A REPORT

ON

<u>PREDICTING CUSTOMER CHURN USING ML</u>

BY

Name(s) of Students	ID No.(s)
1. Sudhanshu Singh	2019A3PS0391G
2. Devesh Praveenkumar Gupta	2019B5A70641G

AT

Celebal Technologies Pvt Ltd (Jaipur)

A Practice School-I Station of

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

(May-July, 2021)

A REPORT

ON

PREDICTING CUSTOMER CHURN USING ML

BY

Name(s) of Students	ID No.(s)
1. Sudhanshu Singh	2019A3PS0391G
2. Devesh Praveenkumar Gupta	2019B5A70641G

Prepared in partial fulfillment of the

Practice School-I Course Nos.
BITS C221/BITS C231/BITS C241

AT

Celebal Technologies Pvt Ltd (Jaipur)

A Practice School-I Station of

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

(May-July, 2021)

ACKNOWLEDGMENTS

We would like to express our gratitude to our esteemed institute, BITS-Pilani, for giving us this unique chance to collaborate with the industry, which has provided us with invaluable experience for our future professions. Prof. Chittaranjan Hota, our BITS-Pilani faculty mentor, has always been by our side and has given his expertise on the subject matter whenever it was required.

We'd also like to acknowledge Mr. Akash Verma and Mr. Sarthak Acharjee, our Celebal Technologies Pvt Ltd (Jaipur) mentors, who have been proactive in assisting us and solving all our project related difficulties by providing relevant study material as well as offering extremely valuable information about the company's operations.

ABSTRACT

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI (RAJASTHAN)

Practice School Division

Station: Celebal Technologies Pvt Ltd

Centre: Jaipur

Duration: 8 weeks

Date of Start: 31st May, 2021

Date of Submission: 23rd July, 2021

Title of the Project: PREDICTING CUSTOMER CHURN USING ML

ID No./Name(s)/Discipline(s)/of the student(s):

1. 2019A3PS0391G/Sudhanshu Singh/B.E. Electrical and Electronics Engineering
2. 2019B5A70641G/Devesh Praveenkumar Gupta/M.Sc. Physics+B.E. Computer Science

Name(s) and designation(s) of the expert(s):

1. Mr. Sarthak Acharjee, Company coordinator.
2. Mr. Akash Verma, ML developer.

Name(s) of the PS Faculty: Prof. Chittaranjan Hota, Department of Computer Science Engineering at BITS-Pilani, Hyderabad.

Key Words: Customer, Data, Python, Matplotlib, Scikit-Learn, preprocessing, Regression, Random Forests, Gradient Boost, Decision Tree, Exploratory Data Analysis(EDA), modelling.

Project Areas: Machine Learning(ML) and Artificial Intelligence(AI), Data Visualization using EDA, Customer behavior analysis using relevant data.

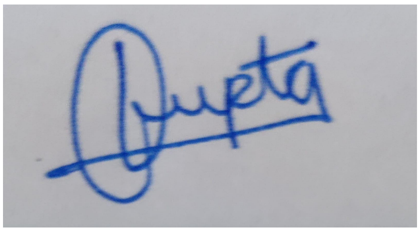
Abstract:

Customer churn (also known as customer attrition) is the phenomenon of customers abandoning a brand and ceasing to be a paying client of a company. Customer churn (attrition) rate refers to the percentage of customers that stop utilising a company's products or services within a given time period.

We need data to work with, much like in machine learning jobs. We determine what data they need to collect based on their objectives. The data is then prepped, preprocessed, analyzed graphically, and converted into a format that can be used to create machine learning models and then their accuracies are tested and compared. Finally we choose the model having the highest accuracy on the test data.

Signature(s) of Student(s):

1. Sudhanshu Singh



2. Devesh Praveenkumar Gupta

Date: 22/07/2021

Signature of PS Faculty:

Date:

INTRODUCTION

When consumers or subscribers stop doing business with a company or service, this is known as customer churn.

Because most businesses have a huge number of customers and can't afford to devote much time to each of them, personalised customer retention is difficult. The higher revenue would be outweighed by the increased costs. If a company could predict which customers are likely to depart ahead of time, it could target its customer retention efforts solely on these "high risk" consumers.

The ultimate goal is to increase client loyalty and grow the company's coverage area. The client is at the heart of this market's success. Customer turnover is a crucial indicator since retaining existing customers is far less expensive than acquiring new ones.

