Building A Model

I am building a predictive model using logistic regression, because it is widely used for target variables which give binary output as either True or False. Since we are analyzing or predicting the churn of customer which give the binary output either 0 for not-churn and 1 for churn, hence it is better to use Logistic regression.

Step 1: Split the data into train and valid;

```
*    Stratified the sample using surveyselect procedures;
proc surveyselect data=work.features_sort noprint samprate=0.66 out=features_start outall;
strata churn;
run;


*    creating training and validation dataset;
data features_train features_valid;
set work.features_start;
if selected then output features_train;
else output features_valid;
run;
```

I have used “proc surveyselect” procedures to split the data into training and valid.

The FREQ Procedure

Table of churn by Selected			
	Selected(Selection Indicator)		
churn	0	1	Total
0	11960 29.02 34.00 85.37	23219 56.34 66.00 85.36	35179 85.36
1	2050 4.97 33.99 14.63	3982 9.66 66.01 14.64	6032 14.64
Total	14010 34.00	27201 66.00	41211 100.00

With the help of “proc freq”, I have displayed how much of percentage of data has been splited. “Selected 1” is for Training, which has 66% of data and “Selected 0” is for validation, which has 34% of data.

Step 2: Train the Model

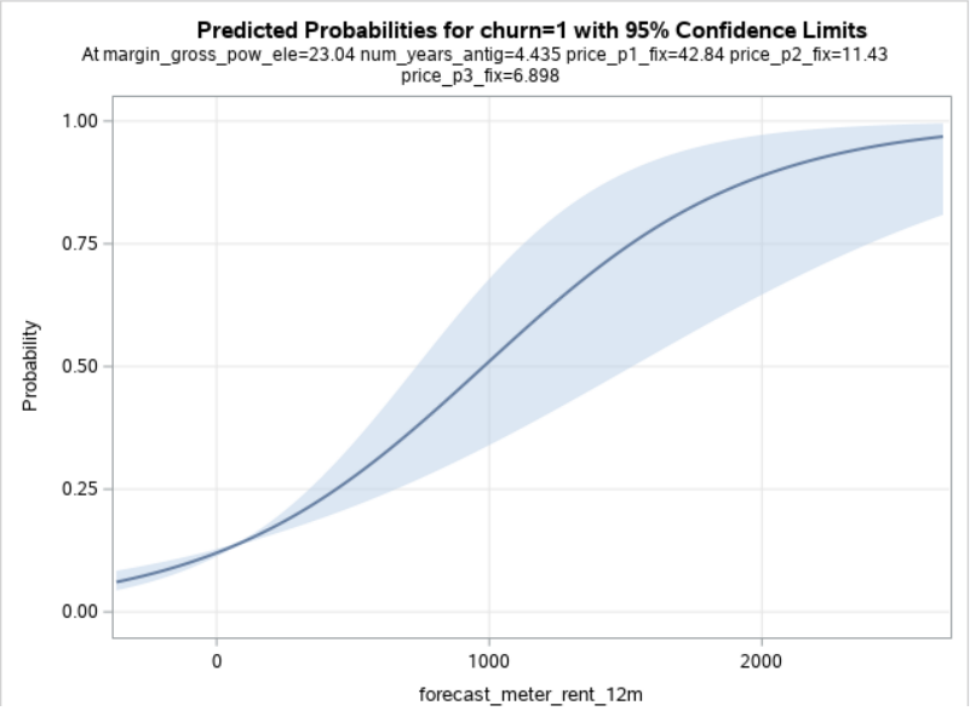
```
*   fitting the model using logistic regression;
proc logistic data=mydata.features_train plots (only maxpoints = none)=
                effect(clband x=(forecast_meter_rent_12m margin_gross_pow_ele num_years_antig
                price_p1_fix price_p2_fix price_p3_fix));
model churn(event='1') =forecast_meter_rent_12m margin_gross_pow_ele num_years_antig price_p1_fix
                        price_p2_fix price_p3_fix / stb clodds=pl slentry=0.5 slstay=0.1 selection=backward;

run;
```

This is very simple “logistic method” where we are using churn event 1, which represent customer leaving the company. I have also plot some graph between the churn (target variable) with other variables to show the relations between them.

