**CUSTOMER CHURNS ANALYSIS**

**Blog by:**

**UJJWAL PRATIK**

## Understanding Customer churn

“For any service company that bills on a recurring basis, a key variable is the rate of churn.” By Harvard Business Review, March 2016

Customer churn is a phenomenon where customers of the business no longer interacting with that business. This simply means that business is losing its customer. It happens when customer of that business is no longer interested in the service or product the business is offering. It may be cause of many reasons like long billing queue, un-courtesy behaviour of the staff, reduced service or product quality etc. There are many other terms which can be used interchangeably with customer churn like customer attrition, customer turnover etc.

A high customer churn indicates that some serious issue in the business process which the business should immediately take care of. This should also true because of acquiring a new customer costs 700% more than retaining an existing customer. Furthermore, if you retain a customer, it will increase their loyalty towards your business and it will help your business grow.

Hence, customer churn can prove to be a roadblock for an exponentially growing organization and a retention strategy should be decided in order to avoid it.

## Problem Definition

A business always wants to keep their customer intact and they always try to take all the measures in order to keep them happy and loyal with the business.

Their life would be much easier if they will know the reason of attrition in advance based on their past data and they would be able to predict who is going to leave their business. This helps them prepare a proper plan to retain their customer like loyalty bonus, proper discount on the services/products etc.

To help them, I have created a machine learning model which helps them predict if a customer will move away from them given a specific set of data. With 80% accuracy, this model takes the load off from being worried of which customer is moving away. They can plan their tempting offers, sometimes even personalized, to keep the customer intact. It will help them prepare a focused marketing strategy.

I have created this model based on telecommunication data. In telecommunication world there are lot of competition among its service provider and a customer usually moving from one provider to another for better network and plan. In this competitive scenario, a vendor will have an edge if they will be able to predict accurately about the customer and they can improve their services to match their competitive.

## DATA ANALYSIS

Before we moved into the details of model I have created, let us understand the data on which the model is created. The data is of customer from telecom sector. The shape of the data are as follows:

|  |  |
| --- | --- |
| Shape:  Columns: 21  Rows: 7043  Nulls: ZERO Exploratory data analysis | Data types:  Object: 17  Integer: 2  Float: 2 |

To get more insight from data let us perform the exploratory analysis. The code snippets are also provided in case if you want to perform some hands on.

The distribution plot of charges (Monthly & Total both) showing skewness towards left. This data requires some correction to take these two charges to normal distribution, only if there are no data losses.

|  |  |
| --- | --- |
|  |  |

The code:



Upon analysing Gender and partner data with count plot, I found that both Gender and partner’s data is evenly distributed. Both Male and Female data is in equal proportion in Gender section. Partner’s data also shows the same pattern here. Unmarried and Married customers are also nearly equally distributed. The number of males is little bit more than female in Gender section and Unmarried customers are a bit more than Married customers.

|  |  |
| --- | --- |
|  |  |

At one hand the data says that it has youngers customers in abundance, on the other hand, it says there are lesser number of Senior Citizen. This simple analysis provides good hints to the telecom company. In the same way dependents and independents analysis also tells a different story to the telecom company. The number of dependents customer is very less than independent customers same as the number of senior citizens is very less than the number of youngsters.

|  |  |
| --- | --- |
|  |  |

EDA for A usages-based says that there are very a smaller number of customers who do not usages phone service. Rest of the customers are using phone services as well. This is also supported by the analysis of multiple lines data. In multiple lines section the number of customers not using multiple lines is more.

|  |  |
| --- | --- |
|  |  |

There are many customers who wants security features form telecom company like backup, antivirus, technical support etc. A deeper EDA of data explains how customer preferences moves around this. A total of 2019 number of customers prefers internet service while 2429 customer wants backup service as their preferences.

|  |  |  |
| --- | --- | --- |
|  |  |  |

Take a deeper look into internet services and we would be able to drill down about the quality of internet services preferred by the customer. As expected, we saw optical fibre connections are in huge demand.

