Rossman Sales

WORD COUNT: 1628

1. Introduction

Historical sales forecasting is a complicated but high-rewarding task, as it equips businesses with the ability to strategically use data insights to improve their revenue and allocate resources efficiently. This report focuses on the preprocessing of time series sales data collected on a large German drug chain – Rossman. The core stages of the analytic approach include data description, cleaning, transformation, feature selection, and extraction, which were tailored to build a reliable and accurate sales forecasting model. The report discusses the linkage of the provided datasets and the implications of missing values, which were handled through model-learning imputation. Necessary variable transformations and problematic variables showing high multicollinearity, were identified, and dealt with. Subsequently, the report focuses on selecting and transforming appropriate features, as well as generating time-series specific predictors, such as cyclical date variables. After completing the preprocessing, Extreme Gradient Boosting Regression was applied to the Rossman sales forecasting problem due to its versatility and ability to handle complex datasets. The relevant modelling steps, such as time series cross validation and hyperparameter tuning are further discussed, followed by assessment of model performance and visualisation of test sales predictions. Lastly, the implications of the results are discussed in relation to Rossman's business strategy, highlighting relevant areas of improvement, such as increasing promotional effort, and suggesting further exploration of location- and product-specific trends.

2. Methodology

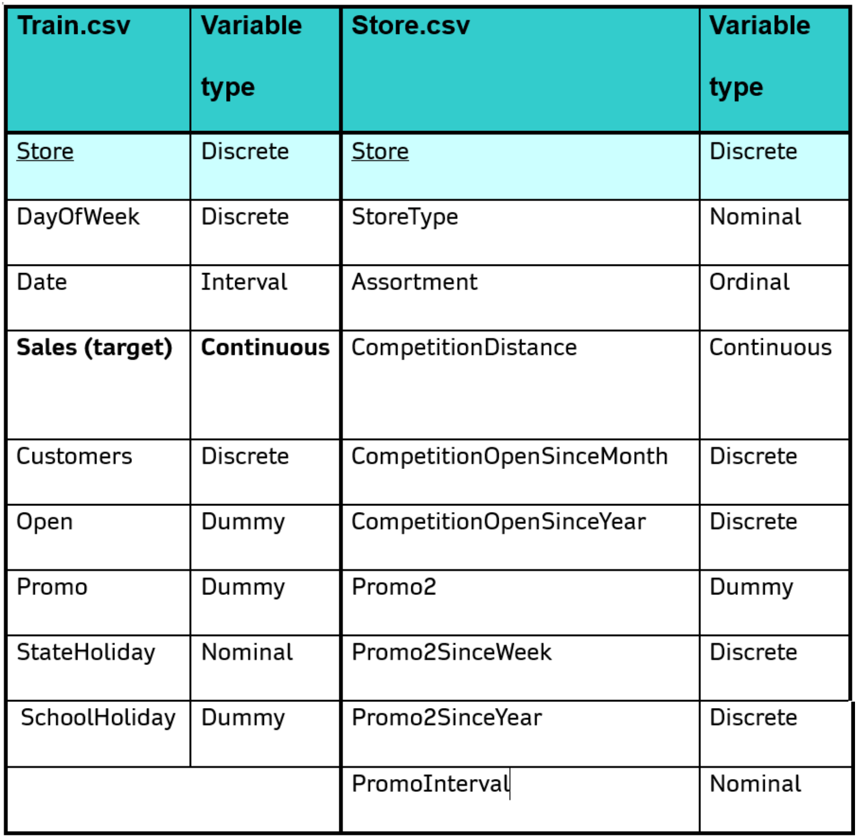
2.1 Data Description

The data available for the prediction of Rossman sales is in three separate datasets and spans from 01/01/2013 to 17/09/2015. The three datasets can be linked on store (id). However, the test set, which would usually be utilised to assess prediction accuracy, does not contain data on *Sales* and *Customers*. Therefore, it will only be used as a template to fill predicted sales values, and the analysis will focus on 'store.csv' and 'train.csv' merged on *Store* (01/01/2013-31/07/2015).

Table 1 displays the available variables for the sales forecasting analysis, as well as their data types.

As the main goal is to achieve the most accurate prediction of Rossman sales, rows which indicate closure of stores, and hence zero sales, have been deleted.

