Gokhale Institute of Politics and Economics

TOPIC :- TELECOM CHURN

PROJECT NUMBER 2

Subject Code – IBEF-A-19

Submitted By
Group No. 9

Anish Dhopeshwarkar	IBEF2006
Indu Dahiya	IBEF2021
Nancy Ahlawat	IBEF2031
Tanu Goyal	IBEF2042
Yash Mehra	IBEF2050

Submitted to IMS ProSchool

INDEX

Topics	Name
Problem Statement, Data description and conclusion	Anish Dhopeshwarkar
Data Cleaning	Indu Dahiya
EDA	Nancy Ahlawat
Model(Decision Tree)	Tanu Goyal
Model and Analysis	Yash Mehra

PROBLEM STATEMENT (ANISH)

- All over the world, numerous telecom companies are present. To keep up in the competition and expand their business client have to invest in the market. But, due to increasing competition, company is facing severe loss of revenue and loss of potential customers. So, the client wants to find out the reasons of losing customers by measuring customer loyalty to regain the lost customers.

- **Aim** :- We want to know what are the reasons that lead people to churn or retain by studying the dataset. This will tell us in which areas we should focus to reduce the churn rate.

DATA DESCRIPTION

- There are 3333 rows and 17 columns in the dataset
- There are 4 categorical, 2 discrete, 10 continuous and 1 Boolean variable in the dataset
- The target variable is Churn, which is a Boolean.
- Churn is the percentage of subscriber to a service that discontinue their subscription to that service in a given time period.

Variable	Type
State	Object
account length	Integer
area code	Integer
phone number	Object
international plan	Object
voice mail plan	Object
number vmail messages	Float
Total day calls	Float
Total day charges	Float
Total eve calls	Float
Total eve charges	Float
Total night calls	Float
Total night charges	Float
Total intl calls	Float
Total intl charges	Float
Customer service calls	Float
Churn	Boolean

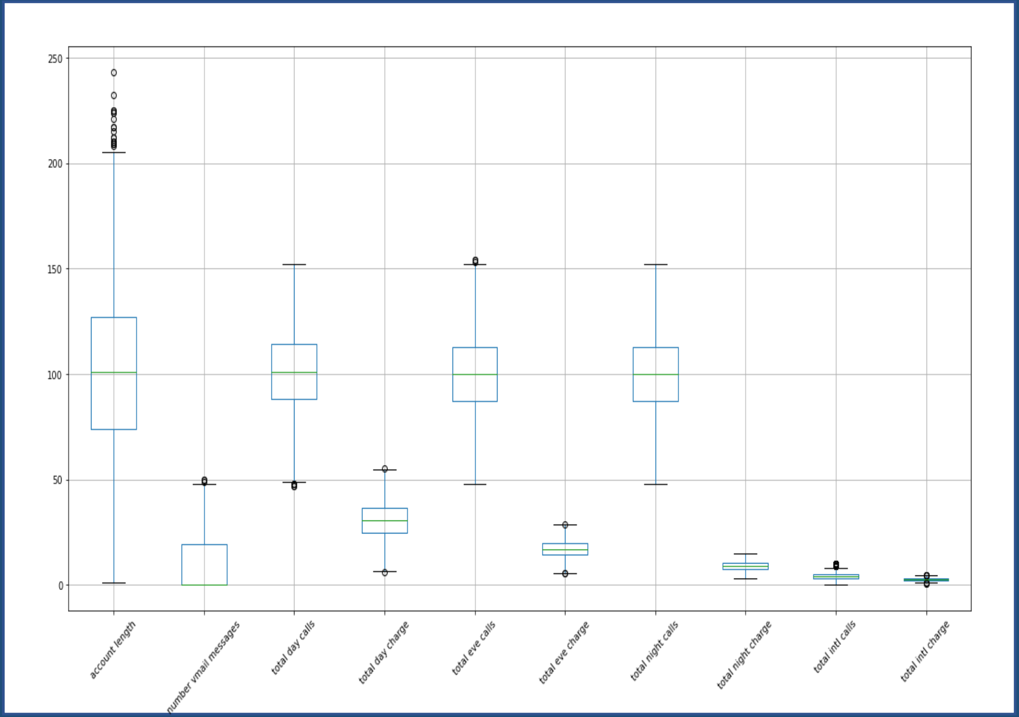
DATA CLEANING (INDU)

- **Treatment of missing values-** The table here shows the number of missing values in the dataset and their respective percentage
- Since these values are less than 5% of the total value of the respective column, thus instead of replacing them, we remove the null values from the dataset.
- Since the variable phone number has no significance, we dropped it.

Column	Number of null values	Percentage of null values
number vmail messages	1	0.0003
Total day calls	2	0.0006
Total day charge	3	0.0009
Total eve calls	2	0.0006
Total eve charge	4	0.0009
Total night calls	3	0.0009
Total night charge	1	0.0003
Total intl calls	2	0.0006
Total intl charge	3	0.0009
Customer service calls	1	0.0003

TREATMENT OF OUTLIERS

- An outlier is an observation that diverges from well-structured data.
- The root cause for the Outlier can be an error in measurement or data collection error.
- We can either delete the outliers or replace them with average values. They are replaced with mean or median in quantitative data and mode in case of qualitative data.
- We are using box plot to identify the outliers in each variable.
- Then we are replacing those outlier values with median value of the concerned variable.