|  |
| --- |
|  |

There are environment savvy customers who is contributing to their bit for environment by making paper less billing as their preferences. Few tech savvy customers also found who take technology very seriously. The choose to electronic payment method or device protection as their choices.

|  |  |  |
| --- | --- | --- |
|  |  |  |

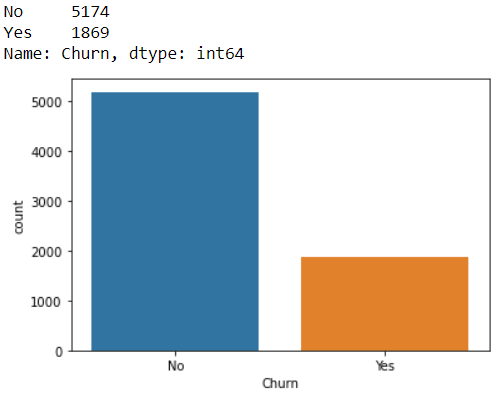
Some prefers entertainment as their choice, and it is well reflected in the EDA where we saw the choices for movie streaming and TV steaming.

|  |  |
| --- | --- |
|  |  |

Major chunk of customers is interested into Monthly contract while there are also few of customers who are interested in yearly or biannual contract.

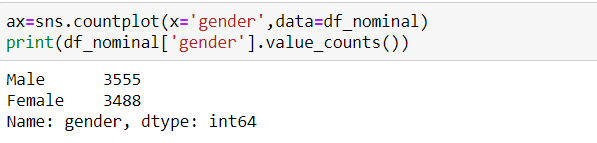
|  |
| --- |
|  |

Preventing customer churn is critically important to the telecommunications sector, as the barriers to entry for switching services are so low. total number of customers have not had chances to leaving is 5174 and the chance of leaving customer is 1869. It means around 265 customers have chances of leaving.



We have done good amount of EDA until now. Let us move to pre-processing pipeline part of the model.

The code here can be used to generate all count plots by just changing the variable name from gender to desired column.



## Pre-processing of the dataset

We will discuss pre processing pipeline of data in a bit detail here. I have done Label encoding, correlation matrix, statistical summary, skewness analysis and removal, and any outlier detections & removal.

Let me explain the label encoding first.

## *Label Encoding:*

As seen in the data analysis section, most of the data is of object type. We have a limitation in Panda library of Python to process object type of the data. Therefore, we have to convert object data into numerical data. This will also help us process data on outlier, skewness, correlations etc using panda package of Python.

Label encoding will help us achieving this.

After label encoding, data for gender which is object type is converted into numerical data type.

