MartinDow

Internship Report - IT Department

Mohib Ali Khan

18th December 2023 - 26th January 2024 *(6 Week Internship)*

Executive Summary:

During my internship at Martin Dow, I worked on various projects in the IT department. I worked on the development of a dashboard using Power BI, read articles about business warehouses, and designed a machine learning model which predicts the quantity of 2024 brick wise and learned about business warehouses through sessions. Overall, my internship was a valuable learning experience that allowed me to develop new skills and gain practical experience in Data Science or IT.

Introduction:

I was excited to have this opportunity to intern at Martin Dow Company, a leader in the field of Pharmaceuticals. My primary responsibilities were to learn or work under the BI team.

Description of Duties:

During my internship, I was responsible for different tasks in the IT department. these include:

- **Designing a Dashboard:**

I was given a dataset of Sales of a supermarket in Myanmar on which I had to create an interactive dashboard using Power BI, at first I loaded the data on Power BI and then used data modelling to transform the data based on my needs. The data had members and normal to identify a customer type hence to make it sound better I transformed the normal value to be as Not Member so that it makes more sense. Then after receiving instructions from Muhammad Imran Khan, I was asked to make different visualisations which could cater for the needs of the Dashboard

- **2024 Product Quantity Prediction(Machine Learning):**

Since our data was extensive and the task was to capture the seaosonal impact of our quantities brick wise so upon researching I found out ARIMA model (AutoRegressive Integrated Moving Average)

ARIMA(p,d,q) forecasting equation: ARIMA models are, in theory, the most general class of models for forecasting a time series which can be made to be "stationary" by differencing (if necessary), perhaps in conjunction with nonlinear transformations such as logging or deflating (if necessary). A random variable that is a time series is stationary if its statistical properties are all constant over time.

The acronym ARIMA stands for Auto-Regressive Integrated Moving Average. Lags of the stationarized series in the forecasting equation are called "autoregressive" terms, lags of the forecast errors are called "moving average" terms, and a time series which needs to be differenced to be made stationary is said to be an "integrated" version of a stationary series. Random-walk and

random-trend models, autoregressive models, and exponential smoothing models are all special cases of ARIMA models.

A nonseasonal ARIMA model is classified as an "ARIMA(p,d,q)" model, where:

p is the number of autoregressive terms,

d is the number of nonseasonal differences needed for stationarity, and

q is the number of lagged forecast errors in the prediction equation.

import pandas as pd: Pandas is used for data manipulation, and it provides data structures for efficient data analysis.

from pandas import to_datetime: This function is used to convert data types, specifically converting date strings to datetime objects.

from statsmodels.tsa.arima.model import ARIMA: Statsmodels is a library for estimating and testing statistical models. ARIMA (AutoRegressive Integrated Moving Average) is a model used for time series analysis.

import numpy as np: Numpy is a library for numerical operations in Python.

import matplotlib.pyplot as plt: Matplotlib is used for creating visualizations such as plots and charts.

from statsmodels.tools.eval_measures import rmse: This imports the root mean square error (rmse) function from statsmodels, which is a metric used to evaluate the performance of a model.

Once I was done with importing libraries the next task is prepare the data before passing it into the model for training and testing

The initial steps of the code focus on preparing and structuring the dataset for subsequent time series analysis using the ARIMA model. The data is loaded from a CSV file, and particular attention is given to the 'DATEOFSALE' column, which represents the timestamp for each observation.

*Converting this column to datetime format is imperative for accurate temporal analysis. Following this, the dataset is grouped by both the 'DATEOFSALE' and 'BRICKID' columns, aggregating the total quantity ('QTY') for each unique combination. The 'DATEOFSALE' column is then converted to a monthly period, facilitating a more granular time-based analysis. To ensure completeness in the analysis, a DataFrame is created to encompass all possible combinations of 'DATEOFSALE' and 'BRICKID'. This is achieved by merging the combination DataFrame with the grouped totals DataFrame, filling any missing value*s with

zeros. The final step involves grouping the merged data by 'DATEOFSALE' and 'BRICKID' once more, aggregating the total quantity for each group.

This structured and aggregated dataset is then ready for utilization in the subsequent ARIMA model. Each of these preprocessing steps is undertaken to ensure the dataset is appropriately formatted, complete, and ready for accurate time series analysis.

The next step is to determine p d and q values which will be passed into our model, there are various ways to choose the values of parameters of the ARIMA model. Without being confused we can do this using the following steps:

Test for stationarity using the augmented dickey fuller test.

If the time series is stationary try to fit the ARMA model, and if the time series is non-stationary then seek the value of d.

If the data is getting stationary then draw the autocorrelation and partial autocorrelation graph of the data.

Draw a partial autocorrelation graph (ACF) of the data. This will help us in finding the value of p because the cut-off point to the PACF is p.

Draw an autocorrelation graph (ACF) of the data. This will help us in finding the value of q because the cut-off point to the ACF is q.

I did this in ACF.py file.

Once I determined the plots I had an idea about the fit, the next step is just trial and error you pass in different combinations of p d and q and once you find the right combination you proceed with predictions.

Some Observations:

- *There are some bricks for which the model predicts abnormal value upon verification that particular brick was an anomaly, there are many bricks with similar behaviour.*
- *There are some bricks to which we have sold nothing in particular month hence during transformation phase just run through the whole file and assume zero for those bricks.*

The final task was to create a dashboard for analysis of our model on Power BI so I used merger.py python file to create CSVs for each year to build our model's comprehensive analysis

- **Business Warehouse:**

The exploration of business warehousing has provided valuable insights into the intricate processes of data management. Delving into on-premises data centres revealed their significance as the foundational infrastructure for securely storing and managing enterprise data. The role of transactional databases within ERP systems became evident, capturing and preserving real-time transactional data critical for day-to-day business operations. Understanding the Extract, Transform, and Load (ETL) processes underscored their crucial role in preparing and optimizing data for analytical purposes. Weekly sessions further enriched this understanding, covering topics such as data modelling and governance, contributing to a comprehensive grasp of the complexities involved in effective data lifecycle management.

Accomplishments:

One of the projects I was most proud of during my internship was developing the ARIMA model for quantity prediction, it was my first time dealing with big data and using it to deploy the model.

Skills Learned:

During the internship, I had the opportunity to learn a variety of new skills. These included dealing with big data, gaining insights into SAP, extensive dashboards and using Power BI.

Conclusion:

Overall, my internship at Martin Dow was a valuable learning experience. I developed new skills and gained practical experience in Data Science. I am grateful for the opportunity to have interned at such a successful and dynamic company.

Acknowledgements:

I would like to thank the entire BI team at Martin Dow Company for their support and guidance during my internship. In particular, I would like to thank my line manager, Muhammad Imran and Ahsan Qamar , for their mentorship and the opportunity to intern at the company.