

MARCH 15, 2019



CUSTOMER CHURN PREDICITON

DATA ANALYTICS

AKSHAY HIRPARA
EDWISOR

Table of Contents

Project Description	2
INTRODUCTION	3
DATA PRE-RPROCESSING	4
VISUALIZATION (TABLEAU)	12
MACHINE LEARNING	14
EVALUATING PERFORMANCE OF MODEL	22

Project Description:

- This project focuses on Churn Prediction (Losing of Customers). Nowadays as competition have increased customers go for different brands for better facilities. The cost of retaining customers by giving them offers is more convenient way as compared to building a new customer base.
- The objective of this case is to predict customer behaviour. We have data which contains customer usage patterns and Churn Data whether customer has Churned Out or not.
- We have two data sets 1.) Train Data 2.) Test Data
- Below are the number of variables in our data set.

S.no	Variable Name
1	State
2	account_length
3	area_code
4	phone_number
5	total_day_minutes
6	total_day_minutes
7	total_day_minutes
8	international_plan
9	voice_mail_plan
10	number_vmail_messages
11	total_day_calls
12	total_day_charge
13	total_eve_calls
14	total_eve_charge
15	total_night_calls
16	total_night_charge
17	total_intl_calls
18	total_intl_minutes
19	total_intl_charge
20	number_customer_service_calls
21	Churn

Chapter-1

INTRODUCTION

The first step we always go for is knowing the data. We will check the data i.e. what are the number of rows and columns (Observations vs Variables), what is the data type of all the variables present and what are the variables associated with the data.

```
> dim(df)
[1] 1667 21
> str(df)
'data.frame': 1667 obs. of 21 variables:
 $ state          : Factor w/ 51 levels "AK","AL","AR",...: 12 27 36 33 41 13 29 19 25 44
 ...
 $ account.length : int 101 137 103 99 108 117 63 94 138 128 ...
 $ area.code      : num 510 510 408 415 415 415 415 408 510 415 ...
 $ phone.number   : num 452 906 1468 1602 1502 ...
 $ international.plan : Factor w/ 2 levels " no"," yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ voice.mail.plan : Factor w/ 2 levels " no"," yes": 1 1 2 1 1 1 2 1 1 2 ...
 $ number.vmail.messages : int 0 0 29 0 0 0 32 0 0 43 ...
 $ total.day.minutes : num 70.9 223.6 294.7 216.8 197.4 ...
 $ total.day.calls   : int 123 86 95 123 78 85 124 97 117 100 ...
 $ total.day.charge  : num 12.1 38 50.1 36.9 33.6 ...
 $ total.eve.minutes : num 212 245 237 126 124 ...
 $ total.eve.calls   : int 73 139 105 88 101 68 125 112 46 89 ...
 $ total.eve.charge  : num 18 20.8 20.2 10.7 10.5 ...
 $ total.night.minutes : num 236 94.2 300.3 220.6 204.5 ...
 $ total.night.calls : int 73 81 127 82 107 90 120 106 71 92 ...
 $ total.night.charge : num 10.62 4.24 13.51 9.93 9.2 ...
 $ total.intl.minutes : num 10.6 9.5 13.7 15.7 7.7 6.9 12.9 11.1 9.9 11.9 ...
 $ total.intl.calls   : int 3 7 6 2 4 5 3 6 4 1 ...
 $ total.intl.charge  : num 2.86 2.57 3.7 4.24 2.08 1.86 3.48 3 2.67 3.21 ...
 $ number.customer.service.calls: int 3 0 1 1 2 1 1 0 2 0 ...
 $ churn            : Factor w/ 2 levels " False"," True.": 1 1 1 1 1 1 1 1 1 1 ...
```

Fig- Dimensions and Structure of Data (Test Data)

- We can see from the figure that there are total 1667 observation in our Test Data and 21 Variables.
- Variables are divided into various data types such as Integer, Factor and Numeric.
- There are few variables which required data type conversion. We have converted those variable's data type. For E.g. Phone Number and Area Code were converted into Numeric Data Type.
- Now we have gathered enough information to go ahead for Data Pre-Processing.

Chapter-2

Data Pre-Processing

Now we will Clean, Analyse and Prepare the data for modelling.

1.) Missing Value Analysis:

- First, we will go for Missing Value Analysis. Missing Values are the values which are often forgotten to be filled during preparation of data. This might happen due to human error or even software error.
- These values are either imputed or deleted as per the industrial standards.
- According to industrial standards We should impute missing values on data which contains $\leq 30\%$ of missing values.
- There are 3 Methods to impute Missing Values.
 - **Central Statistic Method**
 - **Distance Formula- Knn Imputation**
 - **Prediction Method**
- Important Remark:
 - While using KNN Method for treating missing values we need to select K Value, we must make sure that selected K Value number is ODD i.e. 1,3,5,7.....
 - Because If K-Value is EVEN, Algorithm will get confused on which value to be selected as Majority.
- In our Data Set, there are no Missing Values.

2.) Outliers Analysis

- Outliers are the data which are highly deviated from original mean.
 - I have a bank account. My average monthly transactions are 1.5 Lakh Dollars. Now suddenly, I transacted 10.5 Lakh Dollars. Now this is called outlier.
- Outliers Analysis can be used in Fraud Detection, checking MRI Values etc.

What impact Outliers have in a data set?

