Sales Forecasting with XGBoost

Jenna Coffman
Computer Science Department
San Diego State University
San Diego, United States
jcoffman1154@sdsu.edu

***Abstract*—Sales forecasting helps predict the revenue and product sales of a sales unit (this could be an individual, a team, or a company). It is important for making decisions regarding cash flow, marketing, and resource management. Sales forecasting coupled with an exploratory data analysis helps the company understand its clientele, the growth of the company, and the short- or long-term performance. This research predicts the most important features for Rossmann retail stores, determines the best store model and best assortment of products to optimize sales. This research utilizes XGBoost regressor model and data visualization to create a well rounded sales forecast. It was found that the day is the most important feature for this company. Along with this store model b and product assortment b produce the most sales. Applying these methods to the Rossmann store dataset will help ensure confidence in business decisions moving forward and help the company understand its overall performance.**

***Index Terms*—Sales Forecast, XGBoost, Exploratory Data Analysis**

I. Introduction

In retail one of the most important indicators of success for a business are the number of sales, customers, and revenue. These factors can determine whether the business needs to make changes, whether that be its model, the assortment of products, the promotions that are offered, or even the days the business is open. Through different techniques in data science, companies are able to better understand their clientele, the growth of their company, and their short- or long-term performance. This research utilizes exploratory data analysis (EDA) and eXtreme Gradient Boosting (XGBoost) in order to form a sales forecast for Rossmann retail stores. For this company, the EDA will seek to determine the best existing store model as well as the most profitable assortment of products. Along with this, the XGBoost model will help predict the most important features for their company which will help Rossmann stores make decisions in the future.

The structure of this paper will be as follows: section 2 will present the approaches used for the analysis and predictions for the Rossmann Stores, section 3 will discuss the findings from the exploratory data analysis as well as describe the results from the XGBoost, section 4 will discuss works related to sales forecasting and other approaches used for this task, section 5 will present key findings for the analysis.

II. Approach

This section will present the approaches used to perform the EDA and build the XGBoost model. Before any analysis or predictions could take place the data was cleaned. Stores missing information pertaining to being open will be assumed open. Days that stores are closed or that have no sales were dropped from the dataset.

A. Exploratory Data Analysis

This research will use Exploratory Data Analysis to find trends that will help the Rossmann company be more successful. In an EDA, data visualization is one of the most important tools. It helps both the researcher and the client understand what features are affecting the company and gives insight into how to conduct the business going forward. In this paper, I will use data visualization to show the most effective store model and the best assortment of products a store carries. To ensure the most effective graph is being used, different graph styles were tested, but bar plots were found to be the best indicator of these trends.

B. XGBoost Regression Model

The XGBoost model is a regularizing gradient boost framework. It is designed to be flexible, portable, and highly efficient[1]. To show efficiency the time it takes to build the model and predict importance will be shown. To determine the accuracy of the model, the root-mean-square percentage error (RMSPE) will be calculated. This measures the distance between the test data and the predicted data. A sign of a more accurate model is an RMSPE close to 0[2]. RMSPE will be used to determine the accuracy of the XGBoost model used in this research. The model will be trained on 90% of the data and tested on 10%. Having a larger training dataset helps improve the RMSPE. In this research, the XGBoost regression model will be used to predict the most important features for the Rossmann company.

III. Evaluation

A. Data Description

1) Store - a unique ID that is associated with each store in the dataset.
2) Sales - the number of sales at a store on a given day.
3) Customers - the number of customers who visit the store on a given day.
4) Open - specifies whether the store was open(0) or closed(1).
5) StateHoliday - specifies a state holiday (a = public holiday, b = Easter holiday, c = Christmas, 0 = None).
6) SchoolHoliday - specifies if school closures affected the (Store, Date).
7) StoreType - indicates the store model (a, b, c, d).
8) Assortment - indicates the assortment of products sold at a store (a = basic, b = extra, c = extended).
9) CompetitionDistance - distance to the nearest competitor store, in meters.
10) CompetitionOpenSince[Month/Year] - indicates the month and year the competition store opened.
11) Promo - specifies if a store was having a promotion that day (1 = promo, 0 = no promo).
12) Promo2 - indicates a continuous promotion at a store (1 = promo, 0 = no promo).
13) Promo2Since[Year/Week] - indicates the year and week Promo2 started.
14) PromoInterval - indicates the months Promo2 is run.
15) DayOfWeek - indicates the day of the week, starting on Monday.