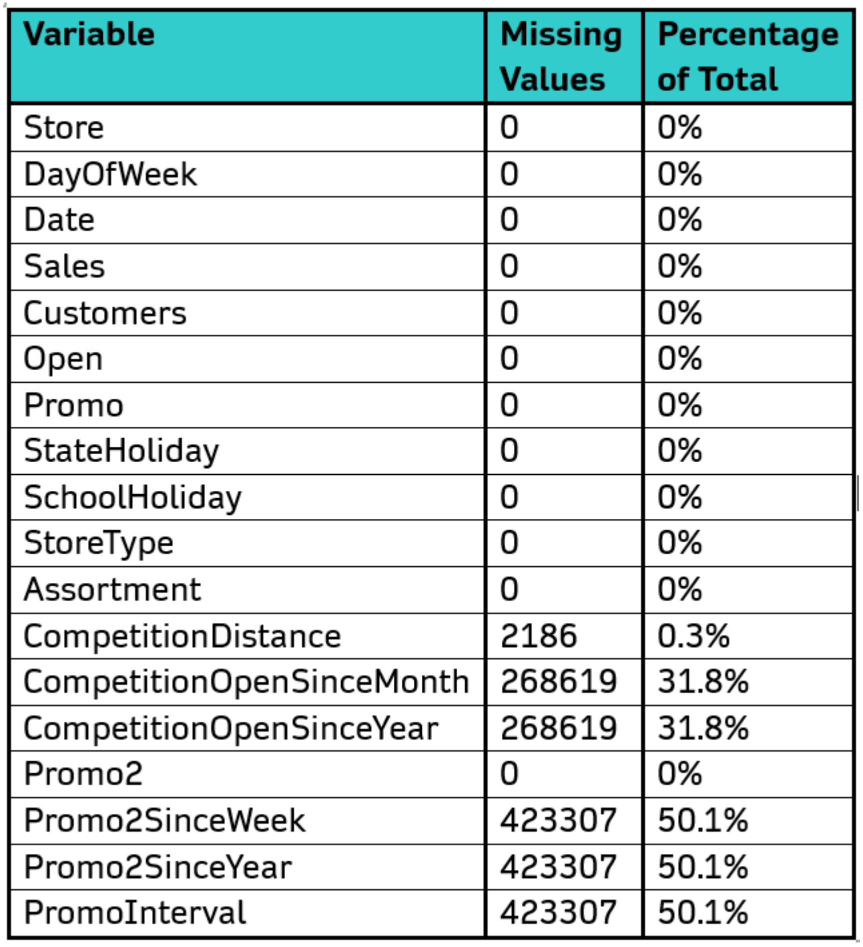
Table 1. Variables in 'train' and 'store'



Train.csv	Variable type	Store.csv	Variable type
Store	Discrete	Store	Discrete
DayOfWeek	Discrete	StoreType	Nominal
Date	Interval	Assortment	Ordinal
Sales (target)	**Continuous**	CompetitionDistance	Continuous
Customers	Discrete	CompetitionOpenSinceMonth	Discrete
Open	Dummy	CompetitionOpenSinceYear	Discrete
Promo	Dummy	Promo2	Dummy
StateHoliday	Nominal	Promo2SinceWeek	Discrete
SchoolHoliday	Dummy	Promo2SinceYear	Discrete
		PromoInterval	Nominal

2.2 Data Cleaning and Transformation

Table 2. Missing Data



Variable	Missing Values	Percentage of Total
Store	0	0%
DayOfWeek	0	0%
Date	0	0%
Sales	0	0%
Customers	0	0%
Open	0	0%
Promo	0	0%
StateHoliday	0	0%
SchoolHoliday	0	0%
StoreType	0	0%
Assortment	0	0%
CompetitionDistance	2186	0.3%
CompetitionOpenSinceMonth	268619	31.8%
CompetitionOpenSinceYear	268619	31.8%
Promo2	0	0%
Promo2SinceWeek	423307	50.1%
Promo2SinceYear	423307	50.1%
PromoInterval	423307	50.1%

Data on distance to nearest competitor is missing for 3 stores, which results in 0.3% missing cases in the merged dataset (Table 1). Distance is likely not missing due to its value (too high/low) and could rather be a result of an input error in the database. Therefore, it is assumed the data are Missing at Random.

However, while some of the missingness in *CompetitionOpenSinceMonth/Year* is due to missing *CompetitionDistance,* a high percentage cannot be explained without additional information. Thus, the values were imputed with MICE (Multiple Imputation by Chained Equations), which is applicable in Missing Not at Random scenarios. Moreover, MICE can model dependencies in missingness between multiple variables, which is potentially relevant in this case.

The variables *CompetitionOpenSinceMonth/Year* were combined into *Months_Passed* to properly reflect the continuous dimension of the variable. However, Figure 1 suggests that *Months_Passed* has no correlation with *Sales*, deeming it a redundant feature, which will be dropped from the analysis.

Additionally, *the missingness of Promo2SinceWeek, Promo2SinceYear* and *PromoInterval* is related to the observed variable *Promo2*, and it is not random, making the data MNAR. The format of the two variables is not suitable for the analys is – thus, they have been transformed into a single variable measuring weeks since *Promo2* started*,* imputed with 0 for missing values, as no weeks have passed due to lack of promotion.

