Using Machine Learning to Predict Grocery Sales

Kamil Samara[1] and Mark Stanich[2]

[1] University of Wisconsin - Parkside, 900 Wood Road, USA
[2] University of Wisconsin - Parkside, 900 Wood Road, USA

Abstract. The purpose of this research is to construct a sales prediction model for grocery stores using linear regression. A forecasting model which can effectively predict the sales of a grocery store will help retailers stock the right quantity needed of their selling items and eventually compete better in the market. The present study uses four years' worth of point-of-sale (POS) data from 54 grocery stores from Corporación Favorita, which is a large Ecuadorian-based grocery retailer to construct a sales prediction model that can predict the items demand for the upcoming year. XGboost (extreme gradient boosting) was used to build the prediction model. The constructed model can predict upcoming sales with an accuracy of 83%.

Keywords: Machine Learning, XGboost, Prediction, Sales, Forecasting.

1 Introduction

1.1 Project background

A traditional retailer had to deal with stocking of inventory based on estimates of his daily or weekly product sales to make available the right product at the right time for his customers. With forecasting techniques, stores can efficiently and accurately predict store-product sales to replenish inventory with minimum capital requirement.

Current subjective forecasting methods for retail have little data to back them up and are unlikely to be automated. The problem becomes even more complex as retailers add new locations with unique needs, new products, ever-transitioning seasonal tastes, and unpredictable product marketing [1].

Corporación Favorita, is a large Ecuadorian-based grocery retailer [1]. The purpose of this project is to build a model that more accurately predicts the unit sales for thousands of items sold at different Favorita stores with training dataset of dates, store, and item information, promotions, and unit sales. The project idea and data were provided through a Kaggle competition [2].

The problem is important to the company because it will help better set them up for their upcoming years. This will provide them with invaluable information to help guide the process of acquiring goods for sales and keeping the company on track to make a profit. This is applicable to any company that sells a product. It will give them information needed to better set their course for the following years.

1.2 Problem description

The main objective of this study is Store Sales - Time Series Forecasting. Kaggle provides the needed data as csv files, which include: train.csv, stores.csv, test.csv, oil.csv, transactions.csv, and holiday_events.csv. The data comes from 54 stores and spans for a period of four years from 2013 till 2017.

The task that is needed to be completed is a prediction of a grocers' sales per id number. Each id number correlates to a different item within the grocers' inventory and the prediction will predict a value for each of these sales variables. Included with the task is multiple sets of data. These sets are as listed, holidays events, oil, stores, train, transactions, test, and sample submission.

The remaining of the paper is organized as follows: Section 2 is related work; Section 3 presents the work done is this study and Section 4 concludes the work.

2 Related Work

A very related work using XGboost to predict sales, is the work of Zhaoweijie and Chenliang in [3]. XGboost was used to build a prediction model, and forecasts the sales volume of Rossmann's (Rossmann is one of the largest daily chemical supermarkets in Europe) 1115 stores in next 48 days. They achieved an accuracy of 89.07%.

In a similar study [4], the authors used deep learning to construct a sales prediction model for retail stores. In their study they used three years' worth of point-of-sale (POS) data from a retail store to construct a sales prediction model that, given the sales of a particular day, predicts the changes in sales on the following day. They were able to achieve a sale forecasting accuracy rate of 86%.

In [5], Yang and Sutrisno they used regression analysis to provide a better demand forecasting in a relatively short period for a bakery franchise business in China. More than 10 million of point-of-sales (POS) data were collected and analyzed. This approach provides a systematic forecasting model to reduce the waste of the supply chain system.

The researchers in [6], used XGBoost to build a forecasting model which can effectively predict the sales of a retail store. The proposed model predicted sales for 10 stores with 50 products. Using R2 as their performance metric, the proposed model achieved a value of 0.76.

Another example of using XGBoost to predict sales can be seen in [7]. In this work, XGBoost was used to develop a predictive model for forecasting sales of Big Mart. The data were obtained from Kaggle.com with a training data of size 8000 examples and a test set of size 5000 examples. In their results the authors concluded that Xgboost regression gives the better prediction with respect to accuracy than the Linear and polynomial regression approaches.

3 Proposed Work

3.1 Background

As mentioned earlier, the main objective of this work is to build a sales predictive model to accurately forecasts product sales for a chain of grocery stores owned by Corporación Favorita, an Ecuadorian-based retailer. Corporación Favorita owns 54 grocery stores. Corporación Favorita posted their problem as a challenge on Kaggle [2].

3.2 Data sets

The task that is needed to be completed is a prediction of a grocers' sales per id number. Each id number correlates to a different item within the grocers' inventory and the prediction will predict a value for each of these sales variables. Included with the task is multiple sets of data. These sets are as listed, holidays events, oil, stores, train, transactions, test, and sample submission.

