Customer Churn prediction

---Model building

Presented by :- M.J.K.Manideep,

pragati engineering college,

Table of contents:-

* Introduction
* **Milestone1**:- data gathering and software installation.

Week 1:- Data gathering

Week2:- Anaconda software installation

* **Milestone 2**:- Data preprocessing with all 8 steps.

Week 3:- Data preprocessing for the dataset.

Week 4:- improving dataset by further preprocessing.

* **Milestone3:**- Model building.

Week5:- Model building using Decision tree.

Week6:- checking accuracy of decision tree model.

* **Milestone4**:- Model building and documentation.

Week7:- Model building using Random

forest and Logistic regression.

Week8:- Documentation and final report.

* **Conclusion**

Introduction:-

**Churn prediction** is a method through which we find whether a person is churned or not i.e; whether a person changed or not.

Building a churn prediction model can help companies forecast revenue for the year and develop strategies for retaining high-risk customers. While it’s impossible to build a model that’s 100 percent accurate, churn prediction is something every company should tackle if it wants to grow.



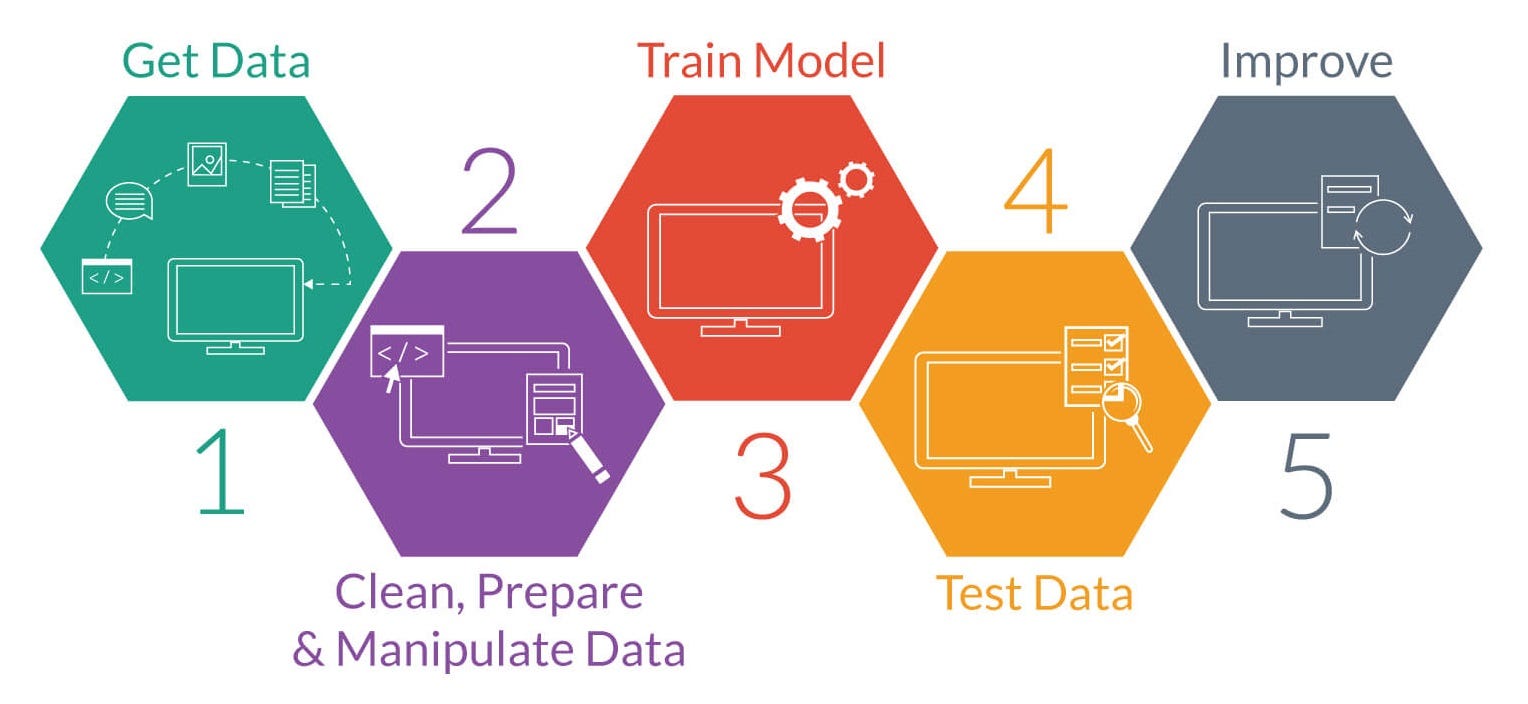
Predicting it helps you get ahead of churn and ideally prevent it, that is what everyone is trying to do .You have to think about it outside of the box , ‘How satisfied are our customers with us?’ and get to that point where we know what we need to have to change inorder to satisfy customers. this can be done by processing data and building models basing on how accurate testing and training datasets are.

