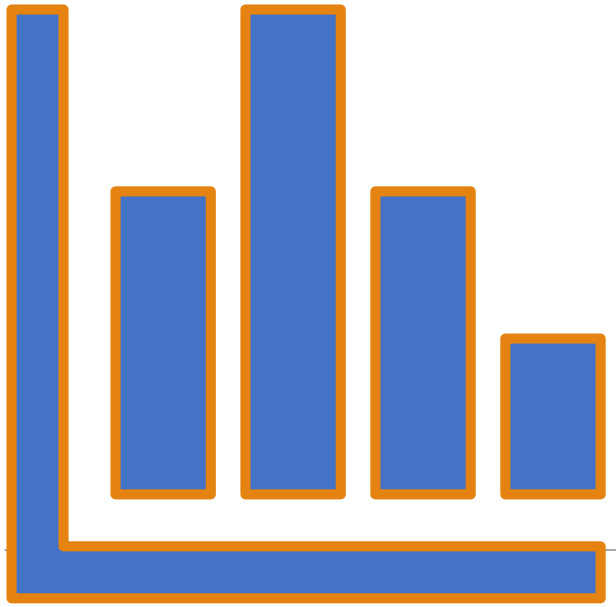


IBM Watson Marketing Data analysis

by
Doddanaik Basavaraj Vakkund
Mercy Ahalya Seelam
Sindhusha Vempati
Vineel Vishwanth Busi
Shiva Ram Kaushil Pabba

Problem statement

- Predicting the behavior of the customer and retain the customers using the customer information.
- There is one dependent variable “response” available in the dataset and few independent variables as categorical and numerical. from observing the response with respect to other features we can analyze whether the customer respond back based on the service he is offered and to also find what features are important to retain customer.

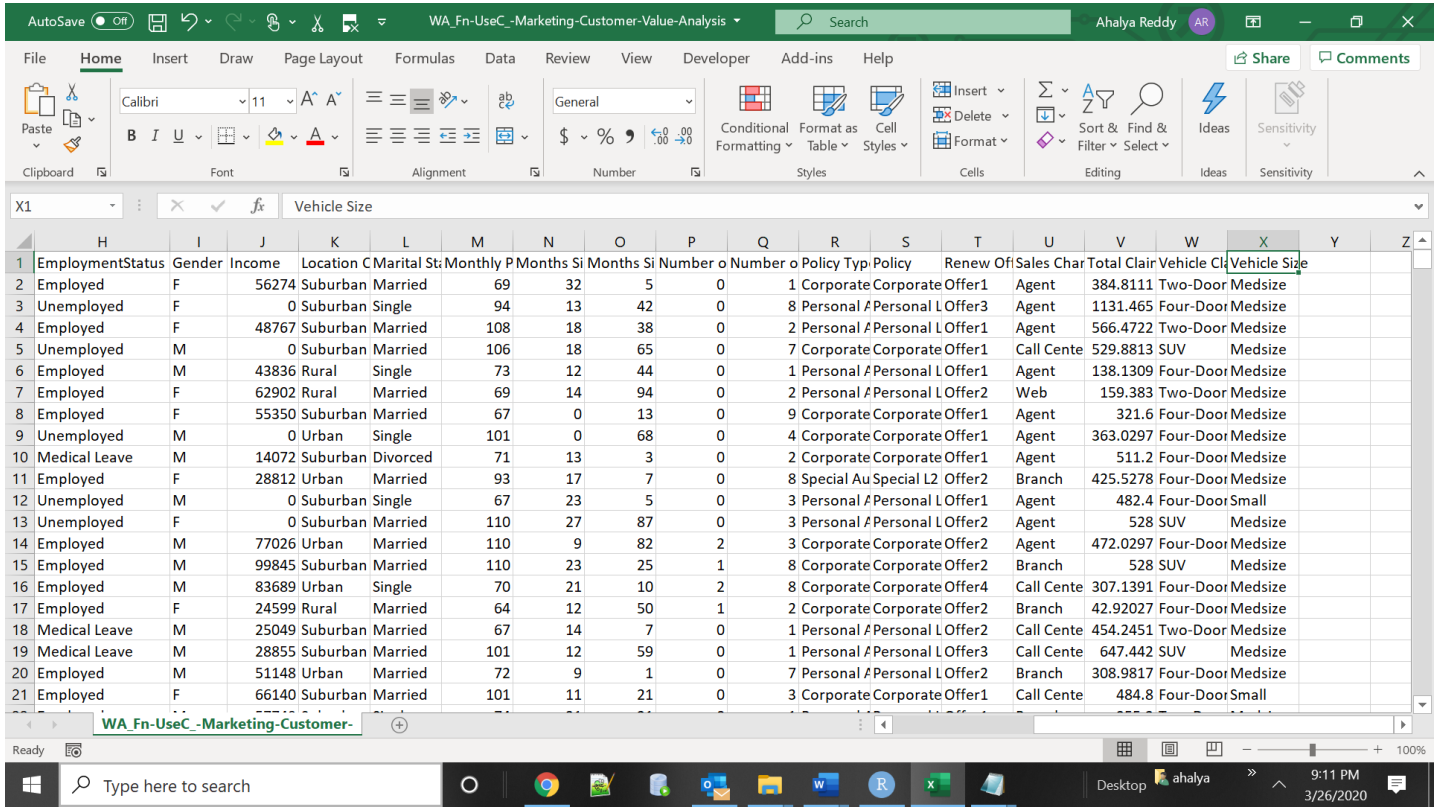
Database description

Source: Taken from the Kaggle source which is owned by Google LLC.