The Rossmann Store data set spans from 01 Jan. 2013 to 31 July 2015 encompassing data from 1115 stores. There are 1017209 entries in this data set.

Fig. 1 shows two graphs. The first graph displays the average number of sales over the given time. The second shows the average number of customers that come into the stores over time. From the similarity in these graphs, we can infer that the number of customers and number of sales is directly related.

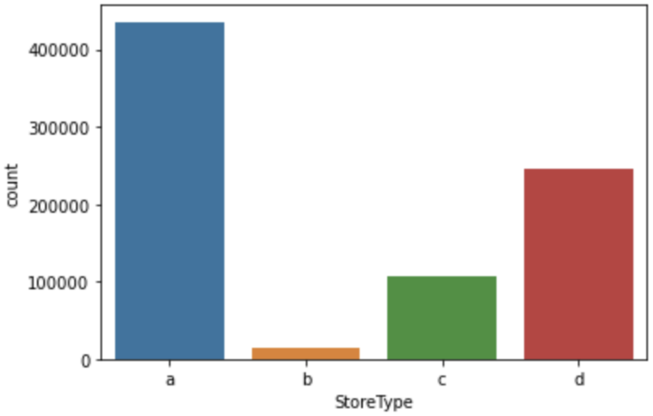

Figure 2: Store Type Count

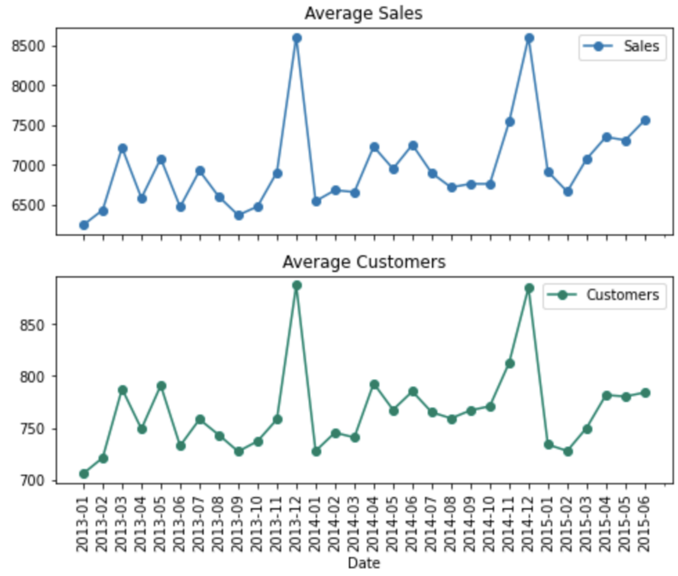

Figure 1: Average Sales and Customers

B. Exploratory Data Analysis

The goal of this evaluation is to determine the store model and the product assortment that produces the highest number of sales. This section will include the bar plots that were used to determine the outcome of this study.

Fig. 2 shows a bar plot of the number of different store models. Fig. 3 displays the number of sales for each store model. From these graphs, we can infer that store model b has the most sales but also has the least amount of locations. This trend could be attributed to model b being the most effective store model or that this model is in the most desirable locations. Due to not having information on how advantageous the locations are it would be recommended to test model b in more locations to determine if store model b will continue to produce the most amount of sales.