Data contains all types of values that contains null values,missing values ,unique values and many other factors that impact on predicting the churning. so, to encounter these issues we need to do following steps:-

1. Data gathering .
2. Software installation.
3. Data preprocessing.
4. Model building.
5. Accuracy Checking.
6. Choosing best model.
7. Checking whether customer churned or not

We have total 4 milestones, each milestone consisting of two weeks ,in total we have 8 weeks for this project.so, we have split it into 4 milestones and made our work simple .

**Process for Churn prediction :-**



* Milestone1 :- Data gathering and

Software installation

Milestone1 consists of first two weeks where we gathered data related to churn prediction in week 1 and downloaded softwares required for churn prediction in week2.

Week1:-

During week 1 ,we gathered information related to churn prediction dataset from reliable sources and prepared a dataset based on the gathered information.

Firstly, we have an introductory session from where we officially started project. From then, we initiated project by gathering data and arranging data . The data we gathered is raw data that consists of unnecessary data with null values,unique values,missing values

While gathering data we have navigated to various sites and finalised our dataset by the end of the week1.



Week2:-

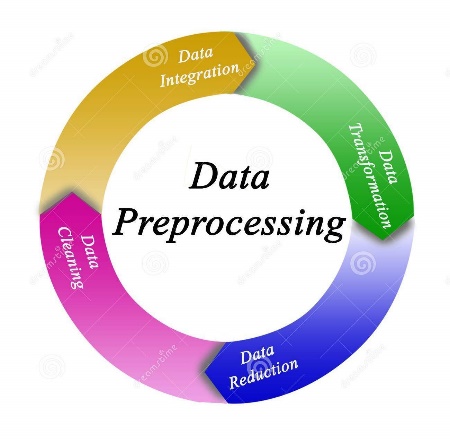
During week2, we have installed the software required for our project and accessed them using steps mentioned below:-

1. Download & Install anaconda,first we have to download & install anaconda software in our local device . 

1. We have to launch anaconda navigator,from there we have to navigate to jupyter notebook and we have to launch it.
2. After launching, we have to select python-3 as an option for kernal and create a new notebook.

* Milestone2 :- Data Preprocessing

Milestone 2 consists of week 3 and week 4 , In which we have done data preprocessing. when it comes to data preprocessing, the data we take is an raw data which needs to be processed inorder to get accurate and desired result. It is similar to the purification of water by removing impurities before drinking.



While doing data preprocessing we have split it into 8 steps:-

1.**converting misclassified datatypes**:-

Datatypes which are misclassified are to be converted into suitable datatypes inorder to make the fit for designing model.

Datatypes that include strings, date & time, Boolean values for some models will come under misclassified data ypes which we classify them into numerical data types, inorder to make the model.

2.**Removing duplicate records**:-

* df.drop\_duplicates() method identifies and removes rows that are considered duplicates based on the specified criteria.
* By default, it considers all columns and removes rows where all column values are identical.
* By using the formula :- df.drop\_duplicates(inplace=True)   
  we can remove duplicate values present in the dataset.

3.**Removing Unique Value Variables**:- unique value variables are variables in dataset which are unique and doesnot occur more than once which doesnot predict any information hence, need to be removed.so, we remove the m with the formula :- df.nunique().

