Project Report: Sales Forecasting Analysis

1. Introduction

In this project, our objective is to forecast sales for the next two years based on historical sales data. We aim to utilize time series analysis techniques to understand sales trends, explore seasonality, and develop a forecasting model. The dataset used for analysis contains information about sales transactions, including order dates, sales amounts, and other relevant attributes.

2. Data Cleaning and Manipulation

We began by inspecting and cleaning the dataset to ensure data integrity and consistency. The following steps were performed:

- Renamed columns to lowercase and replaced spaces with underscores for consistency.
- Checked for missing values and duplicates, and handled them accordingly. The dataset had 11 missing values in the 'postal_code' column, which were addressed appropriately.
- Converted date columns to the appropriate datetime format to facilitate time series analysis.
- Resampled the data to aggregate sales on a monthly basis for further analysis.

3. Exploratory Data Analysis (EDA)

EDA was conducted to gain insights into sales trends over time. Key observations from the analysis include:

- Daily sales trend: Sales were highest on the 31st and 30th of the month, followed by a significant drop on other days.
- Monthly sales trend: November had the highest sales followed by December and September. February had the lowest sales.
- Yearly sales trend: Sales increased from 2015 to 2018, with the highest sales recorded in 2018.
- We also performed seasonal decomposition analysis to identify underlying patterns in the data.

4. Checking Stationarity

To ensure the time series data's stationarity, we performed an Augmented Dickey-Fuller (ADF) test. The results indicated stationarity in the data, allowing us to proceed with modeling.

5. Feature Engineering

Relevant features such as day, month, and year were extracted from the order date to further understand the data and aid in modeling.

6. Model Building

We utilized the Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors (SARIMAX) model for forecasting. The model was trained on the historical data up to a certain point and then used to predict future sales.

7. Model Evaluation

The forecasting model was evaluated using metrics such as Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Mean Squared Error (MSE) on the testing dataset.

- The Mean Absolute Error (MAE) of the model is 16459.69.
- The Mean Absolute Percentage Error (MAPE) of the model is 36.77%.
- The Mean Squared Error (MSE) of the model is 377195390.05.

These evaluation metrics provide insights into the model's accuracy and performance.

8. Results and Conclusion

The forecasting model demonstrated promising results in predicting future sales. However, there may be areas for improvement, such as incorporating additional features or exploring alternative modeling techniques. Overall, the analysis provides valuable insights for sales forecasting and can aid in decision-making processes for business planning and strategy.

9. Future Work

Potential areas for future work include:

- Refinement of the forecasting model to improve accuracy.
- Incorporation of external factors such as economic indicators or marketing campaigns for enhanced forecasting.
- Continuous monitoring and updating of the model to adapt to changing sales patterns and dynamics.