Title

Sales forecasting for a supermarket company

Abstract

Sales forecasting helps predict the revenue and product sales of a sales unit (this could be an individual, a team, or a company) in a given amount of time. It is important for making decisions regarding cash flow, marketing, and resource management. These predictions help understand the clientele, the growth of a company, and the short- or long-term performance. This research seeks to predict sales for a supermarket company for the upcoming *year/quarter/time*. It will utilize supervised learning through the XGBoost Regressor Model. After cleaning the dataset, this model will be trained on 80% of the data and make predictions based on the remaining 20% of the data. This method is highly flexible and can handle missing data with built-in functions for optimized outcomes. Applying this method to this company's sales data will help ensure confidence in business decisions moving forward and help it understand the overall performance of the company.

Evaluation

This research will utilize eXtreme Gradient Boosting (XGBoost), which is an implementation of gradient boosted decision trees designed for speed and performance. Using this library lends tools for data cleaning, splitting data into training and testing, as well as providing a flexible environment. Mean absolute error will be used to estimate the error in the model. Along with this, several different data visualization techniques will be used. A count plot will be used to visualize the number of different types of products sold. A seaborn distplot will show the gross income over the three months and it will be used to show average customer ratings over the three months.

Datasets

The data set will consist of 17 columns and 1000 entries. The entries represent transactions across 3 branches of a supermarket over 3 months (January 2019 to March 2019). The 17 features are invoice id, the branch (either A, B, or C), the city, the customer type, the gender, the type of product, the unit price, the quantity purchased, the tax, the total including tax, the date and time purchased, the payment type, the cost of goods, the gross margin percentage, the gross income, and the customer rating. The branch, the city, customer type, gender, type of product, and payment type are categorical features. The remaining features are continuous. While all columns will be important for this sales forecast, the main columns will be the quantity at which certain product types are sold, the cost of goods, the gross income, and the customer rating.

References

https://www.sciencedirect.com/science/article/pii/S0040162520313068

https://www.hindawi.com/journals/mpe/2019/8503252/