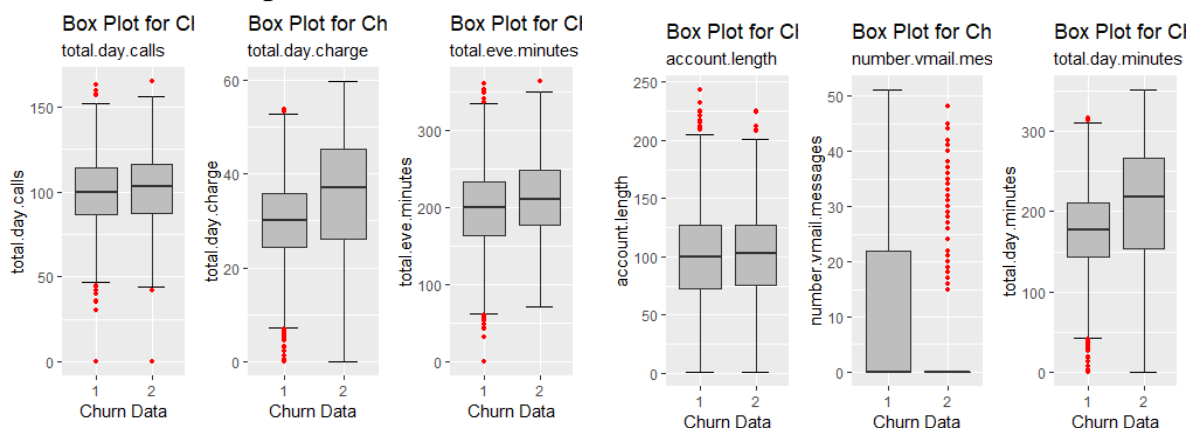
- Suppose we have a data set 1,3,5,7,9. The mean is 5. Now let's introduce outlier 1,3,5,7,9,14. Our mean would be 6.5. we can see that mean is deviated.

- While dealing with large data sets it would create lot of problems and can have major impact.

How to detect outliers?

There are several methods at very high level to detect the outliers.

- Box Plot:
 - Data above the Upper Fence and below the Lower Fence would be considered as outliers.
- Statistical Technique- Grubbs Test
 - It has an assumption that data is normally distributed. Now in real cases, there are rare chances that data would be Normally distributed.
- R-Package Outlier
 - There is a package in R named Outliers. This uses concept of Mean to detect outliers.
 - Those data which are highly deviating from Mean would be considered as outliers.
- Replace with NA
 - In this method we calculate outliers. After calculating all outliers, we will replace them with NA.



- Now all the NA's would be imputed with suitable missing value analysis method.

CUSTOMER CHURN PREDICITONS

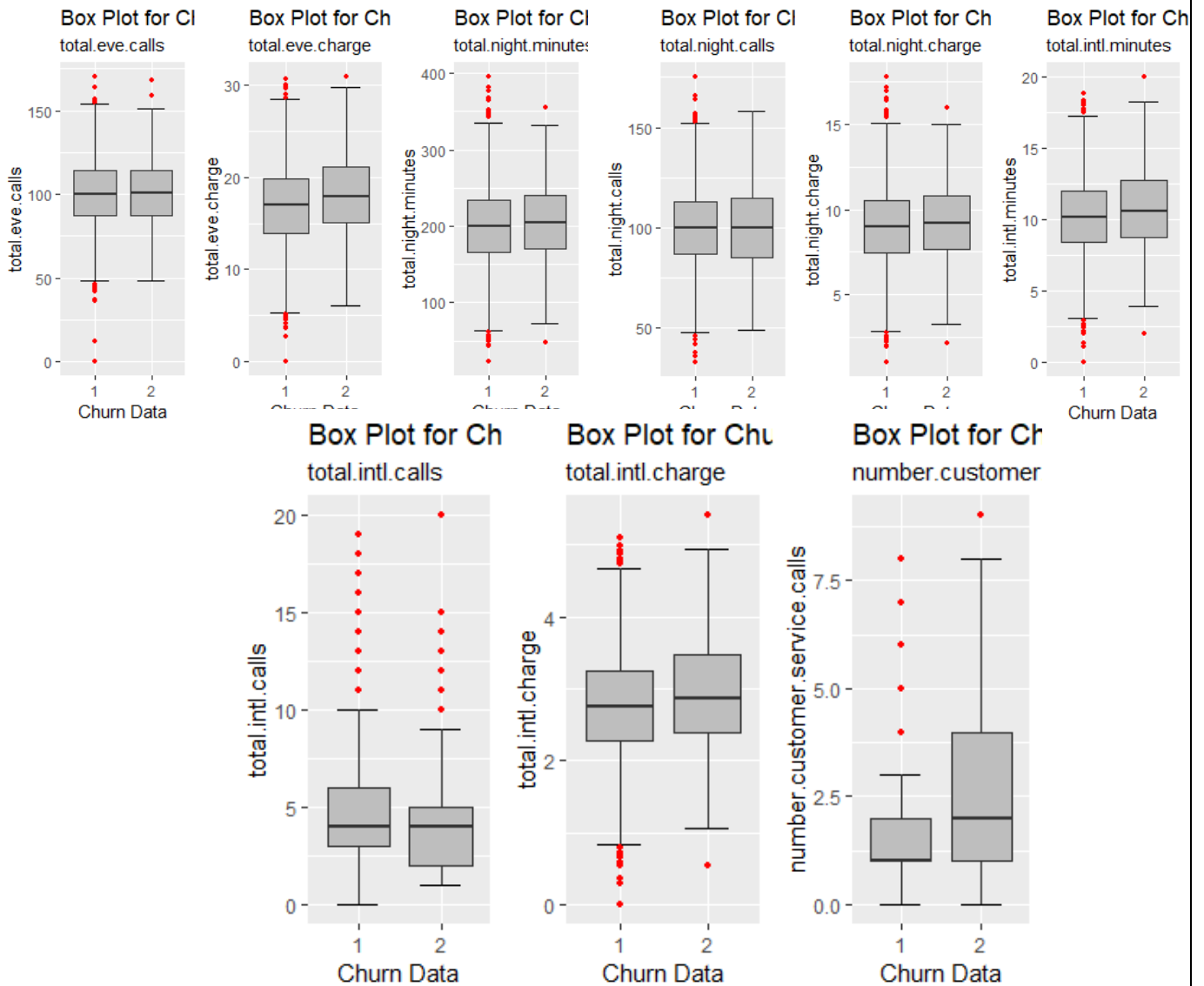


Fig-Box Plots for Outliers Analysis

We won't remove any outliers from our data. As after analysis we are losing data from various variables such as number of customer calls, number of vmail messages and total international calls. The information which would get deleted might carry meaningful information, so we would build our model without removing outliers data.

