INTERNSHIP: INTERIM PROJECT REPORT

Dear Intern

Interim project report is an inherent component of your internship. We are enclosing a reference table of content for the interim project report.

The key objective of this report is for you to capture how far you have got in completing the internship work against milestones expected to be achieved within a specific duration and seek the mentor's feedback. Depending on the internship project and your progress (IT/Non-IT, Technical/Business Domain), you may choose to include or exclude or rename sections or leave some sections blank from the table of content mentioned below. You can also add additional sections. You can refer the project presentation to view the milestones related to your internship project. Please populate milestone# (1 / 2 / 3) and the milestone description in the interim project report based on the milestone for which you are submitting the interim project report.

You can refer the project presentation to view the milestones related to your internship project.

Internship Project Title	*Forecasting System - Project Demand of Products at a Retail Outlet Based on Historical Data.*
Name of the Company	*TCS ion*
Name of the Industry Mentor	*Debashis Roy*
Name of the Institute	*ICT Academy of Kerala*

Start Date	**End Date**	**Total Effort (hrs.)**	**Project Environment**	**Tools used**
20 Apr 2023	*15 Jul 2023*	*63.5*	*VS Code,* *Jupiter Notebook.*	*Excel, Python-3.9.13, Python-3.8.4, Python packages like NumPy, Pandas, Seaborn, Matplotlib, Seasonal_decompose, Auto_arima.*

Milestone	02	**Milestone:**	Student should be able to choose the appropriate forecasting model for the dataset. Students should also be able to fit the model to the dataset and make predictions

TABLE OF CONTENT

ACKNOWLEDGEMENTS

I ***Govind S*** would like to express my sincere gratitude to my Industry mentor ***Debashis Roy*** and ***TCS ioN*** for providing me with all the necessary resources and guidelines for the completion of this project Milestone 01 as part of the RIO-125 internship on Demand Forecasting.

I would also like to thank my colleagues who have helped me in various ways, such as sharing their knowledge and perspective, providing technical supports, and on giving me constant feedback of my works.

Finally, I would like to express my appreciation to the kaggle user ***Pavan Kumar D*** who has provided this dataset.

Thank you all for your support and encouragements.

OBJECTIVE

Create a predict/ forecast model which can produce an ahead weekly sales of the product categories by applying concepts of time series forecasting, quantitative forecasting methods, auto regressor, moving average approaches, ARIMA, etc. so that the stores owners could be prepared of the number of products that are needed in their inventory.

The project milestone 02 involves several steps like: Choose the appropriate and apt forecasting model for the dataset. Fitting the model to the dataset and make predictions of the next week demands.

INTRODUCTION

In today's market of fast-paced service industry, predicting client/ customer demand for products/ requirements is a crucial for inventory management and meeting customer needs.

In this project, I will be developing a forecasting model that can predict the demand of products at a retail store based on their past data of timeline between 2012-2014.

Once the data understanding is completed and preprocessed, the next common procedure would be to select a suitable/ ideal forecasting model for the. respective After selecting the model, I will fit the model with the dataset and make predictions. To validate the accuracy of the predictions, I will visualize them using various visualization techniques after splitting a train and test then comparing the predictions with test also there is diagnostic plot to check the model goodness.

Source of data:

The dataset was downloaded from the open source Kaggle platform.

About Dataset: Predicting the Demand of Products Across Stores of a Retail Chain.

A large Indian retail chain has stores across 3 states in India: Maharashtra, Telangana, and Kerala. These stores stock products across various categories such as FMCG (fast moving consumer goods), eatables / perishables and others. Managing the inventory is crucial for the revenue stream of the retail chain. Meeting the demand is important to not lose potential revenue, while at the same time stocking excessive products could lead to losses.

In this problem you are tasked with building a machine learning model to predict the sales of products across stores for one month. These models can then be used to power the recommendations for the inventory management software at these stores.

The datasets are provided as cited below.

train_data.csv Features:

- *date*: The date for which the observation was recorded.
- *Product-identifier*: The id for a product.
- *Department-identifier*: The id for a specific department in a store.
- *Category-of-product*: The category to which a product belongs.
- *outlet*: The id for a store.
- *state*: The name of the state.
- *sales*: The number of sales for the product.