However, the new variable indicates high correlation with *Promo2* (Figure 1), suggesting redundant information, which necessitates the exclusion of *Weeks_Passed_Since_Promo* from the analysis.

Figure 1. Correlation Matrix

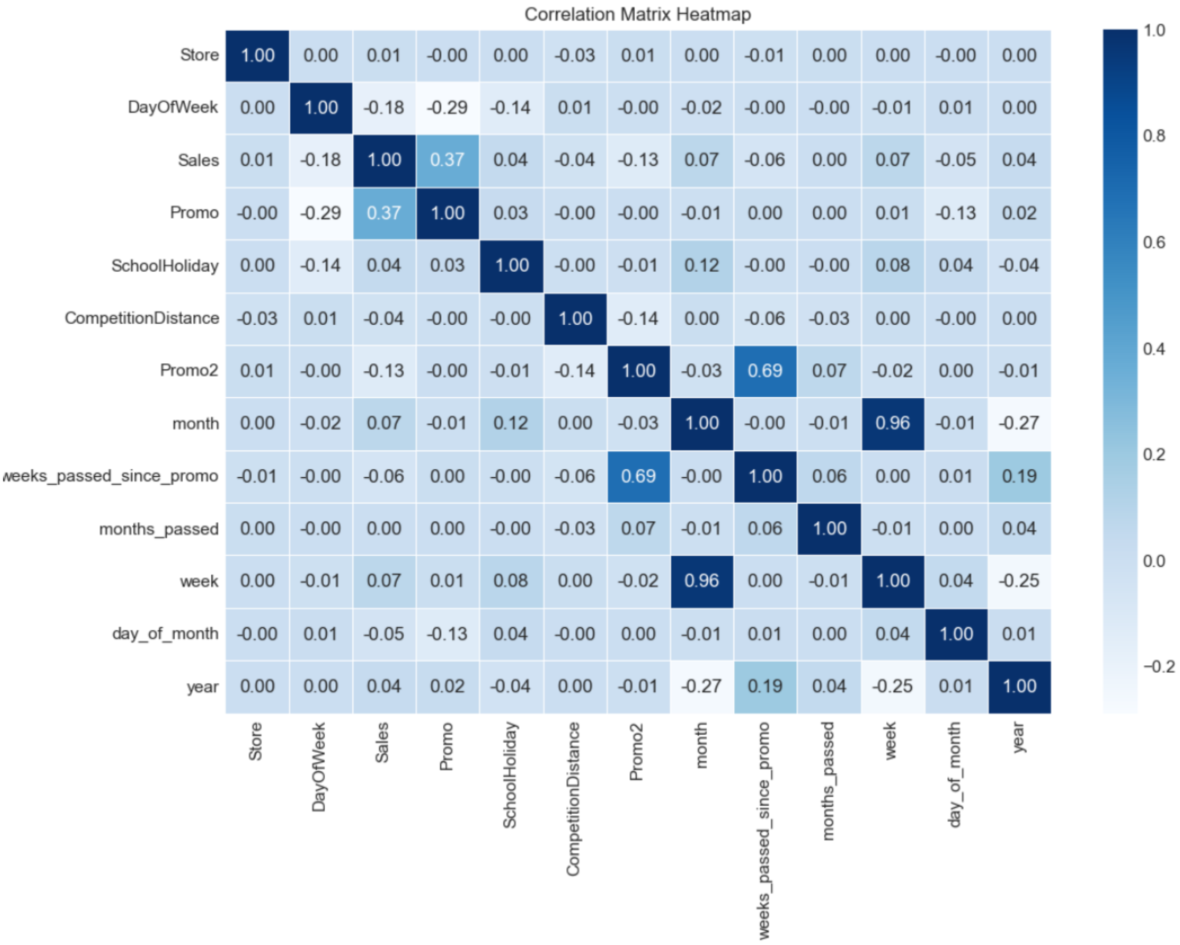
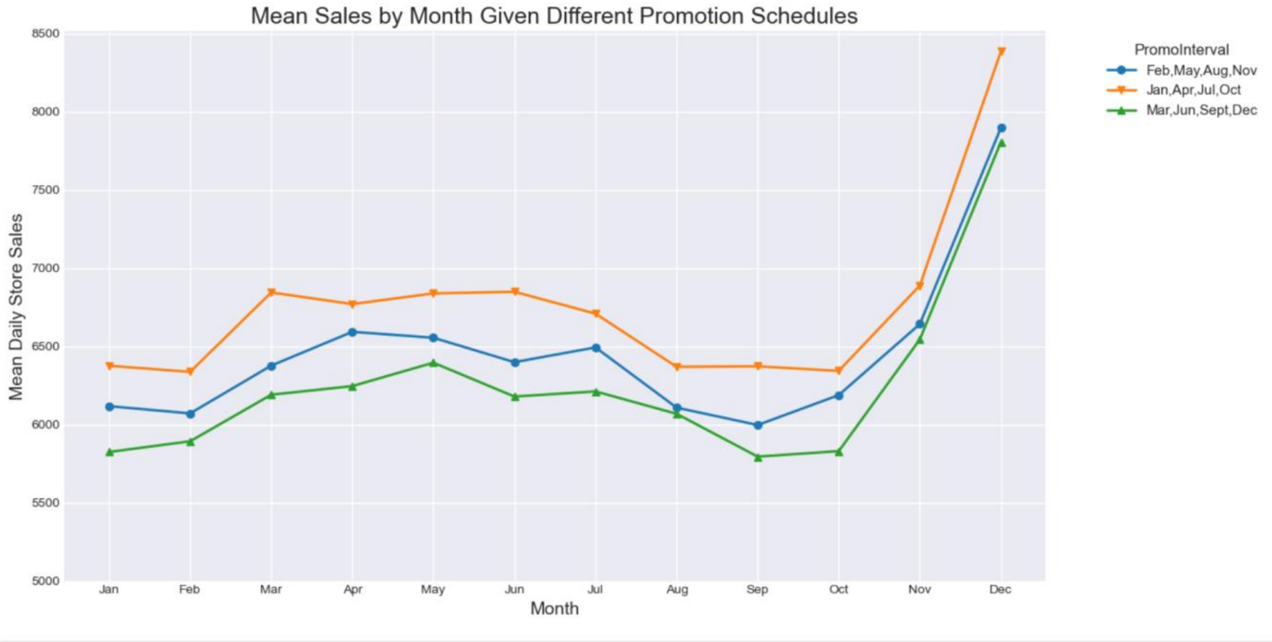