3.) Feature Selection

- This method is used to extract relevant and meaningful features out of data.

How does Feature Selection help to extract meaningful information?

- This concept is used to reduce complexity of data by reducing variables which carries irrelevant information to explain target variable.

E.g. Suppose we received a data set with 10000 variables, now we cannot feed every variable to the model because not every variable can be used to extract meaningful information. Hence, we would use Feature Selection to reduce complexity of data and select only those variables which carries significant information.

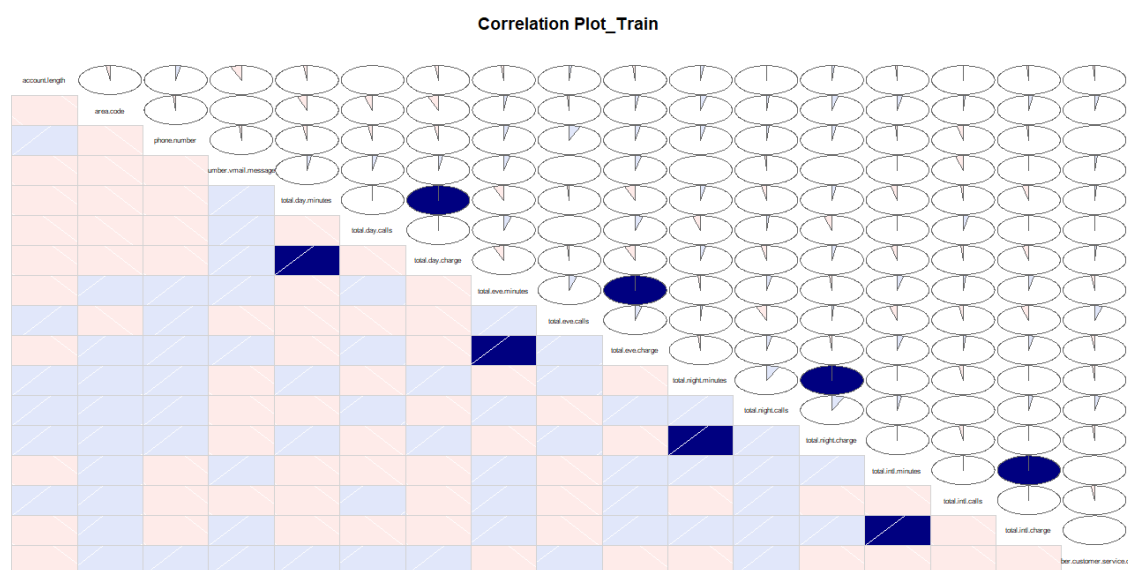


Fig- Feature Selection Graph (Correlation Plot)

Impact of irrelevant variables in a data set!

E.g. In text mining, suppose we extracted a paragraph with 28000 words. Now during analysis each word would act as variable. This means there would be

28000 variables *This would result in increase in size of irrelevant data. This is termed as Curse of Dimensionality.*

We would now go for two methods which are generally used in Feature Selection.

- Correlation Analysis:
 - This method is used for Continuous or Numerical Variables.

$$Cor(X, Y) = \frac{CoV(X, Y)}{SD(X)SD(Y)}$$

- Range of Correlation Values is between -1 to +1
- Chi-Square Test
 - This method is used for Categorical Variables.

We develop Hypothesis on basis of Chi-Square Test.

- Null Hypothesis- Two Variables are dependent
- Alternate Hypothesis- Two variables are independent

$$\chi^2_c = \sum \frac{(O_i - E_i)^2}{E_i}$$

E_i=Expected Values

O_i=Original Values

Degree of Freedom= (number of rows-1) (no of columns -1)

On basis of Chi Square value, we will calculate p-value.

- If p-value is < 0.05, This would mean that Alternate Hypothesis is true that means both variables are independent of each other.
- If p-value is > 0.05, This would mean that Null Hypothesis is True that mean both variables are dependent on each other.

As we can see from the graph that there is high correlation between few variables such as total day charge, total day minutes, total eve charge, total eve minutes, total night charge, total night minutes, total int charge and total int minutes. Hence, we would remove total night charge total day charge, total eve charge and total int charge from our data.

4.) Feature Scaling-

- This is also known as Feature Engineering and Variable Scaling.

E.g. We have a data set with variable named Salary and Age. Now in this case, salary can go to any range 1000\$,10000\$,50000\$ while Age can go maximum till 100. Such type of cases can cause Anomaly in Data.

There are 2 methods named:

- Normalization
- Standardization
 - Normalization:
 - This is used to reduce unwanted variations. This convert our data range in between 0 to 1 and bring to common scale.

$$X_{changed} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Fig- Normalization Formula

- Standardization:
 - This will convert each datapoint to unit of Standard Deviation. This method is also knowna as Z-Score.

Formula to find population mean

$$\mu = \frac{\sum x}{n}$$

Formula to find population standard deviation

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{n}}$$

Formula to find the **z-score**

$$z\ score = \frac{(x - \mu)}{\sigma}$$

Fig-Standardization Formula

Now question is which method to use,

If most of our data is Normally distributed than we will use Standardization or Z-Score.

