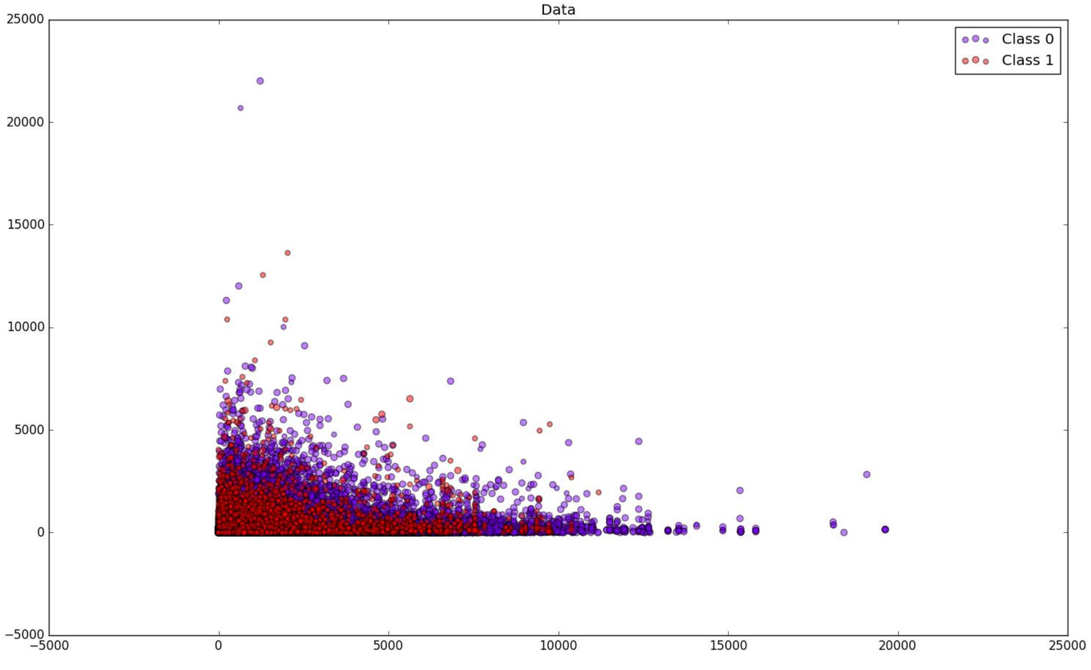


Isotonic Gaussian Probabilities



Comparision of Gaussian, Isotonic, and Sigmoid probabilities

Instances sorted according to predicted probability



Probabilities of the data

5. Output:

Output for train and test data (20% of the data):

Brier scores: (the smaller the better)

```
No calibration: 0.120
With isotonic calibration: 0.099
With sigmoid calibration: 0.101
Accuracy Score of Naïve Bayes: 0.595561467376
Accuracy Score of Naïve Bayes with isotonic calibration: 0.614743769252
Accuracy Score of Naïve Bayes with sigmoid calibration: 0.66182441893
```

The final test set is the entire data set with only Policy ID that are not cancelled. This is because we only need probabilities for the ones that are not cancelled. The Policy IDs and the respective output probabilities are saved to a csv file called 'probabilities.csv' along with the class.

P.S. The results can be improved a lot if given more features such as the reason of cancellation, type of insurance policy, area of residence, age of the pet, species of pet, breed of pet, background of the owner, and so on. These features will definitely improve the accuracy to over 85-90%.