ROSSMANN

Stores Sales Prediction

1 Problem Identification

Rossmann operates over 3,000 drug stores in 7 European countries. The Objectives of this project is to analyze Rossmann's store profile and sales data, build a model to predict their daily sales for up to six weeks in advance.

2 The Data

The list of data and data sources being used in building the model for prediction:

- Source: Kaggle - https://www.kaggle.com/c/rossmann-store-sales/data
  - Store Data – Provides store profile data, including assortment it carries
  - Sales Transaction Data – Sales transaction data down to store and date level. Based data to be used for building the predictive model

3 Data Wrangling

The thought process behind the data cleaning process are as follows:

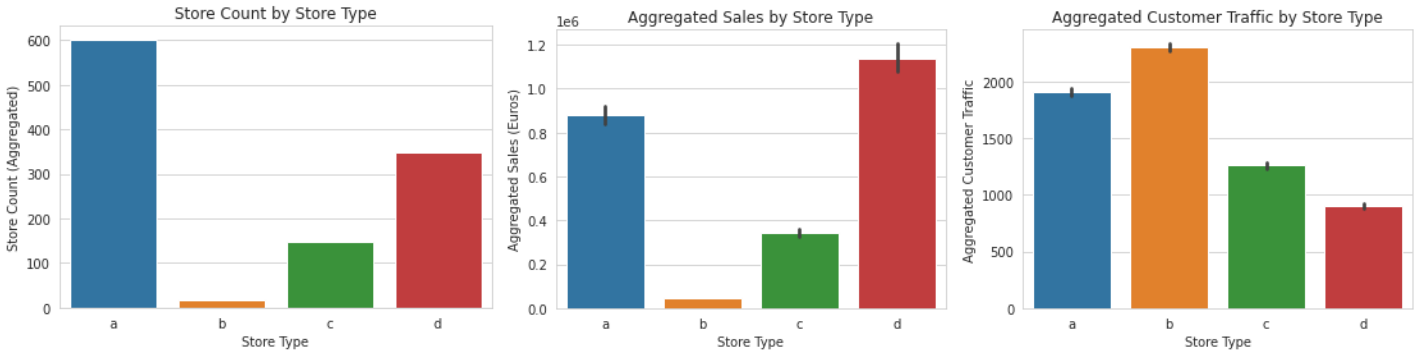
- Step 1: Understanding the profile of the data and nature of the missing data:
  - Store Data contains the profile of a particular store in the network, most missing data is from this dataset.
  - Sales Transaction Data captures sales for each store between 2013 to 2015 Jul. There is no obvious patterns of missing data in this dataset.
- Step 2: Merging the datasets together
  - The store and sales transaction datasets were merged together based on store number
  - No records were dropped before further understanding the nature of the dataset
- Step 3: Checking for inconsistencies in the dataset
  - Checking for inconsistencies between promotion since, marked promotion cadence per annum and actual transaction
  - Checking if all there are sales recorded in a store closure day
  - Identify and drop any duplicate records

4 Exploratory Data Analysis

High level analysis were done to understand the store profile and sales transaction patterns of Rossmann's stores. Data being examined were extracted from 1115 stores, between the period of 1 Jan 2013 to 31 Jul 2015.