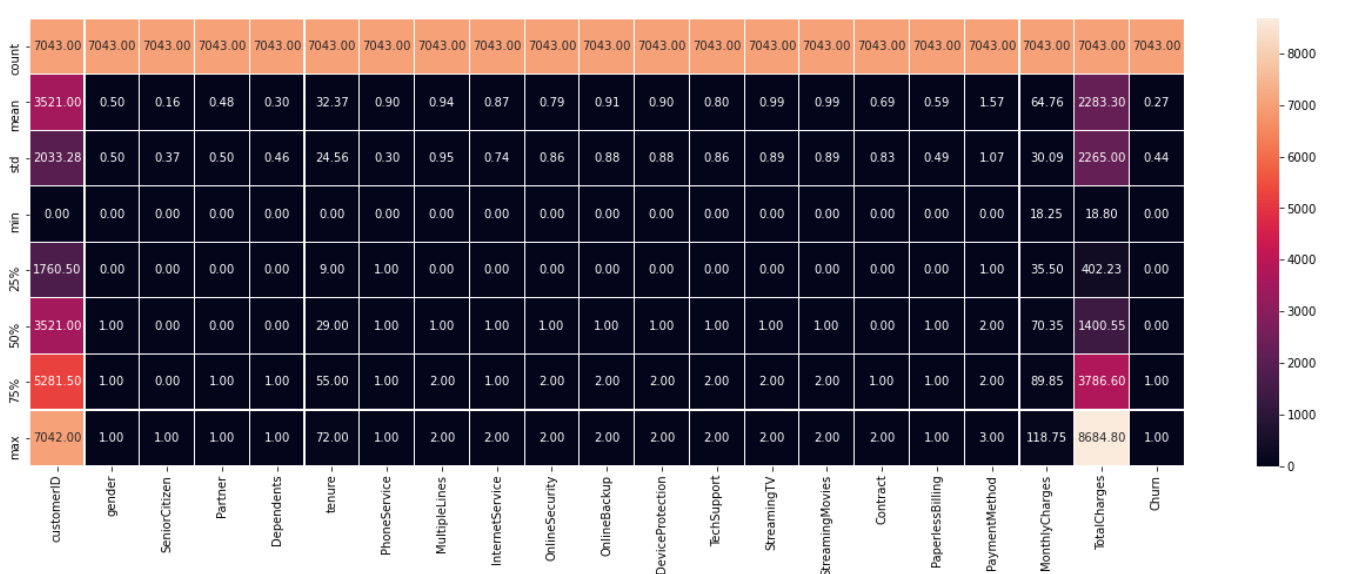


We have applied this label encoding for *Gender, Senior Citizen, Dependents, Partners, Phone services, Multiple Lines, Internet Services, Online Security, Online Backup, Tech Support, Streaming movies, Streaming TV, Device protection, Contract, Paper less billing, Payment Method and Churn*.

We have completed the label encoding, lets move on to statistical summary.

## *Statistical summary:*

Statistics is a pillar of machine learning. You cannot develop a deep understanding and application of machine learning without it.



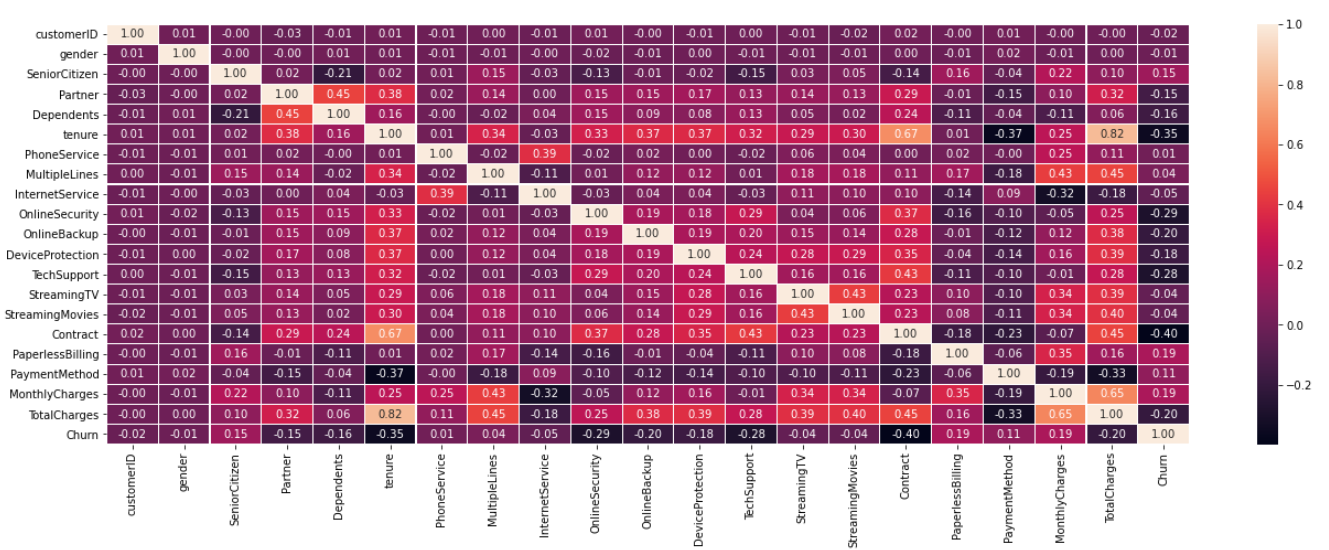
With the statistical summary, we can easily infer that:

* It is showing the number of values is same in every column.
* There is not much noticeable difference between mean and the median (50%) in any of the column.
* Statistical Summary is also showing that data is not much deviated from the path.
* There is not much difference between 75th percentile and maximum in any of the column except Total Charges.

