**Customer Churn Prediction with Python**

**“There’s only one boss.The customer. And he can fire everybody in the company from the chairman on down,simply by spending his money somewhere else.”**

**-Sam Walton**

**A picture containing sign, computer, table, shirt

Description automatically generated**

Customer churn is when a company’s customers stop doing business with that company.Industries that use a subscription-based business model have traditionally focused more on churn than others.Telecom Companies,insurance firms,Banks are among the many types that often use customer churn rates as one of their key business metrics.

**Few reasons why you must care about churn:**

* Churn results is helpful in knowing the company’s health and calculating the lifetime value.
* Churn results in greater marketing also Existing customers will often have a higher volume of service consumption and can generate additional customer referrals.
* The profit rates also depends on the churn.
* Customer retention can be achieved with a good customer service and products.
* Knowing who is most likely to defect means that a company can prioritise focused marketing efforts on that subset of their customer base.

**Overview: Customer Churn Prediction using Python**

Python can be used to make predictions with the help of various libraries for data science and machine learning which are inbuilt in the python.Based on different features of the dataset,it is possible to predict the upcoming events so that we would be aware of the situation.

**The Dataset: Telecom Customer Churn Modelling.**

The dataset you’ll be using to develop the prediction model -customer churn from IBM Sample Data Sets with the aim of building and comparing several customer churn prediction models,You can download the dataset from here.There are 20 independent columns(also known as features set),while the last column is dependent variable (or label/target) that contains binary value(1 or 0).In this case whether the customer churns from the current telecom service (1) or not(0).This is also known as the binary classification problem.

**NOTE:** All the code is executed using jupyter notebook for Python.

The overview of the steps we are going to take:

1. Problem Definition
2. Data Analysis
3. EDA Concluding Remarks
4. Pre-processing Pipeline
5. Building Machine Learning Models
6. Concluding Remarks

Lets’ begin!

**Step 1: Problem Definition:**

Businesses are very keen on measuring churn because keeping an existing customer is far less expensive than acquiring new customer. Here in this problem, you have to predict whether customer unsubscribes from the telecom services based on the features provided.

**Step 2: Data Analysis**

Importing the required libraries and performing the data analysis on the dataset which is loaded from the local CSV file into your python program. Let’s use the read\_csv method of

Pandas library.

A screenshot of a cell phone

Description automatically generated

After reading the dataset, if you open cus dataframe in jupyter notebook, you should see the columns /features as shown below:

A screenshot of a cell phone

Description automatically generated

As you can observe,the dataset consists of 21 columns.

Check for the null and missing values in the dataset and fill them with the proper values so

that ,model building is not effected.

As the picture below shows, there are no null values in this dataset.

A picture containing screenshot

Description automatically generated

The heatmap from seaborn library is used here to visualize the null values.From the above picture ,you can observe that there are no null values present in this dataset.

**Step:3 EDA concluding remarks.**

Check for the correlation and feature dependency on the label so that we can build a strong model with optimum accuracy.For the columns like State,International Plan,Voice mail plan(categorical variables) can be encoded to numeric columns by using **label-encoding.**In this process the categories represent numericals from (0-n-1) columns.

Let’s convert the columns with the help of label encoder.

A picture containing knife

Description automatically generated

Check for the count, categories of the customers who are likely to churn. Here the graph is plotted against the feature to get a better understanding.A screenshot of a cell phone

Description automatically generated

**Step 4: Pre-processing pipelines.**

After dropping several columns which are not helpful for the analysis,the feature set is pre-processed and brought down to the common scale to decrease the complexity in building the machine learning models.

A screenshot of a cell phone

Description automatically generated

The dataset is split into training and testing sets using the inbuilt method from sci-kit learn.

**Step 5: Building Machine Learning models.**

As the problem is binary classification, here in this dataset , performing various machine learning models for classification like Logistic regression model, KNeighbors, DecisionTreeClassifier, RandomForest.

**Step 6: Concluding Remarks.**

Here, in this dataset by considering the classification metrics like *classification\_report,confusion\_matrix and accuracy\_score* ,RandomForestClasssifier has outrun other models and performed best with the accuracy of 94%.

A screenshot of text

Description automatically generated

After fitting the best model, the predictions can be saved into a csv file with the help of (to\_csv) method from pandas and the model can be saved into hard drive with the help of pickle library of python.