Project Report

Rossmann Store Sales

Team: Mavin
Mayur Rajwani (MT2018060)
Ankit Ramtriya (MT2018016)
Vaibhav Kansagara (IMT2016068)

December 2018

Contents

1 Introduction **1**

2 Problem Statement **1**

3 Dataset Description **1**
3.1 Imputing Null Values 3

4 Exploratory Data Analysis **3**

5 Feature Engineering **7**

6 Feature Selection **8**

7 Model Building **9**
7.1 Result Evaluation Metrics 9

8 Results and Analysis **11**

9 Future Work **11**

10 References **11**

List of Tables

1 Training Dataset 2
2 Store Dataset 2
3 Linear Regression Metrics 9
4 Random Forest Regression Metrics 10
5 XGBoost Metrics 10

List of Figures

1 Sales Distribution 3
2 Sales and Holidays Distribution 4
3 Sales-Promo Distribution 4
4 Sales and Customer Trend 5
5 Sales Distribution Varying with Store Type 5
6 Store Assorment with Sales and Customers 6
7 CorrelationMatrix(Store, Sales, CompetitionDistance) 6

1 Introduction

Rossmann operates over 3,000 drug stores in 7 European countries. Currently, Rossmann store managers are tasked with predicting their daily sales for up to six weeks in advance. Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality. With thousands of individual managers predicting sales based on their unique circumstances, the accuracy of results can be quite varied.

2 Problem Statement

Given historical sales data for 1,115 Rossmann stores. The task is to predict the "Sales" for one day. The Dataset is publicly accessible at the following link:

https://www.kaggle.com/c/rossmann-store-sales/data

3 Dataset Description

We are provided with historical sales data for 1,115 Rossmann stores. The Dataset comprises of two parts. One part is sales data of each store from 01/01/2013 to 07/31/2015. This part of data has about 1 million entries with multiple features that could impact sales. Table1 describes all the fields in this training data.

The second part of dataset consists of store information. It provides store information like store type, competitor, promotional info etc. Table2 describes all the fields of store data.

Train Dataset	
id	An Id that represents a (Store, Date) duple within the test set
Store	A unique Id for each store
Sales	The turnover for any given day (this is to be predict)
Customers	The number of customers on a given day
Open	An indicator for whether the store was open: 0 = closed, 1 = open
State Holiday	Indicates a state holiday. Normally all stores, with few exceptions, are closed on state holidays. Note that all schools are closed on public holidays and weekends. a = public holiday, b = Easter holiday, c = Christmas, 0 = None
School Holiday	Indicates if the (Store, Date) was affected by the closure of public schools

Table 1: Training Dataset

Store Dataset	
Store	A unique Id for each store
StoreType	differentiates between 4 different store models: a, b, c, d
Assortment	describes an assortment level: a = basic, b = extra, c = extended
Competetion Distance	distance in meters to the nearest competitor store
Competition OpenSince [Month/Year]	gives the approximate year and month of the time the nearest competitor was opened
Promo2	Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating
Promo2 Since [Year/Week]	describes the year and calendar week when the store started participating in Promo2
PromoInterval	describes the consecutive intervals Promo2 is started, naming the months the promotion is started anew. E.g. "Feb,May,Aug,Nov" means each round starts in February, May, August, November of any given year for that store

Table 2: Store Dataset

3.1 Imputing Null Values

Dataset consists of null-values. We impute null values in following manner:

- **CompetitionOpenSince[Month/Year]**: Since these values indicate from how long a store has a competitor, we impute these values with the oldest record of stores we have.
- **CompetitionDistancce**: The stores for which competition distance is null, we impute these with median value of competition distance. Competition distance is skewed distribution, and thus imputing with mean doesn't makes sense.
- **Promo2Since[Week/Year]**: Theses are imputed with zero, as they are null when Promo2 doesn't apply.
- **PromoInterval**: These values are also imputed with zero, due to mapping to categorical integer values. Else, they signify non-existence of PromoInterval for a particular store, when null.

4 Exploratory Data Analysis

1. **Sales**
   We are primarily interested in predicting sales. Therefore, we take a look at Sales trend over last few years, and plot its distibution.