5.) Sampling Techniques:

- Sampling is a method of selecting a small data from a huge collection of data.
- We select sample in such a way that it represents characteristics of whole data.

E.g. Suppose we got a data worth 7 Crore observation. Now this would consume lot of memory. so, we create a subset which would represent entire data

Let's go through few methods,

1.) Simple Random Sampling-

- This is the simplest method.
 - E.g. Pick any 100 observations out of 26000 observations.
- In this method, each observation would have equal chance of getting picked.

2.) Systematic Sampling/Kth Name

- Select every Kth observation as sample
 - $K = N/n$
 N = Number of observations
 n = Desired Sample

Suppose $N=10000$ and $n=5000$

$K=2$, This means select every 2nd observation.

- There are several disadvantages associated with this method.
- Suppose $K=2$, now in our data every 2nd observation is of same category.

E.g. We have a variable named designation. In designation there are 2 categories Director and Group HR. Now suppose $K=2$ and in our data every 2nd observation is Group HR. so we would only get those data whose designation is Group HR and hence there won't be any data with designation as Director. This would lead to our data being biased.

3.) Stratified Sampling

- In this method Stratus would be created, Stratums is a subset of population which have same characteristics.

E.g. Suppose we have 100 observations, in these 100 observations there are 30-President,30-Vice President and 10-Directors. Now we want only 10 data out of 100 as sample. This method would create sample with equal proportion such as 3 data of President, 3 data of Vice-President and 1 data of Director.

Now question is which categorical data should be selected to create stratums?

For determining suitable categorical variable, feature selection would be used as we need to select that variable which carries relevant information.

Now by performing Data-Pre-processing, our data is now ready to be fed to Machine Learning algorithm.

Chapter-4

Visualizations in Tableau

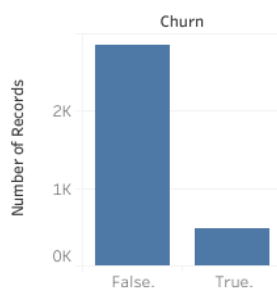
Data Visualization is one of the most integral part of any analysis project. Visualization is used to represent data in such a way that even a non-catalyst person can also understand.

Tableau is one of the fasted BI (Business Intelligence) tool. It is very easy to learn and has advanced concepts associated with it.

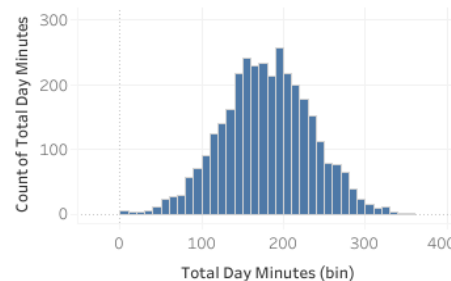
Dashboard and Story Creation are the most salient features of Tableau

- I have created Two dashboards in-order to visualize our data in a speedy and interactive way
- First dashboard contains Churn visualization where you can see the target class imbalance
- Rest other graphs in both dashboards are to check distribution of data.

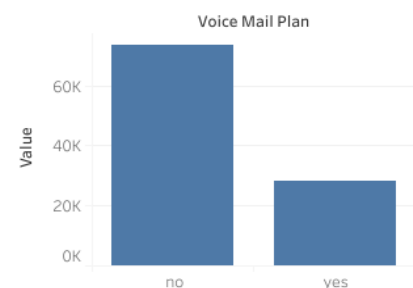
Churn Graph



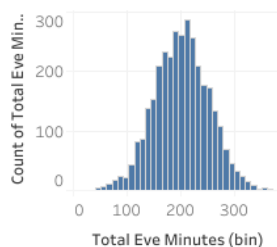
Histogram for Total Day Minutes



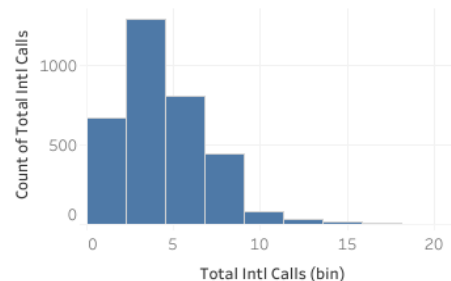
Voice Mail Plan Subscription



Histogram for Total Eve Minutes



Histogram for Total Int Calls



Histogram for Total Day Charge

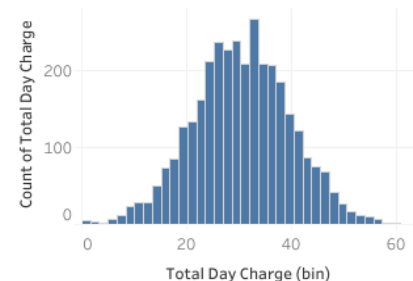
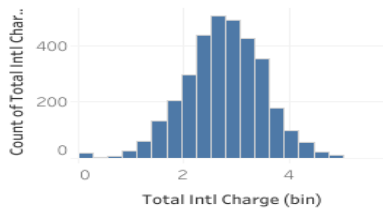
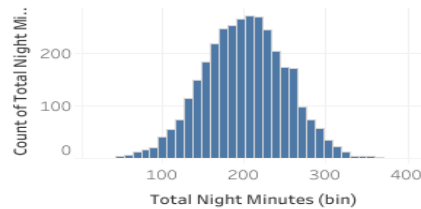