Privacy: This dataset is intended for Public access and use,

IBM Watson dataset is having the details of customers and it is collected to know whether the customer will respond back or not based on the services provided to them.

This dataset is having total of **9134 rows** and **24 variables**.



	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	EmploymentStatus	Gender	Income	Location C	Marital St	Monthly P	Months Si	Months Si	Number o	Number o	Policy Typ	Policy	Renew Of	Sales Char	Total Clair	Vehicle Cl	Vehicle Size		
2	Employed	F	56274	Suburban	Married	69	32	5	0	1	Corporate	Corporate	Offer1	Agent	384.8111	Two-Door	Medsize		
3	Unemployed	F	0	Suburban	Single	94	13	42	0	8	Personal A	Personal L	Offer3	Agent	1131.465	Four-Door	Medsize		
4	Employed	F	48767	Suburban	Married	108	18	38	0	2	Personal A	Personal L	Offer1	Agent	566.4722	Two-Door	Medsize		
5	Unemployed	M	0	Suburban	Married	106	18	65	0	7	Corporate	Corporate	Offer1	Call Cente	529.8813	SUV	Medsize		
6	Employed	M	43836	Rural	Single	73	12	44	0	1	Personal A	Personal L	Offer1	Agent	138.1309	Four-Door	Medsize		
7	Employed	F	62902	Rural	Married	69	14	94	0	2	Personal A	Personal L	Offer2	Web	159.383	Two-Door	Medsize		
8	Employed	F	55350	Suburban	Married	67	0	13	0	9	Corporate	Corporate	Offer1	Agent	321.6	Four-Door	Medsize		
9	Unemployed	M	0	Urban	Single	101	0	68	0	4	Corporate	Corporate	Offer1	Agent	363.0297	Four-Door	Medsize		
10	Medical Leave	M	14072	Suburban	Divorced	71	13	3	0	2	Corporate	Corporate	Offer1	Agent	511.2	Four-Door	Medsize		
11	Employed	F	28812	Urban	Married	93	17	7	0	8	Special Au	Special L2	Offer2	Branch	425.5278	Four-Door	Medsize		
12	Unemployed	M	0	Suburban	Single	67	23	5	0	3	Personal A	Personal L	Offer1	Agent	482.4	Four-Door	Small		
13	Unemployed	F	0	Suburban	Married	110	27	87	0	3	Personal A	Personal L	Offer2	Agent	528	SUV	Medsize		
14	Employed	M	77026	Urban	Married	110	9	82	2	3	Corporate	Corporate	Offer2	Agent	472.0297	Four-Door	Medsize		
15	Employed	M	99845	Suburban	Married	110	23	25	1	8	Corporate	Corporate	Offer2	Branch	528	SUV	Medsize		
16	Employed	M	83689	Urban	Single	70	21	10	2	8	Corporate	Corporate	Offer4	Call Cente	307.1391	Four-Door	Medsize		
17	Employed	F	24599	Rural	Married	64	12	50	1	2	Corporate	Corporate	Offer2	Branch	42.92027	Four-Door	Medsize		
18	Medical Leave	M	25049	Suburban	Married	67	14	7	0	1	Personal A	Personal L	Offer2	Call Cente	454.2451	Two-Door	Medsize		
19	Medical Leave	M	28855	Suburban	Married	101	12	59	0	1	Personal A	Personal L	Offer3	Call Cente	647.442	SUV	Medsize		
20	Employed	M	51148	Urban	Married	72	9	1	0	7	Personal A	Personal L	Offer2	Branch	308.9817	Four-Door	Medsize		
21	Employed	F	66140	Suburban	Married	101	11	21	0	3	Corporate	Corporate	Offer1	Call Cente	484.8	Four-Door	Small		

Data Field	Data Type	Data Field Definition
Customer	Continuous	Unique ID of the customer
State	Categorical	The state location of the customer
Customer Lifetime Value	Numerical	
Response	Categorical	The customer responded for the offer
Coverage	Categorical	The class of the policy
Education	Categorical	Education qualification of the customer
Effective to Date	Date	The start date of customer purchase
Employment Status	Categorical	The employment status of the customer – Employed/Unemployed
Gender	Categorical	The gender details of the customer
Income	Numerical	The Per annum income of the customer
Location Code	Categorical	The location level of the customer – Urban/Rural/Suburban
Marital Status	Categorical	Marital status of the Customer
Monthly Premium Auto	Numerical	Auto Loan monthly premium paid by the customer
Months Since Last Claim	Numerical	The Number of months where cutomer took the gap
Months Since Policy Inception	Numerical	The number of months since the policy taken
Number of Open Complaints	Numerical	Complaints raised by the customer
Policy Type	Categorical	The category of the policy
Policy	Categorical	The category of the policy – L2/L3
Renew Offer Type	Categorical	Which offer type is used to renew the policy
Sales Channel	Categorical	The channel by which the customer took the policy
Total Claim Amount	Numerical	Claimed amount by the customer
Vehicle Class	Categorical	The Vehicle class details
Vehicle Size	Categorical	The size of the vehicle.