Each set has different data within. The first set, holidays events, has date, type, locale, locale_name, description, and transferred. The oil set has date and dcoilwtico. The stores set has store_nbr, city, state, type, and cluster. The train data has id, date, store_nbr, family, sales, and onpromotion. The transactions set has date, store_nbr, and transactions. The test set has id, date, store_nbr, family, and onpromotion. Lastly, the sample_submission set has id and sales.

3.3 XGBoost

XGBoost is a powerful supervised technique used for building regression models. XGBoost uses loss function plus regularization to build its objective function. Loss functions provide the difference between the actual value and the predicted one, which is the distance between the actual value and the model results. XGBoost belongs to the gradient boosting family, which means it builds base learners then uses ensemble learning to combine their week predictions into a one final strong prediction [8].

3.4 Exploring the Data

The task at hand is to clean the data, explore the data, model the data, and create a forecasting. The approach was just that. We took the data provided and cleaned any unnecessary data within the sets. We then went through that data to find out more about what important features would be needed to create a solid model. The data was then modeled for an XGboost (extreme gradient boosting). This gave the information needed for forecasting. It was then saved into the sample data frame and submitted
In this stage, the data was explored to try and identify the features that impact sales the most. Later, these features will be used to train the XGBoost model.

Exploring the Month Feature. The first feature that was explored was the month. To perform this exploratory analysis a bar chart was used to show the relation between average sales and month. The results are shown in figure 1.

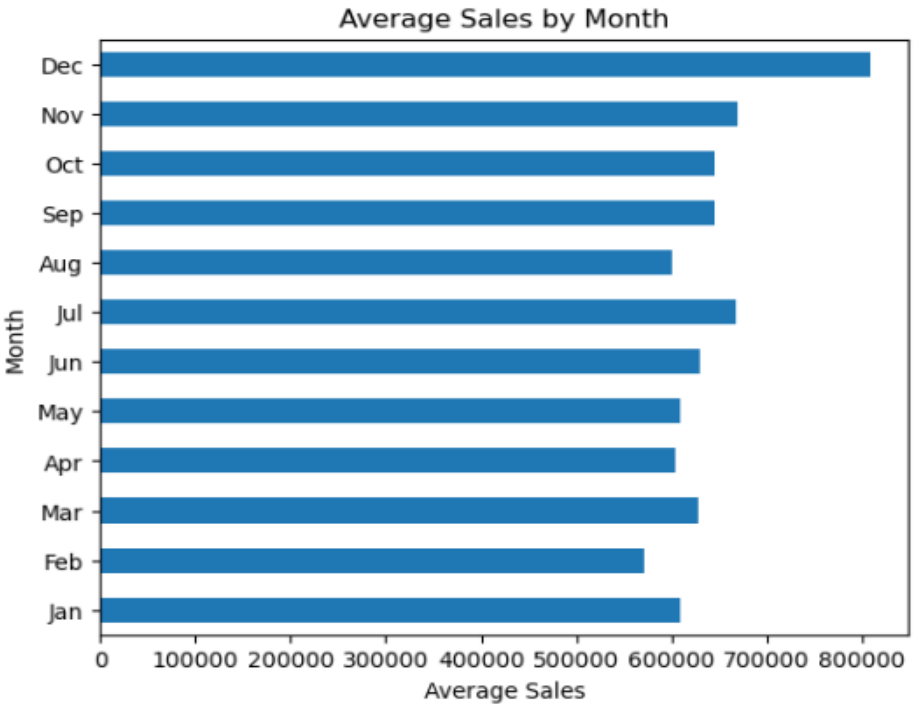

Fig. 1. Average Sales by Month

As shown in figure 1, December has the highest sales compared to the other year months. Based on the differences is sales by month, the month feature could be an effective factor in making stocking predictions.

Exploring the Store Feature. The store number feature was explored to identify which stores have the highest sales. Again, a bar chart was used plotting store against the average daily sales. The results are shown in figure 2.