Once statistical summary is done, I have moved towards correlation matrix.

## *Correlation matrix:*

This technique is very crucial to find out the correlation among the numerical variables.



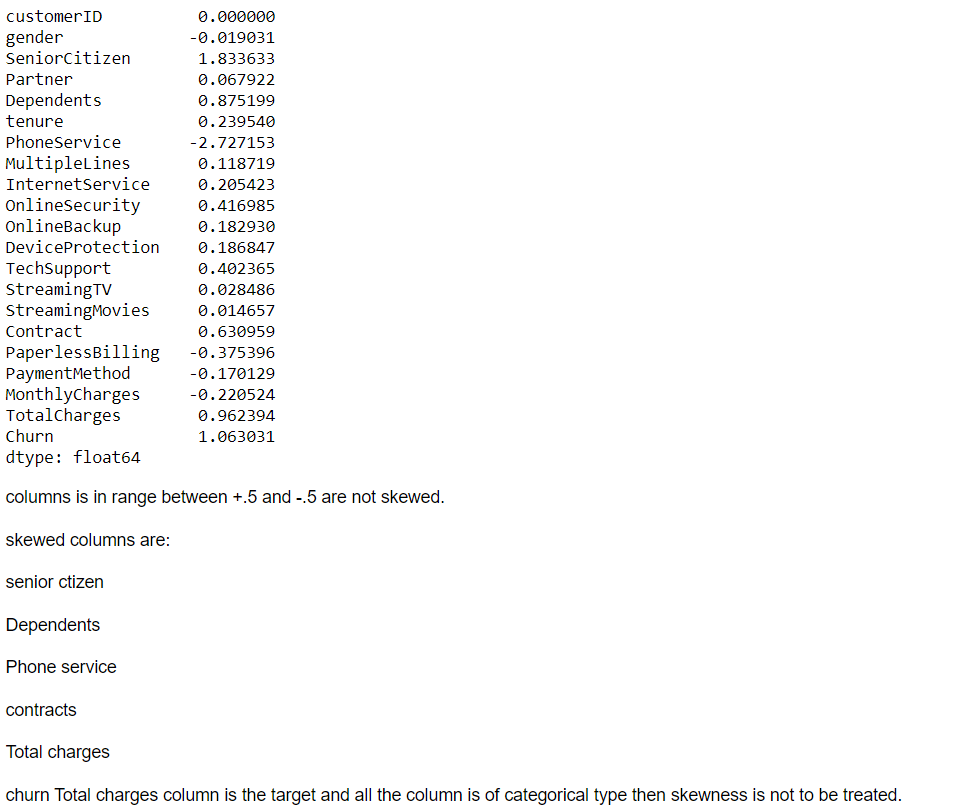
By carefully observing the matrix I found that:

* Light shades are indicating Positive correlations and Dark Shades are indicating negative correlations.
* No columns are showing correlations with target columns churns.
* Monthly charges column is positively correlated with total charges.
* Payment method is negatively correlated with total charges.
* Internet service is negatively correlated with total charges.

## *Skewness in the dataset*

I have also performed analysis on skewness of the dataset. This is the degree of asymmetry observed in a probability distribution. Distributions can exhibit right (positive) skewness or left (negative) skewness to varying degrees. A normal distribution (bell curve) exhibits zero skewness.

With our analysis with this data, we found senior citizen, dependents, phone service, contracts, and total charge are skewed columns where we require to be removal skewness, however, these are object type of columns. We can not perform skewness removal on object type of the data.