Dataset exploration

- Using the predictive analysis, we will predict the behavior of the customer and retain the customers using the customer information. There is only one dependent variable in the dataset which is response which will give information about the customer response based on the service he is offered.

```
summary(IBM_data)
   Customer            State      Customer.Lifetime.Value Response       C
AA10041:   1   Arizona   :1703   Min.    : 1898          No :7826   Basic
AA11235:   1   California:3150   1st Qu.: 3994          Yes:1308   Exten
AA16582:   1   Nevada    : 882   Median : 5780                      Premi
AA30683:   1   Oregon    :2601   Mean   : 8005
AA34092:   1   Washington: 798   3rd Qu.: 8962
AA35519:   1                     Max.   :83325
(Other):9128
               Education    Effective.To.Date      EmploymentStatus Gende
Bachelor            :2748   1/10/2011: 195    Disabled     : 405    F:465
College             :2681   1/27/2011: 194    Employed     :5698    M:447
Doctor              : 342   2/14/2011: 186    Medical Leave: 432
High School or Below:2622   1/26/2011: 181    Retired      : 282
Master              : 741   1/17/2011: 180    Unemployed   :2317
                            1/19/2011: 179
                            (Other)  :8019
    Income        Location.Code   Marital.Status Monthly.Premium.Auto
Min.   :    0   Rural   :1773   Divorced:1369   Min.   : 61.00
1st Qu.:    0   Suburban:5779   Married :5298   1st Qu.: 68.00
Median :33879   Urban   :1582   Single  :2467   Median : 83.00
Mean   :37657                                   Mean   : 93.22
3rd Qu.:62338                                   3rd Qu.:109.00
Max.   :99981                                   Max.   :298.00
NA's   :5                                       NA's   :4
Months.Since.Last.Claim Months.Since.Policy.Inception Number.of.Open.Comp
Min.   : 0.0            Min.   :  0.00                Min.   :-1.0000
1st Qu.: 6.0            1st Qu.: 24.00                1st Qu.: 0.0000
Median :14.0            Median : 48.00                Median : 0.0000
Mean   :15.1            Mean   : 48.13                Mean   : 0.3835
3rd Qu.:23.0            3rd Qu.: 71.00                3rd Qu.: 0.0000
Max.   :35.0            Max.   :640.00                Max.   : 5.0000
```

Pre-Processing

High level summary stats for raw data:

From the summary stats we can observe that this data was collected on 9134 unique customer from 5 states along with other details like income, education, employment status, gender, location code, marital status. from summary stats we can observe that there are null values in features like income, monthly premium auto and number of open complaints we have minimum value as -1.

Summary After Preprocessing

High level summary stats for cleaned-up data:

From the summary stats of cleaned-up data:

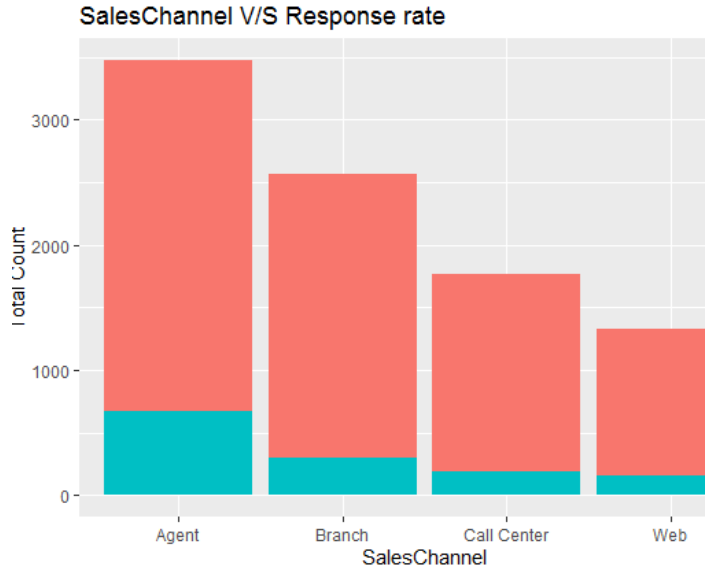
We can observe that there is no null values in features like income replaced with mean, monthly premium auto and number of open complaints we have minimum value as -1 which might have occurred because of human error so we replaced -1 with 1.

