In the last week of the Evaluation Phase, you need to write an article/blog on any 1 project that you have built in minimum 2000 words in MS-Word Document File.

The article should contain the following sub-topics:  
1.      Problem Definition  
2.      Data Analysis  
3.      EDA Concluding Remarks  
4.      Pre-processing Pipeline  
5.      Building Machine Learning Models  
6.     Concluding Remarks

***Evaluation Project 7- Customer Churn Analysis***

**Problem Statement:**

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. However, 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 prioritise 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.

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

1. Problem Definition: Customer Churn or Churn rate is a measurement of how many customers decide to leave or no longer wishes to use the services of the telecom company. For telecom companies churn rate is a matter of concern because acquiring a new customer base is more expensive than retaining one. To prevent the churn, it is essential for the telecom companies to provide good services and exciting offers to their existing customers. One way to prevent customer loss is by understanding of their needs. These telecom companies compile the data and by doing through analysis of that data they can learn about customer behaviour and who is most likely to defect or change the subscription. Churn rates will let you know if your customers are leaving or staying. If customer is staying it means everything is working fine. If customers are leaving, then companies need to re-evaluate their strategies.

By the end of this article let us try to analyse the factors for customer churn (AKA Attrition) and how to reduce them.

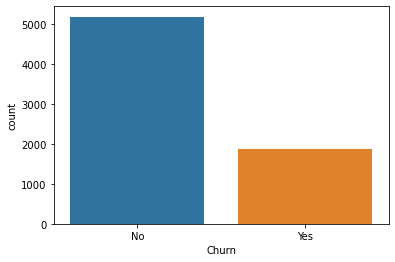
The churn analysis can help companies to understand the behaviour of the customer and understand who is least likely to continue their contract. It can also help in developing the strategies for retaining the high-risk customers.

1. Data Analysis: Through machine learning we can attempt to find out the best solution to tackle the attrition. There could be multiple reasons behind the churn, they can vary from dissatisfaction of the services or some new player coming in the market with newer facilities. To understand this in depth, we start by taking a high-level look at the data provided below. The dataset contains a ‘Yes/No’ field indicating whether the customer has left the telecom company but are not given any information about the departure date.
2. The shape of the dataset is 7043 rows and 21 total columns that describe 18 are Categorical columns and 3 are Numerical columns.
3. Target Variable (y) in this dataset is “Churn”. Below through the count plot you can get better idea of the distribution.

No 5174

Yes 1869

Name: Churn, dtype: int64



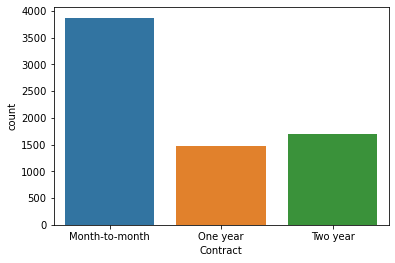
1. Fortunately, the dataset looks clean. We have no Null values in the dataset.
2. Out of the above one feature is redundant because it contained unique information (High Cardinality), which is not relevant with the “churn rate”. Customer Id and Total charges are having High Cardinality (Unique Values). So, we will drop this column, which is not adding any value to our target column.
3. In Total charges we had some blank spaces and we will handle them with replacing it with mean values.
4. Correlation: as per following graphs, we can see two images for correlation. Additionally, we can see which variables are positively and negatively correlated with our target variable. Most of the features provided in this dataset are highly correlated with each other. With the exception customer id, Total charges are uniformly distributed because they have the highest unique values.





1. EDA concluding remarks: The following list highlights the main points that emerged from this analysis

* From total customers, more youngsters opted for the service provider than senior citizens.
* 3875 customers opted for month-to-month contract but only 1695 customers for long-term contract as compared to 1473 customers who opted for one year contract with the telecom company.



* It was noticed that monthly charges had affected the churn more. Because they can leave anytime and were not restricted by the contract (one-year or two-year contract types)
* Customers without internet facilities and without tech support produced higher churn.
* Youngsters were high churners than senior citizens

1. Pre-processing Pipeline: This dataset is not scaled properly, so we will scale it with Min-Max Scaler. In this dataset, we have various categorical columns. Machines use only numerical data to analyse. Hence, to build a machine learning model, we need to convert our categorical columns into numerical with the Label Encoder or one hot encoder. For better accuracy we also have to drop some columns, which we did.

After applying zscore to calculate the Outliers, we will try to balance them for better model building.

After this stage, we need to split the dataset into training and testing sets using train-test-split method. As pre-processing is done, now we can move forward to the next phase of the model building.

1. Building Machine Learning Models: After gaining some insight on the characteristics of this telecom company’s data, we made an attempt to build some models and then carried out the results of several classification models, namely:
2. Naïve Bayes
3. Logistic Regression
4. K-Nearest Neighbours
5. Decision tree classifier
6. SVC
7. These models were then compared and as per their performance and accuracy, we selected one model. Here is the Logistic Regression which performed well as comparison to others. As a first step, the data is cleaned using different steps.

Then we will split the dataset in 80/20 ratio. 80% is for Training purpose and 20% is for Testing purpose.

This data then we will process further for different model building. In all models the dataset is stored in different files.

1. Concluding remarks: In this first endeavour to address the ‘churn’ issue in this company the data provided was analysed and a simple system where data was engineered and transferred into machine learning models that predicted customer churn was developed. Five machine learning models were implemented in this release, resulting in a best prediction of 79%. It can be observed that a decent result was produced by the neural network, when considering the small amount of data that was provided. several hyperparameter configurations were attempted for this classifier and the best results were included in this notebook. There is obviously room for further analysis and improvement and tuning of the models. In addition, more classifiers and different architectures can be tested in future attempts.