Rossman Stores

Amitha Nayak
Computer Science Engineering
PES University ECC
Bangalore, India
nayak.amit.blr@gmail.com

Akshaya J
Computer Science Engineering
PES University ECC
Bangalore, India
akshaya2715@gmail.com

Jigya Shah
Computer Science Engineering
PES University ECC
Bangalore, India
jigyas15@gmail.com

Samyuktha Prakash
Computer Science Engineering
PES University ECC
Bangalore, India
samyukthaprakash.h@gmail.com

***Abstract*—In this project, we applied machine learning techniques to a real world problem of predicting stores sales. This kind of prediction enables store managers to create effective staff schedules that increase productivity and motivation. Given store information, and sales record we applied Multi Linear Regression, Random Forests and Gradient Boosted Trees, and tried to predict sales.**

Keywords— Multi Linear Regression, Gradient Boosted Trees, Recurrent Neural Networks, Hidden Markov Models, Rossmann Sales

I. Introduction

This dataset is about Rossmann Stores, which 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.

A brief overview of the data fields in our dataset :

a) Id - an Id that represents a (Store, Date) duple within the test set
b) Store - a unique Id for each store
c) Sales - the turnover for any given day (this is what you are predicting)
d) Customers - the number of customers on a given day
e) Open - an indicator for whether the store was open: 0 = closed, 1 = open
f) 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
g) SchoolHoliday - indicates if the (Store, Date) was affected by the closure of public schools
h) StoreType - differentiates between 4 different store models: a, b, c, d
i) Assortment - describes an assortment level: a = basic, b = extra, c = extended
j) CompetitionDistance - distance in meters to the nearest competitor store
k) CompetitionOpenSince[Month/Year] - gives the approximate year and month of the time the nearest competitor was opened
l) Promo - indicates whether a store is running a promo on that day
m) Promo2 - Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating
n) Promo2Since[Year/Week] - describes the year and calendar week when the store started participating in Promo2
o) 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.

II. Related Work

A. Assumptions made

In this paper [2] they have discretized the data fields to work with the hidden Markov model and RNN. A smaller discretization leads to the potential ability to be close to the true value. However,it also decreases the probability estimation between states as there are more states.

In this paper [3] the scoring method of the competition differed from the loss function included in the gradient boosted tree library resulting in transforming our target variable or writing our own loss function. So, they used a simple log transformation on the target variable as that still gave similar results as that of the scoring method in the competition.

B. *Approach*

In paper [1], we use a simple Multiple Linear Regression (MLR) model. Linear regression is often used as a model for realizing the correlation between input and output numerical variables.

In paper [2], three different methods for forecasting store sales of Rossmann stores have been examined which need to be applied:
1) Random Forests
2) Hidden Markov Models
3) Recurrent Neural Networks.

Random Forest Regression is a learning algorithm that operates by aggregating many random decision trees to make predictions while avoiding overfitting. Hidden Markov Models and Recurrent Neural Networks predict the target variable using time-series data.

In paper [3], Gradient Boosted Trees algorithm has been implemented. Gradient boosting involves combining a large number of decision trees to produce the final prediction. It is a type of machine learning algorithm that relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error.

In paper [4], a comparative study on sales forecasting using Regression Models is implemented to boost the precision and effectiveness of prediction of sales.
The Regression Models used are: Ridge Regression and LASSO Regression.
Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. It is an extension of linear regression where loss function is improved to minimise the complexity of the model.
Lasso Regression is an analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the model it produces.

C. *Summary*

In paper [1], by using a linear regression ML algorithm, future stock values of APPLE and TSLA were predicted using past values.

In paper [2] the data fields have been discretized to work with the hidden Markov model and RNN. A smaller discretization leads to the potential ability to be close to the true value. However, it also decreases the probability estimation between states as there are more states.

In paper [3], SigOpt Bayesian Optimization was used to modify the hyperparameters of the gradient boosted trees for further reduction of prediction error.

In paper [4], a contrastive analysis on sales prediction involving forecasting is implemented using Supervised Machine Learning.