4.1 Store Type

Examining Rossman's store, there are a total of 4 types -

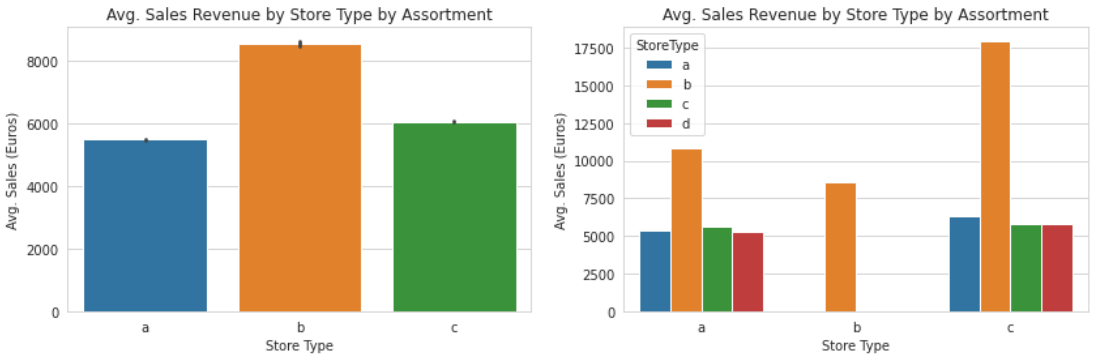


By store type, most stores are either type A or D. While there are no specific information indicating the differentiator between these stores, from the aggregated sales data, and the assortment data, a few things can be deduced:

- Although most stores are type A stores, with the 2nd highest traffic among all store types, on aggregate, it is only the 2nd most revenue generating store type
- Store type b has the least presence, but generated most customer traffic
- While store type d is only around 60% of the amount of store type a, it is the best sales generating store type

4.2 Assortment Profile

The original data illustrated that there are three types of assortment in a store - a = basic assortment, b = extra assortment, c = extended assortment. A further breakdown of the relationship between assortment by sales and by store type:

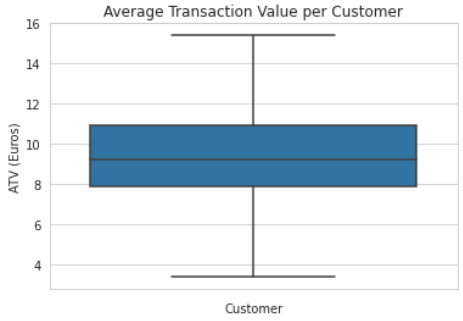


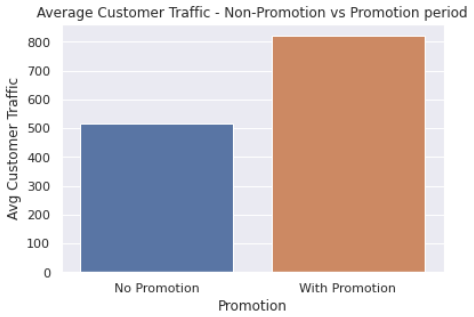
The digrams illustrated that:

- In terms of average sales, extra/ extended assortment generated more sales revenue on average than stores carrying only the basic assortment
- Among all store types, extra assortment is only carried by store type b
- With the exception of store type b, where average sales is significantly more than other store types and assortment types, there is no significant difference on average sales

4.3 Customer & Traffic Profile

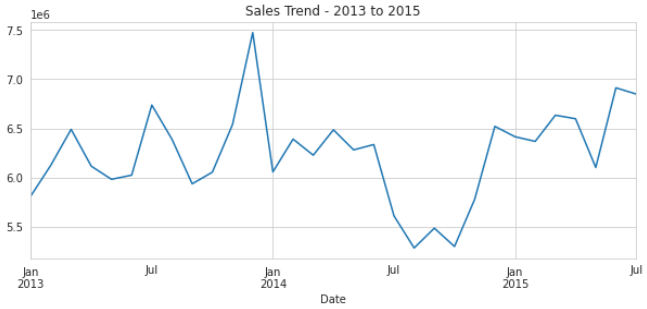
On average, customers are spending around 9.49 Euros per transaction with Rossmann stores. The diagrams indicated that there is a strong correlation between customer traffic and sales. At the same time, when a store is running promotions, the customer traffic is significantly heavier.