EXPLANATORY DATA ANALYSIS (NANCY)

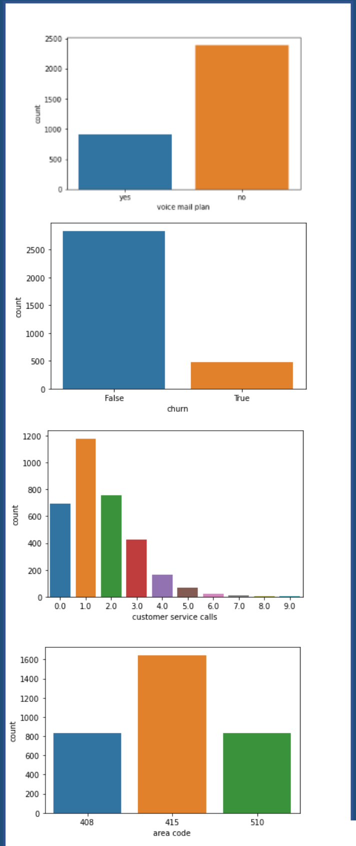
- EDA is a practice of iteratively asking a series of questions about the data at your hand and trying to build hypotheses based on the insights you gain from the data.
- Under EDA we employ graphical tools to analyze the data to identify patterns and relationships.
- In our analysis we have visualization and made histogram, box plot, bar chart and scatter diagram to analyze the data.
- We have done three things for out Explanatory Data Analysis

1) Univariate Analysis

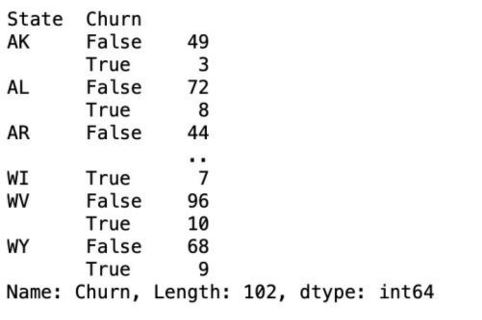
2) Bivariate Analysis

3) Multivariate Analysis

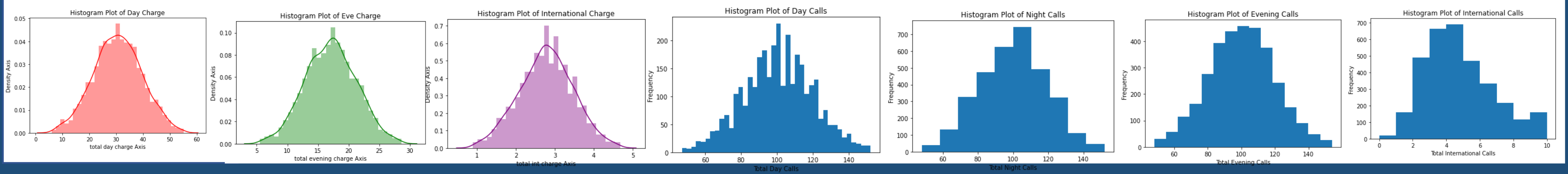
UNIVARIATE ANALYSIS



- Univariate Analysis of all the variables is done by taking each variable separately. We have made bar graph for discrete and qualitative variables and histogram quantitative variables.
- 14% of customers have discontinued telecom services
- 49% of customers belong to Area Code 415
- 36% of customers gave only one customer service call
- 9% of customers use international plan
- 27% of customers use voice mail plan
- Total eve charge, total day calls, total eve calls
- and total day charge columns are uniformly distributed
- Total night calls are moderately skewed and total intl charge are moderately left skewed.
- Total intl calls are rightly skewed.

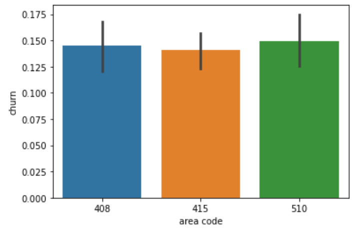


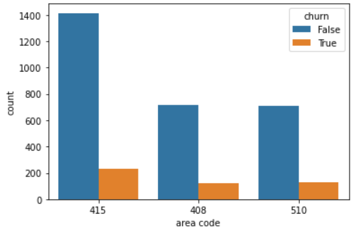
```
State  Churn
AK     False    49
       True      3
AL     False    72
       True      8
AR     False    44
                ..
WI     True      7
WV     False    96
       True     10
WY     False    68
       True      9
Name: Churn, Length: 102, dtype: int64
```

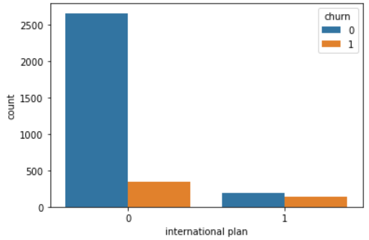


BIVARIATE ANALYSIS

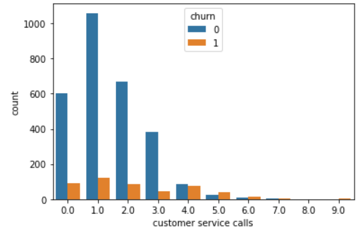
- We have used heatmap to show the correlation of every variable with every other variable.
- To show specific relations between two variables we have use Side Bar Graph and then used different colours to show churn.
- total day charge, total eve charge and customer service calls influence churn.
- account Length, area code, total day calls, total eve calls, total night calls and total intl charge has moderate positive relationship with Churn.
- Number of voice mail messages, total night charge and total intl charge has moderately negative relationship with Churn.