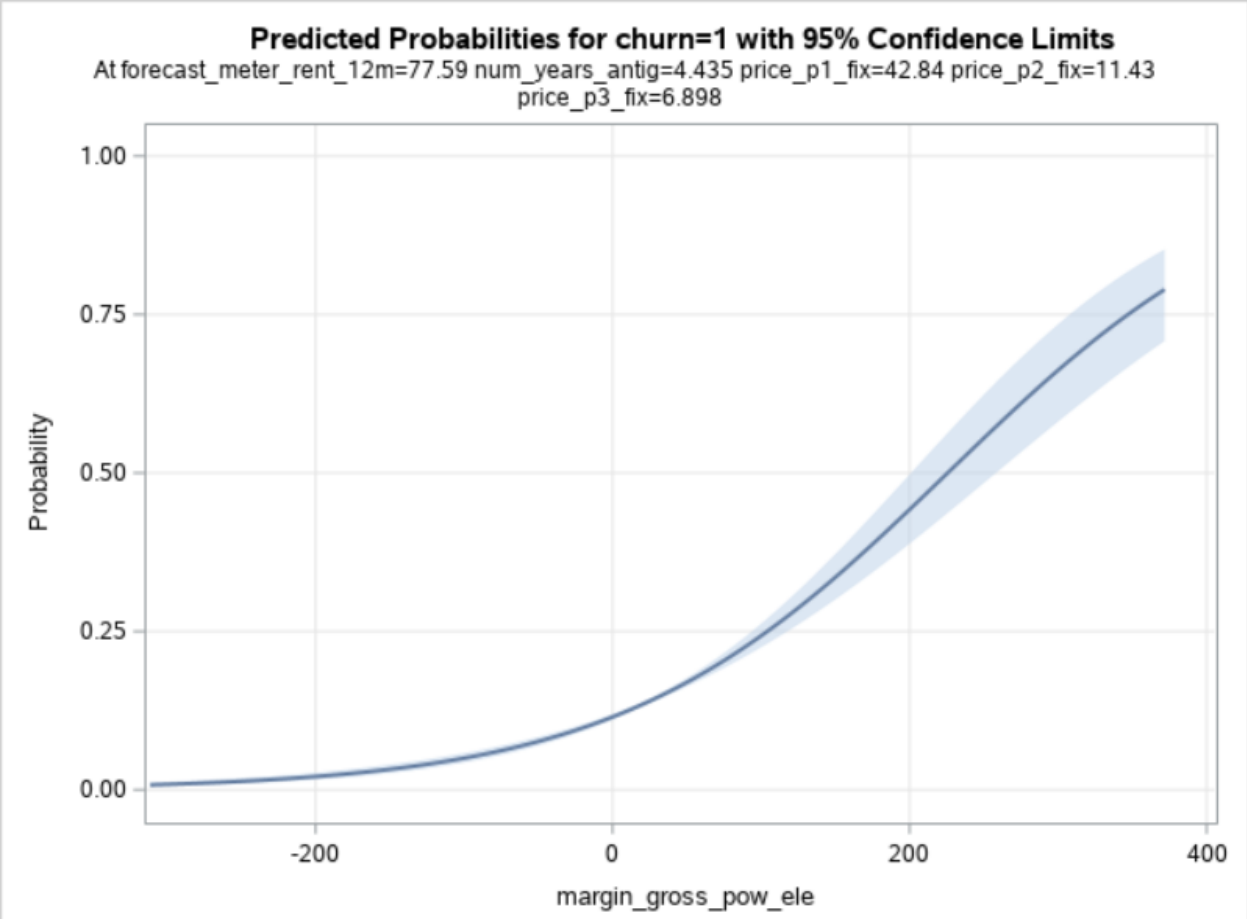
Association of Predicted Probabilities and Observed Responses			
Percent Concordant	62.0	Somers' D	0.240
Percent Discordant	38.0	Gamma	0.240
Percent Tied	0.0	Tau-a	0.060
Pairs	92458058	c	0.620

Note the C-value at bottom right – 0.620



Probability of churn at y-axis and forecast_meter_rent_12m at X-axis.