Figure 2. Mean Sales by *PromoInterval*



Moreover, different promotion intervals display similar trends in sales, which indicates that the variable is unlikely to be an informative feature, deeming it more appropriate to exclude it from the analysis (Figure 2).

2.1. Feature Selection and Extraction

The variable *Customers* was dropped, as it does not exist in the test set and introduces look ahead bias in the validation set, which occurs when we utilise information we would not normally have at the time of prediction. Moreover, *Open* was also excluded, as the dataset was filtered to omit closed stores (0 sales).

Date as an index cannot be explicitly implemented as a feature in the model, and thus needs to be encoded as separate time-based component features: *month*, *day_of_month*, *day_of_week*, and *year*. The *week* feature was not implemented due to evidence of redundancy (high correlation with *month*, Figure 1).

Figure 3. Mean Sales by Date Components

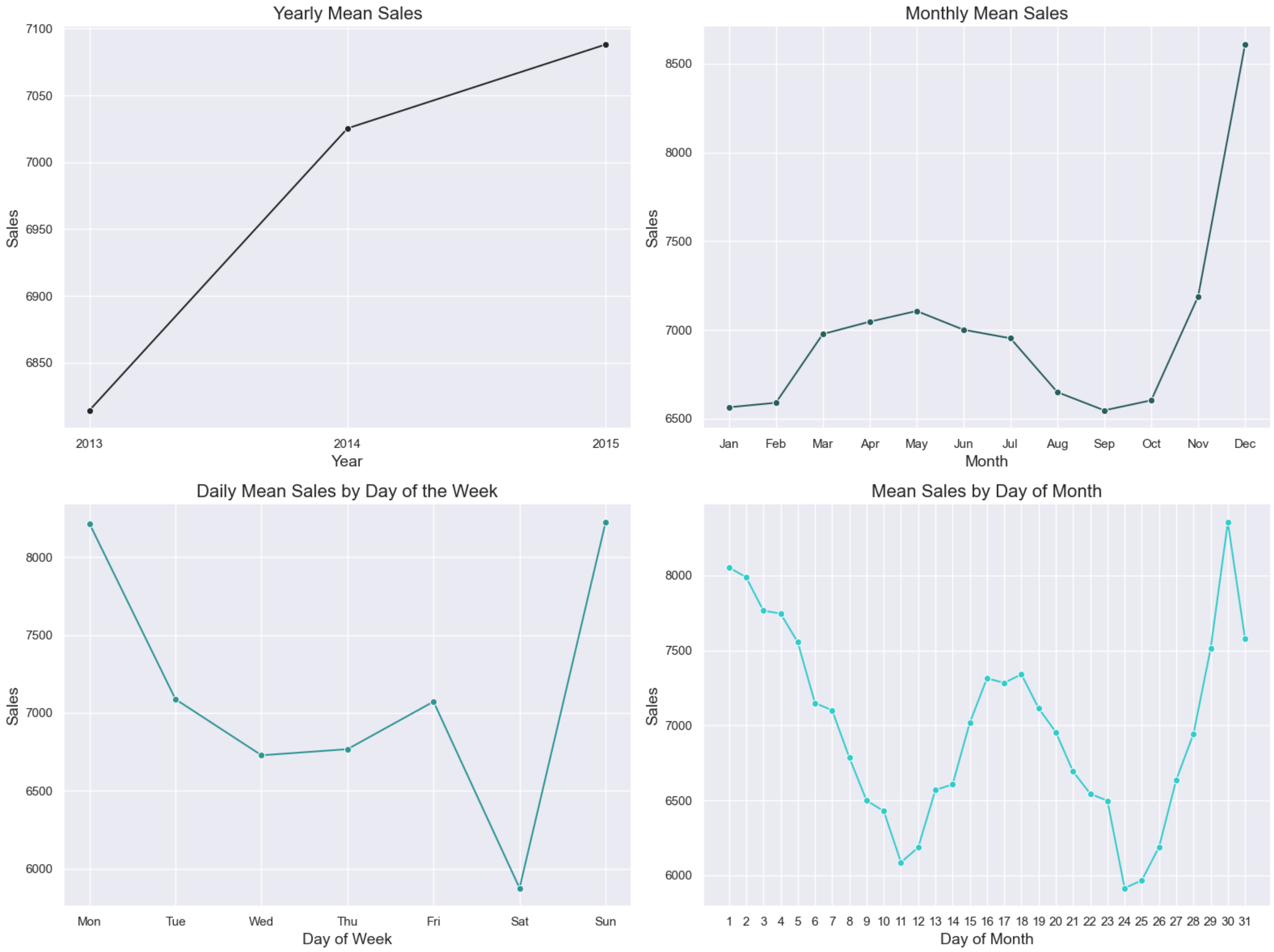


Figure 3 shows monthly fluctuations in average sales with a peak in December, which might be caused by Christmas shopping trends. Furthermore, Figure 3 demonstrates the average variability of sales by day of the month, with a significant spike towards the end, which could be due to a large proportion of people getting paid at that time. Year and day of week also display evident tendencies, which can be utilised in the modelling stage.

Furthermore, sine and cosine transformations were applied to month, day of month and day of week to ensure the model can properly interpret their cyclical pattern.

Additionally, the variables *StateHoliday* and *StoreType* were one-hot-encoded to achieve the desired numeric format, and *Assortment* was ordinally encoded with integers, capturing the stratification of its categories. Lastly, the target variable *Sales* was log transformed to reduce variance and skewness, which helps improve the forecasting accuracy of the model.

3. Sales Prediction

3.1. Model Selection