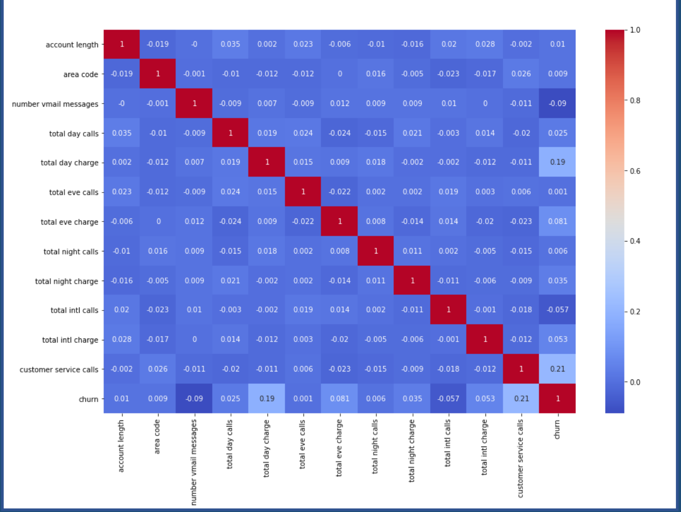




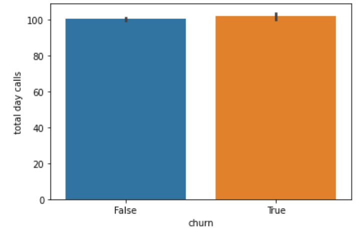


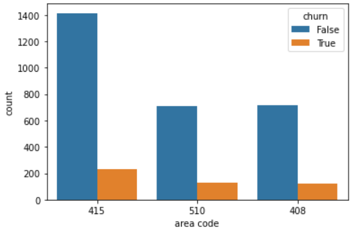
- 90% of customers who have international plan discontinue their telecom services, i.e. they churn.
- This company gets half of it’s revenue from the area code 415.





- Churn rate is approximately same for all three area code(14-15%)
- Proportion for churn rate is very high for more than three customer service calls.

