Sales Prediction of Rossmann Stores

Chetanraj Kadam
Department of Computer Science
Syracuse University

Abstract- In this project I addressed task of predicting daily sales of around thousand stores using various data science strategies and techniques such as data retrieval, pre-processing, exploratory data analysis and machine learning. Also I discussed impacts of particular methods or features on the result.

I. INTRODUCTION

Rossmann is drug store chain operates over thousand stores in 7 European countries. The goal of Rossmann challenge from Kaggle is to predict 6 weeks of daily sales for 1,115 stores in advance.[1]
Daily sales based on many factors such as holidays, location, seasonality, promotions and competition.
The problem here is to create a robust prediction model which will help store managers to focus on important aspects of planning and growth of business.

The reason for selecting topic is wide variety of its application. Sales forecasting is application that can help to almost any business or company. Traditionally, companies need to rely on manual forecasts to predict their sales which are not of good quality and usually not helpful in growth of business. Also for large business the quality of results are not consistent to each other. For example, in case of Rossmann it has 1000 stores and that much store managers. If individual store managers predict sales based on their own parameters then results will not be consistent with each other and also not helpful for deciding strategies as a company. Traditional standard analysis tools are also not much helpful in predicting sales accurately considering many factors impacting sales. So it is very important real world problem which needs to be solved by using Data Science strategies.

The focus of project to carry out detailed analysis of data, selection of important features, building regression models for sales prediction and then analyze results with important features. I am using mainly 3 regression techniques: Linear Regression, Decision Tree Regression and Random Forest Regression.
Evaluation of models will be done using RMSPE metric:

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

Dataset provide in the challenge is of two parts. First one is “Store” data which gives supplemental information about stores and other is “Train” data which includes all historical data including sales.

Training dataset is about 1.02 million entries whereas test dataset is about 41.1k entries. However test dataset is excluding sales column and mainly purpose of challenge. So I have divided training dataset into train and test by using 80:20 split.

Followings are details about given data and features:

Store.csv:

Field Name	Description
Store	a unique Id for each store: integer number
StoreType	differentiates between 4 different store models: a, b, c, d
Assortment	describes an assortment level: a = basic, b = extra, c = extended
CompetitionDistance	distance in meters to the nearest competitor store
CompetitionOpenSinceMonth	gives the approximate year and month of the time the nearest competitor was opened
CompetitionOpenSinceYear	
Promo2	Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating
Promo2SinceWeek	describes the year and calendar week when the store started participating in Promo2
Promo2SinceYear	
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

Train.csv:

Field Name	Description
Store	a unique Id for each store: integer number
DayofWeek	the date in a week: 1-7
Date	in format YYYY-MM-DD
Sales	the turnover for any given day: integer number (This is what to be predict)
Customers	the number of customers on a given day: integer number (this is not a feature. Based on the test data from Kaggle, this feature is not included in test data)
Open	an indicator for whether the store was open: 0 = closed, 1 = open
Promo	indicates whether a store is running a promo on that day: 0 = no promo, 1 = promo
StateHoliday	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
SchoolHoliday	indicates if the (Store, Date) was affected by the closure of public schools: 1 = school holiday, 0 = not school holiday

II. PRIOR WORK

As a prior work on task of predicting sales from historical data I found couple of papers very interesting and also surveyed various techniques used for this kind of prediction tasks.

The novel approach has been discussed in Sales Prediction with Data Mining Techniques paper. [2] The researchers carried out sales forecasting for Chinese online shipping company. Paper discusses step by step methods carried out in the processes of feature selection to building of sales prediction system. Researchers discuss several strategies in experimenting with models to improve its accuracy.

Another research paper I referred is "The Research of Regression Method for Forecasting Monthly Electricity Sales Considering Coupled Multi-factor".[3] This paper addresses sales forecasting techniques on data of electricity sales. The problem is on similar terms of the topic of my project. Detailed discussion on implementation of Regression method carried out in this paper for the task of sales prediction.

I also refereed various articles and tutorials on Random Forest regression method which I was unfamiliar with before the project. It helped me to gain understanding of regression model and implementation of it. I also looked upon interviews of winners of this challenge and learned their ways of approaching this problem. Most of them used more advanced and large number of computing models to carry out this task.

III. METHODS

- **Data Preprocessing**

- I started data preparation with reading of stores and train datasets into separate data frames with including feature for parsing of dates.

- Handling of missing values: I first performed analysis on both to check number of missing values in each data. Features in train data did not have any missing values. Features in stores such as CompetitionOpenSince and Promo2Since contained half of missing values. As these are dates so cannot be replaced by mean or 0, so I removed those features from the data.