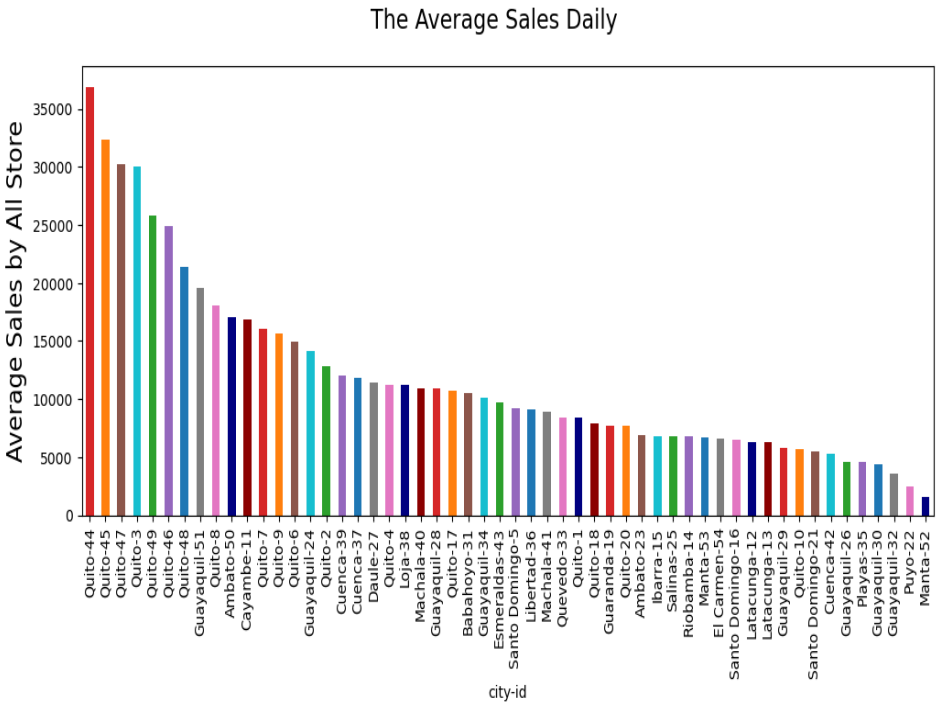

Fig. 2. Average Sales by Store

As shown in figure 2, Some stores have higher sales than others. Identifying busier stores will help in predicting a more accurate stocking.

Exploring the Product Family Feature. The sales items are grouped into families by the retailer. Here we are trying to identify which products' families get the highest demand. The product family was plotted with the average sale for each family in 2017. The results are shown in figure 3.

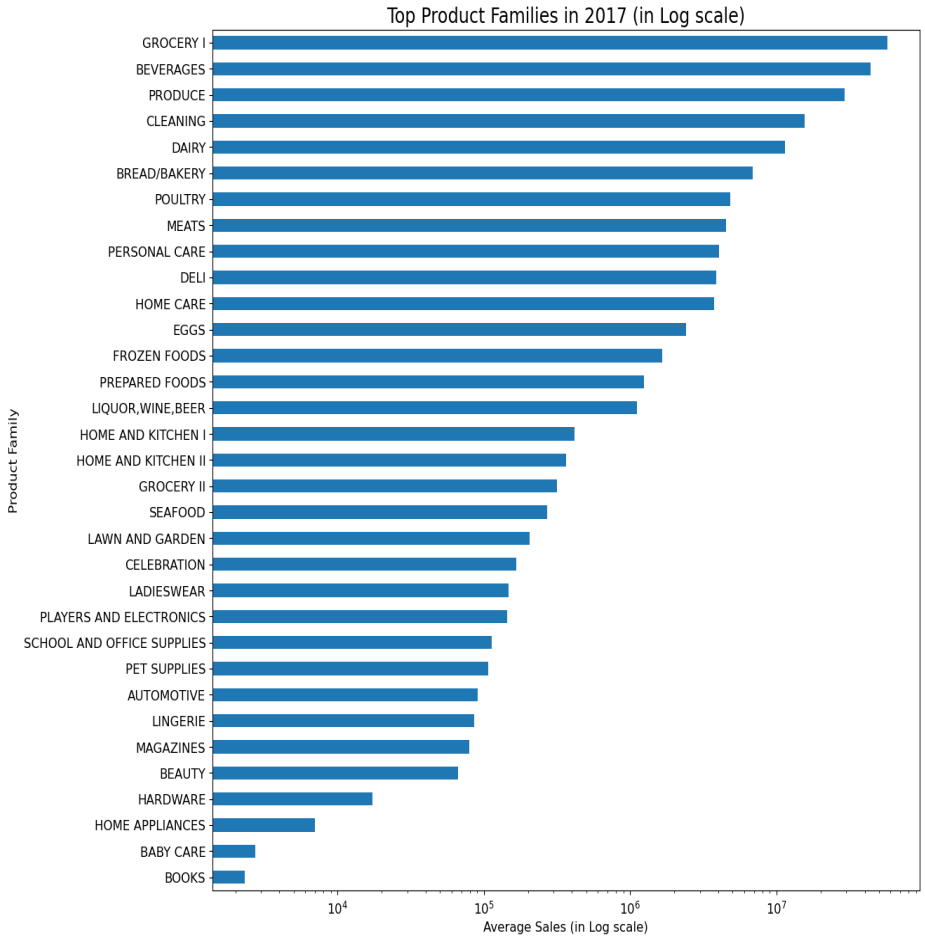

Fig. 3. Average Sales by Product Family

As shown in figure 3, grocery 1 constitutes a third of the sales with beverage doing about a fifth and produce doing a tenth of total sales with some families barely contributing.