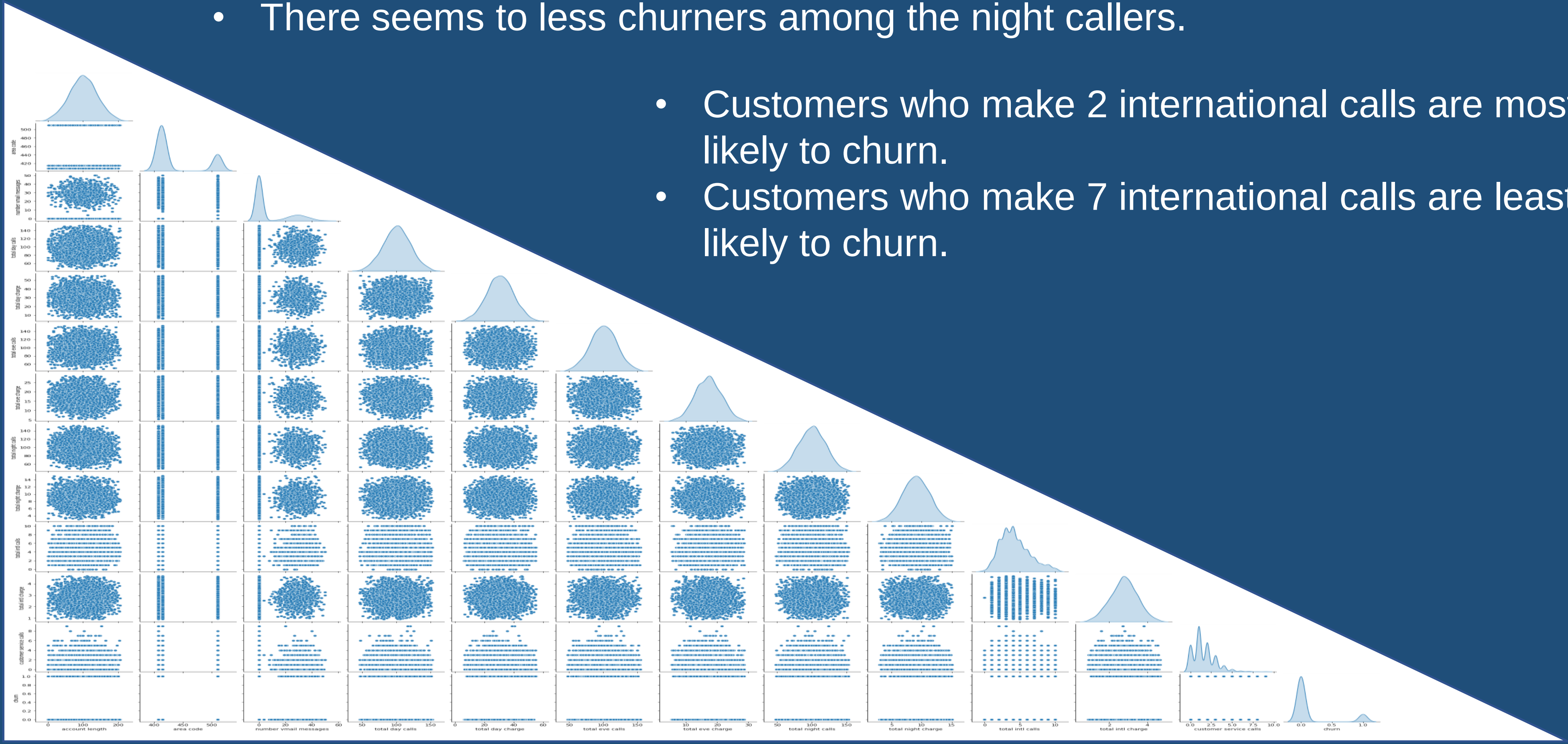


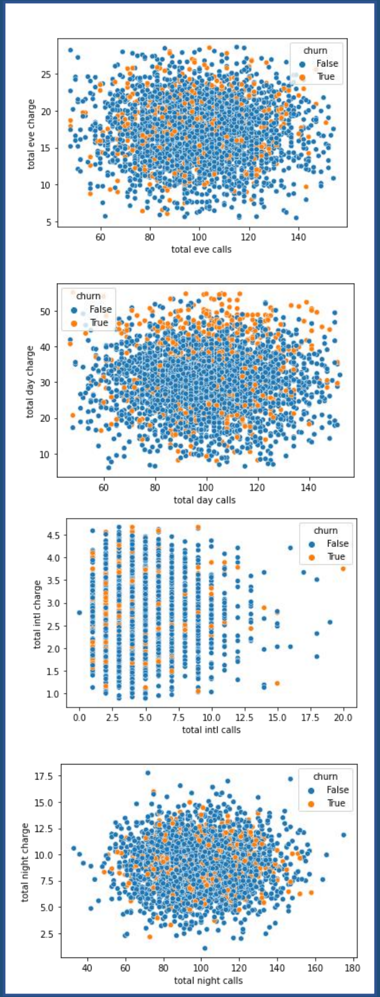


MULTIVARIATE ANALYSIS

- There seems to less churners among the evening callers.
- Churn rate increases at higher rate when the total day charge is more than 40.
- There seems to less churners among the night callers.

- Customers who make 2 international calls are most likely to churn.
- Customers who make 7 international calls are least likely to churn.

