**A Blog on Machine Learning Model Churn Analytics Project**

Author: Amey Prabhune

**INTRODUCTION:**

Machine Learning is method of data analysis which automates analytical model building. It is branch of Artificial Intelligence. Machine Learning algorithms automatically build a mathematical model using sample data (training data) to make decisions without being specifically programmed to make those decisions. Most industries working with large amounts of data have recognized value of machine learning technologies.

Industries use machine learning are:

* Health care
* Government
* Retail
* Financial services
* Transportations

**HISTORY:**

Machine Learning is based on model of brain of cell interaction. The model is created in 1949 by Donald Hebb in the book titled The Organization of Behavior. In 1967, the nearest neighbor algorithm was conceived, which was the beginning of basic pattern recognition. The algorithm was used for mapping routes and used in finding a solution to the travelling salesperson’s problem of finding most efficient route. Currently, much of speech recognition training is being done by a deep learning technique called Long Short-Term Memory (LSTM). A neural network model describe by Jurgen Schmidhuber and Sepp Hochreiter.

**Problem Definition:**

Customer churn is when a company’s customers stop doing business with that company. Businesses are very keen on measuring churn because keeping an existing customer is far less expensive than acquiring a new customer. New business involves working leads through a sales funnel, using marketing and sales budgets to gain additional customers. Existing customers will often have a higher volume of service consumption and can generate additional customer referrals. Customer retention can be achieved with good customer service and products. But the most effective way for a company to prevent attrition of customers is to truly know them. The vast volumes of data collected about customers can be used to build churn prediction models. Knowing who is most likely to defect means that a company can priorities focused marketing efforts on that subset of their customer base. Preventing customer churn is critically important to the telecommunications sector, as the barriers to entry for switching services are so low.

**Problem Statement**

We will examine customer data from IBM Sample Data Sets with the aim of building and comparing several customer churn prediction models.

**Data Analysis:**

In our dataset we have 21 columns 7043 and rows in total we have record of 7043 client. With the help of this information we will train our model and predict the churn. In our dataset there are 21 columns one of them is our target variable. Target variables result is depend upon other remaining columns. First we need to find out our target variable and in this dataset out target variable is Churn column. In the Data Analysis phase we will try to analyze the data. First we will import our data, then we will check our record and shape and other information like null values, any missing values. Using (df.info) command we will check null values from the dataset. Using (df.describe) command we will check count, min, max values and standard deviation of every column. Also we need to drop columns from dataset which are not helping us to predict values.

Incomplete data can occur due to many reasons. Appropriate data may not be persisted due to a misunderstanding, or because of instrument defects and malfunctions.

Noisy data can occur for a number of reasons (having incorrect feature values). The instruments used for the data collection might be faulty. Data entry may contain human or instrument errors. Data transmission errors might occur as well.

**EDA Concluding Remark:**

In statistics, Exploratory Data Analysis is an approach of analyzing data set to summarize their main characteristics, using statistical graphics and other visualization methods. Primarily EDA is seeing what the data can tell us beyond the modeling or hypothesis testing. EDA focuses on checking assumptions required for model fitting.

In EDA we will visualize our data in the graphical form with the help of graph means Univariate analysis (plotting a box plot for single column) then bivariate analysis (plotting a scatter plot for 2 variables) and third one is multivariate analysis (plotting pair plot, heat map for all dataset) using count plot we will find out count of each categorical value for every column. With help of heat map we can check relation between target variable and other future column. Also we need to convert all non-numerical columns into numerical columns with the help of encoding. We will replace the missing values with mean. Also we will drew a graph for visualization. Like count plot, subplot, pair plot.

**Pre-Processing Pipeline:**

Preprocessing is very important step in Machine Learning to yield highly accurate and insightful result. Greater the quality of data, the greater is the reliability of the produced result. Data preprocessing helps in increasing the quality of data filling in missing data, smoothing noisy data.

In our dataset luckily we don’t have any null values, but we have one column (Total charges) which has numerical values but its data type is object means there are some invalid values, so need to convert this column into numeric, we will use one parameter (error = coerce) this parameter will set NaN to any invalid values. After converting this column into numeric data type again we will check for null values. I got 11 null values in total charges column so we will replace those 11 values with mean. We have some more columns which has object data types so we need to convert them into integer, with help of Encoding. Like Customer ID, Partner, Departments, Phone services, Multiple lines, Internet services, Online Backup, Online security, Device Protection, Tech. support, streaming TV, streaming movies, Gender etc. we will use label encoder only for those column who has only 2 category like male female, yes no etc. and for the others we will use one hot encoder. We will drop Customer id column. After cleaning machine learning model we will check skewness and correlation but form my point of view it better it is a god practice to remove skewness after the splitting our dataset. Because I want my target variable untouched for better predication. We will check our columns are normally distributed or not with the help of histplot. It will give us clear vision of our data is skewed or not. We will remove skewness of those columns which are skewed (right or left).

Note : Categorical column doesn’t have any correlation.

**Building Machine Learning Models:**

Building machine learning model consist of algorithm that can automate analytical model building. Building machine learning models that have the ability to generalize well on future data requires thoughtful consideration of the data at hand and of assumptions about various training algorithm.

After dealing with missing values and cleaning our data set we will split out data set into feature columns and target variable. And then we will remove skewness from out data set. After removing skewness, we will build our model with train and test data. For building model we need to find our highest accuracy score on random state. In this model I got best accuracy score of 0.83 at random state 940. We will use random state 940 and will train our model. Then we will use other algorithms to check highest accuracy like. GaussianNB, Random Forest, Decision Tree, KNN, AD Boost, Support Vector Classifier. After that we will cross check all our models using cross validation. And we will compare our both score and check the difference between score. Any model give us best accuracy, we will save that model. Using Random Forest algorithm I got high accuracy score of 0.81. We will save that model. Before choosing our best algorithm accuracy we need to cross validate our model with help of cross validation function

**Technology used:**

* Python 3
* Jupyter Notebook
* Numpy
* Pandas
* Seaborn
* Sklearn
* Matplotlib
* Different types of algorithm
* Min max scaler
* Zscore

**Concluding Remarks:**

After training and testing our data my models accuracy score is 0.81 and we can predict customer churn (is customer leaving or not) with the help of this model.