To cut down on customer churn, telecom companies must be able to predict which customers are most likely to leave. To detect early symptoms of impending churn, create a holistic view of consumers and their interactions across several channels, such as store/branch visits, product purchase histories, customer care calls, Web-based transactions, and social media interactions, to name a few. As a result, tackling churn can help these companies not only maintain their market position, but also grow and thrive. The lower the cost of initiation and the higher the profit, the more consumers they have in their network. As a result, the company's primary focus for success is on lowering client attrition and creating a successful retention plan.

Turnover prediction models are used in predictive analytics to estimate customer attrition by measuring their risk of churn. These models are useful at focusing customer retention campaigns on the customer base that is most prone to churn because they generate a short prioritised list of possible defectors.

METHODOLOGY

The overall scope of our ML based project work, capable to forecast customer attrition may look like the following:

1. Understanding the Data

2. Data Manipulation

3. Exploratory Data Analysis(EDA)

4. Data preparation and preprocessing

5. Modeling and testing

1. Understanding the Data:

We have downloaded the dataset from the Kaggle website. The dataset is of Telco company, a telecommunications company. The quality of the dataset is rich, having very few null values and various features of the customers providing better understanding of the dataset. It is available for downloading at: https://www.kaggle.com/blastchar/telco-customer-churn.

The data set includes information about:

Target Column:

Customers who left within the last month – the column is called Churn

Input Columns:

a. **Services that each customer has signed up for** – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
b. **Customer account information** - how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
c. **Demographic info about customers** – gender, age range, and if they have partners and dependents

Having a look at the glimpse at the dataset using *df.head()* command:

In [4]:
```
df.head()
```

Out[4]:

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	O
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	N
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Y
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Y
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Y
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	N

5 rows × 21 columns

2. Data Manipulation

Since model performance and therefore the quality of received insights depend on the quality of data, the primary aim is to make sure all data points are presented using the same logic, and the overall dataset is free of inconsistencies.

- ***Feature extraction*** aims at reducing the number of variables (attributes) by leaving the ones that represent the most discriminative information. Feature extraction helps to reduce the data dimensionality (dimensions are columns with attributes in a dataset) and exclude irrelevant information.Like in our case customerID was an irrelevant information to predict churn so we drop it from our attrition set.

[10]:
```
df = df.drop(['customerID'], axis = 1)
df.head()
```

- A one-hot **Encoding** allows the representation of categorical data to be more expressive. Many machine learning algorithms cannot work with categorical data directly. The categories must be converted into numbers, like in one of our cases where we have attrition like SeniorCitizen, we map 0 for “NO” and 1 for “Yes”.

[17]:
```
df["SeniorCitizen"]= df["SeniorCitizen"].map({0: "No", 1: "Yes"})
df.head()
```

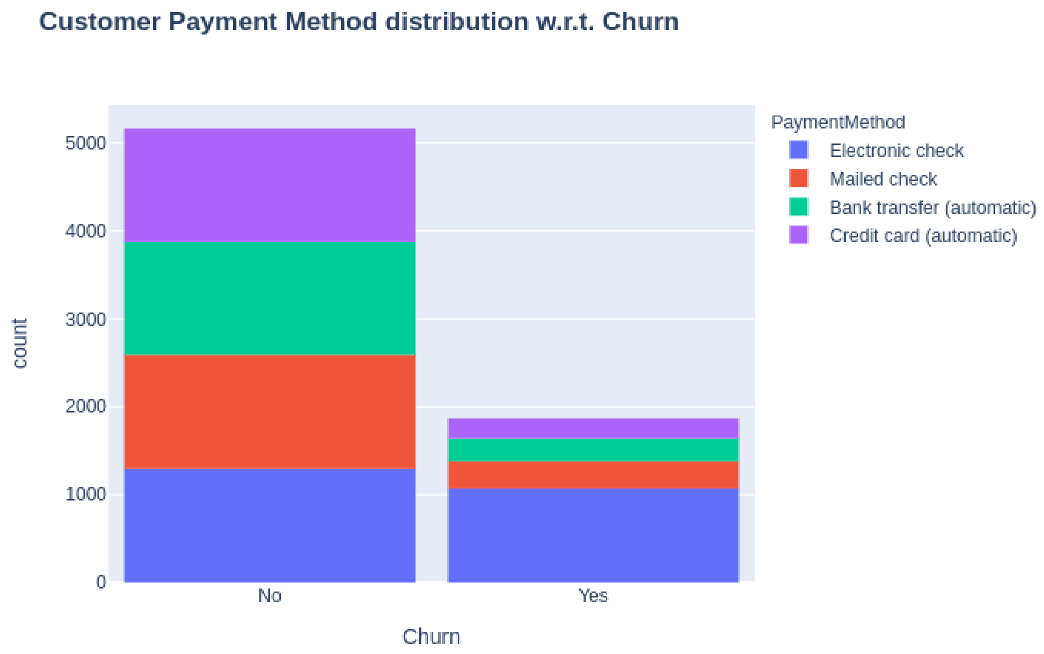
- **Imputing**, The missing values can be imputed with the mean of that particular feature/data variable. That is, the null or missing values can be replaced by the mean of the data values of that particular data column or dataset.We have applied Imputing to solve the problem of missing values in TotalCharges column, we decided to fill it with the mean of TotalCharges values.

```
df.fillna(df["TotalCharges"].mean())
```