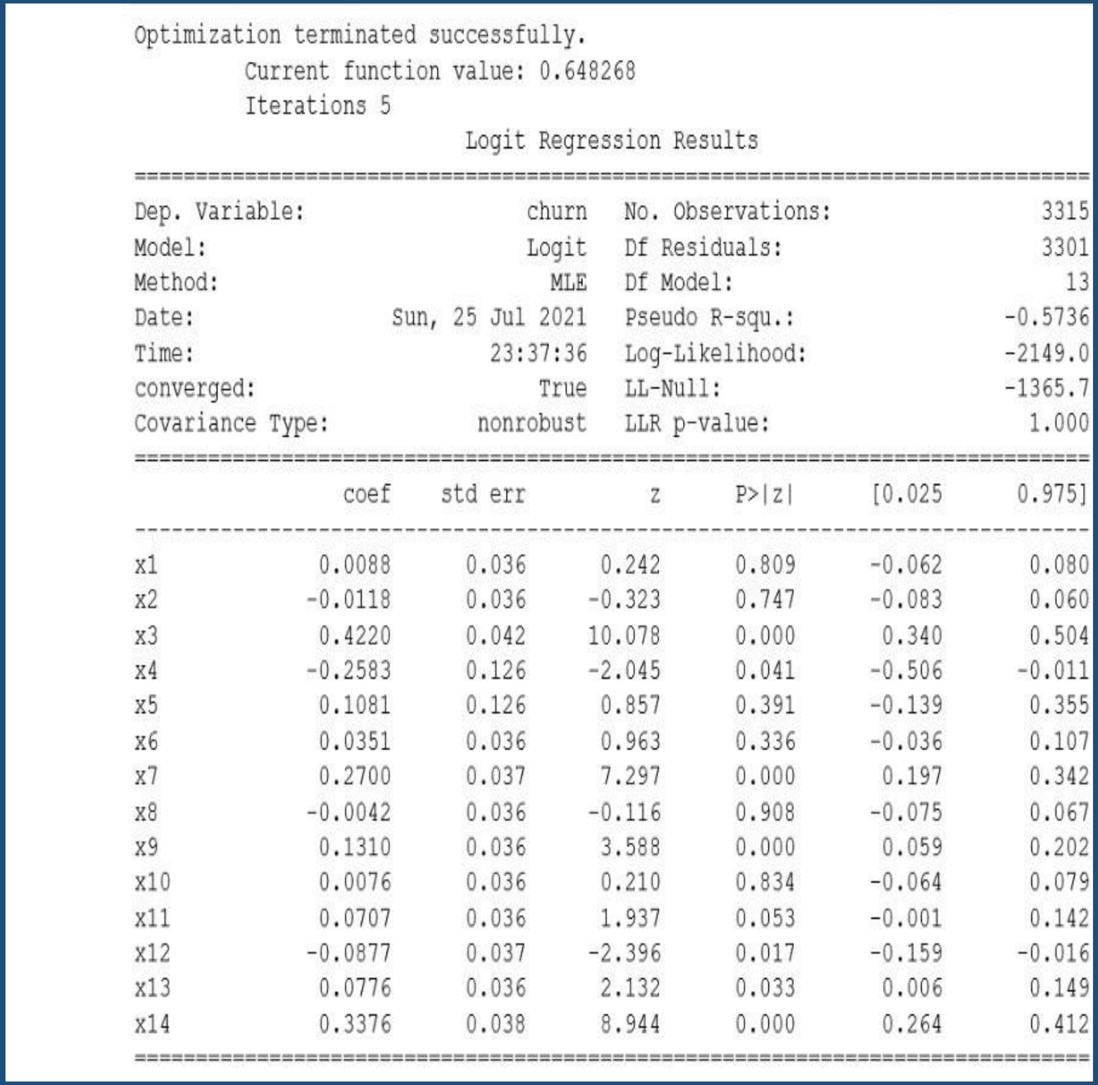




PREPARATION OF MODEL(YASH)

- After all the explanatory data analysis, data visualization and data cleaning, we are left with 3315 data points and 17 columns.
- In telco churn data Churn, Voice mail plan, and International plan in particular are binary features that can easily be converted into 0’s and 1’s. Hence we change it into dummy variables.
- In order to further analyze the data using various Machine Learning models. We split our dataset into train and test in 80:20 ratio.
- This is a two way classification problem. We perform four classification models and compare them on the basis of various matrices which are defined in the python file itself. We use the following models -logistic regression, random forests, decision trees and KNN classifier. And we compare the result of various models on basis of four matrices – Accuracy score, Recall value, Precision score and F1 value.

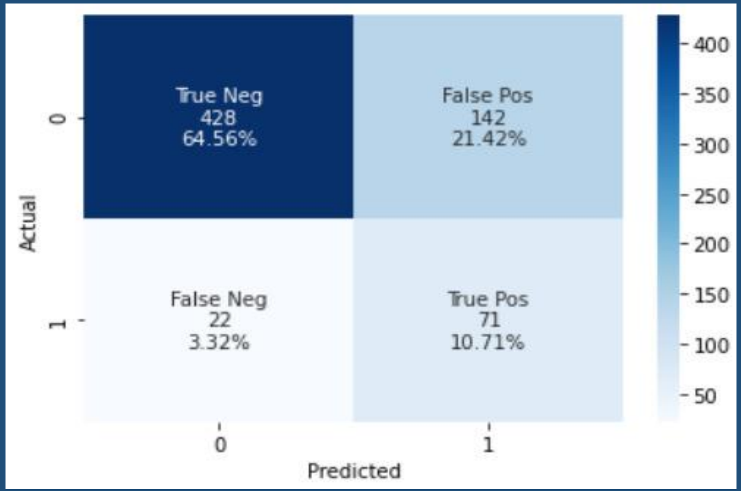
LOGIT REGRESSION MODEL



```
Optimization terminated successfully.
         Current function value: 0.648268
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:                  churn   No. Observations:                 3315
Model:                          Logit   Df Residuals:                     3301
Method:                           MLE   Df Model:                           13
Date:                Sun, 25 Jul 2021   Pseudo R-squ.:                 -0.5736
Time:                        23:37:36   Log-Likelihood:                -2149.0
converged:                       True   LL-Null:                       -1365.7
Covariance Type:            nonrobust   LLR p-value:                     1.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.0088      0.036      0.242      0.809      -0.062       0.080
x2            -0.0118      0.036     -0.323      0.747      -0.083       0.060
x3             0.4220      0.042     10.078      0.000       0.340       0.504
x4            -0.2583      0.126     -2.045      0.041      -0.506      -0.011
x5             0.1081      0.126      0.857      0.391      -0.139       0.355
x6             0.0351      0.036      0.963      0.336      -0.036       0.107
x7             0.2700      0.037      7.297      0.000       0.197       0.342
x8            -0.0042      0.036     -0.116      0.908      -0.075       0.067
x9             0.1310      0.036      3.588      0.000       0.059       0.202
x10            0.0076      0.036      0.210      0.834      -0.064       0.079
x11            0.0707      0.036      1.937      0.053      -0.001       0.142
x12           -0.0877      0.037     -2.396      0.017      -0.159      -0.016
x13            0.0776      0.036      2.132      0.033       0.006       0.149
x14            0.3376      0.038      8.944      0.000       0.264       0.412
==============================================================================
```

X = Independent variable, Y dependent variable. X1= Account Length,X2 = Area Code, X3 = International Plans, X4 = Voice mail plans, X5 = Number Vmail message, X6 =Total Day calls, X7 = Total Day charge, X8 = Total eve calls, X9 = Total eve charge , X10 = Total night calls, X11 = Total night charge, X12 = Total intl calls, X13 = Total intl charge, X14 = Customer Service calls and Y = Churn.

