Rossmann Store Sales Prediction

Jade Lam
Jun 2021

Content

I. Executive Summary

II. Understanding Rossmann's Store & Sales Profile

III. Pre-Processing & Feature Engineering

IV. Modelling & Hyperparameters Tuning

V. Interpreting the Results

VI. Conclusion

I. Executive Summary

Dirk Rossmann (GmbH)

Usually referred as Rossmann, Dirk Rossmann (GmbH) it is one of the largest drugstore chains in Europe with over 4000 stores across Europe. In 2020, Rossmann has recorded 10.35 billion turnover in Euros. It's store network covers Germany, Poland, Hungary, Czech Republic, Turkey, Albania, Kosovo and Spain.

Project Objectives

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.

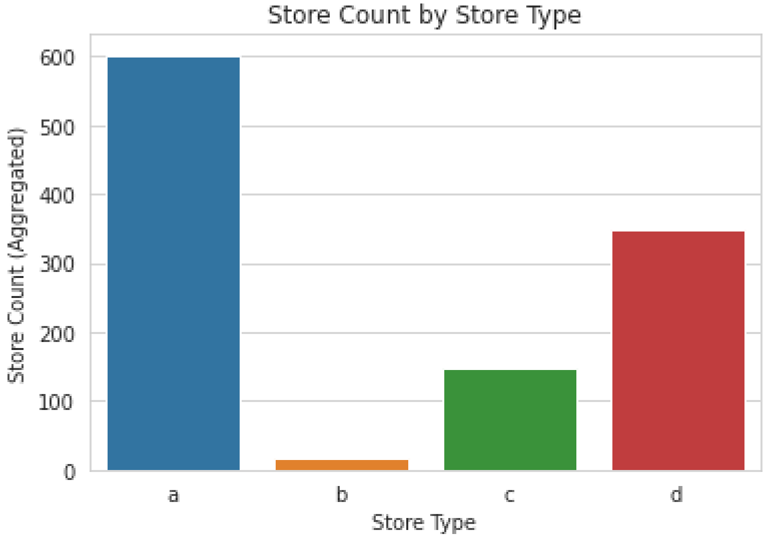
Project Conclusion

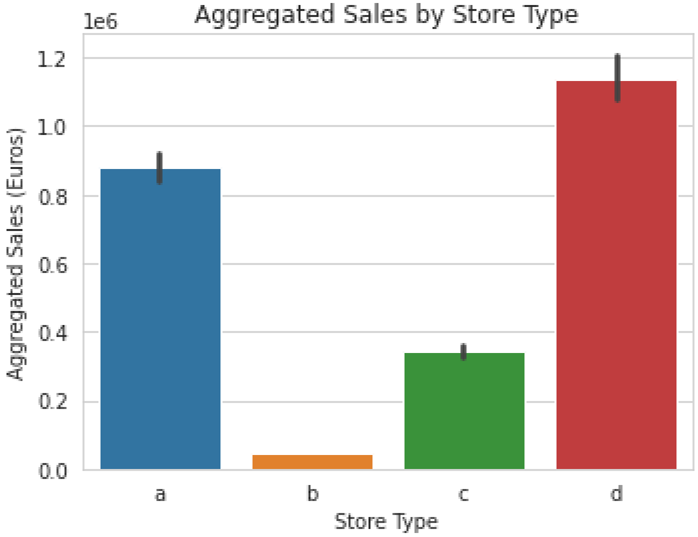
- Number of customers and distance from Competition were by far the most influential factors in predicting Rossmann's store sales
- Having stores carrying a wider assortment and holiday seasons are other key factors affecting store sales
- XGBoost Regressor with parameters of {'n_estimators': 100, 'max_depth': 20, 'learning_rate': 0.1} yields the lowest Residual Mean Squared Error (RMSE) score at 544.44.
- The model is highly applicable to unseen data, with a cross-validation score at 97%

