Capstone Project Report:-

**Sales Forecasting for Retail Stores**

Table of Contents

1. Problem Statement

2. Project Objective

3. Data Description

4. Data Pre-processing Steps and Inspiration

5. Choosing the Algorithm for the Project

6. Motivation and Reasons For Choosing the Algorithm

7. Assumptions

8. Model Evaluation and Techniques

9. Inferences from the Same

10. Future Possibilities of the Project

11. Conclusion

12. References

Problem Statement:

The retail store with multiple outlets across the country is facing challenges in managing inventory efficiently to meet the demand with respect to supply. The objective of this project is to leverage data science and predictive modelling to provide useful insights to improve inventory management and forecast sales for the next 12 weeks.

Project Objective

The primary aim of this project is to leverage data-driven strategies to improve retail management. This objective can be summarized as follows:

* **Data-Driven Insights**: Our goal is to extract valuable insights from the data that will inform and enhance inventory management practices. These insights will be based on a comprehensive analysis of the dataset, shedding light on patterns and trends that can be used to optimize inventory.
* **Sales Forecasting**: We seek to create robust forecasting models capable of predicting sales for the next 12 weeks with precision. This forecasting capability will enable retail stores to proactively manage their inventory, align supply with demand, and improve overall operational efficiency.

Data Description

The dataset used in this project is "walmart.csv," which contains the following columns:

|  |  |
| --- | --- |
| **Feature Name** | **Description** |
| Store | Store number |
| Date | Week of Sales |
| Weekly\_Sales | Sales for the given store in that week |
| Holiday\_Flag | If it is a holiday week |
| Temperature | Temperature on the day of the sale |
| Fuel\_Price | Cost of the fuel in the region |
| CPI | Consumer Price Index |
| Unemployment | Unemployment Rate |

The dataset comprises 6435 rows and 8 columns.

Data Pre-processing Steps And Inspiration:

The pre-processing of the data included the following steps:

* **Data Loading**: The dataset, sourced from the "walmart.csv" file, was loaded into the analytical environment using the pandas library. Dataframes were created to store data for the selected 5 stores (Store 10, 22, 34, 17, and 14). Each row of data was assigned date as the index to facilitate time series analysis. We loaded the date columns as datetime format for our use case
* **Exploratory Data Analysis (EDA):** A critical component of data pre-processing, EDA was undertaken for selected stores to gain a deeper understanding of the data characteristics unique to each. This process enabled the identification of pertinent insights and trends.
* **Data Cleansing**: Missing values were carefully checked in whole data and addressed to ensure the dataset's integrity. Additionally, the presence of duplicate entries was also examined. We found 0 null and duplicate values in our data

Choosing the Algorithm for the Project

1. **ARIMA (AutoRegressive Integrated Moving Average):**

ARIMA, or AutoRegressive Integrated Moving Average, is a widely-used time series forecasting model that combines autoregressive (AR) and moving average (MA) components.

Here's a brief explanation of its key elements:

* **AutoRegressive (AR)**: This component accounts for the relationship between the current observation and past observations, capturing auto-correlation in the time series data.
* **Integrated (I)**: The "I" indicates differencing, which transforms the data to make it stationary. Stationarity is essential for time series analysis.
* **Moving Average (MA)**: The MA component considers the relationship between the current observation and past white noise (random error) terms.

1. **SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors):**

SARIMAX is an extension of ARIMA that includes seasonal components and can accommodate exogenous variables (external factors).

Here's why SARIMAX was chosen for this project:

* **Seasonal Components**: SARIMAX is designed to capture both non-seasonal and seasonal patterns in the data. It accounts for seasonal AutoRegressive (SAR) and Moving Average (SMA) components in addition to the standard ARIMA components. This is crucial when dealing with data subject to seasonal fluctuations, such as retail sales.
* **Exogenous Variables**: SARIMAX allows for the inclusion of exogenous (external) variables in the model. In our case, external factors like temperature, fuel prices, CPI, and unemployment rates can influence sales. By incorporating these factors, SARIMAX makes the model more comprehensive and capable of producing accurate forecasts that consider the impact of these external variables.

**Using Auto ARIMA to Determine Model Order:**

Auto ARIMA is an automated tool that systematically explores different combinations of ARIMA and SARIMA model orders to find the most suitable order for a given time series.

The key steps involved are as follows:

* **Grid Search**: Auto ARIMA conducts a grid search over a range of potential order parameters (p, d, q, P, D, Q, s) for both ARIMA and SARIMA models. These parameters include autoregressive (p and P), differencing (d and D), moving average (q and Q), and seasonal (s) components.
* **Model Evaluation**: For each set of parameters, the algorithm fits a corresponding model to the data and evaluates its performance using a specified metric (e.g., AIC or BIC). It aims to minimize this metric to identify the optimal model.
* **Order Selection**: Auto ARIMA selects the model order with the lowest AIC or BIC, indicating the best trade-off between model complexity and goodness of fit.