4.**Removing Zero variance variables:-** zero variance variables are variables which doesnot have variance in any parameter ,which inturn doesnot give any information hence needed to be removed . these can be removed using formula:- df.columns[df.var() == 0]

5.**Outlier Treatment**:-

a.**Using Boxplot**:-

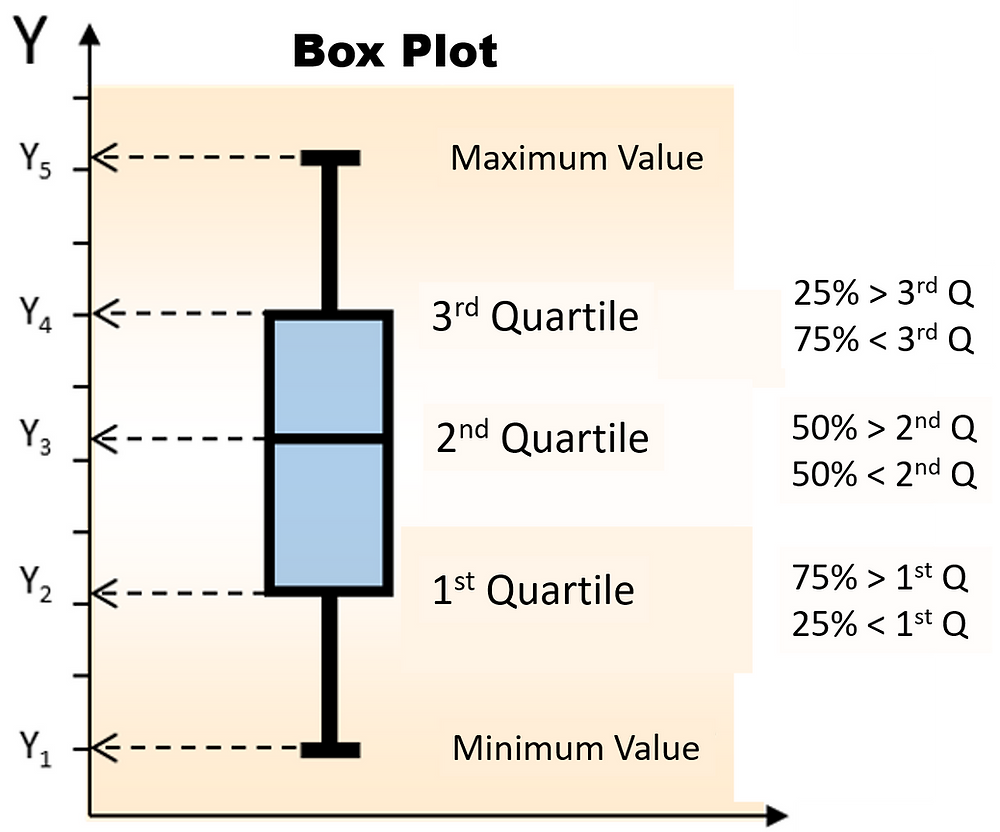
First we Calculate the 1st quartile (q1) and 3rd quartile (q3) using np.percentile. and then,

We Computes the Interquartile Range (IQR) as q3 - q1.

Then we define lower and upper bounds for outliers:

* Lower bound: q1 - 1.5 \* IQR
* Upper bound: q3 + 1.5 \* IQR

Then, we remove columns greater than upper bound and lesser than lower bound . so ,that our dataset removes values

which are not used for model building. 

b. **Standardisation**:-

To use standardisation, first we find mean and standard deviation and then, we calculate  
 (df-mean)/standard deviation.

With this formula ,we can standardise the dataset so that we can treat outliers by standardisation.

c. **Capping & flooring**:-

The cap\_floor\_data function is designed to cap and floor the values in a DataFrame df based on specified upper (cap) and lower (floor) limits. Here's how the function works:

1. Cap the Data: Using the clip() function with the upper parameter set to the cap value, the function ensures that any values in the DataFrame exceeding the specified upper limit (cap) are replaced with the upper limit.

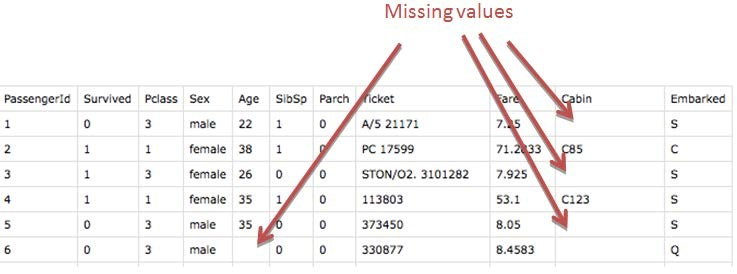
2. Floor the Data: After capping the data, the function applies another clip() operation, this time setting the lower parameter to the floor value. This ensures that any values in the DataFrame below the specified lower limit (floor) are replaced with the lower limit.