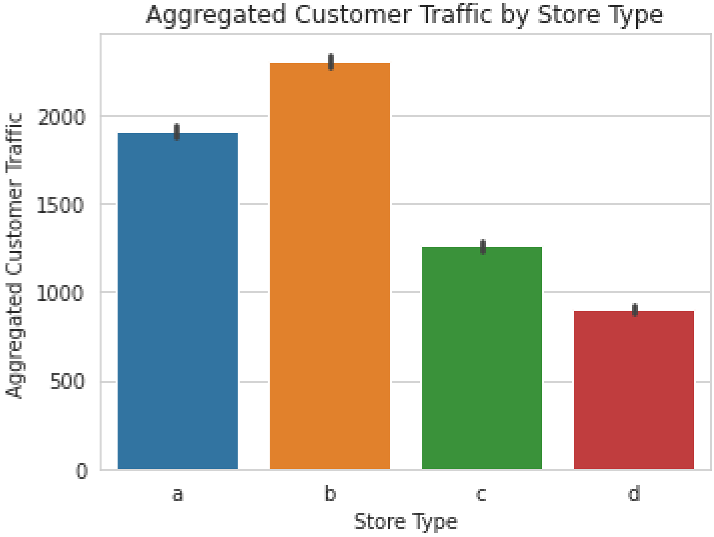
II. Understanding Rossmann's Store & Sales Profile

Rossmann's Store Profile

- There are 4 types of Rossmann's Stores.
- Store type b has the least presence, but generated most customer traffic
- While store type d is only around 60% of the store count for type a, it is the best aggregated sales generating store type