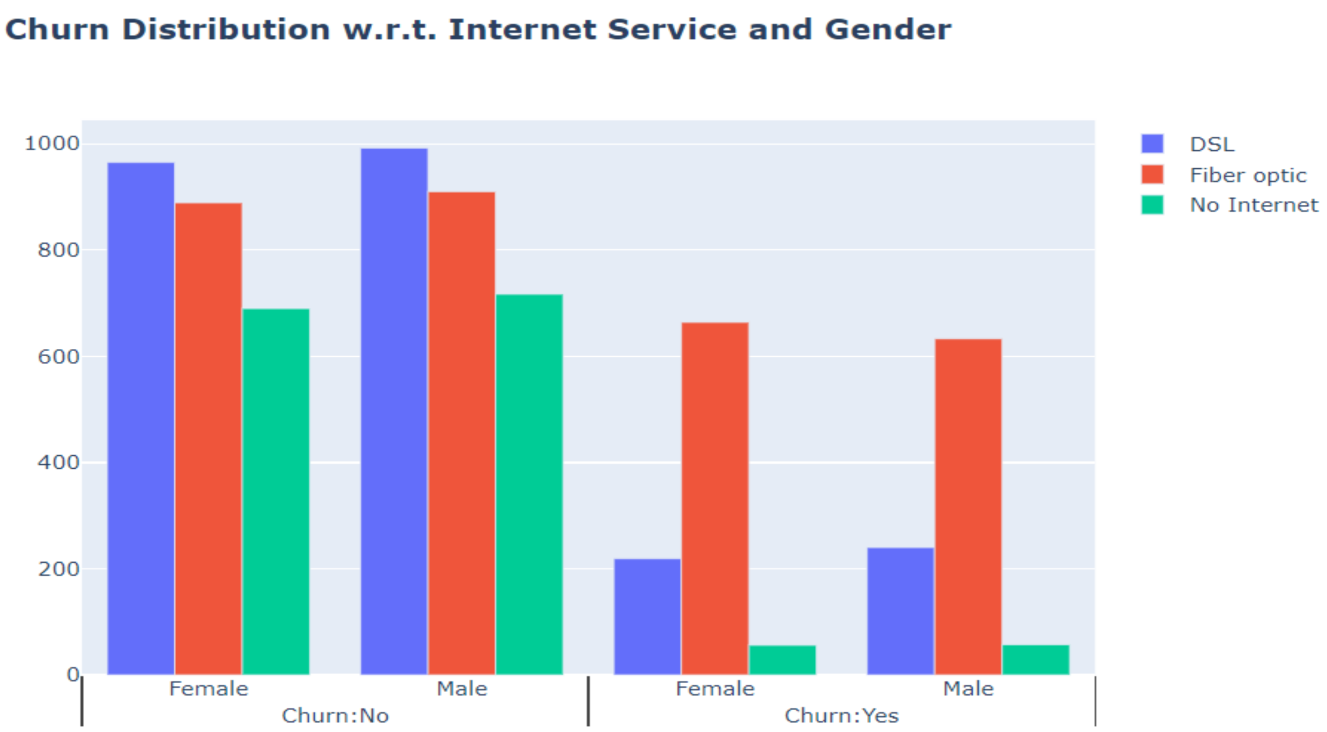
3. Exploratory data analysis

EDA in Python uses data visualization to draw meaningful patterns and insights. It also involves the preparation of data sets for analysis by removing irregularities in the data. Based on the results of EDA, companies also make business decisions, which can have repercussions later. If EDA is not done properly then it can hamper the further steps in the machine learning model building process. If done well, it may improve the efficacy of everything we do next.