Fig-Tableau Dashboard-1

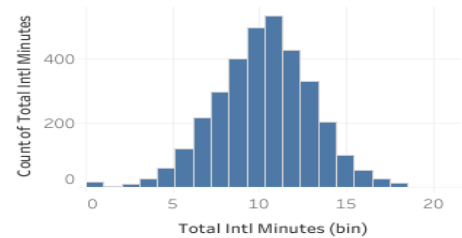
Histogram for Total Intl Charge



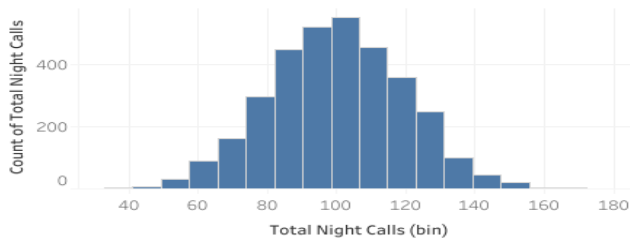
Histogram for Total Night Minutes



Histogram for Total Intl Minutes



Histogram for Total Night Calls



Histogram for Total Night Charge

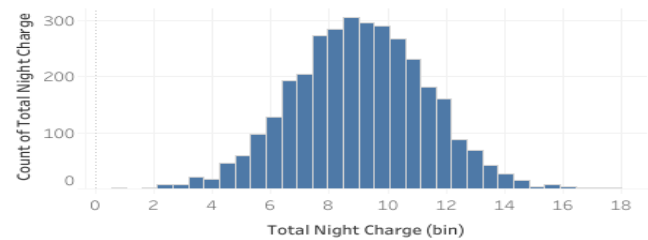


Fig- Tableau Dashboard 2

Link for both dashboards is,

https://public.tableau.com/profile/akshay.hirapara#!/vizhome/Dashboard_Churn_2/Dashboard2

https://public.tableau.com/profile/akshay.hirapara#!/vizhome/Dashboard_Churn/Dashboard1

Chapter-5

Machine Learning

There are three types of Machine Learning algorithms:

- **Supervised Machine Learning**
- **Un-Supervised Machine Learning**
- **Recommendation System**

1.) Supervised Machine Learning-

- We have target variable in this type of machine learning technique.
- Regression and Classification are types of supervised machine learning.

2.) Unsupervised Machine Learning-

- We do not have any target variable in this type of machine learning technique.
- Clustering and Text Mining comes under Un-supervised machine learning.

3.) Recommendation System-

- When we open Amazon Website or application and select a product to buy. It will show 'similar products.'
- When you are using Instagram, it will show you certain advertisements. That advertisements would be related to past searches on browser history.
- When you are using Twitter, you like certain tweets. Now your news feed would be customised in accordance with tweets you like, retweets etc

Above all are examples of Recommendation Systems.

As per our business problem, we need to predict 'Customer Churn'. The data provided contains a Target Variable Churn divided into sub-classes Yes and No. This is a pure classification problem.

- We will use Random Forest, Naïve Bayes, K-Nearest Neighbours and Logistic Regression to build our model.

Random Forest:

- The concept behind Random Forest is to generate 'n' number of trees to make our model accurate.

Why Random Forest is superior to Decision Tree?

- Suppose we have a data with crores of observations. Now a single decision tree will not be able to cover entire data to get proper meaning out of it. In such cases we will use Random Forest.
- Random Forest works on 2 algorithms:
 - Biernann's Bagging Idea- Once we will build a decision tree there would be some errors. Now when next tree is built we will feed that error to next tree to get meaningful decision out of it.
 - Random Selection of Features- It is random selection of data for building a decision tree.

```

F:/Data Scientist/Project_Churn/
Accuracy : 0.805
95% CI : (0.7852, 0.8238)
No Information Rate : 0.7247
P-Value [Acc > NIR] : 1.754e-14

Kappa : 0.4193
McNemar's Test P-Value : < 2.2e-16

Sensitivity : 0.9627
Specificity : 0.3900
Pos Pred Value : 0.8060
Neg Pred Value : 0.7991
Prevalence : 0.7247
Detection Rate : 0.6977
Detection Prevalence : 0.8656
Balanced Accuracy : 0.6764

'Positive' class : 1

> calculate_data(Conf_rf)
[1] "Accuracy :- 80.503899220156"
[1] "FNR :- 20.0892857142857"
[1] "FPR :- 19.4040194040194"
[1] "FPR :- 19.4040194040194"
[1] "precision :- 38.9978213507625"
[1] "recall/TPR :- 38.9978213507625"
[1] "Sensitivity :- 79.9107142857143"
[1] "Specificity :- 80.5959805959806"
> readrules[1:2,]
[1] "account.length<=0.129145024476263 & international.plan %in% c('1') & number.vmail.messages<=-0.037085027908129 & total.eve.calls<=-0.185466335761425 & total.night.minutes<=0.0433320477404131 & total.intl.minutes<=-2.061936101192"
[2] "account.length<=0.129145024476263 & international.plan %in% c('1') & number.vmail.messages<=-0.037085027908129 & total.eve.calls>-0.185466335761425 & total.night.minutes<=0.0433320477404131 & total.intl.minutes<=-2.061936101192"
> |

```

Fig- Output of Random Forest with Rules

- **Key Points:**

Why should we use Random Forest?

- Higher the number of trees, higher the accuracy and parameters of our model. We can build 100,300,500 trees.
- Random Forest can be used for a larger data set with 1000 independent variables without any variable getting deleted.
- This model also tells us what variables are suitable for a classification problem.
- This model can be used for both Classification as well as Regression.
- Random Forest works on concept of Gini Index. Gini Index measures impurity of data. We will calculate Gini Index and select that variable as parent node whose Gini Index is Lowest.
- If Gini Index (GI)=0, This would mean that Independent Variable carries important data to explain Target Variable.