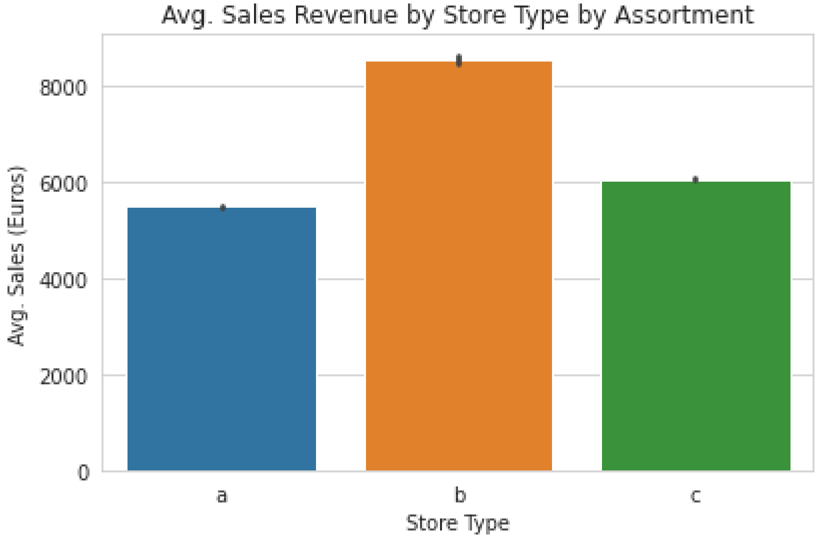


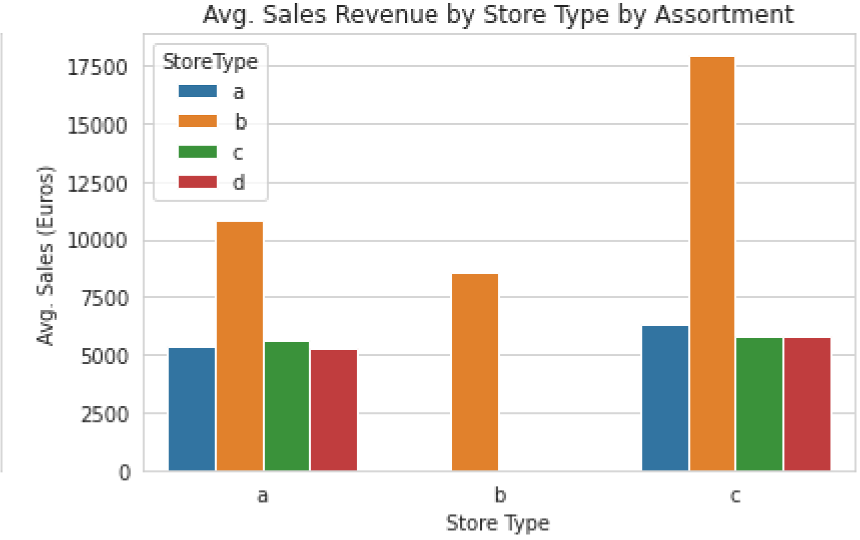


II. Understanding Rossmann's Store & Sales Profile

Rossmann Store's Assortment Profile

- There are 3 types of assortment at a Rossmann store - a = basic assortment, b = extra assortment, c = extended assortment
- Extra assortment is only carried at store type b
- Having an extra/ extended assortment is yielding a positive impact on average sales over stores with only basic assortment