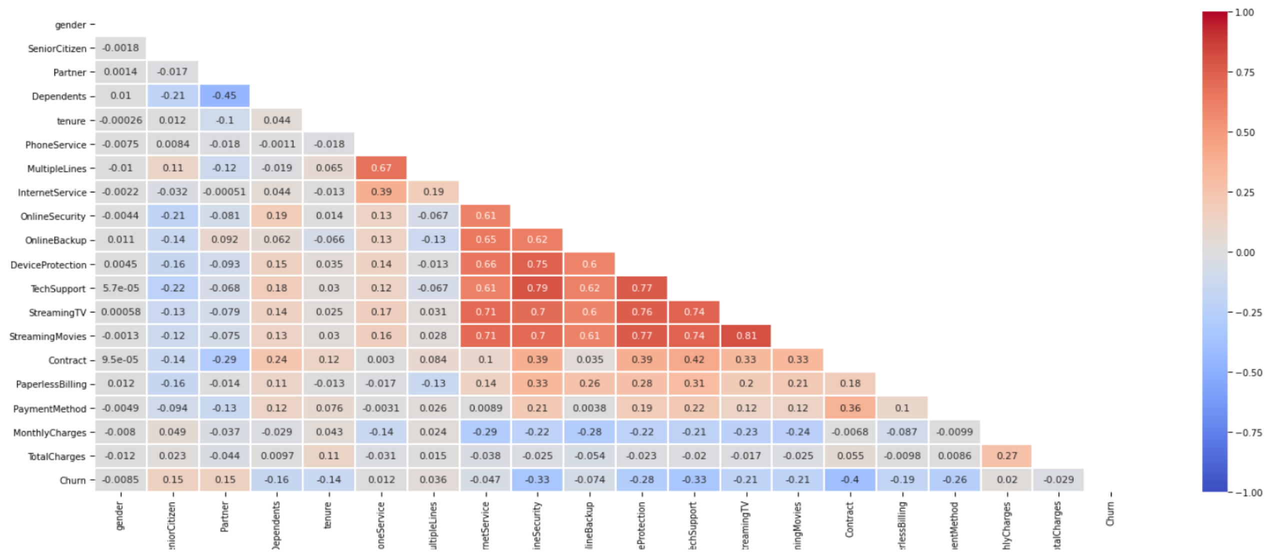
Few of the many EDA we had provided are : -



- If we analyze data over a single variable/column from a dataset, it is known as Univariate Analysis. So in this univariate analysis we can see that the majority of the customers who moved out were having Electronic Check as Payment Method, around 45% of the total Electronic check have churned.Also, Customers who opted for Credit-Card automatic transfer or Bank Automatic Transfer and Mailed Check as Payment Method were less likely to move out.



- If we analyze data by taking two variables/columns into consideration from a dataset, it is known as Bivariate Analysis. From this bivariate analysis we can imply that a lot of customers choose the Fiber optic service and it's also evident that the customers who use Fiber optic have high churn rate, this might suggest a dissatisfaction with this type of internet service.Also customers having DSL service are majority in number and have less churn rate compared to Fibre optic service



- Since we cannot use more than two variables as x-axis and y-axis in

Scatter and Pair Plots, it is difficult to see the relation between three numerical variables in a single graph. In those cases, we'll use the correlation matrix. Based on the Heatmap we can infer that attributes like MonthlyCharges, MultipleLines, PhoneService and Partner are more likely to respond positively to the churn whereas OnlineSecurity, TechSupport, StreamingTV, StreamingMovies contribute negatively to churn.