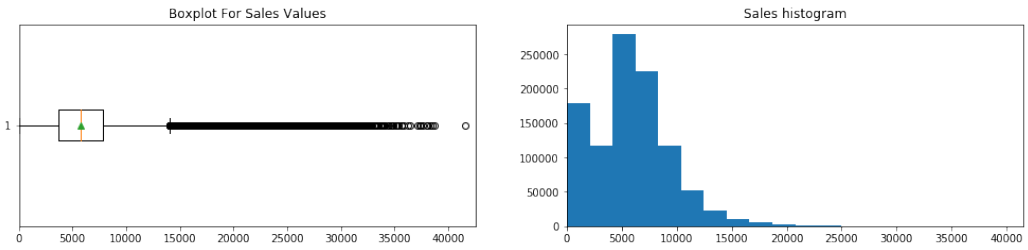
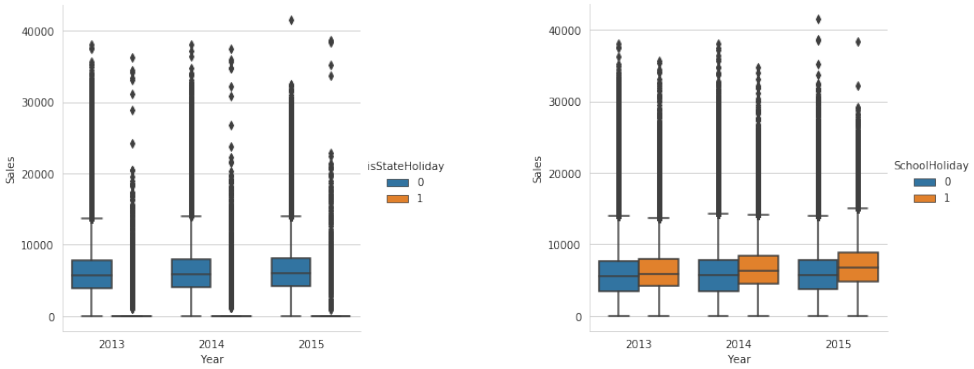


Figure 1: Sales Distribution

   Take-away: Sales data is right skewed. In order to make dataset less skewed and resemble normal distribution we consider price on a logarithmic scale.

2. **Holidays**
   We first take holidays into consideration and check if they have any effect on sales.

StateHoliday: Starting with StateHoliday, we check how StateHoliday (Figure 2) affect the overall sales of the store.



(a) Sales-StateHoliday (b) Sales-SchoolHoliday

Figure 2: Sales and Holidays Distribution

SchoolHoliday We also look at SchoolHoliday (Figure 2) and check if it affect our sales.

3. **Promo**
   We also look at how Promo affects our sales.

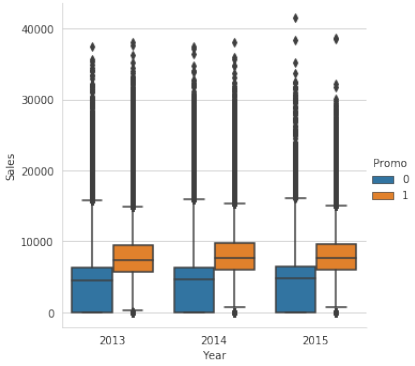


Figure 3: Sales-Promo Distribution

Take-Away: Sales are affected by School Holidays and Promotions and not due to State Holidays.

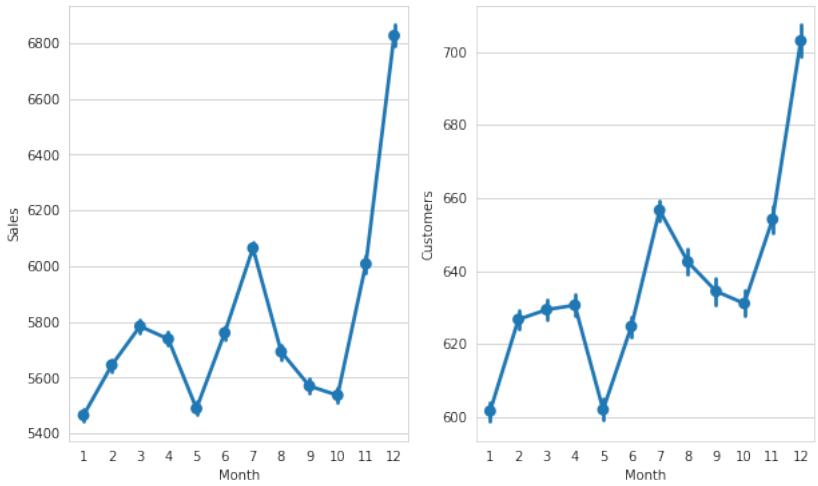


Figure 4: Sales and Customer Trend

4. **Sales and Customer Trend**
   We take a glance at how Customers and Sales are varied over the months (Figure 4).
   Take-away: A sharp rise in customers and sales towards the year end, may be due to christmas. It will be a useful feature in our prediction.

