TELECOMMUNICATIONS

Under the Gudience:

Neha guptha
Kishore.Rajashekaran

Team members

Lekha Sree Yenneti

Jaya Lakshmi

Karavadi

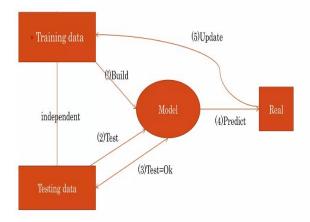Shirisha Muddangula

Vaishnavi Dappu

Bhadra Laxmi Atike

Akshitha Mudhiraj

HemaLatha Gorantla

Project Objective :

- To Predict Customer Churn .

- Highlighting the main variables or factors influencing customer churn.

- Use various ML classification algorithms to build prediction models, evaluate the accuracy.

- And performance of these models.

- Finding out the best model for our business case & providing executive summary.

CHURN PREDICTION MODEL

DataSet Details:

The data file telecommunications_churn.csv contains a total of 19 features for 3333 customers. Each row corresponds to a client of a telecommunications company for whom it has been collected information about the type of plan they have contracted, the minutes they have talked, or the charge they pay every month.

EDA

```
data.info()   # Information about dataframe
```
executed in 23ms, finished 19:58:18 2022-12-27

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3333 entries, 0 to 3332
Data columns (total 19 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   account_length         3333 non-null   int64
 1   voice_mail_plan        3333 non-null   int64
 2   voice_mail_messages    3333 non-null   int64
 3   day_mins               3333 non-null   float64
 4   evening_mins           3333 non-null   float64
 5   night_mins             3333 non-null   float64
 6   international_mins      3333 non-null   float64
 7   customer_service_calls 3333 non-null   int64
 8   international_plan      3333 non-null   int64
 9   day_calls              3333 non-null   int64
 10  day_charge             3333 non-null   float64
 11  evening_calls          3333 non-null   int64
 12  evening_charge         3333 non-null   float64
 13  night_calls            3333 non-null   int64
 14  night_charge           3333 non-null   float64
 15  international_calls     3333 non-null   int64
 16  international_charge    3333 non-null   float64
 17  total_charge            3333 non-null   float64
 18  churn                  3333 non-null   int64
dtypes: float64(9), int64(10)
memory usage: 494.9 KB
```