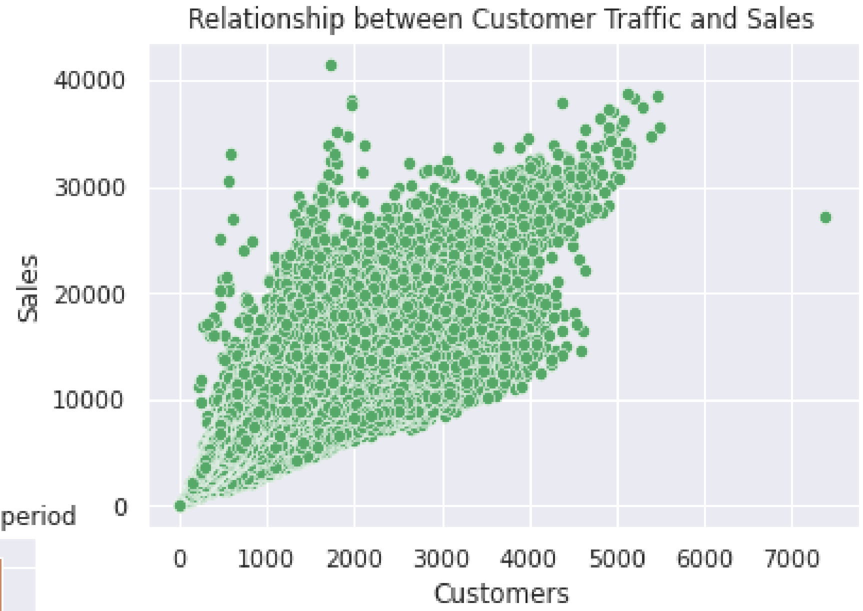


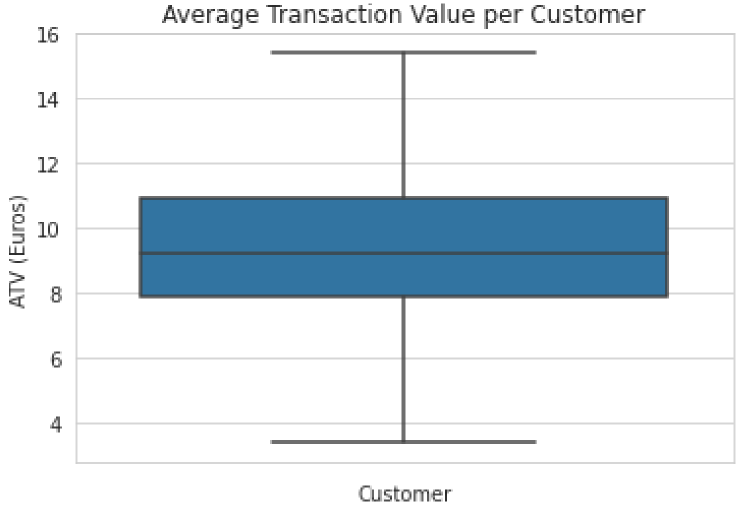


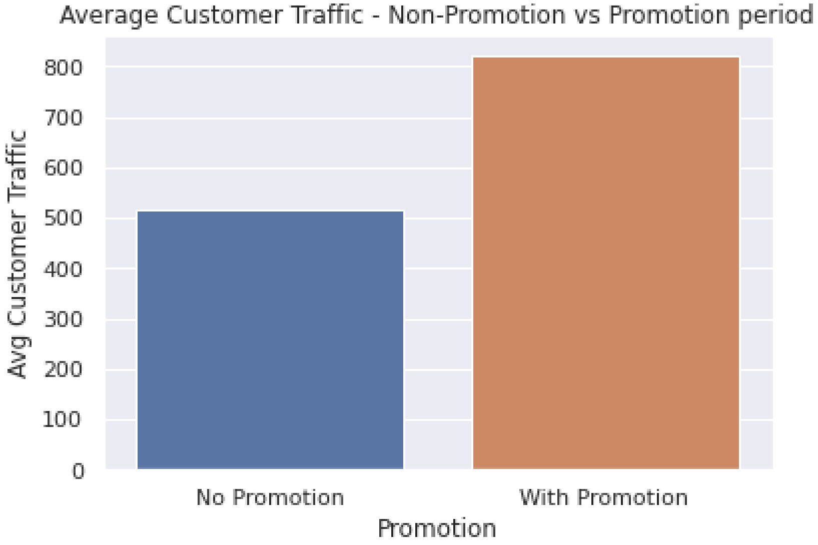
II. Understanding Rossmann's Store & Sales Profile

Rossmann Store's Customer & Store Traffic Profile

- On average, customers are spending around 9.49 Euros per transaction with Rossmann stores
- Whether a store is holding promotions has huge impact on customer traffic in general
- At the same time, there is a clear positive relationship between customer traffic and sales revenue