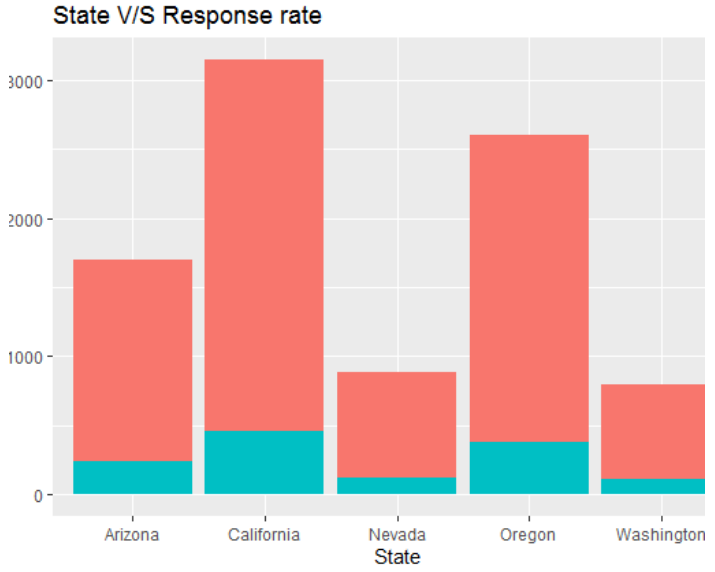
```
> summary(IBM_data)
       State      Customer.Lifetime.Value Response       Coverage
 Arizona   :1703   Min.    : 1898         No :7826   Basic   :5568
 California:3150   1st Qu.: 3994         Yes:1308   Extended:2742
 Nevada    : 882   Median : 5780                    Premium : 824
 Oregon    :2601   Mean   : 8005
 Washington: 798   3rd Qu.: 8962
                   Max.   :83325

                Education    Effective.To.Date      EmploymentStatus Gender
 Bachelor            :2748   1/10/2011: 195    Disabled     : 405    F:4658
 College             :2681   1/27/2011: 194    Employed     :5698    M:4476
 Doctor              : 342   2/14/2011: 186    Medical Leave: 432
 High School or Below:2622   1/26/2011: 181    Retired      : 282
 Master              : 741   1/17/2011: 180    Unemployed   :2317
                             1/19/2011: 179
                             (Other)  :8019
     Income       Location.Code   Marital.Status Monthly.Premium.Auto
 Min.   :    0   Rural   :1773   Divorced:1369   Min.   : 61.00
 1st Qu.:    0   Suburban:5779   Married :5298   1st Qu.: 68.00
 Median :33890   Urban   :1582   Single  :2467   Median : 83.00
 Mean   :37657                                   Mean   : 93.22
 3rd Qu.:62320                                   3rd Qu.:109.00
 Max.   :99981                                   Max.   :298.00

 Months.Since.Last.Claim Months.Since.Policy.Inception Number.of.Open.Complaints
 Min.   : 0.0            Min.   :  0.00                Min.   :0.0000
 1st Qu.: 6.0            1st Qu.: 24.00                1st Qu.:0.0000
 Median :14.0            Median : 48.00                Median :0.0000
 Mean   :15.1            Mean   : 48.13                Mean   :0.3844
 3rd Qu.:23.0            3rd Qu.: 71.00                3rd Qu.:0.0000
 Max.   :35.0            Max.   :640.00                Max.   :5.0000

 Number.of.Policies         Policy.Type            Policy     Renew.Offer.Type
 Min.   :1.000      Corporate Auto:1968   Personal L3 :3426   Offer1:3752
```

Exploratory Data Analysis

Analyzed customer responses over the 5 states : from the plot we can see that California and Oregon have more customer who have responded as compared to other states, but they also have higher number of non- responders.

when we observe the responses with respect to sales channel, we can see that through agent medium we got more responders as compared toother medium.