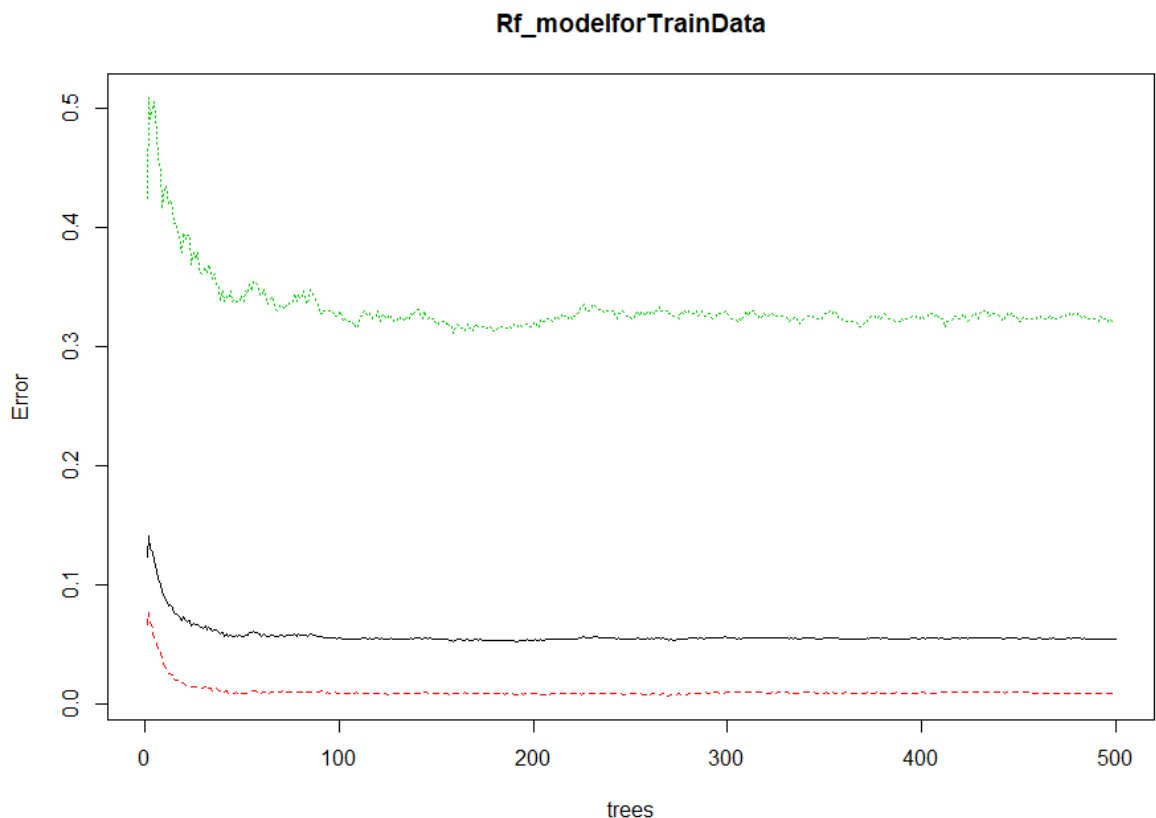


Fig-Trees vs Error Graph

- As we can see from the graph, after building 100 Trees the error rate is almost stagnant there is no significant change found.
- In the range of building 50-100 Trees, there is large variation found in error rate.

Logistic Regression:

Key Points:

- This algorithm can only be used for Classification.
- Data fed to the model can be Binomial, Multinomial or Ordinal.
 - Binomial= Target Variable having two classes e.g. Yes and No.
 - Ordinal=Target Variable have classes in an order e.g. High-Medium-Low.
 - Multinomial=Target variable have more than two classes. E.g. Rich-Medium-Poor

```

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3226  -0.6168   0.0000   0.6417   3.4133

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.054580   0.628746  -3.268  0.001084 **
state2         1.055659   0.717515   1.471  0.141218
state3         2.085018   0.783490   2.661  0.007786 **
state4         2.216615   0.791364   2.801  0.005094 **
state5         2.227110   0.776029   2.870  0.004106 **
state6         0.743437   0.877026   0.848  0.396616
state7         0.743535   0.926726   0.802  0.422365
state8         0.074496   0.871504   0.085  0.931880
state9         1.796991   0.735626   2.443  0.014574 *
state10        0.579313   0.799196   0.725  0.468532
state11       -0.556547   1.036420  -0.537  0.591275
state12        0.287849   0.899189   0.320  0.748877
state13        1.353929   0.917532   1.476  0.140046
state14        0.684254   0.732267   0.934  0.350080
state15        1.562112   0.787235   1.984  0.047222 *
state16        3.867558   0.769006   5.029  4.92e-07 ***
state17        0.808790   0.787943   1.026  0.304676
state18        2.083958   0.721953   2.887  0.003895 **
state19        1.555158   0.811698   1.916  0.055374 .
state20        0.170918   0.796267   0.215  0.830041
state21        1.089480   0.858354   1.269  0.204346
state22       -0.780758   1.058383  -0.738  0.460704
state23        1.202288   0.907736   1.324  0.185340
state24        1.814779   0.743391   2.441  0.014638 *
state25        1.416948   0.878451   1.613  0.106743

```

Fig- Output of Logistic Regression Model

- **Maths behind algorithm:**
 - It uses Logit link function

$$\text{Ln}\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

Fig-Logit Link Function

Above formula will be used to transform probabilities from 0 to 1 to logit score $-\infty$ to $+\infty$.