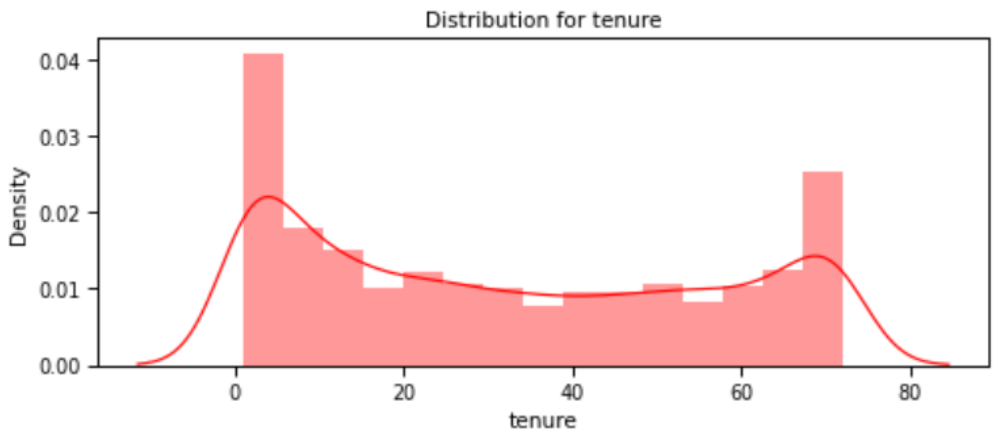
4. Data preparation and preprocessing

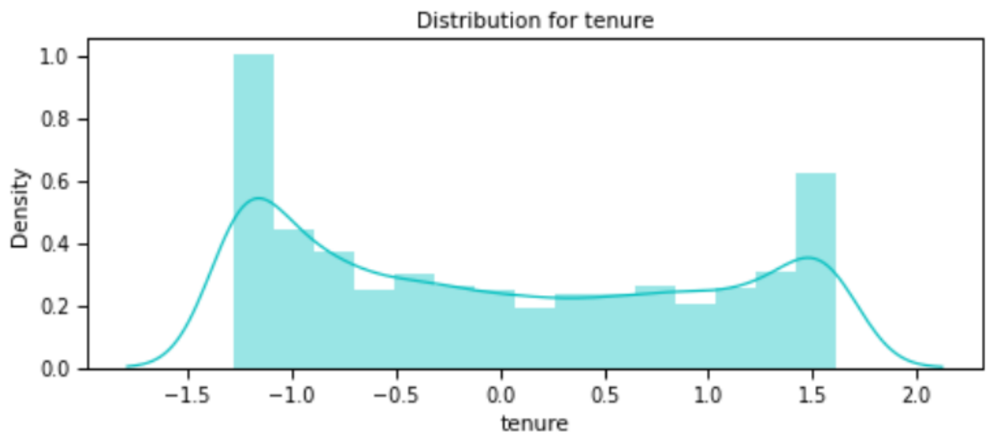
Also known as Feature engineering, is a very important part of dataset preparation. Features are measurable characteristics of observations that an ML model takes into account to predict outcomes (in our case the decision relates to churn probability.) The data(collection of features) which is fed to make the model must be preprocessed such that the overall execution of the model is efficient in RAM usage and is less time consuming.

Some of the steps involved in data preparation and preprocessing are as follows:

a. **Standardizing numeric attributes(Scaling):**

   Here, we scale down the entire numerical data range to a specific small data range. This is because, computer takes less time and storage for processing small data. This therefore, increases the efficiency of the model.





b. Splitting the data into train and test sets:

This is an important step in preprocessing where we divide the entire dataset into two parts: training dataset and testing dataset. Training dataset is used to build the ML model and is therefore usually 75% of the entire data. Testing dataset is used to make evaluations on the model and is usually 25% of the entire dataset.

```
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size = 0.25, random_state = 40, stratify=y)
```

5. **Modeling and testing**:

We have prepared 8 different ML models. They are as follows:

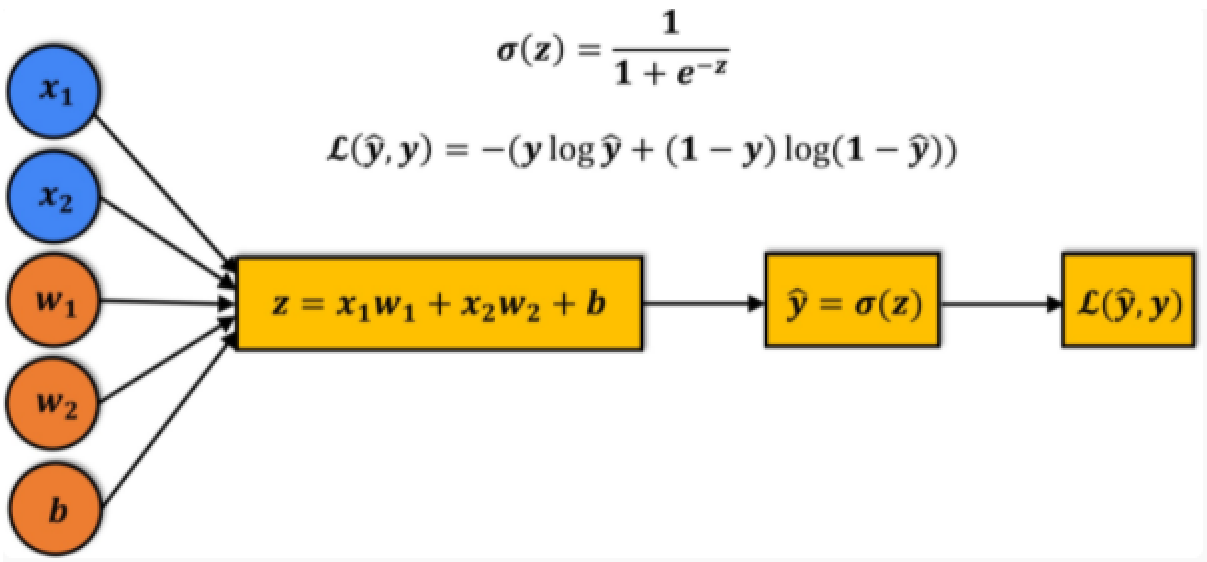
a. **Logistic Regression**:

Logistic regression is a commonly used technique for solving binary classification problems. In a logistic regression model:

a. We take linear combination (or weighted sum of the input features)
b. We apply the sigmoid function to the result to obtain a number between 0 and 1
c. This number represents the probability of the input being classified as "Yes".

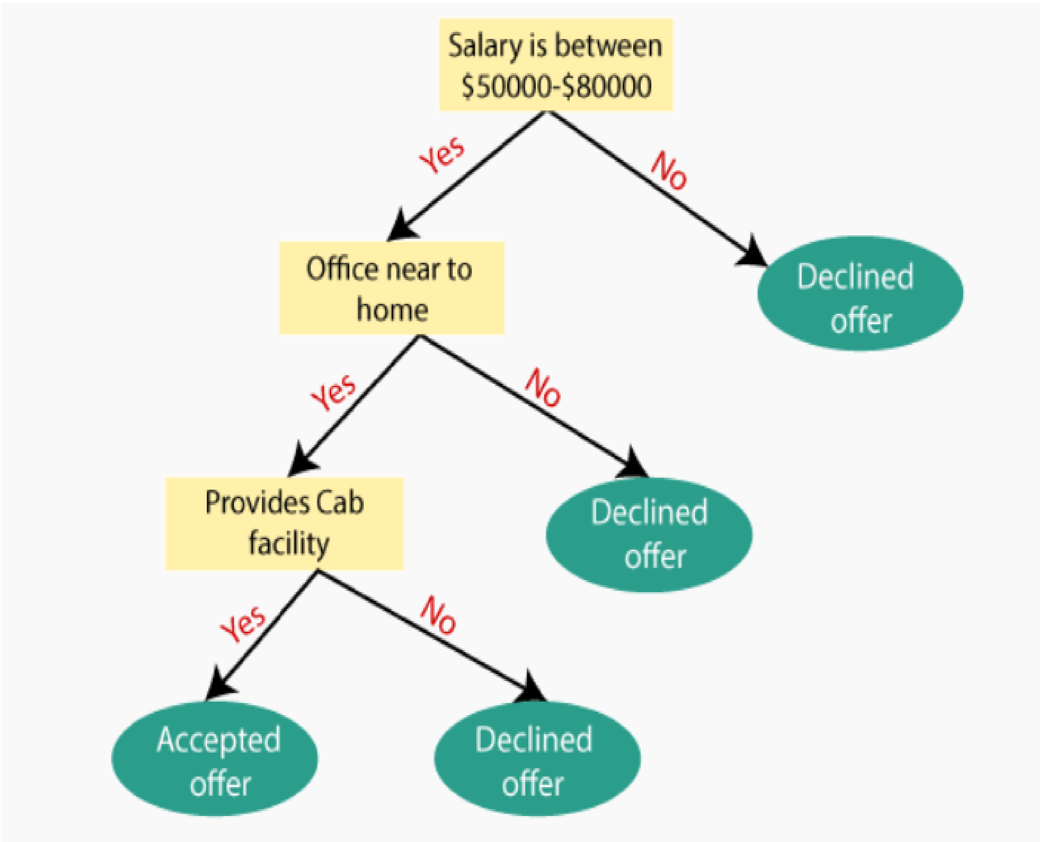
The output of the sigmoid function is called a *logistic*, hence the name logistic regression.