5. **Store Information**
   The following figures describes how sales varies with store type.

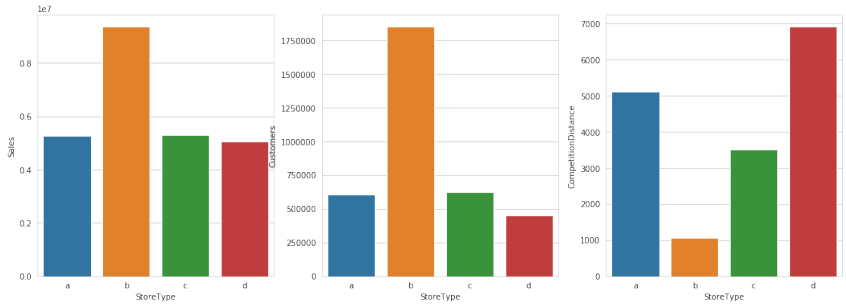


Figure 5: Sales Distribution Varying with Store Type

6. **Store Assortment**
   Stores also contain a feature describing the assortment it is provided with. Figure 6 describes the assortment levels per store type along with customers they have and sales they make.

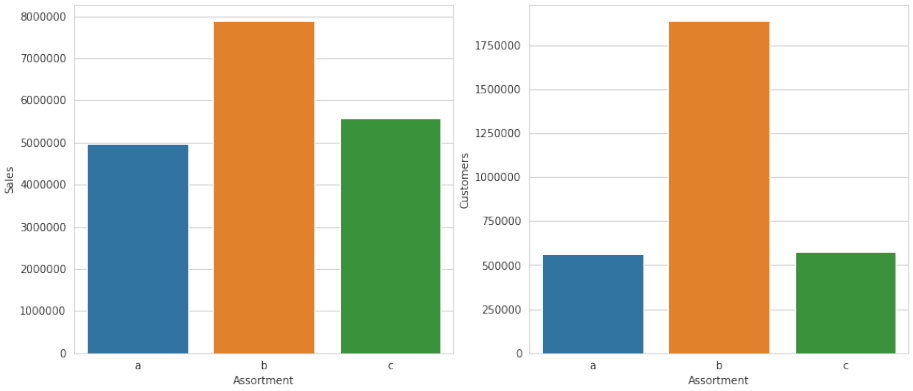


Figure 6: Store Assorment with Sales and Customers

7. **Correlation between Customers and Sales**
   Do Customers and Sales have any correlation? From EDA done before, these two parameters doesn't seem to disagree much on any feature. Lets have a look.

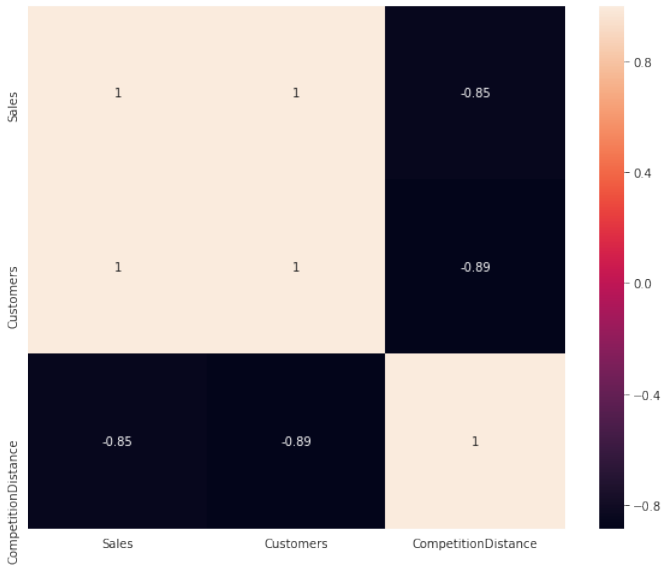


Figure 7: CorrelationMatrix(Store, Sales, CompetitionDistance)

Take-Away: Customers and Sales are directly correlated to each other. Also note that Competition Distance is negatively correlated to Customers and Sales. Hence if competition distance is more, it suggests lower sales and fall in customer count.

5 Feature Engineering

In order to improve the predictions, We merged store dataset with sales dataset. We also created several features to provide deeper insights on the dataset. These includes:

- Disintegrating date parameter to explore our dataset on a daily, monthly, quarterly and yearly basis. It provide the trends that sales follows.
- Features CompetetionOpenSinceYear and CompetetionOpenSinceMonth are merged to create a single feature - *CompetetionOpen*. It indicates, the presence of competitor in months (other measure could be in terms of weeks). That is, how long before (in months) competitor store was opened.
- Similarly, features Promo2SinceYear and Promo2SinceWeek was combined to form *Promo2Open*, which describes the number of weeks promo2 is running at a store.
- We also added timeline features like, *Is_year_end*, *Is_quarter_start* describing if the sale is made in year end or start of the quarter.