- To build a model there are 2 methods:]
 - OLS=Optimum Least Square
 - MLE= Maximum Likelihood Estimator

There are few assumptions of the model:

- **Ratio of Cases to Variables**
 - If I have two Target Class Yes and No. We have 1000 observations. Now the ration of both Yes and No should be equal.
- **Absence of Multicollinearity**
 - There should not be any correlation between
- **No Outliers**
 - Data should be free from Outliers
- **Independence of Errors**
 - Suppose we made a model and we got some errors. Now these errors should not follow same patterns.

K-Nearest Neighbours:

Key Points:

- This algorithm can be used for both Classification and Regression.
- This method is also known as Lazy Learning method. Unlike other algorithms it never stores patterns from training data.
- It is time consuming but very accurate.

Maths behind algorithm:

- **Euclidean Distance Formula:**
 - This formula is used when data contains all continuous variables.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Fig- Distance Formula

- **Manhattan Formula:**

- This formula will be used if our data contains both Continuous as well as Categorical variables.

$$D = \sum_{i=1}^n |x_i - y_i|$$

Fig-Manhattan Distance Formula

- **Weighted Distance Method:**

- Suppose we have 100 TV, now we have good business knowledge and we want to assign weights as per our knowledge. Here weighted distance method would be used.

$$\left(\sum_{i=1}^k (|x_i - y_i|)^q \right)^{1/q}$$

Fig-Weighted Distance Method

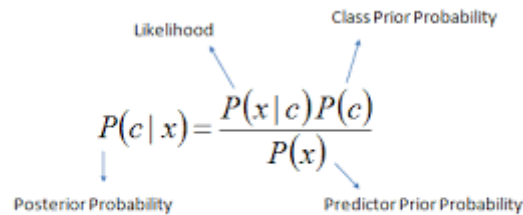
Naïve Bayes:

Key Points

- This algorithm works on probability theorem i.e. Bayes Theorem.
- It assumes that the presence of a feature is unrelated to presence of any other feature in a data.
 - E.g. A mobile is an I-phone if it has half cut apple behind, has SIRI and I-Tunes.
- Naïve Bayes is immune to Multi-collinearity effect. If we are choosing to apply Naïve Bayes than we do not need to perform correlation analysis because it treats all the terms individually.

E.g. GMAIL- If you receive mail and report it to spam folder. Now Naïve Bayes would run behind and learn the key words attached with that mail. Now whenever next email comes with similar keywords. It will automatically transfer it into Spam.

$$P(y|x_1, \dots, x_n) = \frac{P(x_1|y)P(x_2|y)\dots P(x_n|y)P(y)}{P(x_1)P(x_2)\dots P(x_n)}$$



The diagram shows the formula $P(c|x) = \frac{P(x|c)P(c)}{P(x)}$ with four labels and arrows pointing to the terms: 'Likelihood' points to $P(x|c)$, 'Class Prior Probability' points to $P(c)$, 'Posterior Probability' points to $P(c|x)$, and 'Predictor Prior Probability' points to $P(x)$.

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

Fig(a)-Bayes Formula

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Fig(b)-Bayes Formula

- The formula in fig (a) will be used in case we have continuous data.
- The formula in fig (b) will be used when we have both continuous as well as categorical data.

```

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace, type = "class")

A-priori probabilities:
Y
      1      2
0.5074919 0.4925081

Conditional probabilities:
state
Y      1      2      3      4      5      6      7
1 0.017971759 0.028241335 0.021822850 0.012836970 0.014120668 0.019255456 0.014120668
2 0.009259259 0.031746032 0.021164021 0.035714286 0.026455026 0.014550265 0.005291005
state
Y      8      9      10      11      12      13      14
1 0.016688062 0.014120668 0.021822850 0.016688062 0.015404365 0.011553273 0.041078306
2 0.009259259 0.043650794 0.019841270 0.006613757 0.006613757 0.011904762 0.017195767
state
Y      15      16      17      18      19      20      21
1 0.025673941 0.016688062 0.020539153 0.020539153 0.026957638 0.024390244 0.015404365
2 0.015873016 0.030423280 0.023809524 0.041005291 0.009259259 0.011904762 0.023809524
state
Y      22      23      24      25      26      27      28
1 0.019255456 0.016688062 0.029525032 0.008985879 0.021822850 0.025673941 0.010269576
2 0.005291005 0.006613757 0.023809524 0.015873016 0.014550265 0.015873016 0.009259259
state
Y      29      30      31      32      33      34      35
1 0.016688062 0.020539153 0.020539153 0.019255456 0.020539153 0.015404365 0.025673941
2 0.014550265 0.000000000 0.013227513 0.033068783 0.026455026 0.013227513 0.010582011
state
Y      36      37      38      39      40      41      42
1 0.016688062 0.021822850 0.024390244 0.016688062 0.023106547 0.017971759 0.015404365
2 0.019841270 0.029100529 0.025132275 0.000000000 0.009259259 0.010582011 0.011904762
state
Y      43      44      45      46      47      48      49

```

Fig-Naïve Bayes Output

- **Assumptions in Naïve Bayes:**
 - Data should be normally distributed.
 - Assumption of Independent Variables:
 - In real case, there would be rare chance that would get independent predictors i.e. every variable should be independent of each other.
- **Limitations of Naïve Bayes:**
 - There might be cases of zero frequency i.e. new test cases contains variables which are not present in Training cases. In such cases, probability would be zero.
 - This is called Zero Frequency.
 - To solve such instance, we should use SMOOTHING. One of the best methods is Laplace Estimation.