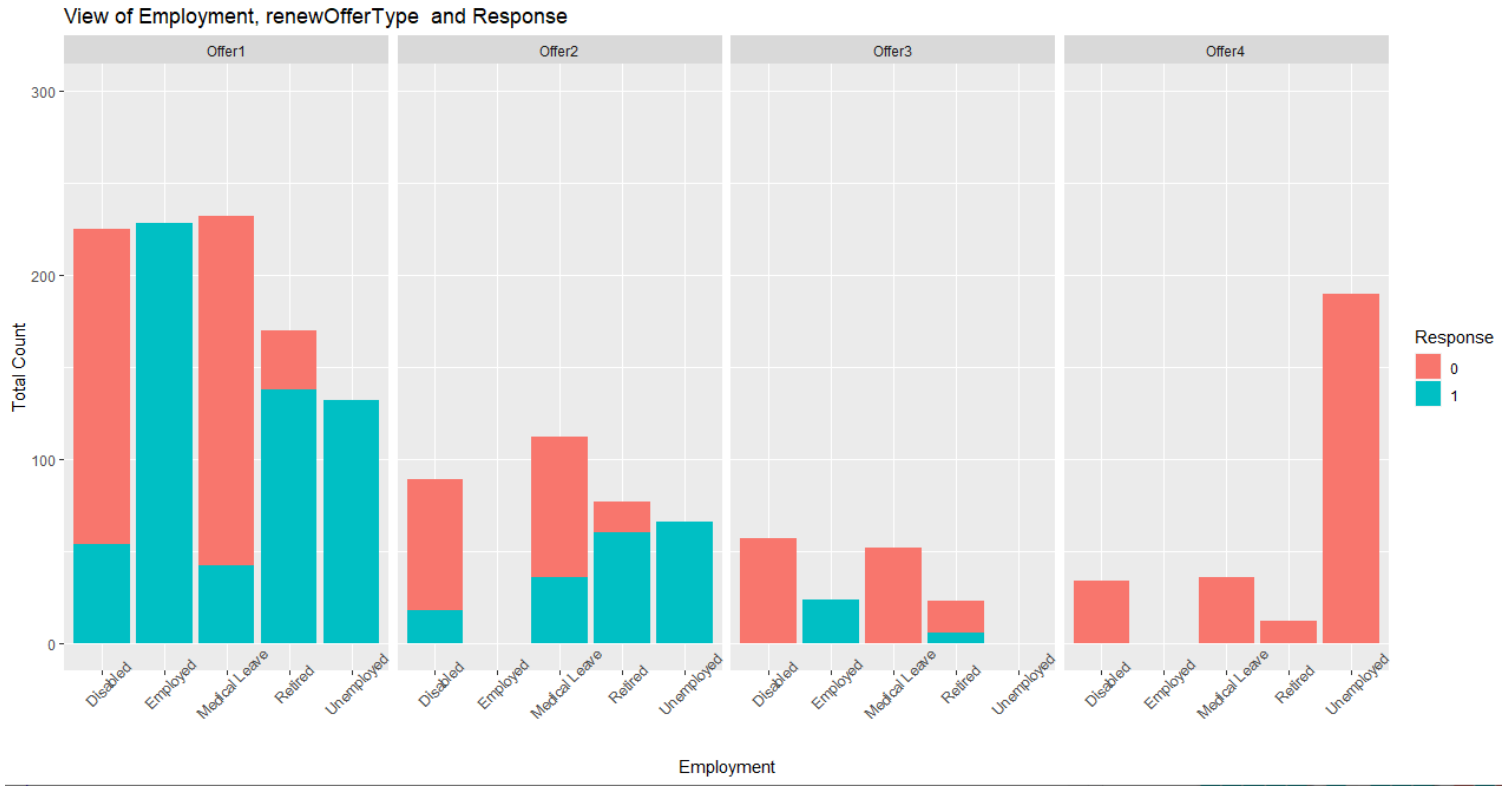




Analysis

- VISUALIZATION:

we tried to analyze how customers responded to 4 offers offered to them based on there employment status. in the plot we can observe most of the employed customers have responded to offer1 and offer 3.and we can also observe that no customer have responded back to offer4.



```
Step:  AIC=4692.26
Response ~ Education + EmploymentStatus + Income + LocationCode +
    MaritalStatus + MonthlyPremiumAuto + MonthsSinceLastClaim +
    NumberofOpenComplaints + NumberofPolicies + RenewOfferType +
    SalesChannel + TotalClaimAmount + VehicleSize

                         Df Deviance    AIC
<none>                        4638.3 4692.3
- NumberofPolicies        1   4640.4 4692.4
- NumberofOpenComplaints  1   4640.6 4692.6
- Income                  1   4643.0 4695.0
- MonthsSinceLastClaim    1   4643.1 4695.1
- Education               4   4653.5 4699.5
- MonthlyPremiumAuto      1   4653.8 4705.8
- TotalClaimAmount        1   4661.1 4713.1
- VehicleSize             2   4665.1 4715.1
- MaritalStatus           2   4668.7 4718.7
- SalesChannel            3   4696.5 4744.5
- LocationCode            2   4763.4 4813.4
- EmploymentStatus        4   4990.4 5036.4
- RenewOfferType          3   5231.0 5279.0
```

Feature Selection

we have used stepwise backward regression to find out important features

we got " education, employment status, income, locationcode, maritalstatus, monthly premium auto, number of open complaints, renew offer type, sales channel, total claim amount, vehicle size".

Modeling And Analysis

We used Four different models in order to meet our problem statement

They are:

- Logistic Regression
- Naïve Bayes
- GBM – Gradient Boosting
- Random Forest

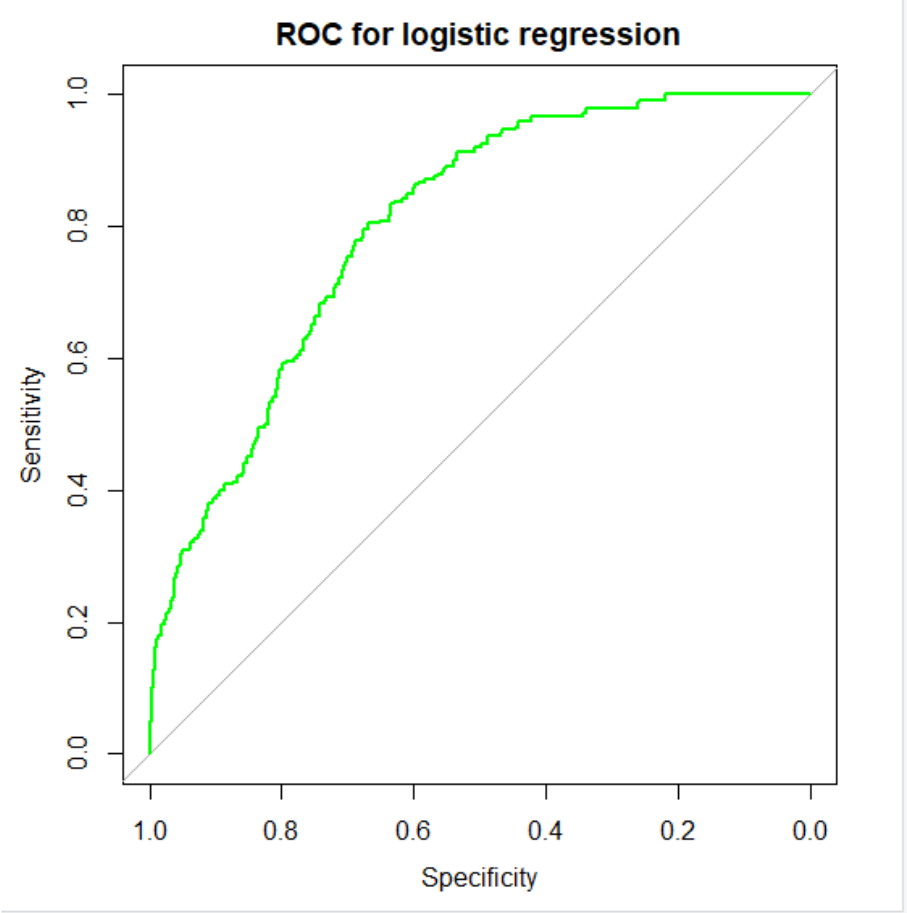
Logistic Regression

Binary Logistic Regression model

- Accuracy : 88.12%
- AUC : 0.799
- Confusion matrix :

	0	1
0	1567	197
1	20	43

- Cross Validation (k=5):
- Accuracy : 88.06%



Naïve Bayes

Naïve Bayes

```
> confusionMatrix(predic, test$Response)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1574  235
         1   11    8

               Accuracy : 0.8654
                 95% CI : (0.8489, 0.8807)
    No Information Rate : 0.8671
    P-Value [Acc > NIR] : 0.5983

                  Kappa : 0.0426

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.99306
            Specificity : 0.03292
         Pos Pred Value : 0.87009
         Neg Pred Value : 0.42105
             Prevalence : 0.86707
         Detection Rate : 0.86105
   Detection Prevalence : 0.98961
      Balanced Accuracy : 0.51299

       'Positive' Class : 0
```