Here's a visual representation of Logistic regression:

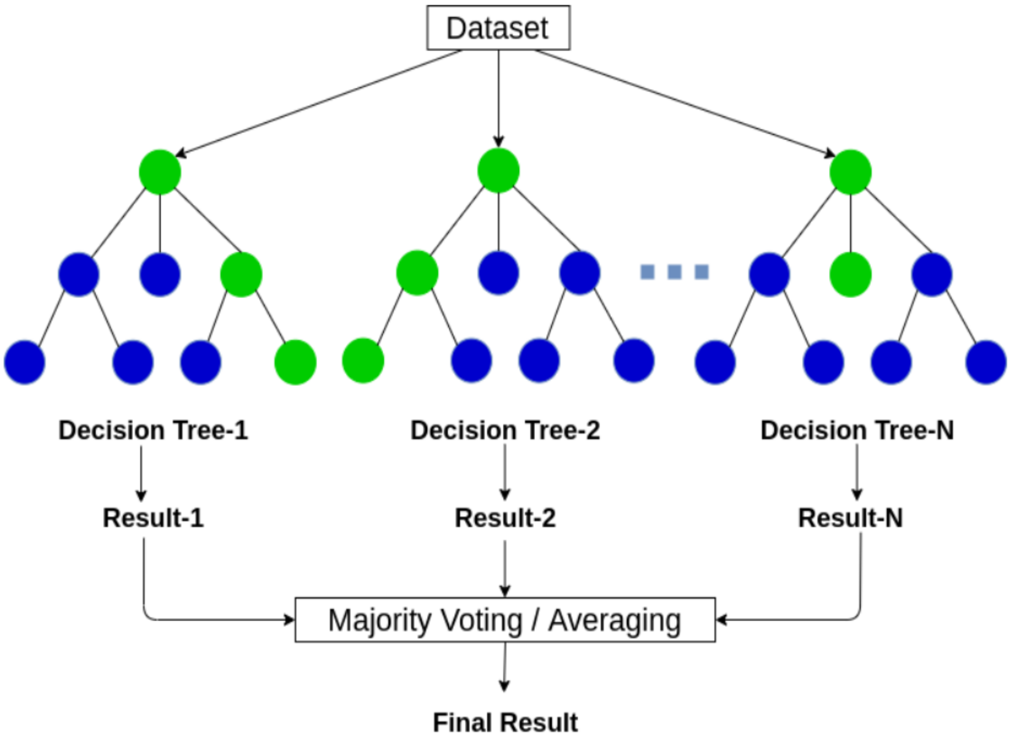


b. **Decision Tree**: A supervised learning algorithm is a decision tree (with a predefined target variable.) A decision tree in general parlance represents a hierarchical series of binary decisions. While it is most commonly used for classification jobs, it can also handle numeric data. To create a prediction, this method divides a data sample into two or more homogeneous sets based on the most significant differentiator in input variables. A component of a tree is formed with each split. As a result, a tree is formed, including decision nodes and leaf nodes (decisions or classifications). A tree begins with a root node, which is the most accurate predictor. Decision tree prediction findings are simple to interpret and depict.

For example, here is an example of a decision tree:



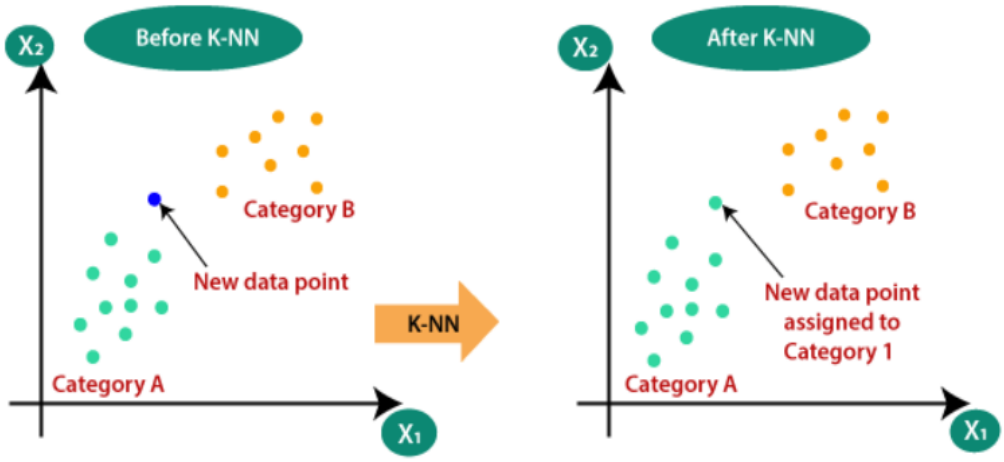
c. **Random Forest**: A Random forest is an ensemble learning method that employs a large number of decision trees to improve prediction accuracy and model stability. While tuning the hyperparameters of a single decision tree may lead to some improvements, a much more effective strategy is to combine the results of several decision trees trained with slightly different parameters. This is called a random forest model.This idea is also commonly known as the "wisdom of the crowd".