```
data.describe(include='all').T   # discreptive statistics of the data (T means tabular form)
```
executed in 91ms, finished 19:58:18 2022-12-27

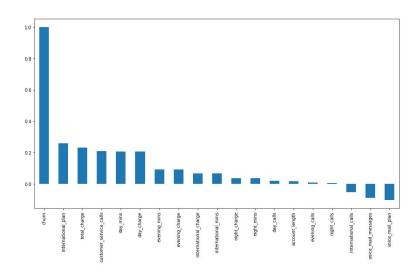|	count	mean	std	min	25%	50%	75%	max
account_length	3333.0	101.064806	39.822106	1.00	74.00	101.00	127.00	243.00
voice_mail_plan	3333.0	0.276628	0.447398	0.00	0.00	0.00	1.00	1.00
voice_mail_messages	3333.0	8.099010	13.688365	0.00	0.00	0.00	20.00	51.00
day_mins	3333.0	179.775098	54.467389	0.00	143.70	179.40	216.40	350.80
evening_mins	3333.0	200.980348	50.713844	0.00	166.60	201.40	235.30	363.70
night_mins	3333.0	200.872037	50.573847	23.20	167.00	201.20	235.30	395.00
international_mins	3333.0	10.237294	2.791840	0.00	8.50	10.30	12.10	20.00
customer_service_calls	3333.0	1.562856	1.315491	0.00	1.00	1.00	2.00	9.00
international_plan	3333.0	0.096910	0.295879	0.00	0.00	0.00	0.00	1.00
day_calls	3333.0	100.435644	20.069084	0.00	87.00	101.00	114.00	165.00
day_charge	3333.0	30.562307	9.259435	0.00	24.43	30.50	36.79	59.64
evening_calls	3333.0	100.114311	19.922625	0.00	87.00	100.00	114.00	170.00
evening_charge	3333.0	17.083540	4.310668	0.00	14.16	17.12	20.00	30.91
night_calls	3333.0	100.107711	19.568609	33.00	87.00	100.00	113.00	175.00
night_charge	3333.0	9.039325	2.275873	1.04	7.52	9.05	10.59	17.77
international_calls	3333.0	4.479448	2.461214	0.00	3.00	4.00	6.00	20.00
international_charge	3333.0	2.764581	0.753773	0.00	2.30	2.78	3.27	5.40
total_charge	3333.0	59.449754	10.502261	22.93	52.38	59.47	66.48	96.15
churn	3333.0	0.144914	0.352067	0.00	0.00	0.00	0.00	1.00

Break down our features

- Account Length : the number of days that this account has been active.

- VMail Plan : whether the customer has a voice mail feature: yes/no.

- VMail Message : presumably the average number of voice mail messages per month.

- Day Mins : the total number of calling minutes used during the day.

- Night Mins : the total number of calling minutes used during the Night.

- evening Mins : the total number of calling minutes used during the evening.

- Intl Mins : the total number of international minutes.

- CustServ Calls : the number of calls placed to Customer Service.

- Int'l Plan : whether the customer has an international calling plan: yes/no.

- Day Calls : the total number of calls placed during the day.

- Day Charge : the billed cost of daytime calls.

- Eve Calls : the total number of calls placed during the evening.

- Eve Charge : the billed cost of evening time calls.

- Night Calls : the total number of calls placed during the night.

- Night Charge : the billed cost of nighttime calls.

- Intl Calls : the total number of international calls.

- Intl Charge : the billed cost for international calls.

- Total Charge : the total cost summing day, eve, night, international charges.

- Churn : whether the customer left the service: true/false.