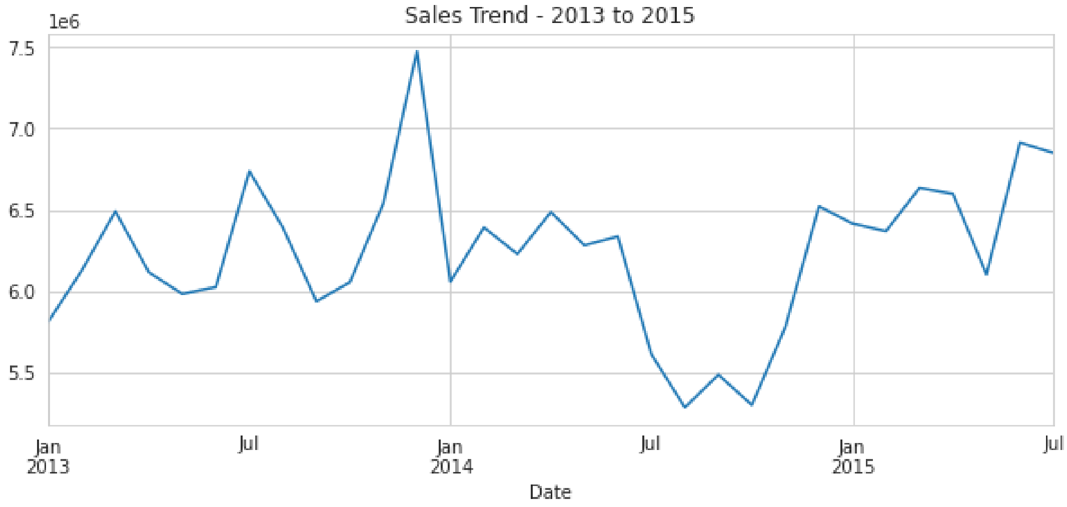


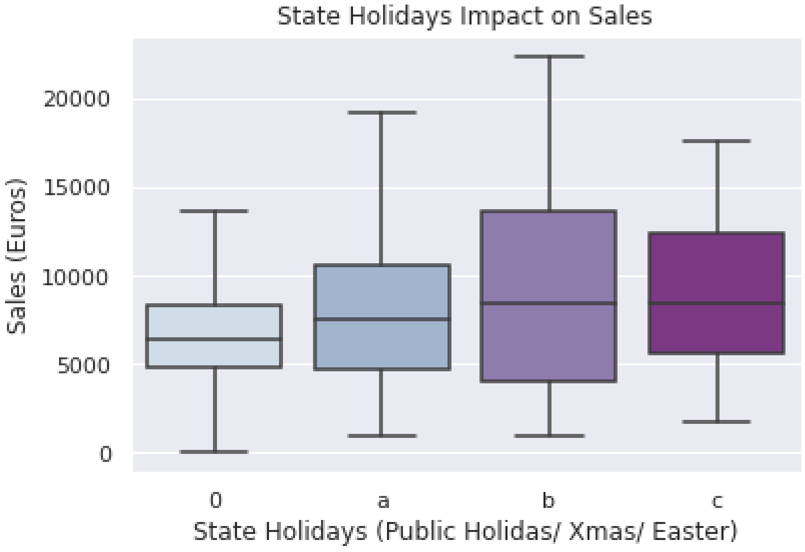


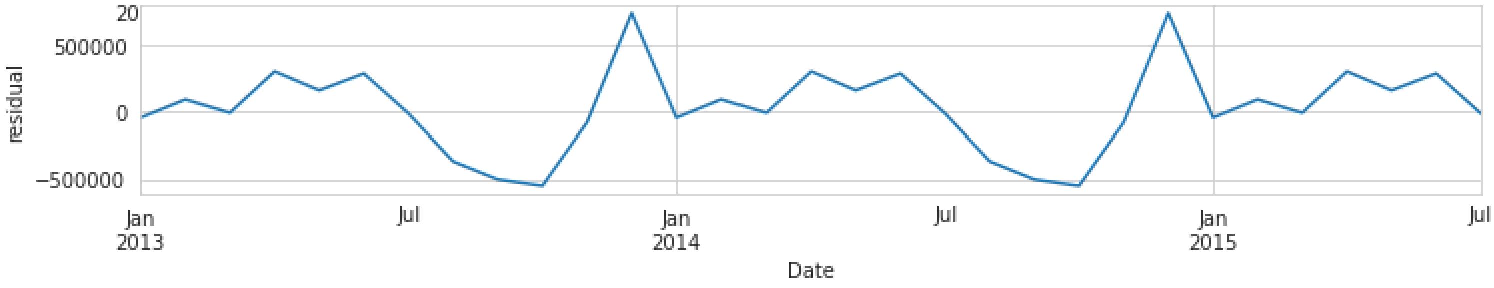
II. Understanding Rossmann’s Store & Sales Profile

Rossmann Store’s Sales Profile

- The aggregated sales indicated that there were no exponential growth during the period
- Further deep dive into the seasonality suggests that quarter four (near Dec) has the highest sales while Q3 is trending the lowest
- While average sales are similar across holiday and normal days, the range of sales are indicating customers are spending more during state holidays particularly Christmas and Easter







III. Pre-Processing & Feature Engineering

Issues Addressed in Pre-Processing

- Extracting additional features from "Date" column. Created additional "Quarter of Year"
- Removed records during store closure day where sales is zero
- One-hot-coded "Day of Week", "Quarter of Year" and other categorical features