6 Feature Selection

Before we begin predicting sales, we need to identify key features that affect the sales of stores. These features are selected on the basis of Exploratory Data Analysis carried out before. We breifly enumerate the key features we picked for model building.

- **DayOf[Year/Month/Week]**: Since we need to predict on a daily basis, these should be our first key features.
- **[State/School]Holiday**: As can be inferred from Exploratory Data Analysis StateHoliday do not affect the sales much. Also, SchoolHoliday affect the sales marginally. Thus, this feature might be considered but with low priority.
- **Promo**: We saw promotions do affect the sales. Thus, this would be another key feature.
- **CompetitionDistance**: Since CompetitionDistance is negatively correlated to sales, we consider this as feature in prediction.
- **CompetitionOpen**: CompetitionOpen combine CompetitionOpenSin- ceYear and CompetitionOpenSinceMonth to proved a single variable that helps in reducing number of decisions made for each feature. As a matter of fact, this feature resulted in improving the accuracy of our prediction. Hence, a key feature to include.
- **Promo2Open**: This feature was created by merging Promo2SinceYear and Promo2SinceWeek to provide single quantitative measure of how long a promo2 is running at store, similar to store CompetitionOpen.
- **Other Features**: In order to predict sales for rossmann 1,115 stores, we need to take into account various store related information like **StoreType**, **PromoInterval**, **Assortment**. These are included in the feature set used to train data.
- **Extracted Features**: Features like `Is_year_end` appears to help in predicting the sales. Thus, we include features extracted from the date like `Is_year_end`, `Is_quarter_start`, `Is_quarter_end`

7 Model Building

We used, Linear Regression, Random Forrest, and XGBoost as our baseline models. The Following tables summarises the prediction accuracy

7.1 Result Evaluation Metrics

RMSPE or Root Mean Sqaure Percentage Error is used to evaluate the prediction quality. It takes percentage error of predicted sales to actual sales, and then calculates standard deviation. The RMSPE is calculated as

$$RMSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2} \tag{1}$$

where `y_i` denotes the sales of a single store on a single day and `yhat_i` denotes the corresponding prediction.

Linear Regression		
	Training	**Testing**
Basic	0.42053858	0.46269652
Removing Features [State/School]Holiday	0.42051035	0.46270961
Adding feature YearEnd QuarterStart QuarterEnd	0.41859202	0.46098522

Table 3: Linear Regression Metrics

Random Forest Regression		
	Training	**Testing**
Basic	0.13516165	0.58107987
Removing Features [State/School]Holiday	0.1345911	0.58620371
Adding feature YearEnd QuarterStart QuarterEnd	0.13488188	0.58589789

Table 4: Random Forest Regression Metrics

XGBoost		
	Training	**Testing**
Basic	0.30316623	0.47435655
Removing Features [State/School]Holiday	0.29828193	0.52706864
Adding feature YearEnd QuarterStart QuarterEnd	0.30019115	0.47583333

Table 5: XGBoost Metrics

8 Results and Analysis

We did training of dataset using various models. The test error feedbacked from the ongoing competetion is 0.07696. Another model that improves on validation test set accuracy fetched an score of 0.22 on kaggle.
From our study we found that features does affect the quality of training model, provided features are chosen and formatted carefully. As an example, including PromoInterval instead of categorising it in ordered manner resulted in bad accuracy of our model.

By adding few features like `is_year_end`, the dataset didn't reflect on validation set. But those features improved our score on kaggle leaderboard. As a matter of fact we considered models, that includes our best model, the other one providing better features of importance than previous one (It perfomed better on our training and validation data, but resulted in lower kaggle score).

9 Future Work

We believe the sales do get affected by holidays especially christmas. The only fact is lack of enough evidence (2-3 days of christmas over entire dataset) suggests, it would be a key feature in future models. Also, dataset from various sources providing list of holidays in a particular state might help in improving predictions.

10 References

- Data Source: *www.kaggle.com*
- Machine Learning hands on tutorials: *www.datacamp.com*
- GEN511 Machine Learning lecture notes.
- The Elements of Statistical Learning by Trevor Hastie, Robert Tibshirani, Jerome Friedman
- The pickled model files can be downloaded from *https://bit.ly/2zMYaes*