Sales Forecasting

Machine Learning and Related Applications

COBScDSFT231F – 006

BSc (Hons) in Data Science

National Institute of Business Management

04TH NOVEMBER 2023

Introduction

Sales forecasting is the process of estimating future sales for a business or organization. Sales forecasting in the retail industry is essential for optimizing inventory, enhancing supply chain efficiency, and managing labor costs. It involves using historical sales data and market trends to make predictions about the quantity and revenue of products that will be sold in the future. Accurate sales forecasting enables organizations to allocate resources efficiently and respond to changing market conditions, leading to improved performance and profitability.

1.1 Statement of purpose

The primary objective of this report is to develop a reliable sales forecasting model for the top 25 fast moving items in various supermarket departments, excluding liquor/tobacco and miscellaneous. With precise predictions, retailers can ensure the right stock levels, minimize overstock and understock. Accurate sales forecasts play an important role in enhancing the overall operational efficiency.

1.2 Document Structure

The structure will be organized into distinct sections. These sections will include data analysis, data preprocessing, model development, hyperparameter tuning, evaluation metrics and the best practices for model management. This structure aims to provide a comprehensive presentation of the approach for forecasting sales of specific items and departments.

Methodology

The programming language that was used for this case study was Python. A few of the libraries embedded within Python were used for data analysis, visualization, and machine learning. Libraries like Pandas and NumPy were used for data manipulation, Matplotlib and Seaborn aided in data visualization. The scikit-learn library enabled the tools for the machine learning aspect of this case study, including data preprocessing, model building (XGBoost), and hyperparameter tuning (GridSearchCV). The utilization of these libraries collectively facilitated the data analysis and model prediction.

Implementation

3.1 Data sources

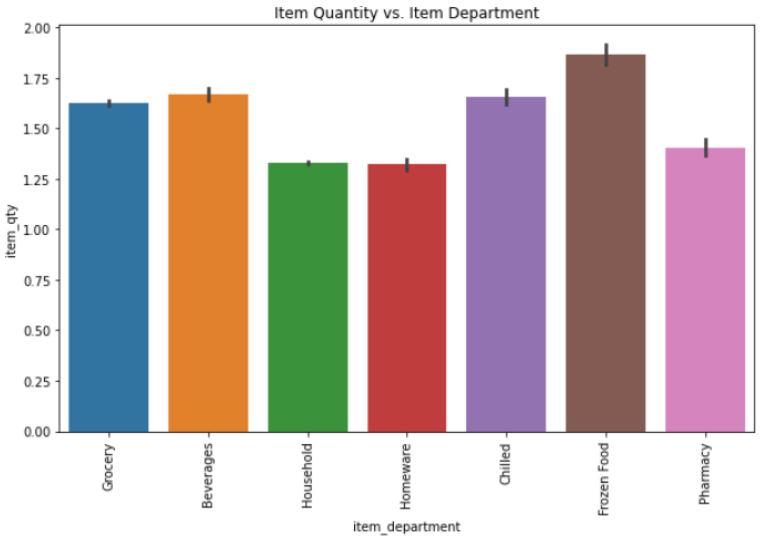
The two datasets provided contain the necessary records of items and transactions, where both are merged in order for the contents to be explored and analyzed for comprehensive research.

3.1.1 Data preparation

An essential stage in the preprocessing of data is data cleaning. This step is vital to guarantee the quality and reliability of a dataset. Both datasets were rid of any present null values and the 'Liquor/Tobacco' and 'Miscellaneous' columns in the items dataset were dropped, before they were ultimately merged into a singular dataframe.

Conducting descriptive analysis is important as it provides an initial understanding of the dataset. Statistical analysis and data visualizations were carried out in order to gain a better understanding of the data available.

	item_qty
count	349633.000000
mean	1.574346
std	2.629919
min	0.022000
25%	1.000000
50%	1.000000
75%	1.104000
max	400.000000



3.1.2 Target Construction

The dataframe was then grouped by the item_code and item_department, along with the sum of the item_qty, in order to find the top 25 fast moving items from each department, with a total value count of 175. The top 25 items were then grouped hourly.

To create the target variable, 'target_qty', for every item_code that was included in the top25_hourly dataframe, the item_qty values are shifted by one hour to predict the following hour's quantity, as demonstrated below:

```
target = pd.DataFrame()
for x in top25_hourly.item_code.unique():
    dummy = top25_hourly[top25_hourly.item_code==x]
    dummy.loc[:, 'target_qty'] = dummy.item_qty.shift(-1)
    target = pd.concat([target, dummy])
target = target.fillna(0).reset_index(drop=True)
target
```

	item_code	invoice_time	item_qty	target_qty
0	898	2022-01-13 18:00:00+00:00	48.0	18.0
1	898	2022-01-13 19:00:00+00:00	18.0	3.0
2	898	2022-01-13 20:00:00+00:00	3.0	1.0
3	898	2022-01-13 23:00:00+00:00	1.0	14.0
4	898	2022-01-14 08:00:00+00:00	14.0	16.0
...	...	...	...	...
44822	1107943	2022-03-15 16:00:00+00:00	1.0	1.0
44823	1107943	2022-03-16 15:00:00+00:00	1.0	1.0
44824	1107943	2022-03-17 13:00:00+00:00	1.0	6.0
44825	1107943	2022-03-17 17:00:00+00:00	6.0	1.0
44826	1107943	2022-03-18 11:00:00+00:00	1.0	0.0

