Milestone Report

Sales Forecasting Model

- **Problem**

  The aim is to build a forecasting model for sales of items' demands across multiple stores.

- **Client**

  Retail stores including brick and mortar as well as online stores are always looking at the demand forecasting. It could be easier to keep the stores and or warehouses always over filled so as not to run into out of stock situations. However that approach will turn out to be potentially cost prohibitive in cases when certain items demands are low or very high depending on the seasons and markets trends etc. To overcome this age old problem, we would be looking into forecasting the item demand well in advance to help the retailers. In this time series forecasting problem, we would be looking at historical data of an item and its sales over a period of 5 years and then forecast for the next quarter.

- **Data Acquisition**

  The training and test data for the model will be acquired from kaggle.
  https://www.kaggle.com/c/demand-forecasting-kernels-only/data

Data Wrangling Steps

- **Data Acquisition**
  The training and test data for the model is acquired from kaggle.
  https://www.kaggle.com/c/demand-forecasting-kernels-only/data

- **Storage and access**
  The data is stored locally as well as in the project's Github repository. Through the jupyter notebook, the data file is opened and data is loaded in pandas dataframe.

- **Investigation and findings**
  - The data is given in csv format. The training set has little less than a million records.
  - The shape of data is (913000, 4)
  - All the columns are integers except the date
  - Sales values are given as integers, so one unit probably corresponds to one item.
  - sales is our target column.
  - There are 10 stores. Each store has 50 items each.
  - Sales values are by store, by item per day for a period of 5 years.

- **Missing Values**
  - In further exploration of data we find that there are no missing values in any of the columns. If there were any columns with missing percentage >50, we would have dropped those features as it will be difficult to project the missing values of that magnitude.

- **Outliers**
  - Out of our dataset, if we go feature by feature the date column has values for 5 years by day. There are no outliers there. Next is store number and item id. That we need not worry about since these are id columns. However closer look shows that these values range from 1-10 and 1-50 respectively. We use the Inter quartile range(IQR) approach to find out if there are any outliers in the sales numbers. And figure out there are none!

Exploratory Data Analysis

Next, We will inspect each feature individually.
We will perform the univariate and bivariate analysis using the following steps-

1. Summary statistics
2. Plot individually
3. Plot against our target variable- sales to see if and how both the features are related
4. Modify the feature to make it ready for modelling, if necessary

For individual summary statistics we will use

- .describe()
- .sample(5)
- .nunique()
- Group_by if necessary

1. Store

This shows there are 10 unique stores in our dataset. Each store has 91300 individual sales recorded in the dataset. The column has no missing values and no special characters etc in the data values. The values are 1-10. So we will not need to do any pre-processing.
After plotting we see that-

a. Store 2 and 8 have high highest sales
b. Store 5 and 7 have lowest sales
c. There is a consistent increase in sales numbers every year at all the stores!

2. item

There are 50 items in each store. The item column values range from 1-50.

3. date

The dates are from Jan 1 2013 to Dec 31 2017. The dates are for 5 years period with the values ranging by the day in this period.

Seasonality:

Next we plot the sales to look at the seasonality for- -Yearly -monthly -weekly. After plotting we observe that

- the sales go up year to year from 2013 to 2017

- there is a steady upward trend.
- there is also a monthly seasonality where sales are going up around middle of the year
- the sales follow the same trend year after year where sales peak up at the middle of the year and taper off with one exception in the month of November. Over the years the sales have gone up basically from year to year but follows the same trend.