This graphs predicts the churn probability comparing with the variable forecast_meter_rent_12m and it shows very positive relations between them.



Now this is another graph showing positive relation between the churn and margin_gross_pow_ele. As the margin_gross_pow_ele increases, probability of of churn also increases just like above graph.

Now, we have to consider how much is the accuracy of this model. For that we have to validate the model using the valid dataset that we created the we have to campare the “C” statistic value. If the value of both train and valid dataset is almost equal, then we can assume that model is fine or generalize the unknown data well, otherwise modle is not good.

Step 3: Validating the model

Here we have use the same model but with different data to see how similar result produce by the model when dealing with new data.

```
proc logistic data=mydata.features_valid plots (only maxpoints = none)=
                effect(clband x = (forecast_meter_rent_12m margin_gross_pow_ele num_years_antig
                price_p1_fix price_p2_fix price_p3_fix));
model churn(event='1') =forecast_meter_rent_12m margin_gross_pow_ele num_years_antig price_p1_fix
                price_p2_fix price_p3_fix / stb clodds=pl slentry=0.5 slstay=0.1 selection=backward;

run;
```

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	63.0	Somers' D	0.259
Percent Discordant	37.0	Gamma	0.259
Percent Tied	0.0	Tau-a	0.065
Pairs	24518000	c	0.630

If we examine the C statistics at bottom right, we have value of 0.63 where we got 0.62 in train data set with only 0.01 different of value. Since there is very low difference in value, we can say that this logistic regression model generalizes unknown data good and model is working fine.