Auxiliary Datasets:

- *product-prices.csv*: The prices of products at each store for each week.
- *date-to-week-id-map.csv*: The mapping from a date to the week_id.
- *sample-submission.csv*: The format for submissions.

INTERNSHIP ACTIVITIES

- ✓ Plotted several plots like boxplot, correlation plot, etc. to understand the data as part of EDA.
- ✓ Separated/ selected the data's based on the categories of products to daf (drinks and food), fmcg (fast moving consumer goods) and ot (others).
- ✓ Feature Engineering done on the above saved data frames to include only the required features/ attributes for Time Series Analysis and Forecasting.
- ✓ Used Group-by function and sum to the daily basis data as there may be one or more of sales of the same product for the same date that have happened and recorded or entered as two different instances.
- ✓ Default index of data is changed to its respective sales data for the ease of time series analysis.
- ✓ Analyzed each group of data's mean separately for individual behavior by resampling with both MS (month starting) and W (weekly).
- ✓ Found the outliers and anomalies their reasons behind behaving as such. (Probably it was due to the store closure on 25 Dec of every year.)
- ✓ Conducted decomposition process of each sales data of the respective categories to analyze the seasonality and stationarity of the data by using seasonal_decompose plot.
- ✓ Gathering insights/ information from seasonal decomposition plots about Sales, Seasonality, Trend and Noise of the data which are useful in understanding the data.
- ✓ Checked the stationarity of data by using Augmented dickey-fuller test.
- ✓ Plotted Auto correlation (ACF) and Partial Correlation (PACF) Plots to understand the parameters required for Forecast Model (ARIMA & SARIMAX Model).
- ✓ Converted the non-stationary dataset (fmcg, ot) into stationary dataset by differencing using the diff () function to use this in ARIMA Model or without this it can be passed to SARIMAX.
- ✓ Used Auto ARIMA model to obtain the optimal set of parameters and values needed to get a neat prediction/ forecast for the next week.
- ✓ Split the dataset to test and train set compares the models with different set of parameters.
- ✓ Trained the model on the entire dataset to predict the sales value of the future month.

*Note: The above activities/ steps are which I have done after Day 5 which is 23 Apr 2023. So, the link to the previous steps and information is embedded just below.

Day05_Milestone#01__Project_Interim_Report (click to see).

CHALLENGES & OPPORTUNITIES

The biggest challenge of all was the time and correct proven resources required to find and make a suitable forecast/ prediction model and to optimize/ fine tune that model for best forecast. During this process I had the opportunity to explore the world of forecasting and its influence, effects and impacts and the areas where such models are widely used and required.

METHODOLOGY

Data Mining: Extracted data from open source 'Kaggle' platform.

Data Analysis: Data analysis involves several steps identifying and understanding each data (attributes and its instances). What's its stance or represents for, what each have impact on the next thing.

In this project, imported the several basic python packages like NumPy, pandas, matplotlib to basic necessary data science projects into the Python environment and obtained basic details about it using info(), describe(), dtypes(). Then started with pre-processing of the data, checking for null values, duplicated values, and outliers. Found that there were no null values in the dataset, allowing us to move on to other steps of feature engineering.

Feature Engineering: The data is separated or classified into three data frames based on the category of products by this way forecasting can be done on each product category. Removing other unnecessary data's that does not play a role in Time Series Analysis.

Time Series Analysis: Each data frame (categorically separated ones) is separately analyzed for better understanding of each product categories sales which gives an idea of the purchase trend of the product and its demand.

Checking for Seasonality: All the three category of data was checked for seasonality and all of them turned out to be seasonal which was a result after seasonal decomposition plot, a rough idea of the sales trend can also be identified from this plot.

Checking for Stationarity: A Hypothetical Statistic test called Augmented Dicky-Fuller Test was conducted to determine whether the sales data were stationary or not. Results were Drinks and Food was a Stationary type whereas the Fast-Moving Consumer Goods and Other were Non-stationary.

Making Stationary: Certain Differencing methods like diff(), shift(), rolling mean() or by applying sqrt(), log10(), etc. could be implemented to manipulate the data's to stationary.

REFERENCES

Link to the Dataset/ Data Card:	Predicting the Sales of Products of a Retail Chain
Link to Code/ Executable File:	Work file.ipynb