44827 rows × 4 columns

3.1.3 Master Table Creation

In order to create the master table, the hourly time intervals must be continuous for each item_code. Therefore, by creating a continuous time range and joining it with the original data, any missing hours are filled with a value of -1. This is done to distinguish between null values and actual data values, since the use of 0 would imply that the value is intentionally zero. This step ensures consistent time intervals for further analysis preventing gaps in the data that could affect the model accuracy.

3.1.4 Features Construction

Feature engineering is crucial to help a model predict, since the more synthetic features created, the more information it provides to the model. Different features can capture patterns, trends and it in turn improves the model's ability to understand and make accurate predictions. The features that were created include:

-average sales per department

-hour

-average hourly sales

-maximum sales per department

-minimum sales per department

-average sales on weekdays

-average sales on weekends

3.2 Choosing the Algorithm

After thorough analysis to select an appropriate algorithm, the XGBoost regression algorithm was chosen due to its ability to process complex, nonlinear relationships. The presence of both item and time related features were considered in suit of this decision.

3.2.1 Model Development

The selected XGBoost forecasting model, was trained on the training data using a pipeline that included data preprocessing steps such as imputation, and scaling, for numerical features, and one-hot encoding for categorical features. The models were fitted to the training data, while the inbuilt hyperparameters were used to establish the basic model.

3.2.2 Hyperparameter Optimization

Hyperparameter tuning was conducted to optimize the model performance. With the use GridSearchCV and cross validation, the best hyperparameters were found, which helps enhance the predictive accuracy and reduce overfitting the model.

3.2.3 Feature Importance

The feature importance analysis was carried out in order to identify which variables had the most influence on the model's predictions. This analysis provides a better understanding of which item or time related features play a more significant role in forecasting the sales.

```
The feature importance for invoice_time is: 0.002483460819348693
The feature importance for item_code is: 0.5768918991088867
The feature importance for item_qty is: 0.0018109522061422467
The feature importance for item_sub_segment is: 0.0017453667242079973
The feature importance for item_segment is: 0.027061978355050087
The feature importance for item_category is: 0.000321648723911494
The feature importance for item_sub_department is: 0.0
The feature importance for item_department is: 0.0
The feature importance for avg_dept_sales is: 0.0
The feature importance for hour is: 0.00015623742365278304
The feature importance for hourly_sales is: 0.0010042899521067739
The feature importance for dept_max_sales is: 0.008271722123026848
The feature importance for dept_min_sales is: 0.0
The feature importance for avg_weekday_sales is: 0.0
The feature importance for avg_weekend_sales is: 0.0006977611337788403
```

3.2.4 Model Accuracy Measures

The evaluation metrics that were employed to assess the model's performance were the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the R-squared value (R^2).

```
RMSE: 5.619903018664598
MAE: 0.3993305465577739
R-squared: 0.4070646090051664
```

The RMSE, with a value of roughly 5.62, represents the average magnitude of predicted errors. If the RMSE value is lower, it indicates the model's capability to make accurate predictions.
MAE, with a value of approximately 0.39, measures the average absolute difference between the predicted and actual values. A low MAE implies better predictive accuracy.
With the R^2 value being around 0.41, it signifies the amount of variance explained by the model, meaning it indicates a slight ability to capture the sales variance. These metrics collectively provide insight on the model's performance.

3.3 Risks and Assumptions

Transparency is required when dealing with risks, such as the potential impact of external factors on sales. The assumptions include data being stationary and that past sales patterns are meant to continue or stay consistent. It is important to remain unbiased during the entire process of analysis, forecasting in order to make reliable predictions.

3.4 Pipelines

In this particular forecasting problem, two separate pipelines streamline the process. The first pipeline transforms and combines the data sources, generating a master table with engineered synthetic features. The second pipeline fits the XGBoost regression model. This ensures a systematic and convenient workflow, which simplifies the data preparation and modeling, which in turn enhances the model forecasting process.

Findings and Conclusions

The findings suggest that the XGBoost model, provides a reasonably good predictive capability. The model's performance represented by RMSE and MAE, demonstrate the ability to provide accurate sales predictions. The R-squared offers insights for inventory management and demand planning. However, the R-squared value, only being a value of 0.41, suggests that the model is not quite strong enough in terms of explaining variance. To remedy this, more synthetic features should be engineered and a different model such as a time series model could be used to compare and contrast the performances.