- A higher number of calls placed to customer service in a given period indicates that a customer is facing many problems, and hence, there is a high potential of churning.

- Customers with high total billed cost are more likely searching for another operator if they are unhappy with the current service.

- There are no Null values

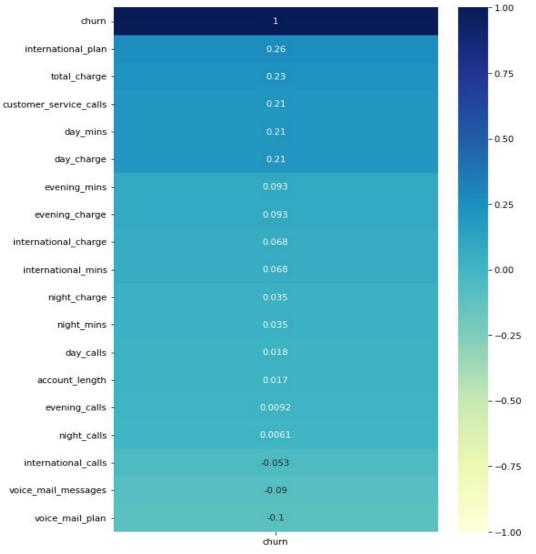- There are no duplicated values.

Correlation:

correlation is used to measure the relationship between variables how strongly associated or how weakly associated or no association
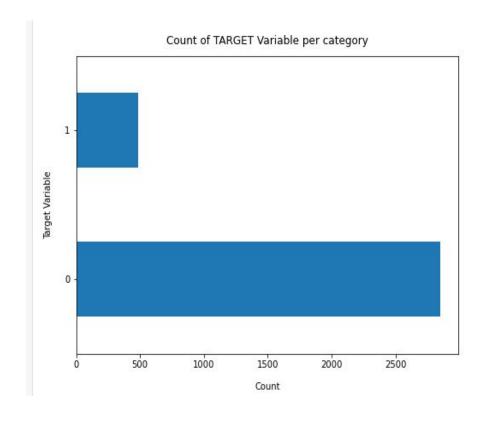
(Except international calls, voice mail messages and voice mail plan everything is positively related to churn.)
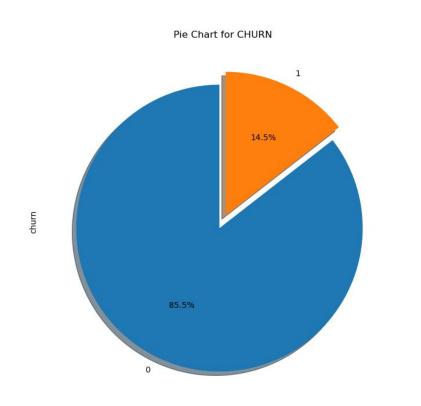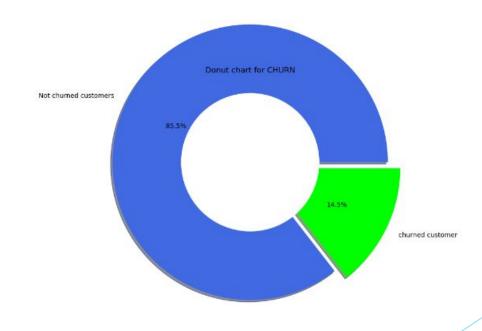

Features Correlating with Churn

	churn
churn	1
international_plan	0.26
total_charge	0.23
customer_service_calls	0.21
day_mins	0.21
day_charge	0.21
evening_mins	0.093
evening_charge	0.093
international_charge	0.068
international_mins	0.068
night_charge	0.035
night_mins	0.035
day_calls	0.018
account_length	0.017
evening_calls	0.0092
night_calls	0.0061
international_calls	-0.053
voice_mail_messages	-0.09
voice_mail_plan	-0.1

COUNT Plot to analyze CHURN

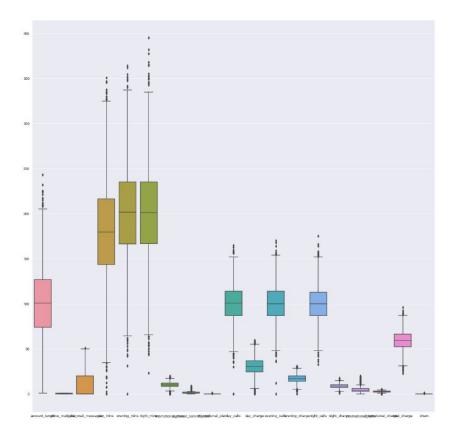

Count of TARGET Variable per category

- Count Plot is used to show the count of observations in each features bin using bars.

- Here Y is our target variable churn(1,0)

- 0 indicates NO CHURN

- 1 indicates CHURN

PIE & DONUT Chart to analyze CHURN

OUTLIERS(BoxPlot)

- ► Account length(18),
- ► Voice mail plan(0),
- ► Voice mail messages(1),
- ► Day mins(25),
- ► Evening mins(24),
- ► Night mins(30),
- ► International mins(46),
- ► Customer service calls(267),
- ► International plan(323),
- ► Day calls(23),
- ► Day charge(323),
- ► Evening calls(22),
- ► Evening charge(24),
- ► Night calls(22),
- ► Night charge(30),
- ► International calls(78),
- ► International charge(49),
- ► Total charge (27),

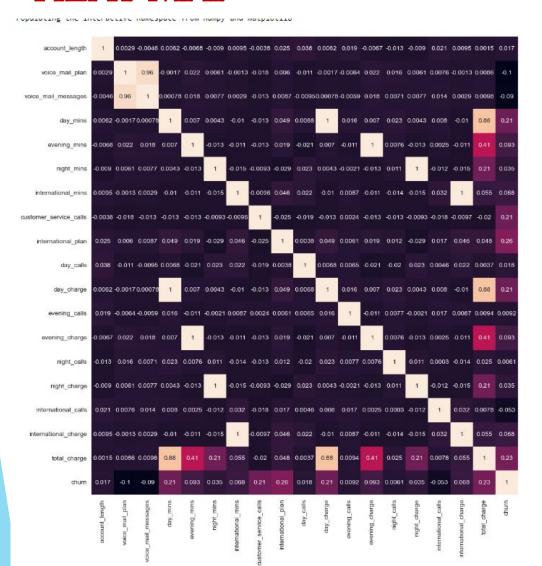**TOTAL OUTLIERS(1332).**

OUTLIER DETECTION AND REMOVAL

Outliers are nothing but values . Those are distinctly different from other values.

(extream values of our data.)

Outliers will gives the correct idea about central value of given dataset.

Here We Use IQR Method For Detecting Outliers and removed

HEAT MAP

The Heat map is used to find the highly correlated values of given dataset

- By using correlation threshold method we got Voicemail messages as a highly correlated feature

Feature Engineering

► By Analysing the Dataset , We had drawn some Insights Required for Model Building.

► In feature engineering we used correlation threshold method for dropping highly correlated columns or less correlated columns.
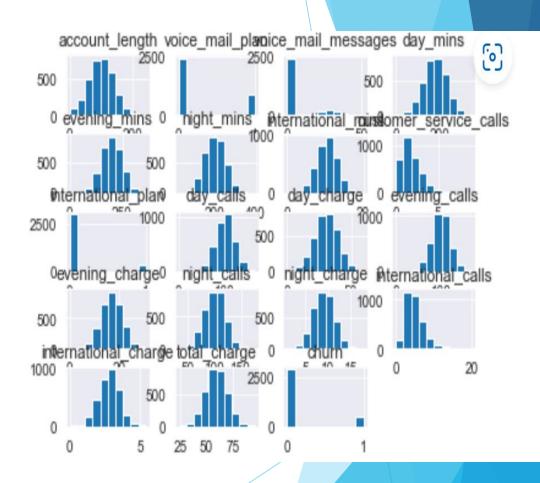
• In the given dataset , we dropped voice mail message and churn in X input variable and Y consist of our target variable Churn.