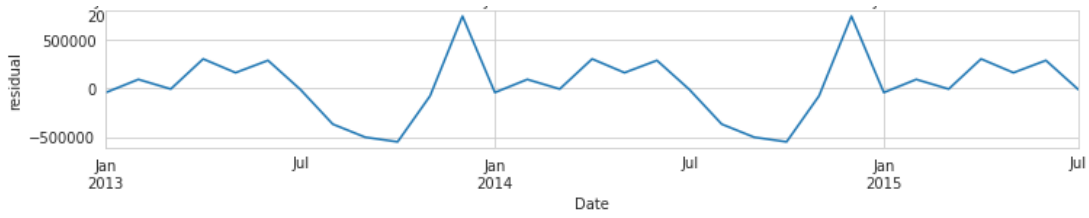


4.4 Sales Transaction & Seasonality Profile

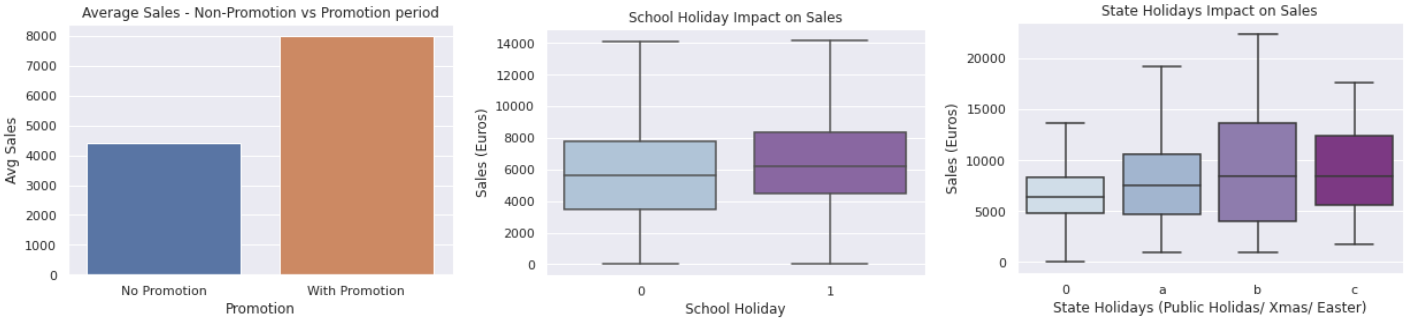
Sales transactions across stores has been aggregated to examine the overall sales trend for the period of 2013 to 2015. The graph indicated that there are no exponential growth in sales during the period.



A further deep dive into the seasonality of the sale, it is apparent that quarter four, close to December records the highest sales transactions, while quarter three is trending-wise the lowest.



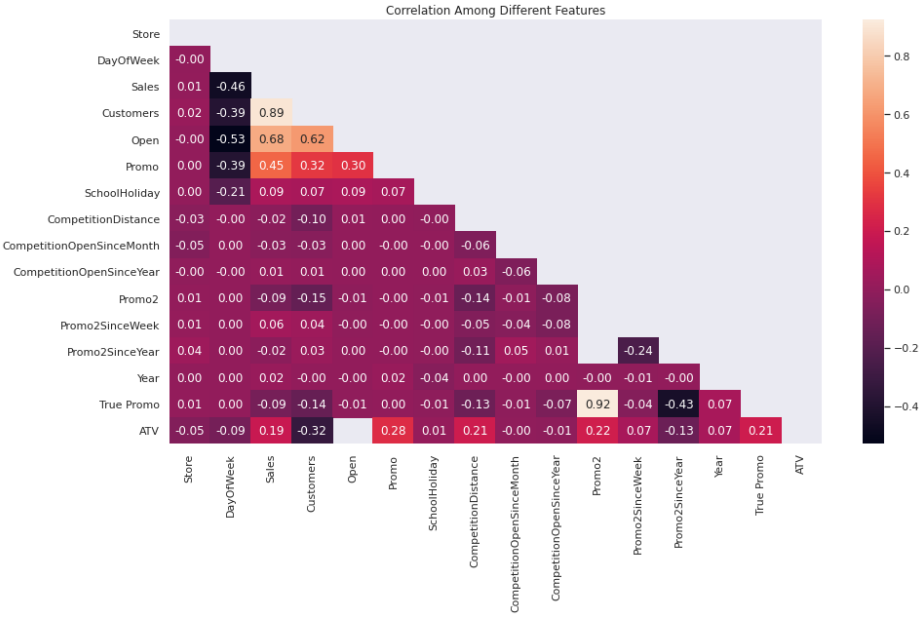
Whether the store is on promotion or not is also an important factor impacting average sales. As illustrated from the figure on far left, with promotion, average sales is significantly higher than days without promotion.