- Feature CompetitionDistance having very small number of missing values. So I first analyzed distribution of its values and found that lot of values are around its median. So I replaced missing values with median.

- Then I removed number of entries for stores those were closed as sales for them was zero and will not helpful in prediction task. On similar lines I removed rows for which sales was 0. This is because it will make our models biased towards outliers.

- Extracted 3 new columns for month, year and day from date column and dropped date column.

- To train our models we need features from both of stores and train data. So I joined both of the table using left join.

- **Addition of features:**

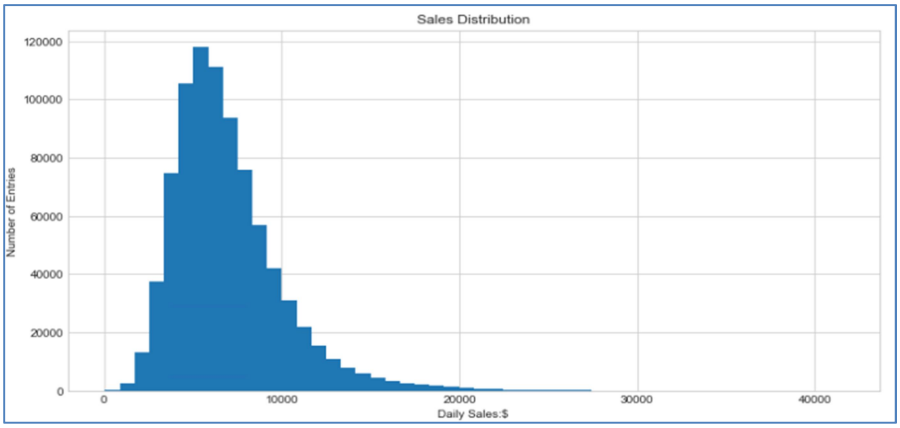
- Added “AvgSales” feature which is calculated by taking monthly averages from each store. These proved important when I visualized relation between average sales and actual sales for each store.

- Also added “AveCustomer” number which gives monthly average of customers visited to the store. This can be used to identify popularity of store.

- Performed label encoding for categorical attributes such as StateHoliday, Assortment and StoreType. This also important to prepare data before giving input to our regression models.

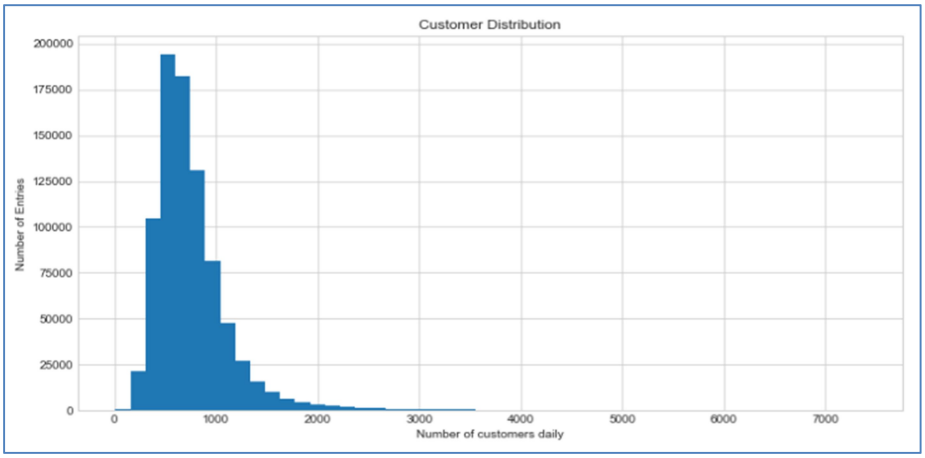
- **Exploratory Data Analysis:**

- I started with plotting number of sales over all open stores from given entries and obtained histogram to get a visual idea.