3. Return Capped and Floored Data: The function returns the DataFrame with both capping and flooring applied.

4. Cap and Floor the Data: Finally, the cap\_floor\_data function is called with the DataFrame df, and the resulting DataFrame with capped and floored data is assigned to capped\_floored\_df.

Hence,this function is useful for dealing with outliers or extreme values in your dataset by setting upper and lower bounds for the data. For example, in the provided code snippet, any values greater than 4 are capped at 4, and any values less than -4 are floored at -4. Adjusting the cap and floor parameters allows for flexible handling of outliers based on the specific characteristics of the data.

6. **Missing value treatment**:-



a. **Remove records if NA’s are less than 5%:-**

if percentage of NA’s are lessthan 5% .those records can be removed safely.  
these can be removed safely with the help of python code:-

df[na\_counts / len(df.columns) < threshold]

b. **Remove records if NA’s are 50% in any variable.**

Missing value records where NA’S are 50% can be removed using the formula:-  
 df.dropna(thresh=0.5\*len(df.columns), axis=1)

c**. Imputing missing values with mean/median,if variable is numeric and with mode if variable is categorical**.:-

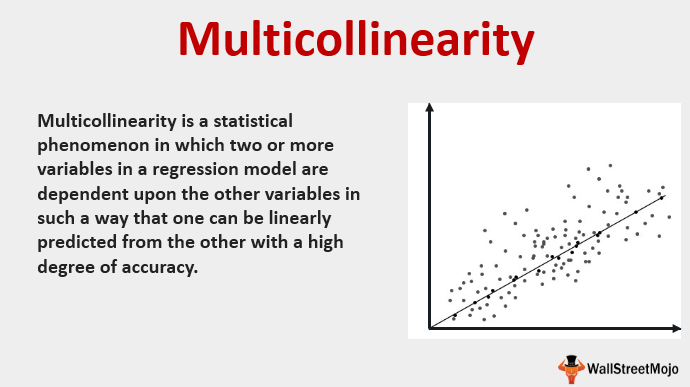
we have imputed missing values with mean or median if the variable is numeric and if variable is categorical we replace it with mode

7.**Removing highly correlated variables:-**

We have removed highly correlated variables by calculating correlation matrix and created s mask for identifying highly correlated numeric colums .so,that we drop those columns with high correlated variables with threshold 0.8/0.9.

8.**Multicollinearity(VIF>5):-**

For this we have calculated vif for all the numeric columns. And identified all the columns with VIF>5 and dropped variables with high VIF.



Data preprocessing is a peculiar step needed to be done as data that we get to build model is some raw data taken from public pulse or the other scenario. So, what must be done is to clean the raw data and make it ready for model building .

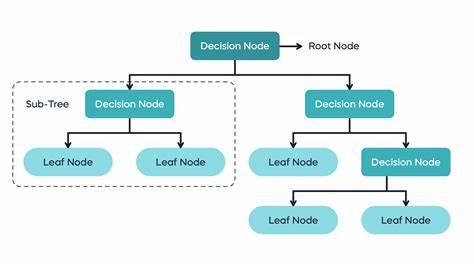
In **week3**, I have completed all the 8 steps of data preprocessing .so , that I have cleaned the dataset inorder to make it ready for model building .

In **week 4** , I have done data preprocessing again and improved dataset by removing unnecessary data from dataset. With these steps that we have done in week 3 and week 4 we have cleaned the dataset in fullfledge so,that it is useful for bulding different kinds of models.

* Milestone3:- Decision tree Model building

In milestone 3,

I have designed a model so that we can predict whether a customer is churned or not churned. I have three mostly used models Decision tree,random forest and logistic regression. Among them I have choosen decision tree model first inorder to build the model.



Inorder to build the model,

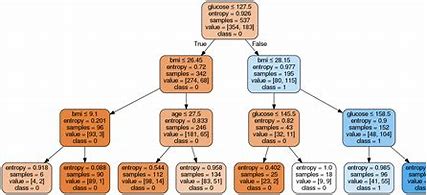
Firstly, we need to split the dataset into train data and test data . then, we need to assign a variable for both of them.while splitting data we can split data in 80:20 ratio or 70:30 ratio.inorder to build decision tree model we need to import certain libraries like pandas,scikit learn, decision tree classifier and others to build the model and visualise the model.