SARIMAX provides a comprehensive framework to capture the complexity of retail sales data. By considering both the inherent time series dynamics and the influence of external factors, it results in more robust and accurate sales forecasts.

In summary, the use of SARIMAX, determined through Auto ARIMA, is well-suited to handle the seasonality and exogenous variables in the retail sales dataset, making it a powerful tool for forecasting sales in this context

Assumptions

* **Stationarity**: We assumed that the time series data, specifically the weekly sales, is stationary or can be made stationary through differencing. Stationarity is a fundamental requirement for time series analysis and forecasting.
* **No Outliers**: It is assumed that the dataset does not contain significant outliers that could unduly influence the results. Outliers, if present, could impact the accuracy of the models and insights.
* **No Structural Breaks**: We assume that there are no significant structural breaks or changes in the data generation process during the observed time period. Structural breaks can introduce complexities in time series analysis.
* **Linear Model Adequacy**: The project assumes that the selected ARIMA and SARIMAX models are adequate for capturing the underlying patterns in the sales data. More complex modelling approaches were not explored.
* **Stability of Model Parameters**: The project assumes that the parameters of the ARIMA and SARIMAX models do not change significantly over time. Model parameters that vary over time can indicate model instability.

Model Evaluation and Technique

The following techniques and steps were involved in the evaluation of the model

1. We first checked for stationarity of data for each store.
2. Auto ARIMA models are used to find order for ARIMA and SARIMAX for each store's sales forecasting
3. Seasonal decompose helped to Identify the seasonal order to be passed In our model
4. We used that order from auto-arima and seasonal decompose to build two models using ARIMA and SARIMAX and compared the forecast through graphs.

The evaluation report suggests the following:

1. We observed that the SARIMAX model had better forecasting than the ARIMA
2. It was due to the fact that our data Is seasonal and SARIMAX deals with seasonal data better

Inferences from the Project

1. Sales Trends across Stores:

- Different stores exhibit varying sales patterns. Some stores show consistent sales, while others display more fluctuation over time.

2. Seasonality Influence:

- Seasonality plays a significant role in retail sales. Holiday weeks tend to have higher sales, and the temperature can impact sales, especially in the case of seasonal products.

3. Correlation:

-From our data we observed that: (i) Fuel Prices show highly positive correlation with CPI. (ii) CPI shows highly negative correlation with Unemployment rate.

4. Accuracy Differences:

- The forecasting accuracy of the models varies among different stores. Some stores show better forecasting performance than others, possibly due to the influence of exogenous variables.

5. ARIMA for Simplicity:

- ARIMA models are suitable for capturing the underlying time series patterns when exogenous variables have less impact on sales. They are computationally less intensive and provide good forecasts for certain stores.

6. Store-Specific Recommendations:

- The analysis can inform store-specific strategies. Stores with strong seasonality might need to adjust inventory levels before the holiday season. Stores influenced by external factors like temperature can plan promotions accordingly.

Future Possibilities

1. Advanced Forecasting Techniques: Consider exploring more advanced time series forecasting techniques beyond ARIMA and SARIMA models. This could involve machine learning models such as LSTM (Long Short-Term Memory) networks for more accurate and dynamic predictions.

2. Ensemble Models: Implement ensemble forecasting models, which combine the predictions of multiple models to improve accuracy. Techniques like model stacking or blending can be explored.

3. Hyperparameter Tuning: Optimizing the hyperparameters of the forecasting models can achieve better performance. Tools like Grid Search or Random Search can be used to fine-tune model parameters.

5. Spatial Analysis: If applicable, consider analysing the spatial aspects of sales. Explore whether the location of the store, proximity to competitors, or demographics of the area affect sales patterns.

6. Demand Planning: Extend the project to include demand planning and inventory optimization. This could involve developing a system for automatically adjusting inventory levels based on sales forecasts.

These future possibilities are designed to enhance the project's capabilities, ultimately leading to more informed decision-making and operational excellence for the retail store. Implementation of these strategies will be a crucial step towards achieving advanced and sophisticated sales forecasting and inventory management.

Conclusion

We were successfully able to implement time series forecasting model ARIMA and SARIMAX on our selected stores data and forecast the 12 week sales

References

* Kaggle (https://www.kaggle.com/): Kaggle is a prominent platform that offers various datasets and kernels related to time series analysis and retail sales. It serves as a valuable resource for accessing real-world data and insights.
* UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php): The UCI Machine Learning Repository provides a wide range of datasets that can be utilized for time series analysis and retail sales forecasting.
* Analytics Vidhya (https://www.analyticsvidhya.com/): Analytics Vidhya is a prominent platform for data science and machine learning enthusiasts. It provides articles, tutorials, and datasets related to time series analysis and predictive modelling.
* Medium (https://medium.com/): Medium is a publishing platform where many data scientists and analysts share insightful articles, case studies, and tutorials on various data science topics, including time series forecasting in retail.
* Towards Data Science (https://towardsdatascience.com): "Towards Data Science" is a well-regarded publication on Medium that offers a wealth of articles and insights in the field of data science, including time series analysis and machine learning.