Examining the impact of holidays towards sales:

- Impact on sales due to school holidays is less apparent than impact on state holidays
- While average sales are similar across holiday and normal days, the top range of sales are indicating customers are spending more during state holidays particularly Christmas and Easter

Overall, most features are not directly correlated. Collinearity test will be conducted at a later stage to remove variables which will skew the machine learning model.



5 Feature Engineering and Pre-Processing

5.1 Feature Extraction

Further work is being done in preparing the data for machine learning. Given dates is still an important factor in the sales prediction. Additional features have been extracted from the year/month/date column, and to be one-hot-coded. Day of week has already been provided as part of the dataset. An additional list of features has been considered for feature extraction:

- Week of Year
- Month of Year
- Day of Month

However, given each of these features will expand into an additional 10-52 columns, “Quarter of Year” has been extracted to ensure the dataset will be manageable running on a local machine.

5.2 Feature Engineering

To increase the performance of the machine learning model, additional features has been engineered. These features are selected based on observations in the EDA process. These are the final additional features built:

- Average sales by store type
- +/- 7 days average sales, time lagged features

In addition, all categorical data has been transformed and one-hot-coded, and all records where store is closed and with store sales recording zero for the day are dropped, to prevent any distortion on the model.

5.3 Removing Multicollinearity

Variance inflation factor (VIF) has been used to identify and remove collinearity issues among the list of features. Below are the list of final features list to be used in modelling.

	VIFactor	features
0	4.512008	Customers
1	2.204749	Promo
3	2.063080	Assortment
17	1.850237	Quarter of Year_2
18	1.800604	Quarter of Year_3
15	1.729734	DayOfWeek_6
10	1.680120	StoreType_3
11	1.666691	DayOfWeek_2
14	1.640080	DayOfWeek_5
12	1.638188	DayOfWeek_3
13	1.607300	DayOfWeek_4
19	1.555562	Quarter of Year_4
4	1.537738	CompetitionDistance
21	1.501966	PromoInterval_2
2	1.411003	SchoolHoliday
8	1.361426	StoreType_1
9	1.247917	StoreType_2
20	1.206819	PromoInterval_1
22	1.193028	PromoInterval_3
16	1.128624	DayOfWeek_7
5	1.011579	StateHoliday_1
6	1.006103	StateHoliday_2
7	1.005207	StateHoliday_3

5.4 Train/ Test Split

A 75:25 train test split has been built, with a specific random state to ensure the results can be re-produced at a later stage.

6 Modelling

Various models have been experimented in the process:

- Linear Regression
- Lasso Regression
- Bayesian Ridge Regression
- Random Forest Regressor
- XGboost Regressor

These algorithms are chosen as the nature of the issue we are resolving is regression problem. Given the nature of the problem, Root Mean Squared Error (RMSE) has been chosen to illustrate how well the model fits. Initial results as follows:

Metrics/ Model	Linear Regression	Lasso Regression	Bayesian Ridge Regression Model	Random Forest Regressor Model	XGBoost
RMSE (Test data)	1309.089	1921.056	1309.091	1373.851	1044.527

Among all models being experimented, XGBoost Regressor is generating the lowest RMSE scores. As such, the model will be further fine turned and to identify what are the optimal parameters.

7 Model Optimization & Evaluation

7.1 Hyperparameters Tuning

In order to identify the optimum parameters for the model, RandomSearchCV has been used to find the best optimized combination of:

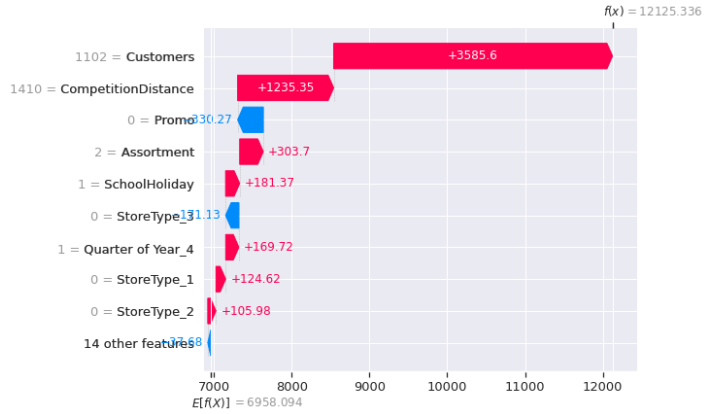
```
'max_depth': [10, 20],
'n_estimators': [25, 50, 100]
'learning_rate': [0.01, 0.1]
```

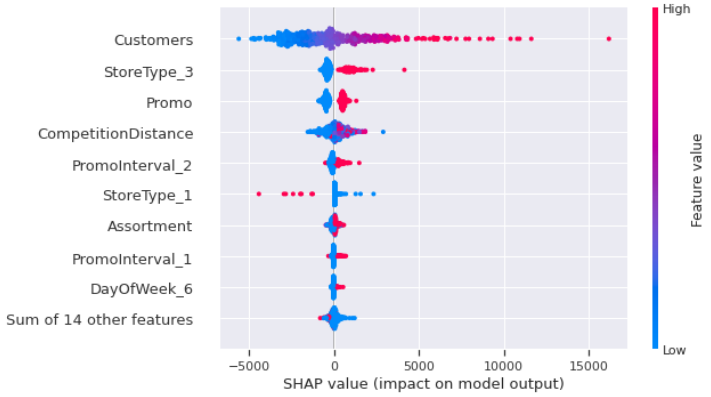
RandomSearchCV has been used instead of GridsearchCV after considering the run time and the number of scenarios the processor needs to handle on a local machine. The suggested best parameters for the model is - `{'learning_rate': 0.1, 'max_depth': 20, 'n_estimators': 100}`

Using the optimized parameters to re-run the XGBoost Regressor model, the RMSE score has improved from 1044.52 to 544.44. Further to the RandomSearchCV, a 5 fold cross validation has been done. Result shows the accuracy of the model with cross validation is at 97%, indicating the model performs well even with data outside of the training set.

7.2 SHAP Analysis

In order to further interpret and understand the model, SHAP analysis has been used to outline which features holds the most important key in helping with predicting Rossmann's store sales.





From the chart, and to further explain the XGBoost Regressor model on the various features which contributed to the prediction of store sales, it is illustrated that:

- Number of customers (or customer traffic) is one of the most important factors in predicting store sales
- Second most important feature is how close a competitor exists next to a Rossmann store
- Other features which contributes to the predictability of store sales
    - having an extended assortment at store
    - having a school holiday date-wise being in quarter 4
    - stores that carriers a more extended assortment

While promotion was initially highlighted as a key factor influencing store sales it does not aid as much in predicting store sales.

8 Conclusion & Future Improvements

Among all the models used to predict Rossmann's store sales, XGBoost Regressor, with parameters - `{'learning_rate': 0.1, 'max_depth': 20, 'n_estimators': 100}`yields the best predictive results. The final XGBoost model yields an RMSE score of 544.44 with cross-validation accuracy score at 97%.

However, the model does have potential limitations, as it assumed:

- The 1115 stores is a good representation of the sales pattern for the other Rossmann stores (which there are over 3000 in the total store network)
- The business environment remains largely similar to the period between 2013-2015 where data is sampled
- In using the regression approach to model this store prediction, the comp store growth may not have been fully modelled out. Comp store growth is defined as sales of an existing store over a certain period, compared to an identical period in the past, usually the previous year

For future areas of improvements or subsequent phases to improve the model, other approaches or data could be considered as an addition:

- Conduct a time series analysis to understand the potential compound growth of store sales on a year-on-year basis
- Further feature engineering using other data sources/ external data, such as – customer profile data from CRM database, competitor's sales or promotional data etc.

Appendix

Github link for the detailed work

- https://github.com/jade-lam/Springboard-DSC-Capstone-3/blob/main/Capstone%203%20-%20Rossmann%20Store%20Sales%20Prediction%20(Regression).ipynb