**Index**

1. Understanding the data
   * 1. Data Cleaning
     2. Data type

* Working with Categorical variable
  + 1. EDA

1. Working with CLRM
2. Alternative to prediction
3. Time series forecasting
   * 1. Univariate time series
     2. Multivariate time series
     3. Evaluation of the models

Firstly, we try to **understand the data** and what it is trying to tell us.



In the Data **Cleaning** part we check if there are any missing values in the data or the data contains any irrelevant duplicate values or columns. For e.g., name of customers, salesperson and products in sales data. In any case we won’t be using them, if needed we will use their Id’s. And if the unit price of the product is not mentioned and we have sales, quantities and discounts we can get the price of a particular item.

Then we move forward to checking the **type of the variables**. For e.g., In sales data Profit will be quantitative variable and categories will be qualitative variable. Now we will see what if our data involves categorical variables. Then let’s understand this by an example. Say e.g. In Sales data say we have 3 categories of the products under the column of categories namely “Technology”, “Office Supply” and “Furniture”. Then the following are the ways.





Now in **Exploratory Data Analysis** we will talk about relationships between the variables. As of now we won’t bother about feature selection much as we are not fitting the model right now we just want to know their behavior with each other. That can be achieved by either scatter plot or correlation between the variables. And to make it presentable we can use scatter matrix and correlation matrix. Which can easily be done in python. From this we can tell if the data is linear or not and it will be helpful in deciding models and methods to be used in prediction.

**Working with CLRM:**

Now we will see if we can work with the Classical Linear Regression Model. CLRM has few assumptions that need to be checked. Following are the steps to be done to achieve that.















If we see some level of linear relationship among them then we can apply our simple linear regression. Now to do so first we have to decide which features we want to have in the model and need to check the assumptions of that. So first we select the features by forward regression method and we have done that in python so it tells us which feature we should prioritize. After getting that we have to check the assumptions of CLRM (Classical Linear Regression Model). Software R provides us a package called “gvlma” to check the assumptions of CLRM. We just have to provide a model which we want to fit and a single line code will tell us which assumption is satisfied and which is violated. And if the assumptions are violated. we have to see if any remedial methods are providing satisfactory results or not.

**Alternative to Prediction**

Now the alternative way to get information from the data without modeling it, is ABC-inventory analysis and Market Basket analysis.

Market Basket Analysis is a technique which identifies the strength of association between pairs of products purchased together and identify patterns of co-occurrence (that is when two or more things take place together). Market Basket Analysis creates If-Then scenario rules, for example, if item A is purchased then item B is likely to be purchased.

ABC analysis divides an inventory into three categories ("A items" with very tight control and accurate records, "B items" with less tightly controlled and good records, and "C items" with the simplest controls possible and minimal records) to determine levels of importance.



While working with time series data we have to make sure that the time stamp in the data is equidistant. i.e. Hourly, Daily, Weekly, Monthly etc.

For e.g. We have data of sales, which contain dates on which the order has been booked and we have data of 3 years; 2016, 2017 and 2018. So, we have to make this data equidistance; Monthly or Quarterly to have enough meaningful time points. So, if we decided to make them monthly data then we will have data of 36 months.

As we begin working with endogenous data and start to develop forecasting models, it helps to identify and isolate factors working within the system that influence behavior. The decomposition of a time series attempts to isolate individual components such as error, trend, and seasonality (ETS).  Stats models in python provide a seasonal decomposition tool we can use to separate out the different components. This lets us see quickly and visually what each component contributes to the overall behavior.

Then we test stationarity in our time-series data. Stationarity means that the statistical properties of a time series (or rather the process generating it) do not change over time. Stationarity is important because many useful analytical tools and statistical tests and models rely on it.

Now as we know Time-Series data can be handled by both Univariate modelling as well as Multivariate modelling. So, let’s understand them one by one.



In the Univariate time series model, there are only two variables. One is an independent variable which contains time points and the other is dependent variable which depends on the time and we want to predict.

If the time series is stationary then we can just apply the simplest models Auto Regressive models of any order, say p AR(p) and Auto Regressive Moving averages ARMA.

If data are non-stationary then we can apply the ARIMA, SARIMA and Holt-Winter Methods as per the type of the variation contained in the time series.

ARIMA: ‘Auto Regressive Integrated Moving Average’, is a forecasting algorithm based on the idea that the information in the past values of the time series can alone be used to predict the future values. It is useful when time series exhibits a trend.

SRIMA: ‘Seasonal Auto Regressive Integrated Moving Average’, is a forecasting algorithm. It is useful when the time series exhibits Seasonal variation.

Holt-Winter Method: It is Triple Exponential Smoothing – used for forecasting data with trend and/or seasonality.



In the Multivariate time series model, there are more than two variables. One is an independent variable which contains time points and other independent variables. Another is a dependent variable which depends on the time and other independent variables, and we want to predict.

First is VAR. Which is the most commonly used method for multivariate time series forecasting – Vector AutoRegression (VAR). In a VAR model, each variable is a linear function of the past values of itself and the past values of all the other variables. Another is Random Forest Regression. It can also be used to forecast TS data.

**Evaluation of the models:**

Now if applicable, one may try several models from these models listed, including Univariate as well as Multivariate models. Then to choose which model is best from this; That is to evaluate these models one may use AIC (The Akaike information criterion) or RMSE (Root Mean Square Error). For both of them, as the value is less the better model is.