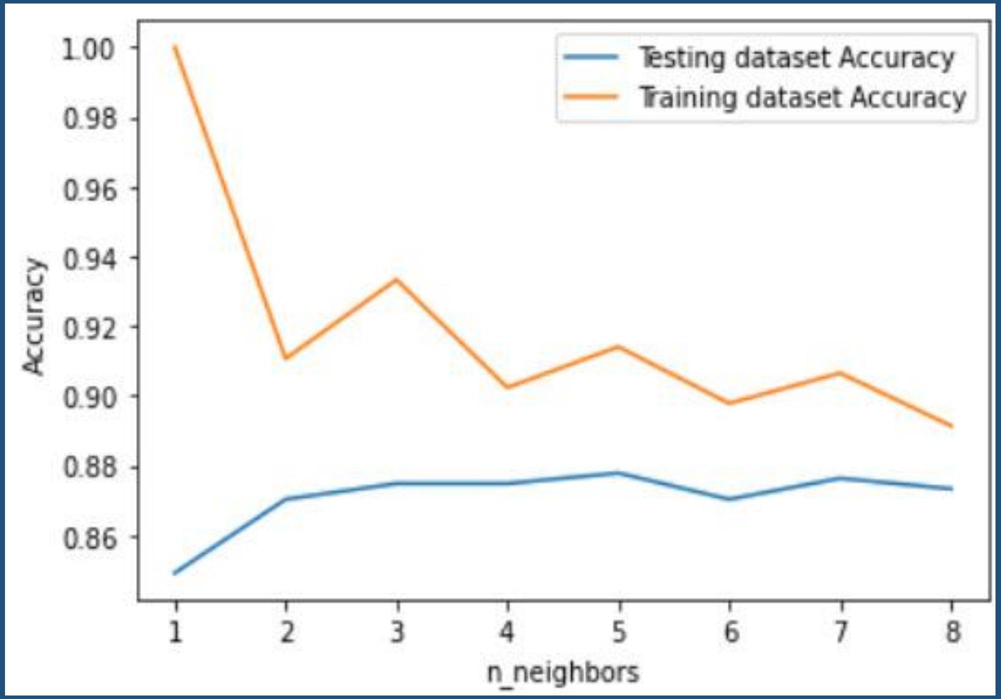
Confusion Matrix



KNN CLASSIFIER

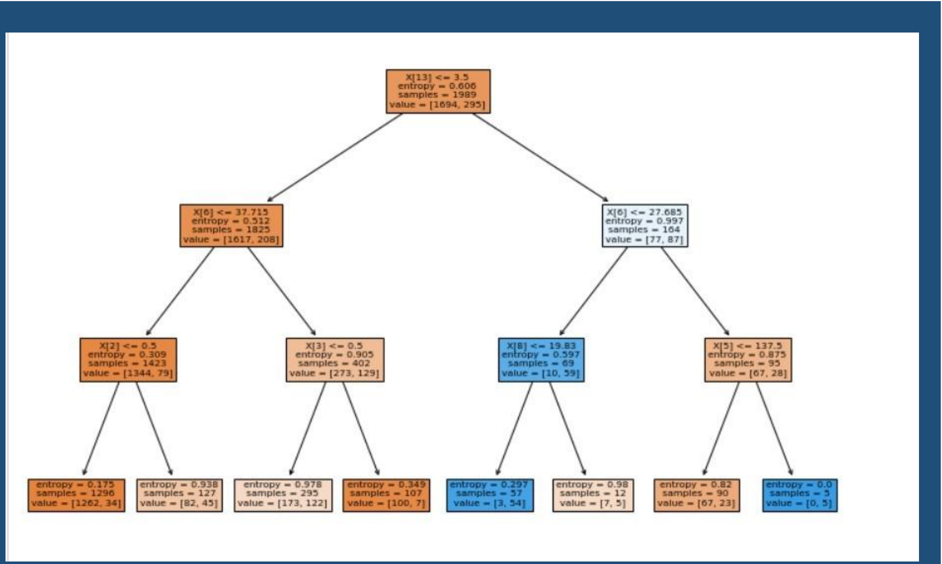
Performance Analysis of Test Data

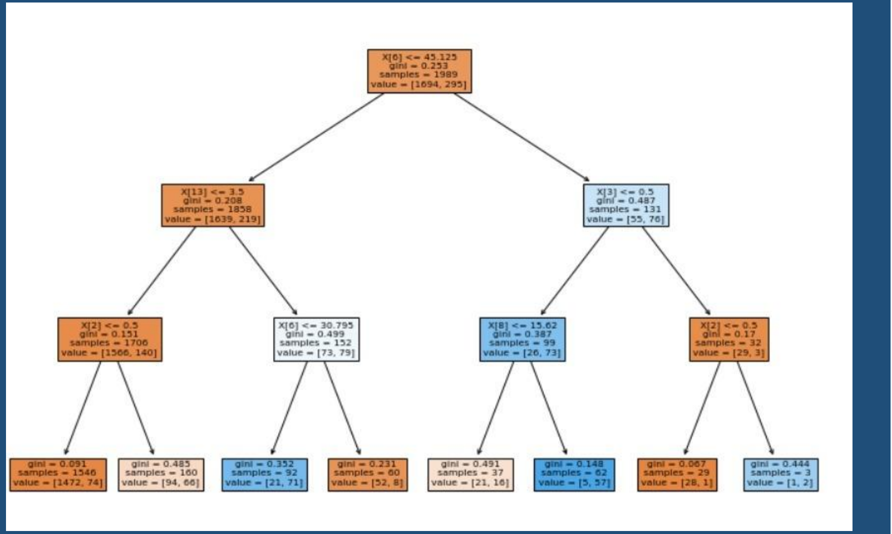
Matrices	Score
Accuracy	87%
Precision Score	76%
Recall	15%
F1	25%



DECISION TREE(TANU)

- In Decision-Tree Classifier two models are used ,one with criterion **Gini Index** and second with **Entropy**.



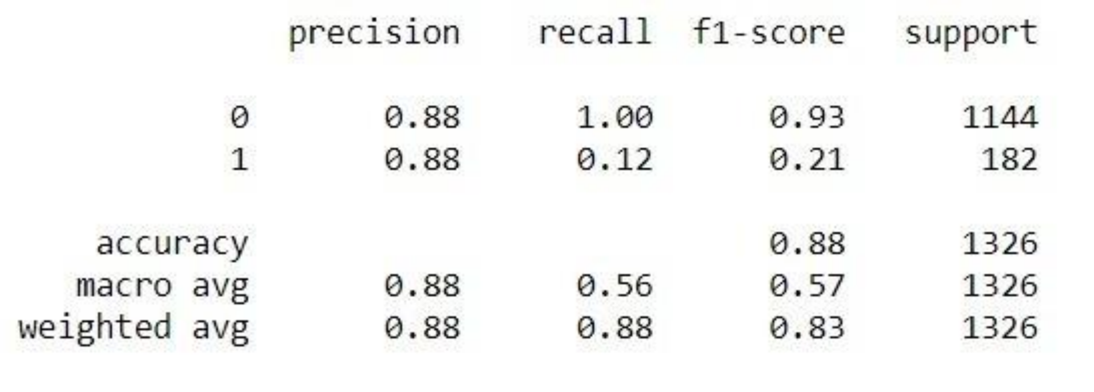


Entropy
Model Accuracy = 0.8771

Gini Index
Model Accuracy = 0.9005

Important Observations

- No overfitting of model was observed.