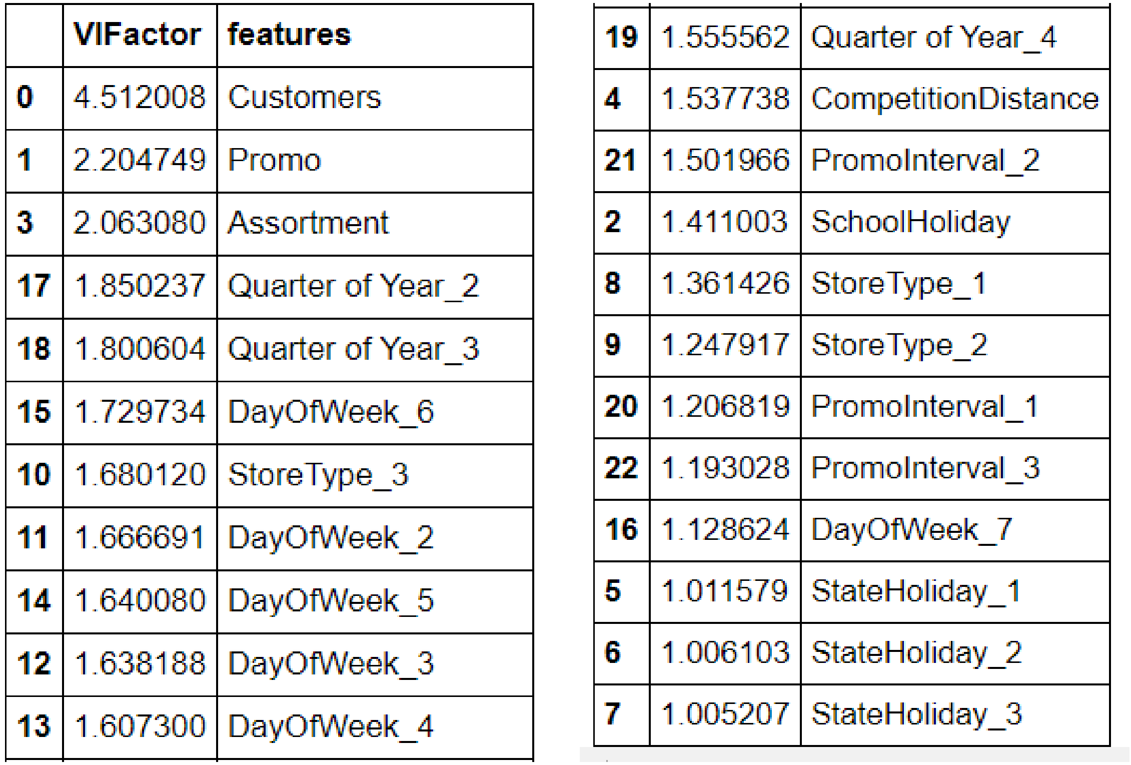
Additional Features Created

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

III. Pre-Processing & Feature Engineering

Removing Multicollinearity & Features Selection

- Variance Inflation Factor (VIF) is used to test for collinearity among features
- These are the final set of features to be used in the predictive model



	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

IV. Modelling & Hyperparameters Tuning

Models Experimented:

Below list of models were chosen as it is suitable for regression type problems:

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

Interpreting the Results:

- The best model would be the model with least RMSE score
- Among all models, XGBoost Regressor is generating the lowest RMSE scores. As such, the model will be further fine turned

Initial Results:

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

IV. Modelling & Hyperparameters Tuning

XGBoost Parameters (Initial Model):

'n_estimators': 25,

'max_depth': 10

‘learning_rate’: -

Cross Validation Score (Initial Model):