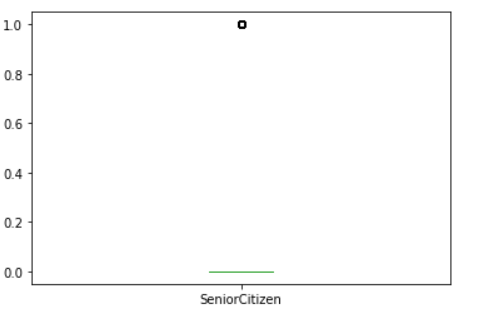


Furthermore, total charges can be a candidate of skewness removal, however, we do not have opportunity for the same as we are losing data while removal of skewness.

## *Outliers:*

Outliers are data points that are far from other data points. In other words, they are unusual values in a dataset. Outliers are problematic for many statistical analyses because they can cause tests to either miss significant findings or distort real results.

However, with the dataset we have, the scope of outlier removal is little as data are of object type. As a best practice towards building a good model, I have verified the outliers in some columns. This is also for the better understanding of the dataset.



Outlier is present in the columns but due to categorical column outlier cannot be removed.

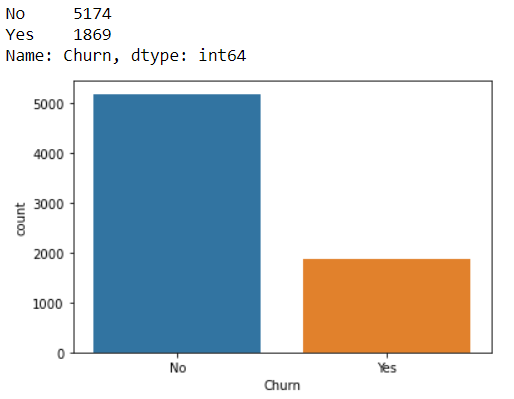
## Building Machine Learning Models for our dataset

With all the analysis and pre-processing we have done so far, its time to use the understanding of the data and build the data model.

For Building Machine Learnings Model, we have to split our dataset into features and target variables that is x and y.

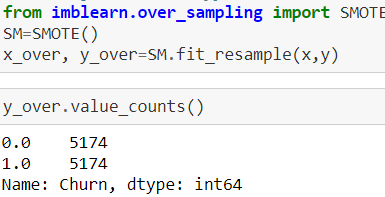


During Exploratory Data Analysis I noticed that our Target variables churns have class imbalanced data means there is much difference between yes and no,



We can see in above plot Number of No is more than Yes .so, if we build our model by taking this data then we will not get accurate prediction from the model built on such data.

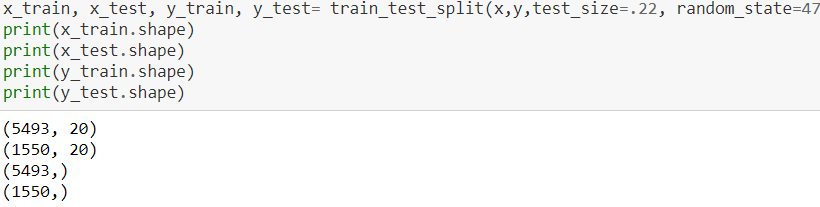
Hence, I am using a method to overcome from this situation. With the help of imblearn library I have imported SMOTE which is used to handle “Class Imbalance Problem by Oversampling The Minority Class” and followed the steps shown here.



Its time to split data into training and test data.

Train test split is a function in Scikit-learn model selection for splitting data arrays into two subsets: for training data and for testing data. With this function, you do not need to divide the dataset manually.

By default, Scikit-learn train test split will make random partitions for the two subsets.



With the given code, I am splitting by taking test size 22 and random state 47.

## *Importing Libraries:*

In python, we require some libraries to be imported in our code in order to perform specified functions. The libraries we needed for our model building are:

## 

We have done all required things until now. Its time to build the data model now. We will test multiple algorithms to find our best model which has capability to predict accurately. We have taken all supervised learning algorithm here.