Time series data rarely displays perfect linearity and independent data points, which makes linear models less reliable for forecasting. The Rossman dataset contains multiple categorical predictors (encoded), which introduce increased model complexity and possibly non-linearity, complicating the task of modelling time series. From a business perspective, a model which provides a low error would be the most desirable, since supply and revenue significantly depend on the accuracy of forecasting. Ensemble-based algorithms have been shown to perform better on time series data compared to linear models (Papadopoulos and Karakatsanis, 2015). Thus, Extreme Gradient Boosting (XGBoost) Regression was chosen as the most suitable modelling technique for Rossman sales. XGBoost is robust to overfitting, which is frequent result of high dimensionality data, and displays feature importance, which is particularly imperative for optimising forecasting models. Additionally, it is a computationally inexpensive algorithm, which minimises modelling constraints on time and resources.

Before fitting the XGBoost model, the data was split into training (01/01/2013-11/06/2015) and validation (12/06/2015-31/07/2015) sets to maintain its chronological order and avoid data leakage. Time series cross-validation (TSCV) was performed to ensure the temporal consistency of the data, as well as its coverage of different subsets of periods. Additionally, the model was tuned with a randomised search to ensure the optimal hyperparameters, which yield the lowest error, while preventing overfitting.

3.2. Model Evaluation

The XGBoost forecasting model was evaluated with the RMSE (Root Mean Squared Error) metric, since it is in the same units as the target variable (sales). This aids model interpretability, which is critical to the business context of sales forecasting. Table 3 demonstrates that the RMSE of the 5 different time series subsets is relatively stable, which indicates that the model performs well across the training dataset. The validation set RMSE is a measure of how well the model generalises to unseen data, and in this case, it performs slightly worse compared to the TSCV folds. However, the RMSE scores are not significantly different, which points to an adequate level of forecasting accuracy.