- This model suggests that total day charge is the most significant independent variable in the prediction of dependent variable (churn).
- The training-set accuracy score is 0.9057 while the test-set accuracy to be 0.9005. These two values are quite comparable. So, there is no sign of overfitting in the Gini model.



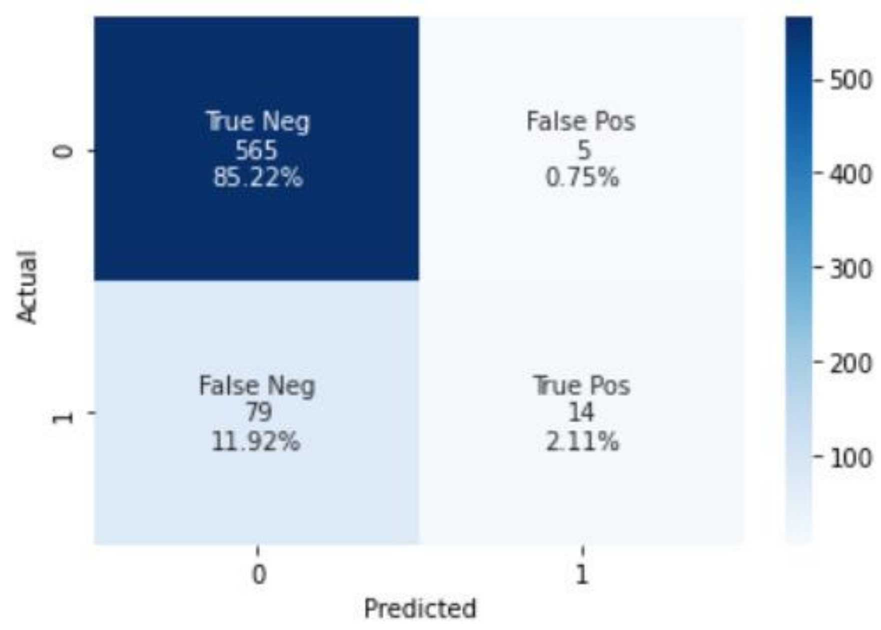
```
              precision    recall  f1-score   support

           0       0.88      1.00      0.93      1144
           1       0.88      0.12      0.21       182

    accuracy                           0.88      1326
   macro avg       0.88      0.56      0.57      1326
weighted avg       0.88      0.88      0.83      1326
```