Now let's see how to evaluate the Machine Learning Model based on Classification problem,

Chapter-6

Evaluating Performance of Model

Confusion Matrix:

A confusion matrix is a technique to summarize performance of a Classification Model.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Fig-Confusion Matrix

- True Positive (TP): We predicted Positive and It is True
- True Negative (TN): We predicted Negative and It is True
- False Negative (FN): we predicted Negative and It is False
- False Positive (FP): We predicted Positive and its False

Terms:

- **Accuracy:**
 - It indicates how correct is our model
 - $TP+TN / TP+TN+FP+FN$
- **Mis-Classification Error:**
 - It indicates how incorrect is our model
 - $FP+FN / TP+TN+FP+FN$

- **Specificity (TNR-True Negative Rate):**
 - Negative cases which are correctly identified.
 - $TN/TN+FP$
- **Recall (True Positive Rate):**
 - Positive Cases which are Correctly identified.
 - $TP/TP+FN$
- **False Positive Rate:**
 - Positive Cases which are in-correctly identified.
 - $FP/FP+TN$
- **False Negative Rate:**
 - Negative cases which are in-correctly identified.
 - $FN/FN+TP$

Let us also understand AUC and ROC?

ROC

- Receiver Operating Characteristics
- This curve tells us about how good a model can distinguish between two things. (Example- If a customer will churn out or not)

AUC

- Area Under the Curve
- This score gives us a good idea of how well a model is performing.

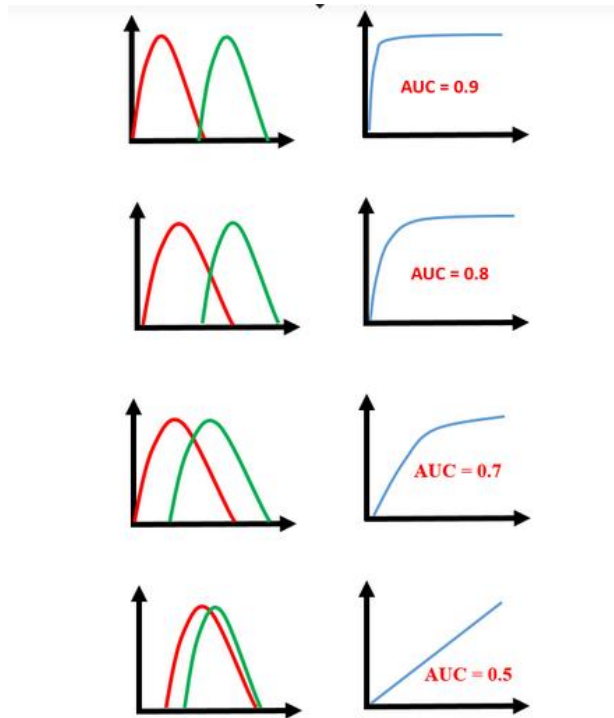


Fig-ROC Curve and AUC Score relationship

Let us see our Churn Project curve,

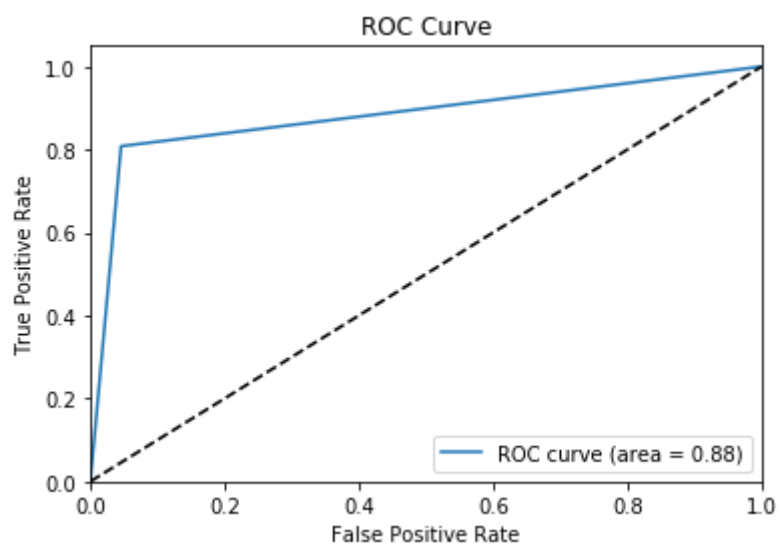


Fig-ROC Curve

Below I'm listing Confusion Matrix scores of every Models,

- **Random Forest:**

```
In [39]: CONFUSION_MATRIX(Y_Test,Predictions_rf)

Accuracy :- 95.74

Specificity // True Negative Rate :- 99.51

Sensitivity // True Positive Rate // Recall :- 71.43

False Negative Rate :- 28.57
```

Fig- Confusion Matrix for Random Forest

- **Logistic Regression:**

```
In [76]: CONFUSION_MATRIX(Y_Test,Predictions_LM)

Accuracy :- 87.1

Specificity // True Negative Rate :- 97.64

Sensitivity // True Positive Rate // Recall :- 19.2

False Negative Rate :- 80.8
```

Fig-Confusion Matrix for Logistic Regression

- **Naïve Bayes:**

```
In [81]: CONFUSION_MATRIX(Y_Test,Predictions_NB)

Accuracy :- 86.02

Specificity // True Negative Rate :- 93.07

Sensitivity // True Positive Rate // Recall :- 40.62

False Negative Rate :- 59.38
```

Fig- Confusion Matrix for Naïve Bayes

- **K Nearest Neighbors:**

```
In [85]: CONFUSION_MATRIX(Predictions_KNN,Y_Test)
```

```
Accuracy :- 89.08
```

```
Specificity // True Negative Rate :- 89.83
```

```
Sensitivity // True Positive Rate // Recall :- 75.0
```

```
False Negative Rate :- 25.0
```

Type *Markdown* and LaTeX: α^2

Fig- Confusion Matrix for K Nearest Neighbors**Conclusion:**

We would finalize Random Forest model as it has shown better accuracy and FNR rates and Recall Rates.