Table 3. Performance Metrics

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Validation Set
RMSE	799.1093	813.0339	776.3237	823.0385	880.5784	1014.6087

Figure 4. Feature Importance

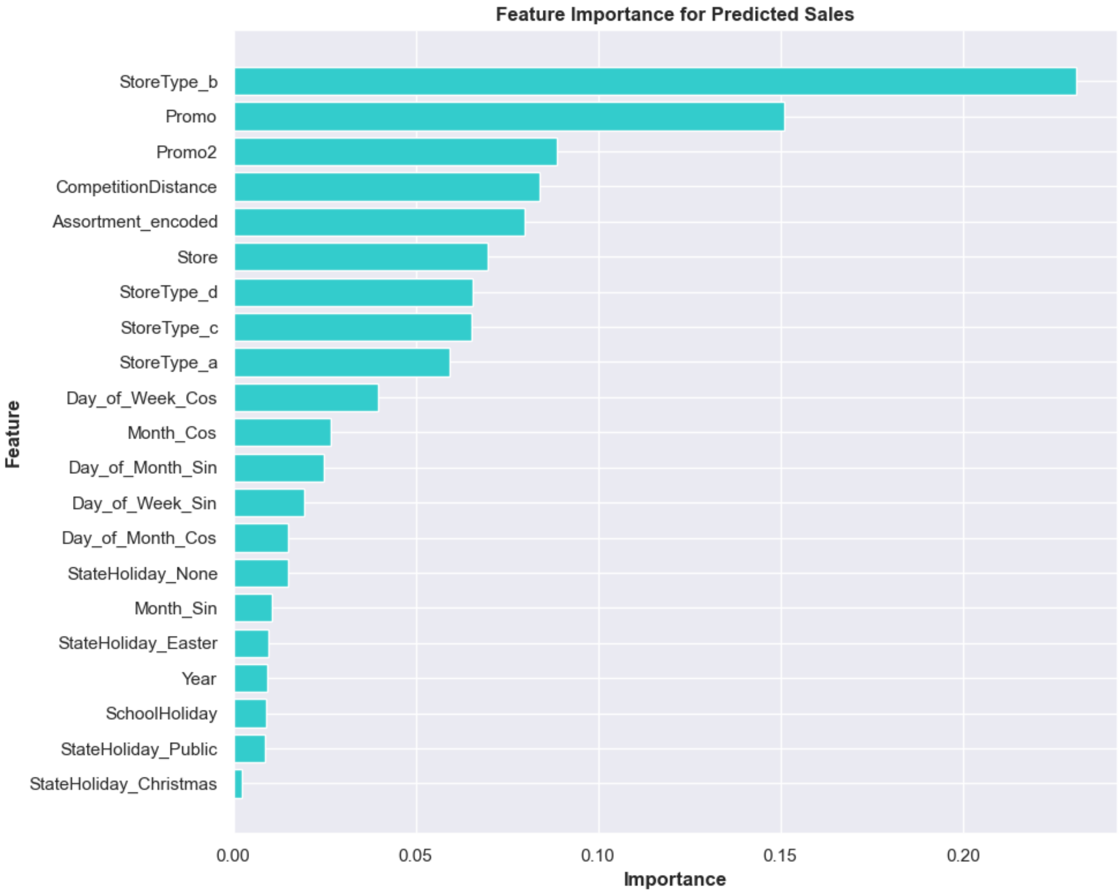


Figure 4 displays the ranked feature importances, with *StoreType_B* and *Promo* being the most powerful sales predictors. However, *StoreType_B* is negatively associated with sales (Figure A-1), while presence of *Promo* positively affects sales (Figure A-2). Unsurprisingly, Christmas was the least important feature, which is logical given that the validation period does not cover this holiday.

3.3. Predicting Sales

After evaluating the final XGBoost model, the 'test.csv' dataset was utilised to forecast sales between 1/08/2015 and 17/09/2015, which entailed matching all features present in the training set to the test set.