SMOTE+ENN Technique

To change imbalanced data to balanced data we use SMOTE+ ENN Technique.

Imbalanced
learn

Model Building :

Decision Tree :   A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.

Support Vector Meachine :   A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they're able to categorize new text.

Random Forest :   Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned.
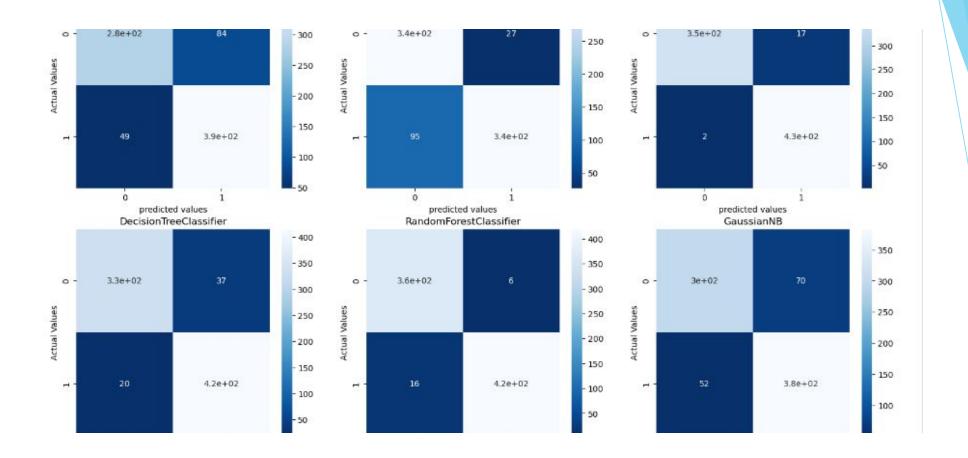
AdaBoost Classifier :   An AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

 AdaBoost Algorithm is also known as Adaptive Boosting is an Ensemble modelling technique used in Machine Learning to find the best model.

KNeighborsClassifier :   KNeighborsClassifier is for K nearest neighbor. Standardization of datasets is a common requirement for many machine learning estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data.

Naive Bayes :   Naïve Bayes is a probabilistic machine learning algorithm based on the Bayes Theorem, used in a wide variety of classification tasks.

Logistic Regression :   Logistic regression is commonly used for prediction and classification problems. Some of these use cases include: Fraud detection: Logistic regression models can help teams identify data anomalies, which are predictive of fraud.

- *A confusion matrix is a table that is used to define the performance of a classification algorithm.*
- *A confusion matrix visualizes and summarizes the performance of a classification algorithm.*
- *Confusion Matrix is a useful machine learning method which allows you to measure Recall, Precision, Accuracy, and AUC-ROC curve.*

Conclusion :

In the above 7 models,

Random Forest Classifier is given the best accuary values.

Finally move on the deployement which model gives the highest Accuracy value,to move on deployement.

To observe the all models random forest classifier given the highest accuracy value.

	Model	Accuracy	AUC
0	LogisticRegression	0.728019	0.69
1	SVC	0.698710	0.63
2	KNeighborsClassifier	0.928488	0.91
3	DecisionTreeClassifier	0.880422	0.87
4	RandomForestClassifier	0.960141	0.96
5	GaussianNB	0.762016	0.75
6	AdaBoostClassifier	0.844080	0.84

Deployment :

Challenges Faced :

✔ **First of all we faced the challenges to understand the the data and we have learn only class room teaching classification methods.**

✔ **To search different classifications models and to understand the model concepts very to difficult but to need to concentrate the concept is easy.**

✔ **Local URL: http://localhost:8501**

✔ **Network URL: http://100.126.94.145:8501**

THANK YOU

We team welcomes yours doubts ;
Reach us @

berries.lekha@gmail.com

kjljayalakshmi@gmail.com

muddangula.shirisha179@gmail.com

vaishuvaishnavi728@gmail.com

bhadraatike@gmail.com

akshithamudhiraj3@gamil.com

gorantlahema2426@gmail.com