Pethulla Naveen Reddy

BT17GCS074

Case Study: Telco Churn

**Data Preprocessing and Exploratory Data Analysis:**

**Problem Statement**

The objective is to predict the churn of a telecom company.  
The model should be able to identify which user is going to churn in the next time period.

**Method of Analysis:**

1.Convertion of data set from text file to csv is done  
2.Understanding the subject and data set provided by some commands like .shape, .head, and .columns  
3.Cleaning the data that involves finding the missing values and filling them with attribute mean or most probable value  
4.Redundant has been removed.  
5.Relationship analysis is done using scatter plot to find if there is any trend followed by the target value on basis of predictors provided.  
6.Finding the relations between predictors and target value.  
7.Concluding the problem.

1. Business Understanding – Goals & Success Criterion:

The goal is to find a Classification model which can predict the users who are

about to churn

1. Data understanding:
2. Missing value identification: It is observed that the columns “logtoll, ”logequi”, ”logcard, “logwire” are the columns with containing missing values. There are significant large number of missing values in the ‘logwire’ column.
3. There are no outliers in the data set.
4. Most of the variables are having right skewed data, while some of them are having normal distribution

**Findings:**

1. It is observed that users between the income range 0 and 280 are likely to churn compared to the income range 281 and 1750.
2. User’s with addresses other than 60 and above 70 are likely to churn a lot.

**Limitations of the report:**

1. Heat map understanding (font size is too small).
2. Relation between target values and predictors using scatter plot.
3. Clear understanding of variables that could be impacting churn.

Case Study:Retail Forecasting

**ARIMA model:**

* ARIMA model is a popular and widely used technique for time series forecasting.
* ARIMA stands for Autoregressive integrated moving average
* AR: (Autoregression) A model that uses the dependent relationship between an observation and some number of lagged observations.
* I: (Integrated) The use of differencing of raw observations (e.g. subtracting an observation from an observation at the previous time step) in order to make the time series stationary.
* MA: (Moving Average) A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

Each of these components is explicitly specified in the model as a parameter. A standard notation is used of ARIMA(p,d,q) where the parameters are substituted with integer values to quickly indicate the specific ARIMA model being used.

The parameters of the ARIMA model are defined as follows:

1. p: The number of lag observations included in the model, also called the lag order.
2. d: The number of times that the raw observations are differenced, also called the degree of differencing.
3. q: The size of the moving average window, also called the order of moving average.

A linear regression model is constructed including the specified number and type of terms, and the data is prepared by a degree of differencing in order to make it stationary, i.e. to remove trend and seasonal structures that negatively affect the regression model.

A value of 0 can be used for a parameter, which indicates to not use that element of the model. This way, the ARIMA model can be configured to perform the function of an ARMA model, and even a simple AR, I, or MA model.

Adopting an ARIMA model for a time series assumes that the underlying process that generated the observations is an ARIMA process. This may seem obvious, but helps to motivate the need to confirm the assumptions of the model in the raw observations and in the residual errors of forecasts from the model.

* All the possible values for p, d, q are printed with the respective values in the end of the document and graph is plotted for the same as p=0, d=2, q=4 has the lowest arima value.