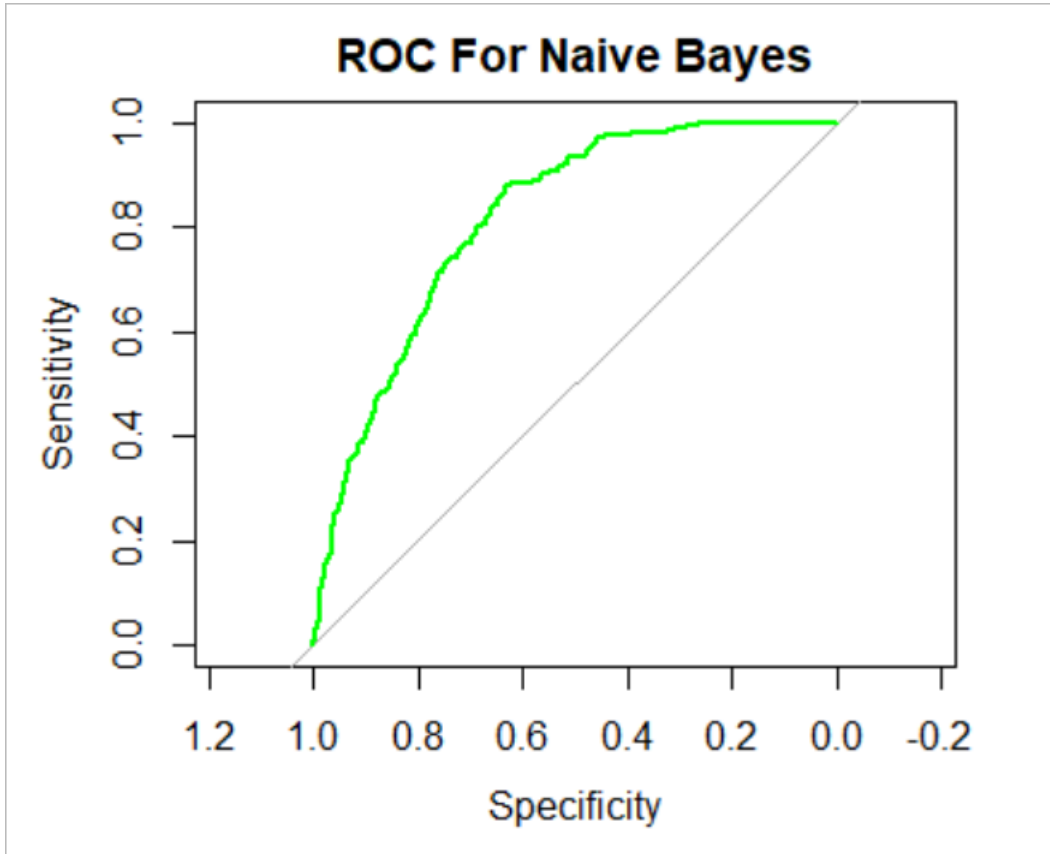
After tuning →

← Before Tuning

Parameters used are train control where we defined cross folds, and also tune grid parameters as search grid using the kernel and also preprocessing as box-cox.

```
> confusionMatrix(pred, test$Response)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1572  234
         1   13    9

               Accuracy : 0.8649
                 95% CI : (0.8483, 0.8802)
    No Information Rate : 0.8671
    P-Value [Acc > NIR] : 0.6246

                  Kappa : 0.0469

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.99180
            Specificity : 0.03704
         Pos Pred Value : 0.87043
         Neg Pred Value : 0.40909
             Prevalence : 0.86707
         Detection Rate : 0.85996
   Detection Prevalence : 0.98796
      Balanced Accuracy : 0.51442

       'Positive' Class : 0
```

ROC curve for Naïve Bayes



Area under curve is 0.815

Accuracy % of Naïve Bayes
0.86

Gradient Boosting

Gradient Boosting

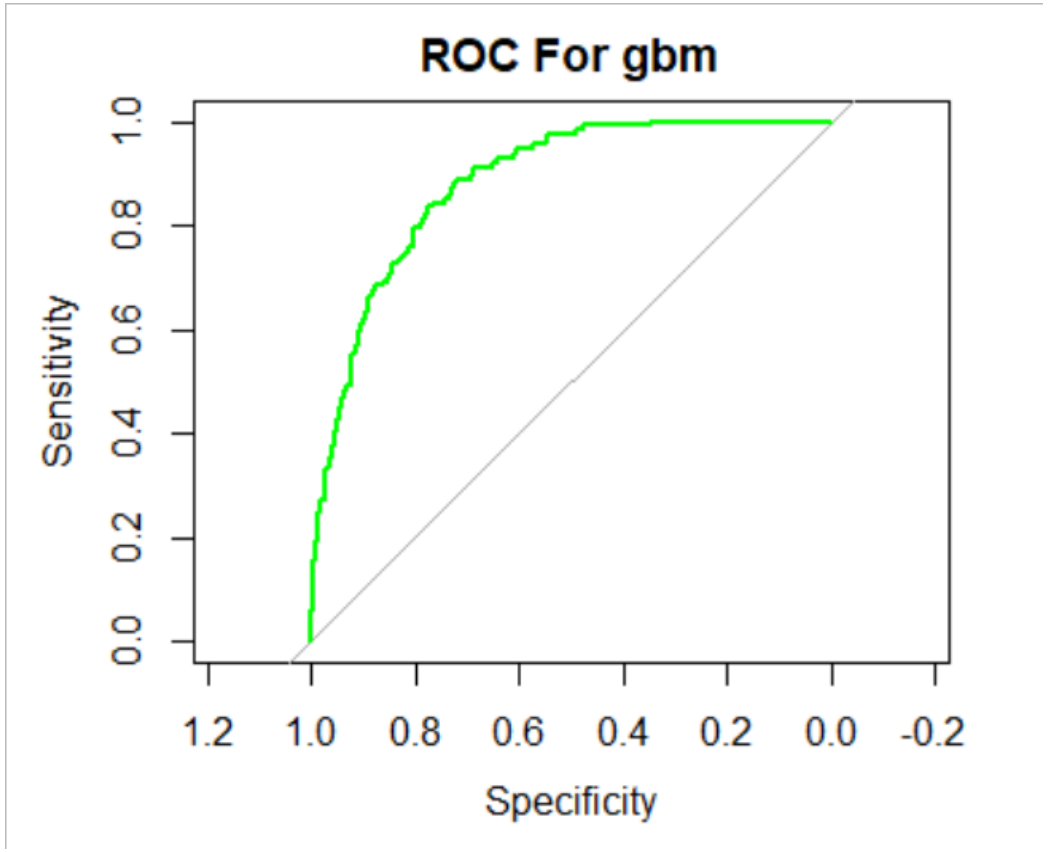
Modeling:

Confusion Matrix:

gbm_ITV2	0	1
0	1574	198
1	11	45

Accuracy % of GBM 88.56

AUC under curve = 0.8827



Random forest model

Random Forest

Modeling:

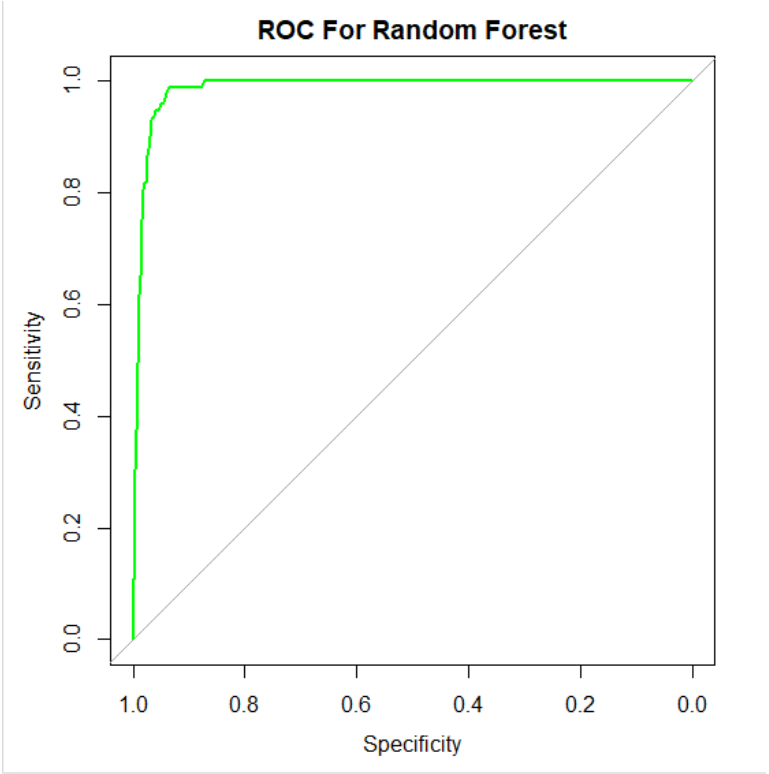
- Number of Trees : 100
- Node size : 25

Model Evaluation:

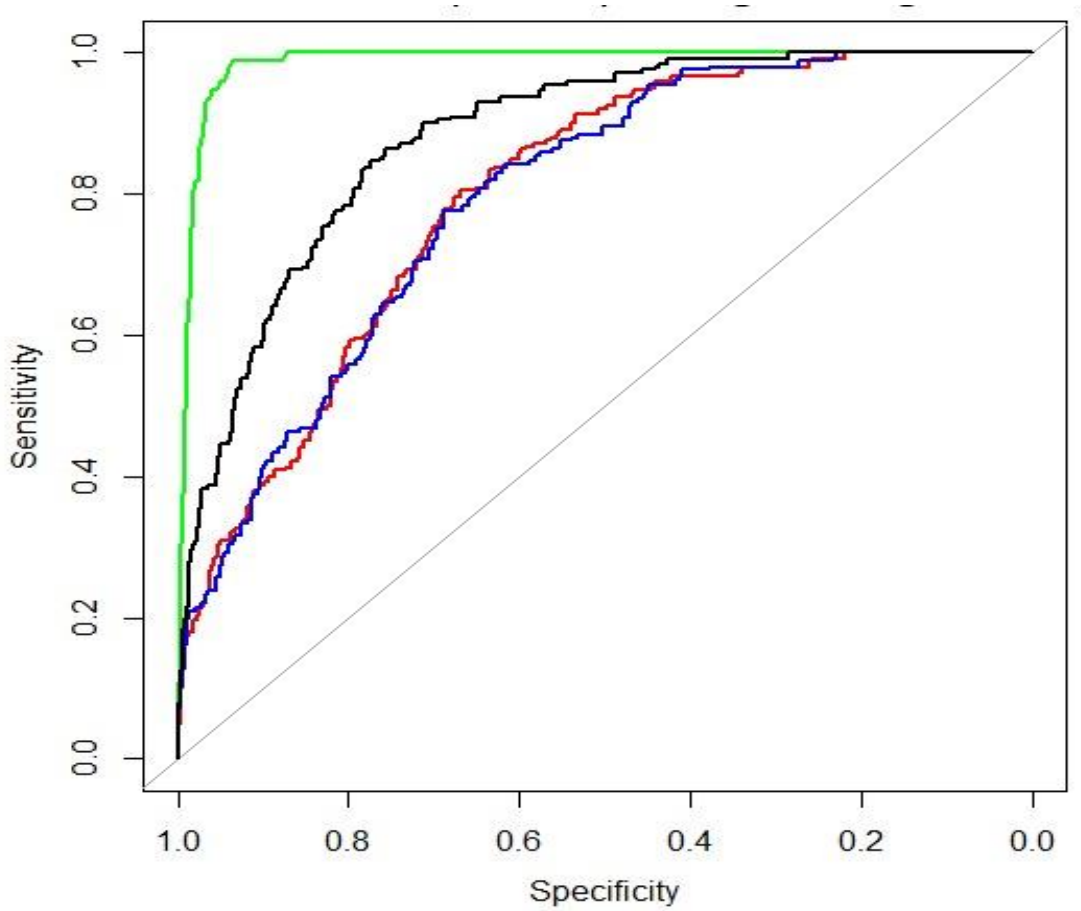
- Accuracy : 93.37%
- AUC : 0.9857
- Confusion matrix :

```
     0   1
0 1569 103
1   18 137
> 
```

ROC Curve for Random Forest Model :



ROCs for Random Forest(Green) vs Logistic Regression (red) vs Naïve Bayes(Blue) vs Gradient Boosting(Black) Models



Comparison of all Models

Models	Accuracy	AUC under
Binary Logistic Regression	88.12%	0.799
Naïve Bayes	86.54%	0.815
Random Forest	93.37%	0.985
Gradient Boosting	88.56%	0.882

Conclusion

We were able to find the factors that are more significant for the customer positive response by using random forest where we are considering it as best model because of its higher accuracy.

Thank you!!!