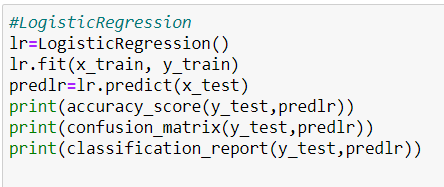
***Logistic Regression:***

Logistic regression is one of the most popular Machine Learning algorithms which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic Regression gives 78% accuracy. With cross validation applied on it by cross validation score method to avoid overfitting, it actually improved by 1% taking the accuracy to 79%. The confusion matrix and ROC curve for this regression model:

|  |  |
| --- | --- |
|  |  |

The code to generate the Logistic regression model.



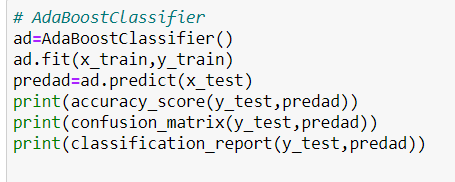
***AdaBoost Classifier:***

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.

This model gives accuracy of 79% and post cross validation by cross validation score method, the accuracy improved to 80%. The confusion matrix and ROC curve for this model:

|  |  |
| --- | --- |
|  |  |

The code to generate the Ada boost classifier model.

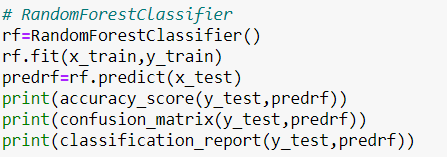


***Random Forest Classifier:*** A random forest is a meta estimator that fits several decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

This model gives accuracy of 79.29% and post cross validation by cross validation score method, the accuracy come down to 79.11%. The confusion matrix and ROC curve for this model:

|  |  |
| --- | --- |
|  |  |

The code to generate the Random Forest classifier model.

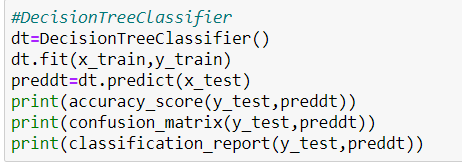


***Decision Tree Classifier:*** Decision Tree Classifier is a simple Machine Learning model that is used in classification problems. It is one of the simplest Machine Learning models used in classifications, yet done properly and with good training data, it can be incredibly effective in solving some tasks.

This model gives accuracy of 73% and post cross validation by cross validation score method, the accuracy come down to 72%. The confusion matrix and ROC curve for this model:

|  |  |
| --- | --- |
|  |  |

The code to generate the Decision Tree classifier model.

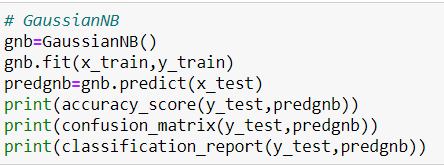


***GaussianNB:*** A Gaussian Naive Bayes algorithm is a special type of NB algorithm. It’s specifically used when the features have continuous values. It is also assumed that all the features are following a gaussian distribution i.e., normal distribution.

This model gives accuracy of 75.93% and post cross validation by cross validation score method, the accuracy come down to 75.37%. The confusion matrix and ROC curve for this model:

|  |  |
| --- | --- |
|  |  |

The code to generate the GaussianNB classifier model.



## Hyper parameter tuning

Finally, I have done the hyper parameter tuning which is being done for fine tuning of the model’s accuracy score. After this fine tuning of the models, Ada boost classifier comes out very accurate and it gives us the accuracy of 80%.

## Conclusion

With all analysis, pre-processing, model building and hyperparameter tuning, I recommend Ada boost classifier as the best model of prediction with highest accuracy of 80%.

This model will help a lot to the business for predicting the customer churn accurately and they will be able to prepare on focused marketing strategy for the specific segment of the customer. For example, if a customer is looking for a backup solution, telecom company can offer some storage on cloud for free backup which will help business retaining the customer. It will also help to identify their loyal customers for which they can create and offer or loyalty program points etc.