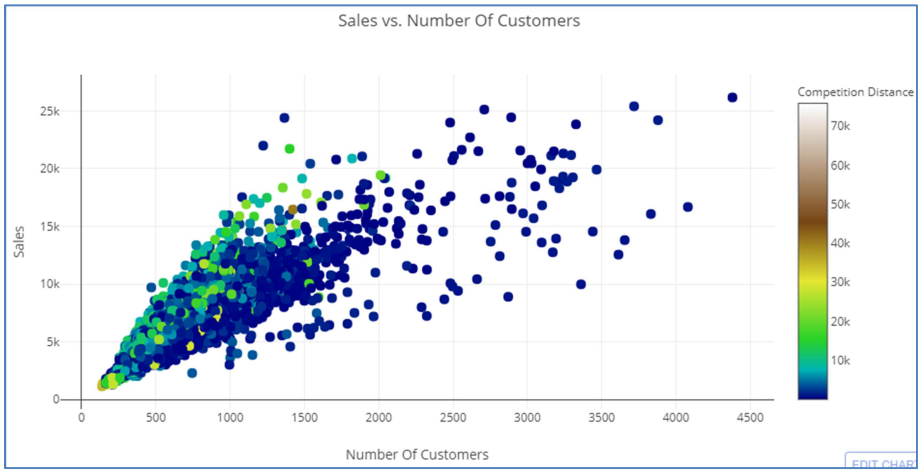


From histogram it can be observed that high number of stores having daily sales of around 5/6 thousands.

- Then I obtained similar kind of histogram for customer distribution.



- Then I performed analysis of how sales are related to number of customers with including distance from competition in terms of color shades.
-



So here we can observe that number of customers and sales are highly correlated and so number of customers will have maximum impact on sales. As in challenge in test data number of customers is not given as feature, we will not use this in our model.

- I also plotted Sales vs Competition distance plot and Sales vs Promotional offeres.

- **Regression Models:**

- I first started with splitting of data into train and test subsets. Then I defined functions for evaluation using Root Mean Square Percentage Error (RMSPE) and Mean Absolute Percentage Error (MAPE).

- I used three different regression models for the sales prediction task. Basic idea behind using 3 different regression models is to compare performance of different techniques on this task and then selection of best model to extract results.

1. **Linear Regression:**
   I implemented linear regression using Scikit-Learn library. Model was trained using training dataset and then evaluated against test dataset. This model used as baseline model for performance comparison.

2. **Decision Tree Regression:**
   This method used to evaluate performance of decision trees in sales prediction. Only parameter tuning is required for 'min sample leaf'. Both of linear regression and decision tree regression models take very less time to train and evaluate.

3. **Random Forest Regression:**
   To implement random forest I used all data including categorical values with encoding. Used Scikit-learn's ensemble model. The parameter tuning required for 'n_estimators' parameter which tells number of trees in the forest. Training time required by random forest was much greater than above two models and which was expected.

Apart from these I also wanted to use gradient boosting regression method but due constraint in time and resource setup, I not implemented that. It will be my future work on this project.

IV. RESULTS & FINDINGS

Lot of time in this project spent on selection of appropriate features and then experimenting with regression models. Implementation of models took less time as compared to process of experimenting with it.

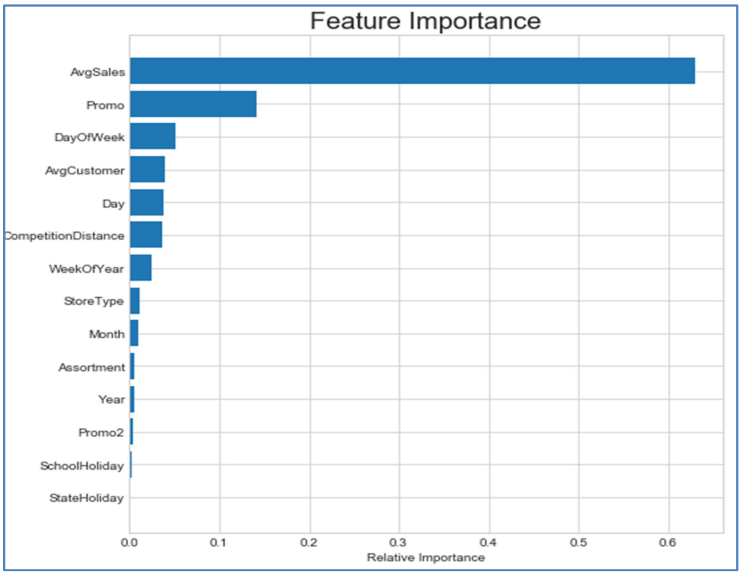
I used Root Mean Square Percentage Error (RMSPE) and Mean Absolute Percentage Error (MAPE) to evaluate 3 models. Performance I got out best models using best feature selection is as following:

	RMSPE		**MAPE**	
Model	**Training**	**Test**	**Training**	**Test**
Linear Regression	0.235	0.235	17	17.03
Decision Tree Regression	0.1312	0.1516	9.6655	11.03
Random Forest Regression	0.047	**0.1187**	3.359	8.828