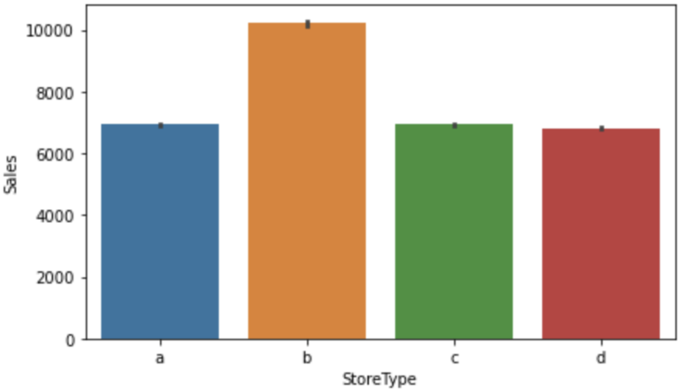

Figure 3: Store Type Sales

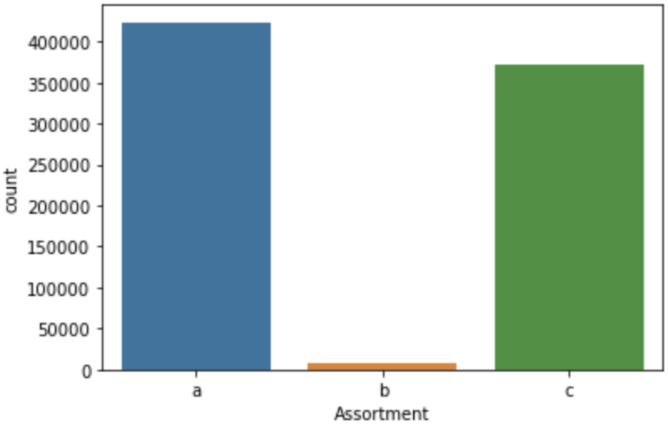

Figure 4: Assortment Count

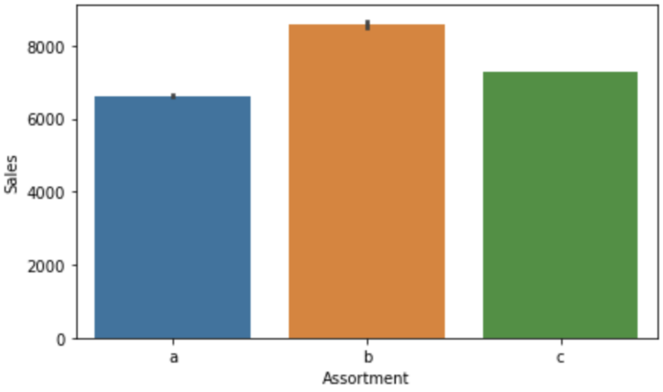

Figure 5: Assortment Sales

Fig. 4 shows a bar plot of the number of stores that carry different assortments of products. Fig. 5 displays the number of sales for each assortment type. In both Fig. 4 and Fig. 5, the assortment types are the following: a is basic, b is extra, and c is extended. From these graphs, we can infer that having an extra assortment of products procures the most amount of sales. Assortment b is held in a significantly less amount of stores than the other two options. This could be due to extenuating circumstances, such as the cost to hold these products in many locations, but it would be recommended to include this assortment type in more stores.

C. XGBoost Regression Model

This section will analyze the results of the XGBoost regression model used to predict feature importance for the Rossmann Store dataset. The analysis will include the RMSPE value to determine the accuracy of the model and the time it took to run the model to determine efficiency.