Step 4: 20% discount applied

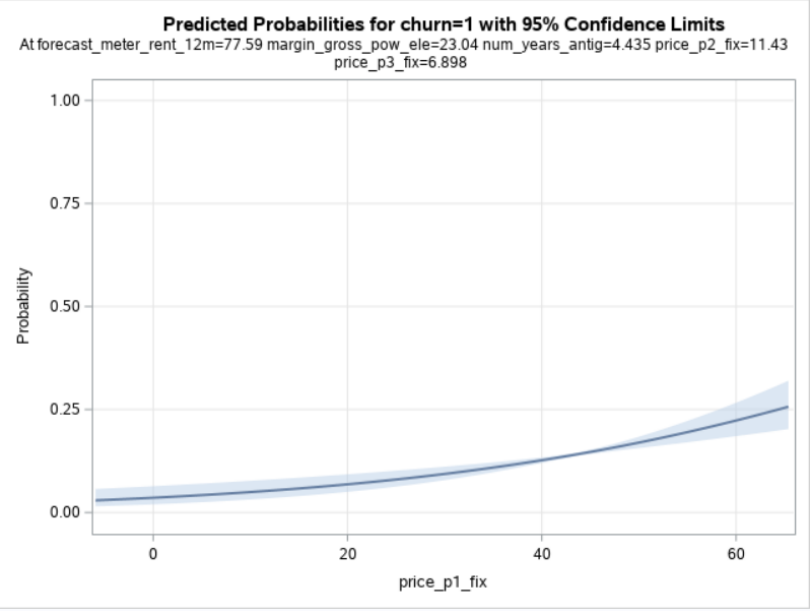


Fig: probability of churn with price_p1_fix

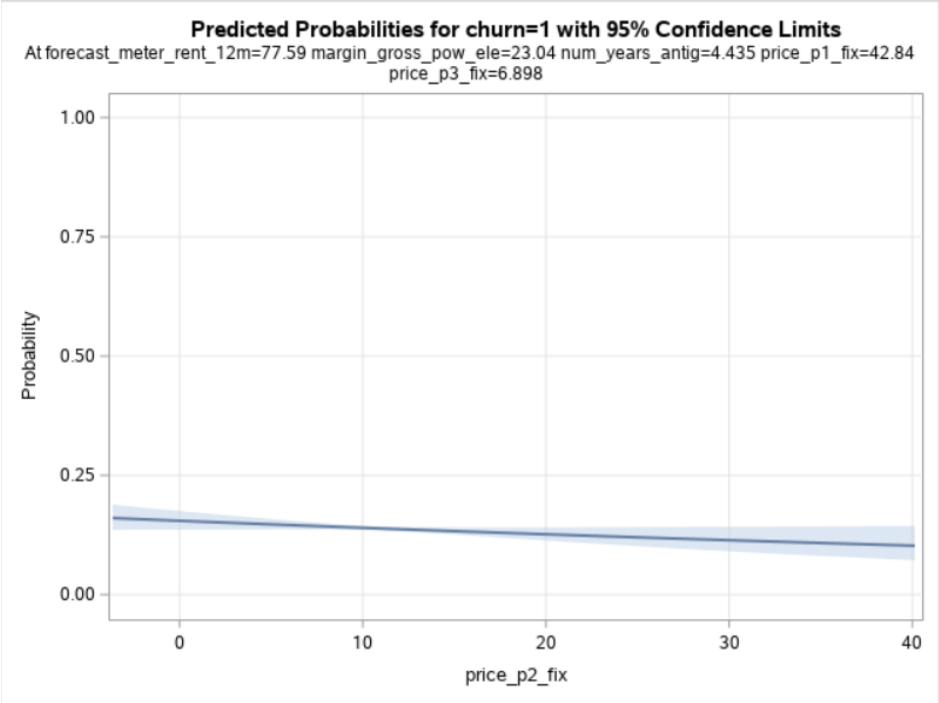


Fig: probability of churn with price_p2_fix

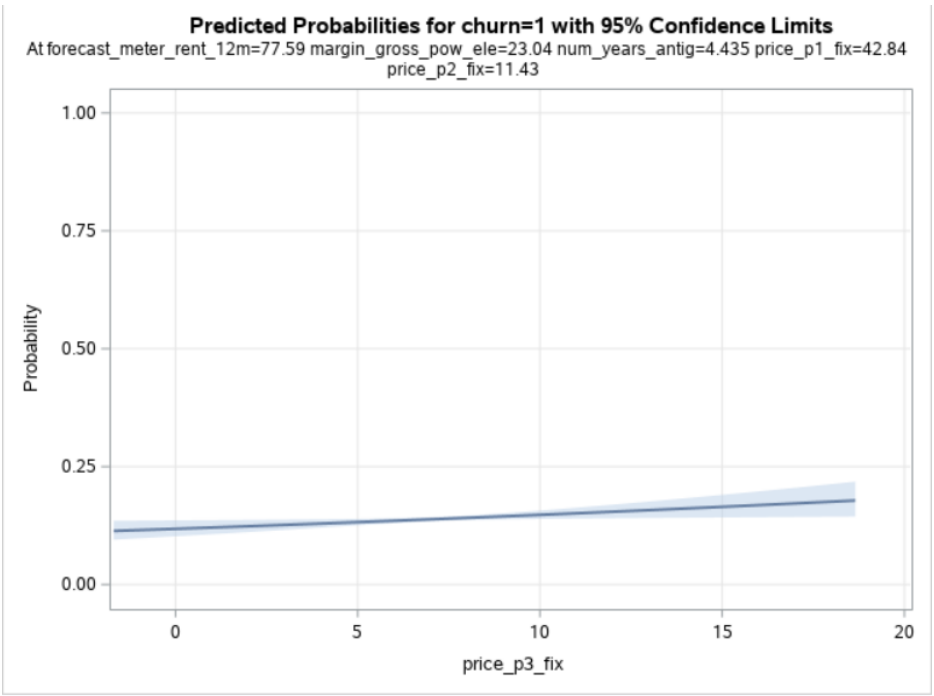


Fig: Probability of churn with pric_p3_fix

By looking at the graph, we can see that there isn't much relations between the different prices and the churn.

Conclusion:

In other word, there is not significant relations of customer leaving the company with price hence there is no point of give 20% discount offer.