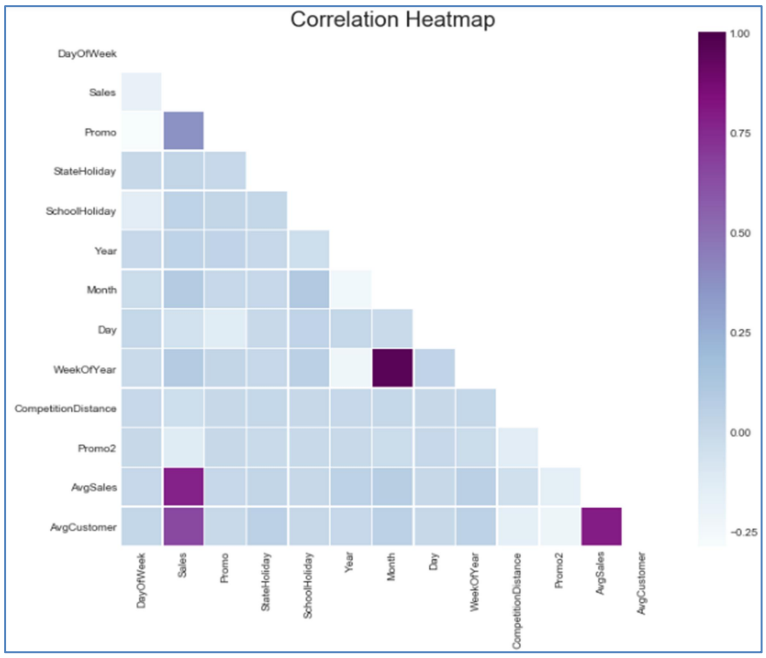
Table1. Performance comparison

By using proper features from dataset and model hyperparameters I achieved RMSPE score around **0.1187** which can be in top 10% result got using random forest on this challenge. From table 1, we can also look out for performance of decision tree regression which is far better than baseline linear regression model.

As random forest model performed best, I choose it to carry out analysis of result in perspective of features. So I extracted feature importance considered by random forest regressor as follows:



We can observe from this plot that our newly added feature "AvgSales" proved to very important in terms of predicting daily sales. From this we can observe fact that if we know history of average sales for particular store then we can predict that store's sales accurately. The same thing can be observed from correlation matrix between these features. AvgSales, AvgCustomer and Promo are highly correlated to Sales.



Common observation drawn from result is also that some stores always perform better than others. This might be because of factors unique to that particular store and not represented in given data.

After AvgSales second best feature was Promo. Sales are highly correlated to promotion from store. Sales are observed to be increased around 30% in the period of promotions. So regression models learned this feature so whenever there is promotion for store it is predicted high sales for that store. This can be important thing to note for store managers as a future strategy to improve business.

From feature importance we observe that 'DayOfWeek' is also very important in prediction of sales. In data analysis step it was observed that Sunday and Monday are day when sales were highest (> 8000$) and Saturday is at lowest (< 6000$). The difference of sales is almost 40% between these days. So this feature proved important in regression model. This can also very important in terms of planning for store. As sales are predicted to be up on Sunday and Monday, it is clear that number of customers will also be large, so managers can employ more staff on these days.

From results we not observed much of seasonality even small increase of sales in December. But which can also be valid considering fact that these are drug stores. So will not be much affected by festival season. But it is observed that AvgSales for Rossmann stores are increasing year by year from 2013-2015. So as a company it is doing well so AvgSales show growth of 3-5% each year. This is important in predicting future sales.

While performing experiment addition and removal of features such as StoreType, Assortment, Month and Year in training data to models not shown much impact on results. Changes in RMSPE score very less in each scenario and even not consistent. So we can say that that these features are not impacting sales that much.

In conclusion this is very interesting Data Science problem to solve where feature selection is important part of prediction accurately. I was able to successfully forecast daily sales for these stores using Linear Regression, Decision Tree Regression and Random Forest Regression models. I also analyzed trends and useful insights from results which can be prove to be important for Rossmann. In future I can work on other regression methods such as gradient boosting to improve results further. Also there is scope to add external real world features about these stores to predict future sales more accurately.

REFERENCES

[1] https://www.kaggle.com/c/rossmann-store-sales

[2] A Novel Trigger Model for Sales Prediction with Data Mining Techniques, Wenjie Huang1, Qing Zhang1, Wei Xu1, Hongjiao Fu1, Mingming Wang1 and Xun Liang1

[3] The Research of Regression Method for Forecasting Monthly Electricity Sales Considering Coupled Multi-factor, Jiangbo Wang1, 2, Junhui Liu1 , Tiantian Li1 , Shuo Yin1 and Xinhui He3.

[4] Decision Trees and Random Forests for Classification and Regression, https://towardsdatascience.com/decision-trees-and-random-forests-for-classification-and-regression-pt-1-dbb65a4