Figure 5. Predicted Total Sales Test Period

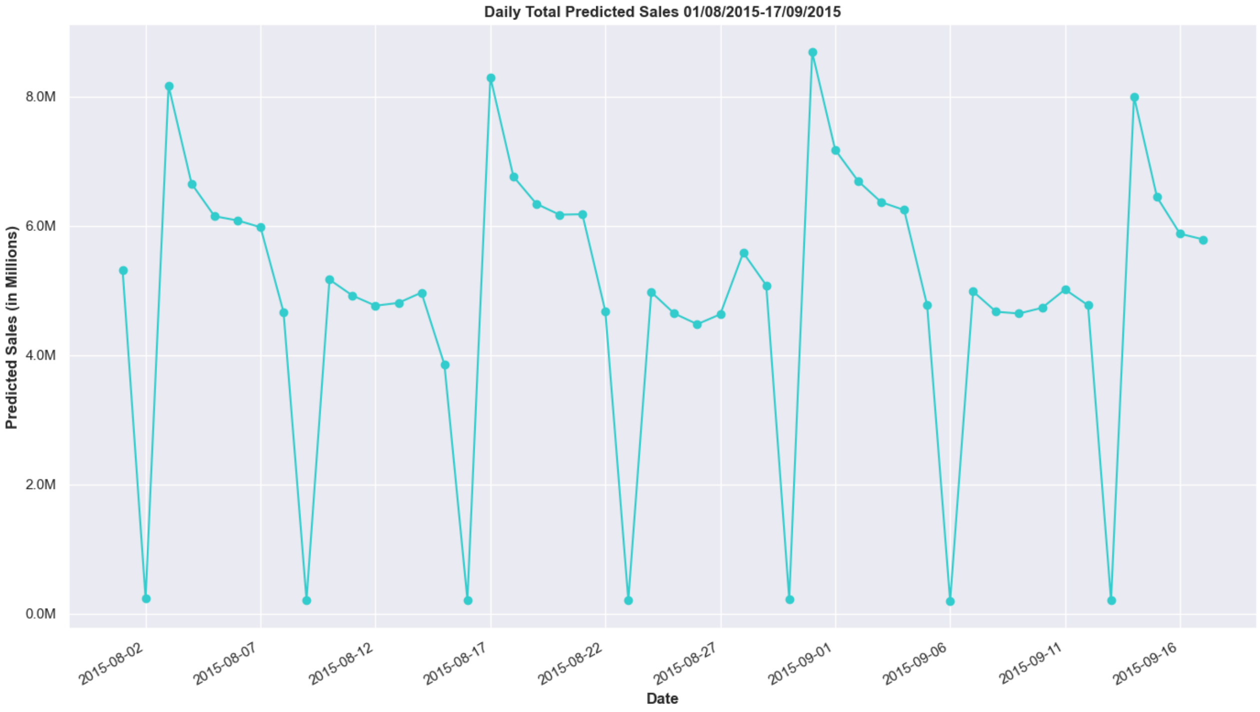


Figure 5 visualises the predicted total sales per day in millions for the test period. The sharp drops in total sales on Sundays are due to most stores being closed, which results in significantly less overall sales. This is to be expected as Sundays are typically non-working days for German stores, however, open Rossman stores generate substantial profits on average on that weekday (Figure 3). Therefore, Sunday store closures are negatively impacting sales, which can be reversed and leveraged by Rossman in the future to maximise their profits.

4. Conclusion, Implications and Recommendations

The findings of this report suggest that *Promo* is an important predictor of sales, which Rossman can utilise to their advantage by implementing more store-specific promotions. Additionally, sales can be maximised by introducing working hours on Sunday, which performs poorly in the model due to the number of closures but shows high average sales among opened stores. Sales in December and at the end of each month tend to increase, which can be help Rossman stores ensure adequate stock supply during these periods.

However, this forecasting model is limited to the predictors provided in the available datasets, as well as the data quality and computational resources available. It could be helpful for future sales forecasting techniques to explore additional information on:

- Location, e.g. which federal state the shop is in, as policies and sales trends can differ vastly.
- Information on specific products, which could provide a more tailored approach to sales prediction.
- Economic indicators such as inflation, interest, and unemployment rates, which might help capture relevant large-scale patterns in sales.

Bibliography

Papadopoulos, S. and Karakatsanis, I. (2015). Short-term electricity load forecasting using time series and ensemble learning methods. 2015 IEEE Power and Energy Conference at Illinois (PECI). doi:https://doi.org/10.1109/peci.2015.7064913.

