Programming for AI – TABA

Project Journal

Mohammad Shehnaj

x23305762

x23305762@student.ncirl.ie

Project Objective:

This project covers the time series forecasting techniques to predict the sales trends using the "Avocado Sales" dataset in retail sector. The main objective is to analyze the past data of avocado sales and its prices and implement the time series models to derive insights which can help in decision-making. This project uses various time series models such as ARIMA, SARIMA and Exponential Smoothing which helps in predicting the future trends of average price and total volume columns of the given dataset.

Project Discussion:

Time Spent – 3 days

For this project my team and I have chosen the topic of Time-series analysis and forecasting in Retail industry. All the team members discussed the structure and approach of the project so that we all are on same track.

Data Collection and Finalizing:

Time Spent – 2 days

The project began by collecting and checking datasets from various sources related to retail industry. Later, the Avacado Sales and Prices dataset was found on Kaggle which was more relavent to the project's objectives. The avocado dataset icnludes date, Average columns, Total Volume, PLU 4046, PLU 4225, PLU 4770, region, type, year and more columns. It has 18249 entries with 14 columns. This dataset correctly aligns with project objectives of sales forecasting.

Challenge: Initially identifying the right dataset and ensuring its relevance was time-consuming. Few datasets lacked key details like dates, less entries, less features, no variation in date making it unsuitable for this project.

Outcome: The avocado dataset was finalized as it provided nescessary data for analysis.

Data Cleaning and Preprocessing:

Time Spent – 5 hours

I have loaded the dataset in jupyter notebook and performed initial manipulation like checking for missing values and duplicate values, renaming columns and dropping irrelavant columns. There were no missing values and duplicates found. The column “Unnamed” was removed as it was not relevant for analysis. This step helped the dataset for analysis and free from errors. The statistics of the dataset was also shown. The outliers were also checked, many outliers were identified in numerical columns, these were plotted with box plot for better understanding and were managed by applying the IQR method as these outliers could impact the model performance. The resulted in reducing the dataset from 18249 to 6725 entries. The data was segmented to weekly data for better analysis. This has reduced the dataset further to 169 entries and the other column values were calculated as average or sum depending on the type of dataset.

Challenges: The presence of outliers required careful handling to avoid losing critical trends. The reduction of dataset has impacted loss of critical data from the dataset. Many methods to handle outliers can be done but I went ahead with IQR method. The segmentation of dataset to consistent pattern was required, initially I started with year but the data was too less, then checked for monthly appoach but still was less for further analysis so I went ahead with weekly segmentation.

Outcome: A cleaned and structured dataset was prepared with relevant features for analysis and modelling.

Exploratory Data Analysis (EDA):

Time Spent: 3 days

EDA visulaizations uncover trends and patterns in data. Many visulaizations such as boxplots highlighting the outliers, plotted line graph for weekly trends for “Average Price” and “Total Volume” shows the fluctuations over time, seasonal decomposition shows the trend, seasonal and residual components of data, correlation heatmap shows the relationship among the feature, scatter plots and bar graphs, horizontal bar graphs region-wise were plotted to see important information like trends and patterns. This step help us identify the seasonality and trends for forecasting for next steps. It also helps in understanding the patterns whether the avocado sales were higher during certain seasons.

Challenges: Identifying the seasonal patterns even though the data had outliers was challenging. Line plots helped to solve this issue.

Outcome: It was observed that there was a clear seasonal trend in sales volume and stable price trend with minor fluctuations.

Data Transformation:

Time Spent: 1 day

I have applied the ADF test to check the stationarity of the data and found that the data was not stationary. Non-stationary data can lead to poor model performance as mostly the time series data are stationary. In order to stablize the variance, log transformation and differencing techniques were applied and have rechecked the stationarity using ADF test. After applying transformation techniques the data was stationary.

Challenges: Choosing the correct transformation method and differencing order was challenging as it required trail and errors.

Outcome: The data was properly transformed and was ready for time series modleing and forecasting.

ACF and PACF Analysis:

Time Spent: 1 day

I have plotted the ACF and PACF plots with original dataset and transformed dataset to identify the significant lags for ARIMA model. This helped to identify the lag values (p,d,q) which is required in model implementation.

Challenges: Understanding the plots was confusing but focusing on first significant spike helped in understanding the process.