Then, we will train an decision tree model and do predictions on the model .so,that we can calculate accuracy of testing and training model separately and we usally compare both and the model that has training data and testing data with same accuracy is used.For this model I got 71% training accuracy and 72% testing accuracy.

**Training Accuracy:** 0.71

**Testing Accuracy:** 0.72

Finally, inorder to increase accuracy I have done hyperparameter testing by considering entropy or ginni index .with this accuracy increased by 0.1 – 0.2 %..so, that our model looks more accurate. In week5, I have created my model and I have checked accuracy . in week6, I have modified my code and I have again done data preprocessing steps and I have done hyper parameter testing .so that accuracy increases.



**Visualisation of Decision tree model:-**

* Milestone4:- Model building &

Documentation.

Week7:-

In Milestone4, As I have already done model building using Decision tree .I aimed for a high accuracy so I have tried other models that include **Random forest** and **Logistic regression .**

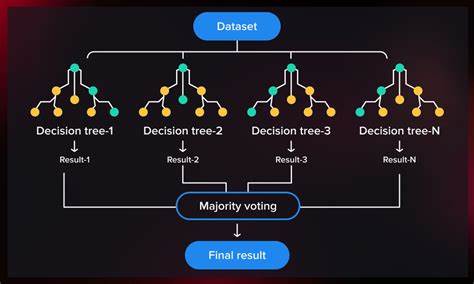
The reason why we need to check for other models is we need to choose a model that gives more accuracy i.e, the model that has nearly same accuracy of training data and testing data could be used for model building as that particular model can predict the particular dataset that you have better than other models.As data is dynamic , trail and error is the only option that we have .so, we need to do all the models and choose the better one.

With the building of all three models, among them I found decision tree model has a balance in such a way that our data can predict whether a customer has churned or not perfectly with the help of decision tree rather than other two models.

**Outcomes of Random forest model I have seen** :-

While working with random forest the main advantage is it is the simplified version of decision tree as it contains a bunch of decision trees inorder to build a randomforest ,so in this randomforest the fundamental idea behind random forests is to introduce randomness into the tree-building process.  they are more complex and less interpretable than decision trees..so, for this dataset decision tree is more likely to be accurate .

**This is how our Random forest looks like:-**

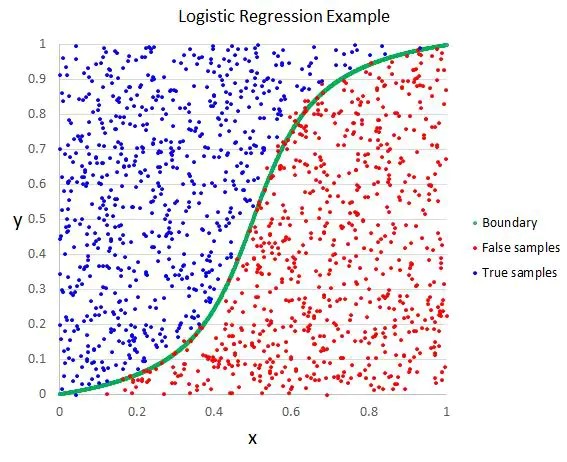


Week8:-

**Outcomes of Logistic regression I have observed:-**

While working with Logistic regression model , the main difference I have seein is it is less interpretable as it used a thin line that separates the categorical values whether yes or no but there are several parameters need to be considered apart from a coustomer is churned or not .it will be like whether a customer is churned because of changing location or network down or the other reason.so, for this reason I choose decision tree over other model inorder to choose for better model building .

**This is how our Logistic regression looks like:-**



**Outcomes of Logistic regression I have observed:-**

Apart from these models I have also tried **k-fold method** but the average accuracy in it also not matching the balaned accuracy. So,I finally been with decision tree model for predicting results.

Conclusion:-

Basing on our observations, A proper conclusion would be building a model and predicting a model nothing other than this is accurate as data is dynamic, predicting is the only way of achieving the desired output.so, in building models the first step could be understanding data ,operations on data for data cleaning , model building and choosing a better model which gives a balanced answers basing on the need of the client. For this dataset, the desired model that we can prefer is decision tree because of its predictive nature and the main reason of selecting decision reason model is having nearly same accuracy of training and testing dataset which gives a proper balance along with predictability factor and better interpretability factor compared to logistic regression and random forest.