d. **KNN(K- Nearest Neighbours)**:

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. KNN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.



e. **SVC**:

Support vector machines (SVMs) are powerful yet flexible supervised machine learning methods used for classification, regression, and, outliers' detection. SVMs are very efficient in high dimensional spaces and generally are used in classification problems. SVMs are popular and memory efficient because they use a subset of training points in the decision function. It is a C-support vector classification whose implementation is based on libsvm. The module used by scikit-learn is sklearn.svm.SVC. This class handles the multiclass support according to a one-vs-one scheme.

f. **Gradient Boosting Classifier**:

Gradient boosting is a machine learning technique for regression, classification and other tasks, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. Gradient boosting involves three elements:

1. A loss function to be optimized.
2. A weak learner to make predictions.
3. An additive model to add weak learners to minimize the loss function.

g. **AdaBoost Classifier:**

AdaBoost algorithm, short for Adaptive Boosting, is a Boosting technique that is used as an Ensemble Method in Machine Learning. It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights to incorrectly classified instances.

Adaptive Boosting is a good ensemble technique and can be used for both Classification and Regression problems. But in most cases, it is used for classification problems. It is better than any other model as it improves model accuracy, one can check this by going in sequence. First try decision trees and then go for the random forest, next apply to boost and finally go for AdaBoost. We can see that the accuracy keeps increasing as we follow the above sequence. The weight assigning technique after every iteration makes the AdaBoost algorithm different from all other boosting algorithms. And that’s the best thing about the AdaBoost algorithm.

h. **Voting Classifier:**

A Voting Classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of chosen class as the output.

It simply aggregates the findings of each classifier passed into Voting Classifier and predicts the output class based on the highest majority of voting. The idea is instead of creating separate dedicated models and finding the accuracy for each of them, we create a single model which trains by these models and predicts output based on their combined majority of voting for each output class.

CONCLUSION

In our project, the accuracies obtained by all the 8 models are as follows:

- KNN - 77.54%
- SVC - 80.76%
- Random forest - 81.37%
- Logistic Regression - 80.90%
- Decision tree - 72.51%
- AdaBoost Classifier - 80.76%
- Gradient Boosting Classifier - 80.80%
- **Voting Classifier - 81.71%**

Here, it can be easily observed that the Voting Classifier is the most accurate model among them.
But it's not always the case. Depending on the dataset, the most accurate model may vary.

For subscription-based businesses, the churn rate is a key metric. Customers who are dissatisfied with offered solutions can be identified, allowing organisations to learn about product or price plan flaws, operational challenges, as well as consumer preferences and expectations, allowing them to proactively reduce reasons for churn.

To have a complete picture of the history of consumer interactions, it's crucial to specify data sources and observation periods. The most important features in a model's selection would have an impact on its prediction performance: Forecasts are more precise when the dataset is more qualitative.

Companies with a large customer base and numerous offerings would benefit from customer segmentation. The number and choice of ML models may also depend on segmentation results.

REFERENCES

- Curriculum provided by the Celebal Technologies Private limited
- https://www.kdnuggets.com/2019/05/churn-prediction-machine-learning.html
- https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0191-6
- https://towardsdatascience.com/churn-prediction-with-machine-learning-ca955d52bd8c
- https://jovian.ai/learn/machine-learning-with-python-zero-to-gbms
- https://scikit-learn.org/stable/

GLOSSARY

Modeling : Use (a system, procedure, etc.) as an example to follow or imitate(here, a dataset).

Regression : In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables.

Preprocessing : Subject (data) to preliminary processing.

Customer Churn : When customers or subscribers stop doing business with a company or service

Segmentation : Division into separate parts or sections.

Kaggle : Allows users to find and publish data sets, explore and build models in a web-based data-science environment.

Extraction : Taking out relevant information from data.

Matplotlib : Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy.

Scikit-Learn: Free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, *k*-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

Labels: attributes / features present in the column of dataset.

Demographic: relating to the structure of populations.