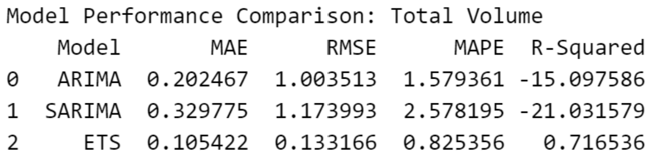
Outcome: Lag Values were plotted and identified for model implementation.

Model Implementation:

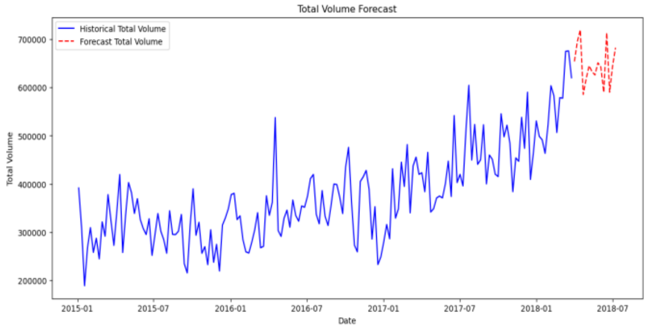
Time Spent: 4 days

The models ARIMA, SARIMA and Exponential Smoothing were used to forecast the Average Price and Total Volume. The data was split into training and testing where 80% of the data was training data and 20% was testing data. Each model was evaluated using metrics like MAE, RMSE, MAPE and R-squared. This determines to try different forecasting approaches and identify the most accurate model.

I have also run the models directly without splitting it as the data was less and ETS gave a good R-squared score with less MAE, RMSE and MAPE for Average Price and Total Volume. But as per the good practice, the models are to split and run to get better forecasting results. Below are the results for it:



```
Model Performance Comparison: Total Volume
    Model       MAE      RMSE      MAPE  R-Squared
0   ARIMA  0.202467  1.003513  1.579361 -15.097586
1  SARIMA  0.329775  1.173993  2.578195 -21.031579
2     ETS  0.105422  0.133166  0.825356   0.716536
```



Challenges:

1. ARIMA gave poor R-squared value in negative.
2. SARIMA was slightly better than ARIMA but wasn't the best model.
3. ETS was a strong fit for Average Price but not for Total Volume as the R-squared values were negative.
4. ARIMA was a best fit for Total Volume

Outcome: ETS was a best-performing model for Average Price but not for Total Volume as the R-squared values were negative and errors were also identified. ARIMA was astrong fit for Total Volume. This suggests that different models excel in predicting different aspects of avocado data. More advanced time series techniques can give us good forecasting results.

Model Evaluation:

Time Spent: 1 day

The Model Evaluation was done by comparing the metrics MAE, RMSE, MAPE and R-squared values of all models for Average Price and Total Volume. The R-squared values of Average Price was 0.68 and low MAPE score of 2.08 shows ETS is best fit for Average Price. But in case of Total Volume, the R-squared value was negative which shows the variability in data. ARIMA came out to be a best fit for Total Volume. More advanced time series techniques needs to applied for this for better forecating and finding the right model.

Challenges: As the data values were it was difficult to run the model. Different variations of yearly, monthly and quarterly were tried earlier but due to insufficient data, we couldn't proceed further.

Outcome: From the given three models, ETS was a strong fit for Average Prices and ARIMA for Total Volume. This suggests that different models excel in predicting different aspects of avocado data.

Forecasting Future Trends:

Time Spent: 4 hours

As the ETS model was comparatively better than others, I have forecasted the Average Price and Total Volume for next 15 weeks. The forecasts were done using the past data and forecasted data. This step helps in decision making of improving the business plan, inventory management and pricing strategies by understanding the future demand.

Challenges: The plotting was initially done for 12 weeks but the prediction pattern was not so clear so different combinations of weeks were applied and finalized with 15 weeks which gave a better forecasting view.

Project Report:

Time Spent: 3 days

The base of the project report was prepared by all the team members and then each member added their part of contributions to the project along with flow charts of the process followed, their results and findings.

Challenges: Ensuring clarity and simplicity and consistency between all the members of the team. There were many iterations done in the report to make it consistent.

README file:

Time Spent: 1 day

The Readme file was created by following the sequence of steps done. It has information about how to run code, project overview, libraries required and step by step execution.