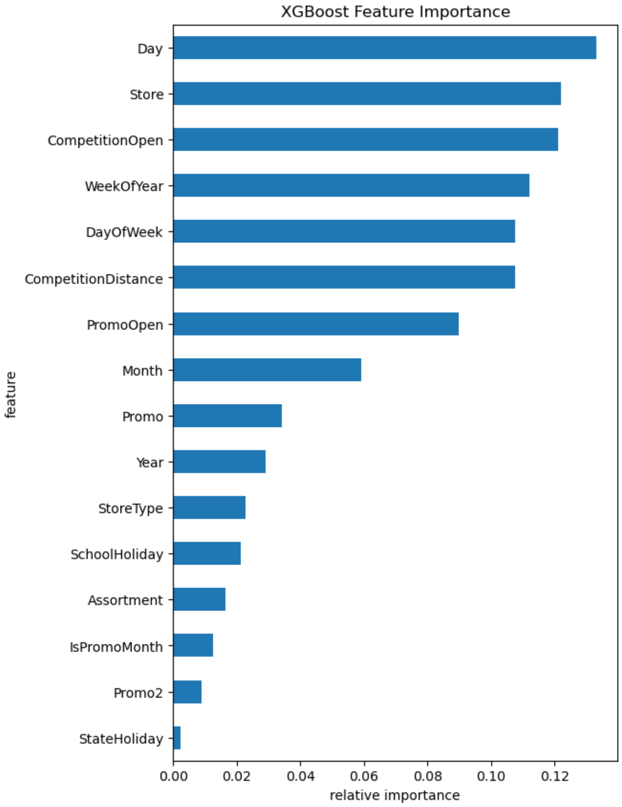

Figure 6: Feature Importance

Fig. 6 shows the output from the XGBoost regression model. From this, we can conclude that the day is the most important feature. A suggestion with this information would be opening promotions at the beginning of the week could promote an increase in sales throughout the week, increasing sales across all stores. The second most important feature is the store. If looked into more deeply, this factor would help determine which stores should stay open and which should consider being closed or changing the store model. For those that are succeeding, there would be no need to make any changes, but those that lack in sales present the opportunity to try different store models or product assortment. By experimenting with these factors, the company would be able to determine what store model and assortment most improve sales. This model has an RMSPE of 0.105. This score indicates that the model is accurate. It runs in 3m46s, showing the efficiency of the model.

IV. Related Work

1. **ForeXGBoost: passenger car sales prediction based on XGBoost [3]:** This research sought to predict the sales of passenger cars using a prediction model. They trained seven models on a small dataset to narrow down which would be most effective for this task. The candidates for their models included (1) Linear Regression model (LRM), (2) Light Gradient Boosting (LGB), (3) Logistic Regression (LR), (4) Gradient Boosting Decision Tree (GBDT), (5) Decision Tree (DT), (6) Support Vector Machine (SVM), and (7) Extreme Gradient Boosting(XGBoost). From this set, they were able to narrow down their selections to LRM, GBDT, and XGBoost using logarithmic difference square root (LDSR) to determine accuracy. These three models were then tested on a larger dataset and analyzed using LDSR, from which XGBoost was determined to be the best candidate due to prediction accuracy. XGBoost reduced the LDSR on the training set by 58% when compared to a linear regression model and by 9.4% when compared to the GBDT. The test set saw a 65% reduction in the LDSR when comparing the XGBoost model and the linear regression model and a 6.2% reduction when compared to the GBDT model. In addition to analyzing the difference in LDSR, this research also determined that the XGBoost model takes less time to run than the GBDT model and the linear regression model. This research determines that the XGBoost model is more accurate and time-efficient in predicting sales of passenger cars.

Hierarchical time series forecasting via Support Vector Regression in the European Travel Retail Industry[4]: This research creates a sales forecast for travel retail stores using bottom-up SVR, top-down SVR, and middle-out SVR models. This approach compares the three models in order to determine which has more accurate predictions. These models were also compared to an ARIMA model and Holt-Winters model. This research notes a disadvantage of the SVR model is the potential for losing information in the node branching process and as the branching process continues more information could be lost. This could result in a higher RMSPE and deem the model inaccurate. This research did not face any issues in this regard. The advantage of this approach is that during aggregation there is no information loss. To calculate accuracy, mean absolute percent error (MAPE) was used. The models were then ranked on a scale from 1(best) to 9(worst) in terms of their performance level. It was found that the bottom-up SVR had the best score, making it the most accurate on this dataset.

V. Conclusion

This research sought to determine which features are most important to Rossmann store success with an XGBoost regression model. An EDA was also performed to determine which store model and product assortment are the most effective. It was found that store model b and carrying an extra amount of products (assortment b) could increase sales, but more analysis and testing would be required to say definitively. The XGBoost model found that the day of the week is the most important feature. In the future, an analysis should be made to determine which days impact this finding. Furthering this research would help determine how other features could be paired with this information to enhance sales.

VI. Works Cited

[1] Chen, T., He, T. xgboost: eXtreme Gradient Boosting. *R package version* 0.4-2, 1(4), 1-4 (2015).

[2] Chai, T. and Draxler, R. R.: Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature, *Geosci. Model Dev.*, 7, 1247–1250 (2014).

[3] Xia, Z., Xue, S., Wu, L. et al. ForeXGBoost: passenger car sales prediction based on XGBoost. *Distrib Parallel Databases* 38, 713–738 (2020).

[4] Karmy, J. P., Maldonado, S. Hierarchical time series forecasting via Support Vector Regression in the European Travel Retail Industry. *Expert systems with applications*. 13759–73 (2019).