Exploring the Day of the Week Feature. The sales were grouped by the day of the week then the average sales for each day was calculated. The results are shown in figure 4.

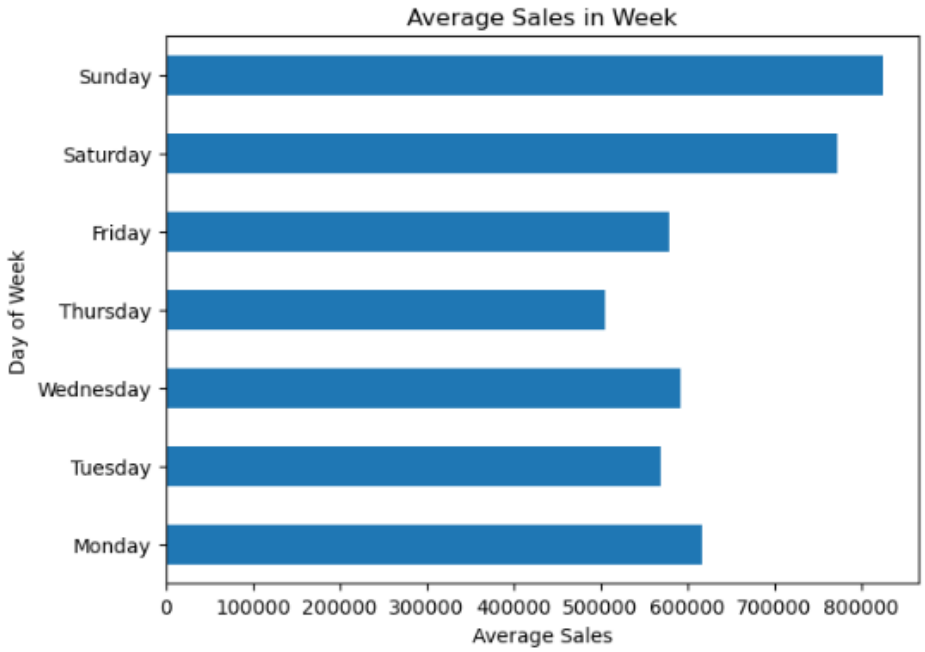

Fig. 4. Average Sales by Weekday

As shown in figure 4, the weekend days (Saturday and Sunday) are the busiest days of the whole week. People tend to do most of their shopping when they are off work.

A similar analysis was performed on the other features, like onpromotion (reflects the number of promotions), oil prices, year, city, state. After preforming the exploratory analysis, the following features were found to affect sales the most: store, day of the week, month, promotions, and product family.

3.5 Training & Prediction

After cleaning and exploring the data, the XGBoost regression model was built and trained using features that showed a strong correlation with sales.

The train.csv and test.csv files provided by Corporación Favorita was used to train and test the XGBoost model. The r2_score metric was used to test the accuracy of the trained model. The model scored 0.83623 in the r2_score test.

4 Conclusions

In this study, we employed machine learning through the use of XGBoost regression in effort to predict products' sales for an Ecuadorian-based retailer. The regression

model was implemented on a data sets provided by the retail company that consist of four years' worth of point-of-sale (POS) data from 54 grocery stores.

Using r2-score the regression results predicted products' sales for the next year with a .83 accuracy.

These findings should help retail stores to stock the right quantities of products which will result in better competition and higher customer satisfaction.

References

1. Corporación Favorita. (2022, August 25). www.corporacionfavorita.com/en/
2. https://www.kaggle.com/competitions/store-sales-time-series-forecasting/
3. Zhaoweijie, Chenliang and Hujiangmin, "Forecast Rossmann Store Sales Base on Xgboost Model," 2020 2nd International Conference on Economic Management and Model Engineering (ICEMME), 2020, pp. 521-525, doi: 10.1109/ICEMME51517.2020.00110.
4. Y. Kaneko and K. Yada, "A Deep Learning Approach for the Prediction of Retail Store Sales," 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), 2016, pp. 531-537, doi: 10.1109/ICDMW.2016.0082.
5. C. -L. Yang and H. Sutrisno, "Short-Term Sales Forecast of Perishable Goods for Franchise Business," 2018 10th International Conference on Knowledge and Smart Technology (KST), 2018, pp. 101-105, doi: 10.1109/KST.2018.8426091.
6. M. Spuritha, C. S. Kashyap, T. R. Nambiar, D. R. Kiran, N. S. Rao and G. P. Reddy, "Quotidian Sales Forecasting using Machine Learning," 2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), 2021, pp. 1-6, doi: 10.1109/ICSES52305.2021.9633975.
7. R. P and S. M, "Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 1416-1421, doi: 10.1109/ICICCS51141.2021.9432109.
8. XGBoost Documentation. (2022, November 12). xgboost.readthedocs.io/en/stable/