Appendix

Figure A-1. Sum of Sales by Store Type

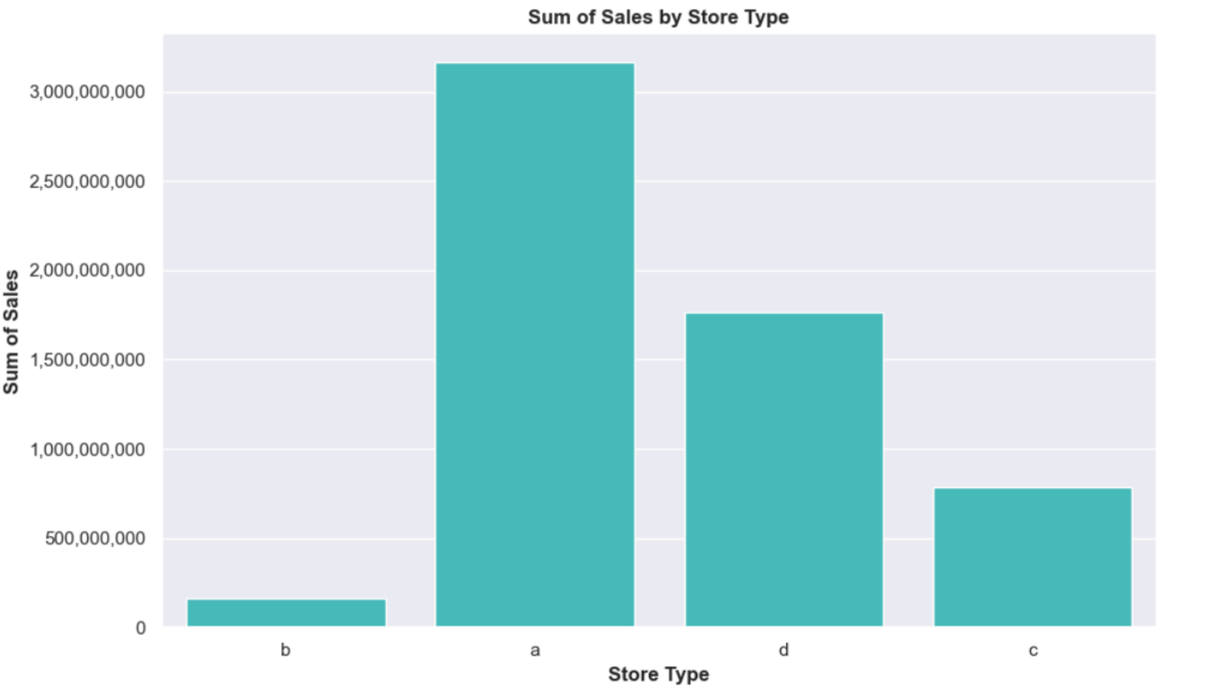
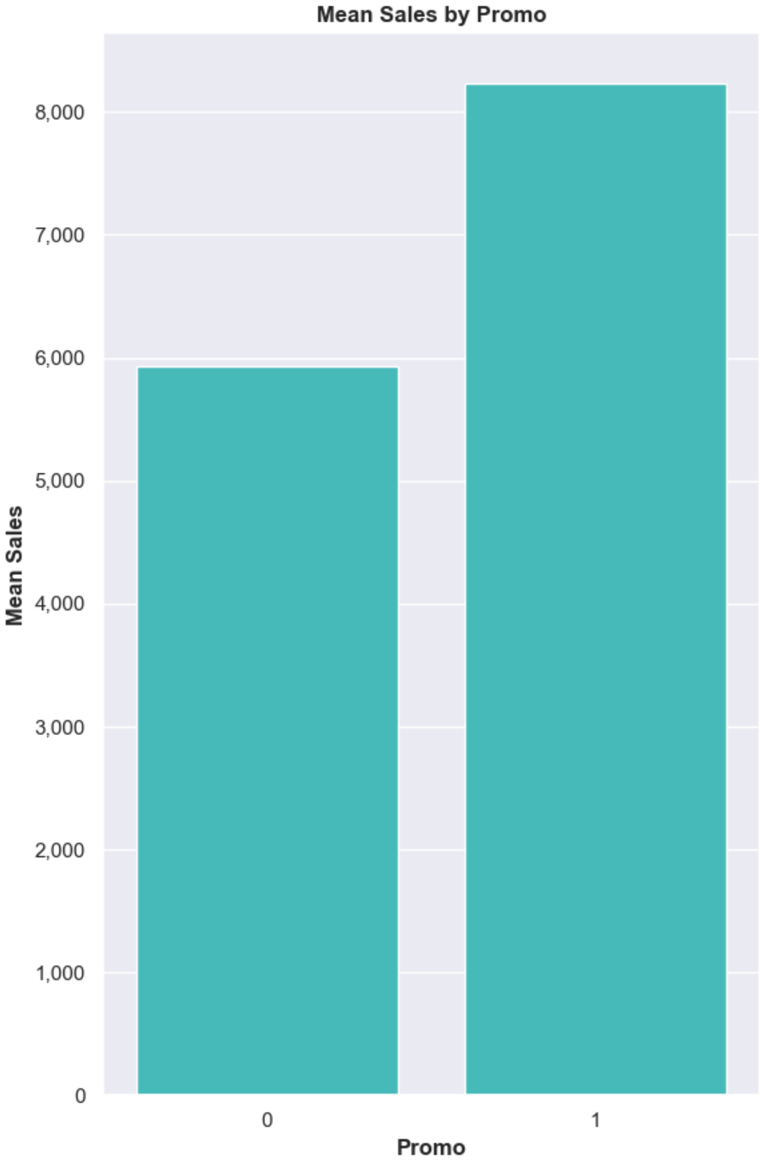


Figure A-2. Mean Sales by Promo


Mean Sales by Promo
Mean Sales
Promo