RANDOM FOREST CLASSIFIER(YASH)

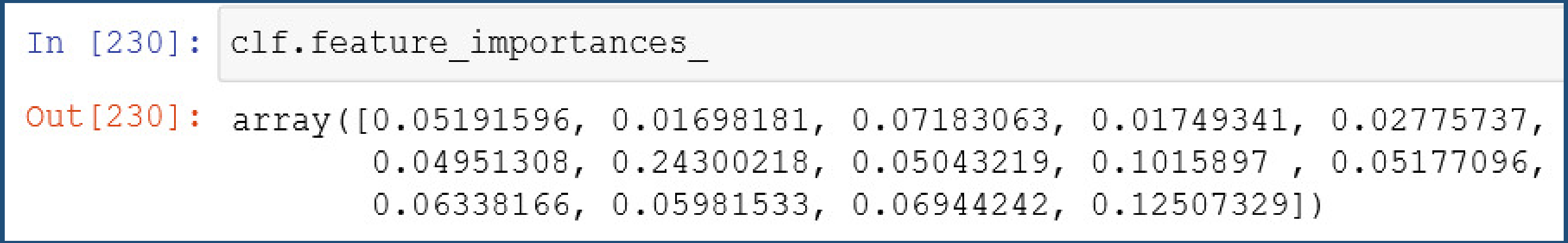
Performance Analysis of Test Data

Matrices	Score
Accuracy	95%
Precision	95%
Recall	66%
F1	78%

Confusion Matrix



Predicted Important Variables



```
In [230]: clf.feature_importances_

Out[230]: array([0.05191596, 0.01698181, 0.07183063, 0.01749341, 0.02775737,
                 0.04951308, 0.24300218, 0.05043219, 0.1015897 , 0.05177096,
                 0.06338166, 0.05981533, 0.06944242, 0.12507329])
```

EVALUATION OF THE MODELS

	Accuracy Score	Recall value	Precision score	F1 score
Logit model	75%	76%	33%	46%
KNN Classifier	87%	15%	73%	25%
Decision Tree Classifier	87%	56%	88%	57%
Random Forest Classifier	95%	66%	95%	78%

- From the above table we can observe that among all 4 classifiers random forest classifier gave us better accuracy score of 95% and recall value as 66% which is low. This implies that the model making more mistake in predicting churn customer as non churn customer. In logistic model the recall value is 76% which is highest among all the 4 model. So this model is predicting churned customers as churned customers 76 times out of 100 and churn as non-churn just 24 times of 100.But precision value score is very low in logit model as compared to random forest classifier which has a precision value of 95%, highest of all. While KNN model has only 15 % recall value that means it predicts non-churn customers as churn customers, so we wont consider this model as appropriate.
- *No model can predict the churn factor accurately, so we have to depend on different model for predicting the variable of churn and make a suggestion based on that for the client.

ANALYTICS

- The question arise what factors or variables out of the given dataset are responsible for user to churn.

Analysis from Logit Model

- From Logit model we can conclude that the statistical findings here ties back nicely to our own observations in the initial exploratory analysis phase. We noted that users on the International plan had a significantly higher churn than customers on other plans.
- From our explanatory analysis phase as well as machine learning model we also noticed that churn spiked among customers who exceeded 3 customer service calls.
- Higher the total day charge, more likely it is for the users to churn.
- Similar is the case for eve charge, higher total eve charge, more likely the users to churn.
- We also noted that users on the voice mail plans had a significantly higher churn than customers without voice mail plans.

Analysis from Random Forest

- Higher International Charge leads to higher churn rate.
- Higher night charge more likely the user is to churn.
- Higher international calls had higher possibility of churning.
-

Analysis from EDA

- While California is the most populous state in the U.S, there are not as many customers from California in our dataset. Arizona (AZ), for example, has 64 customers, 4 of whom ended up churning. In comparison, California has a higher number (and percentage) of customers who churned. This is useful information for a company.

CONCLUSION AND RECOMMENDATION(ANISH)

- From our analytical finding we can suggest that a good strategy to address churn can be enhanced focus on these drivers
- International Plan, Customer Service Call, charges which includes – day, evening, night, and international, voice mail plan, international call and State.

Recommendations

- Survey International plan customers to understand pain points and identify root causes for churn intent. Then take steps to address those concerns and ensure that appropriate service is provided.
- Escalate all calls beyond the 1st customer call to ensure that any issues that the customer faces is fixed before the next call. Proactively check on customers to confirm that their issue is fixed.
- Survey customers whose call minutes/charges are above average to check for churn intent. Identify root causes for churn intent and take steps to alleviate those concerns.
- Survey voicemail messages plan customers to understand pain points and identify root causes for churn intent. Then take steps to address those concerns.
- Client must focus on states like California, which is the largest state in US, but has very less customer base as well highest churning rate as compared to other states. Client must focus on improving its service and reputation in California.