Capstone Project Report

- **A. Sai Gowtham Babu**

- **Main objective of the analysis that also specifies whether your model will be focused on a specific type of Time Series, Survival Anlaysis, or Deep Learning and the benefits that your analysis brings to the business or stakeholders of this data.**

My dataset consists of sales of medicines in a particular year. For my project, I need to forecast the future sales of medicines using time series analysis. So that the company can order that quantity od medicines. I tried to create a model that can best fir the data by tuning down with different parameters and hyperparameters.

- **Brief description of the data set you chose, a summary of its attributes, and an outline of what you are trying to accomplish with this analysis.**

My dataset contains the following attributes

1) Date: - Date of medicine sold
2) Customer name: - Represents the name of the customer
3) Quantity sold: - Number of medicines sold to a particular customer
4) Medicine ID: - Unique ID for medicines
5) Free Quantity: - Free quantity that are given to customer who bought medicines in bulk.
6) Order Quantity: - The quantity the distributor has ordered
7) Order free: - The quantity the distributor got for free

- **Brief summary of data exploration, actions taken for data cleaning or feature engineering.**

Coming to the dataset, my dataset consists of stationary data. So, there is no problem in stationary. Since, this being a medicines sales dataset, there is no seasonality and there is variance in dataset. But, decreasing variance is leading to complete damage of the model. So, I trained the data with some variance fluctuations.

```
In [19]: from statsmodels.tsa.stattools import adfuller

adf, pvalue, usedlag, nobs, critical_values, icbest = adfuller(df2)
print (pvalue)

if pvalue < 0.05:
    print('Stationary')
else:
    print('Not- Stationary')

df2.plot()
```

```
1.4856538538590908e-24
Stationary
```

- **Summary of training at least three variations of the Time Series, Survival Analysis, or Deep Learning model you selected. For example, you can use different models or different hyperparameters.**

Even though there is no seasonality, I tried fitting my model with an auto_arima model, so that my machine can choose the best parameters for my model. My machine chooses a model the best fits my model. So, I have chosen that model as my final model.

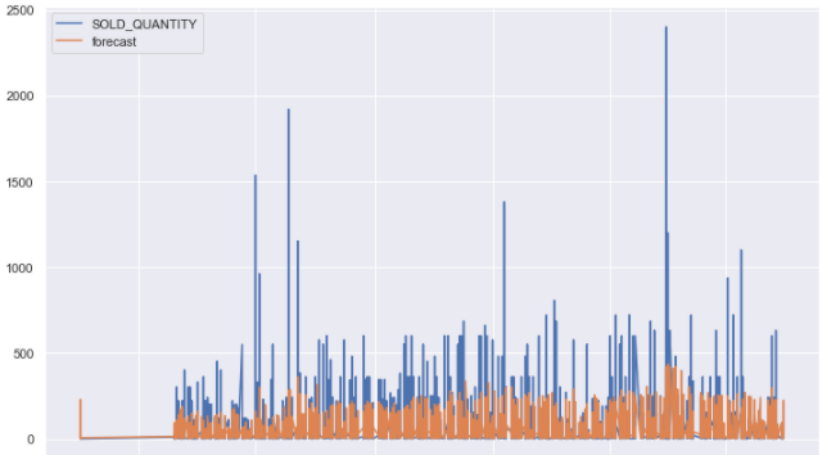
```
In [81]: import statsmodels.api as sm
         model=sm.tsa.statespace.SARIMAX(df2['SOLD_QUANTITY'],order=(27, 1, 1))
         results=model.fit()
         df2['forecast']=results.predict()

         from sklearn. metrics import mean_squared_error
         sqr_error = mean_squared_error(df2.SOLD_QUANTITY, df2.forecast)
         np.sqrt(sqr_error)

Out[81]: 72.46744419967166

In [84]: df2[['SOLD_QUANTITY','forecast']].plot(figsize=(12,8))

Out[84]: <matplotlib.axes._subplots.AxesSubplot at 0x1d301b1c8e0>
```



- **A paragraph explaining which of your models you recommend as a final model that best fits your needs in terms of accuracy or explain ability**.

According to the graph, the results that I got are not as good. But there is a reason behind decreasing the future orders. (explained in next part). The above shown model is the best model that I could probably use for prediction of the data.

- **Summary Key Findings and Insights, which walks your reader through the main findings of your modelling exercise.**

1) The data is following particular order with respect to orders. The distributor was correctly able to guess how many medicines the customers are going to answer
2) Out of the total year over 50,000 medicines were present in the left-out stock.
3) There were some huge spikes in the graph. They can be removed but included for better understanding in the data. (According to my analysis I found that those are the customers who bought medicines in the start of a new season.)
4) The stock is predicted already considering the left over 50,000 units. Since, the left-out stock needs to be sold first. (This is place where my model comes to place since, there is already left out stock. So, that is the reason why there are some less high spikes in my data.)

- **Suggestions for next steps in analyzing this data, which may include suggesting revisiting this model or adding specific data features to achieve a better model.**

The model is pretty good. But there is always room for improvement and this can be done by bringing in more data. Since, we only have one year of data in the model. The model can be improved by testing it on real data that occurs in the future. We can also check whether there is some day to day trend going on by further diving the data into date, week and month columns.