- The initial cross-validation score for the model is at 89%

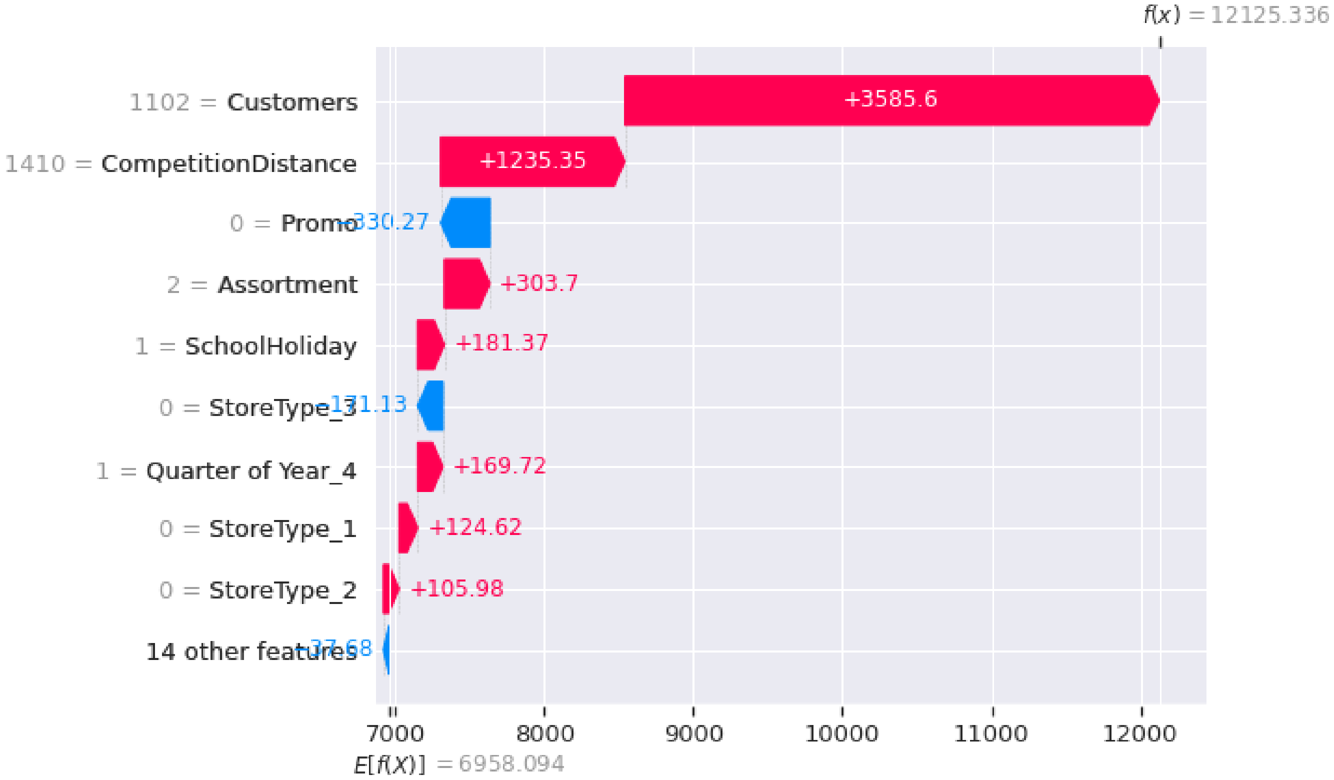
Searching for Optimized Parameters:

- To ensure the operation can be run on a local machine, RandomSearchCV has been used
- Parameters used in search were across number of estimators, max depth and learning rate.

Re-running the Model with Optimized Parameters & Final Results:

Model	Parameters/ Metrics
XGBoost Regressor	{‘n_estimators’: 100, ‘max_depth’: 20, ‘learning_rate’: 0.1}
RMSE Score (Optimized Model)	544.44
5 fold Cross Validation Score (Optimized Model)	97%

V. Interpreting the Results



Interpreting the Model & the Results

- SHAP analysis was performed on the final model to understand the importance of the features in contributing to the model's predictability
- The biggest 2 factors in predicting the store sales is the number of customers (or customer traffic), followed by the distance away from a competitor
- Other features which positively contributed to the store sales prediction:
  - having an extended assortment at store
  - having a school holiday date-wise being in quarter 4
  - stores that carriers a more extended assortment

VI. Conclusion

Limitations/ Other Considerations

- 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

Future Scope

- 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.