III. Problem Statement

The aim is to minimize prediction error for a data science competition involving predicting store sales.

A. *Approach*

After going through the research papers, Multi Linear Regression, Recurrent Neural Networks, Hidden Markov Models, Gradient boosted Trees and Lasso Regression models have been used. These models have been proven to show the best results for prediction of Rossmann Sales. For Rossmann Sales dataset Gradient boosted trees [3] have shown the least prediction error.

For Linear regression and Recurrent Neural Networks, the features which show a positive correlation as per the heatmap done using EDA are used for analysis.

For Recurrent Neural Networks and Hidden Markov Models, the data column ‘Date’ has been split into various related features (Day, Month, Year, Season..etc) to create a time series related analysis.

Machine learning datasets of this scale usually work well with weak learners (tree stumps). The concept of ensemble learning was used to combine multiple trivial decision tree stumps to create a gradient boosted tree.

IV. Exploratory Data Analysis (EDA)

CSV files

● train.csv - historical data including Sales

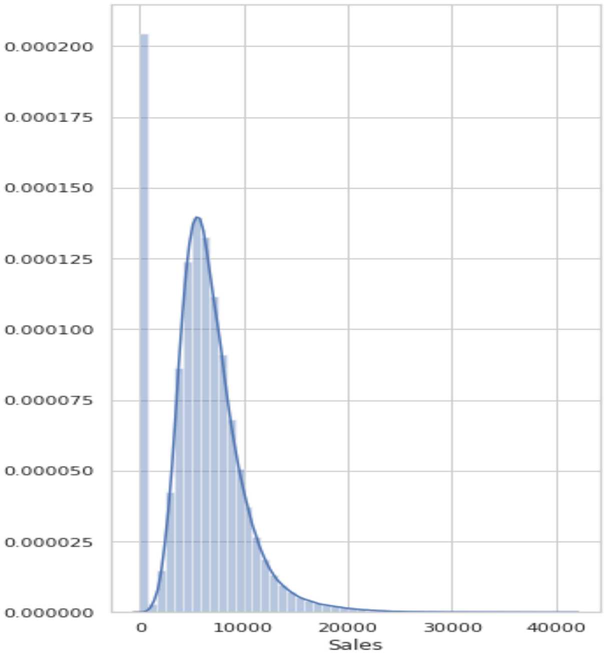
● test.csv - historical data excluding Sales

● store.csv - supplemental information about the stores

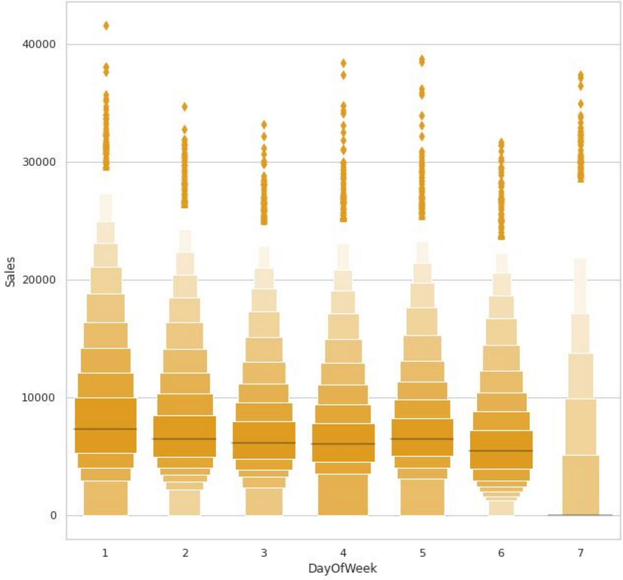
Two of these files are used for EDA, i.e train.csv and store.csv Firstly the two datasets were merged based on the ‘StoreId’ to do the visualization and analysis.

The dataset was then cleaned by getting rid of columns that have more than 30% missing values. Whereas for the columns which had a lesser amount of data missing, the 'mode' of the very same column was used as a substitution.

Outlier Detection using distplot

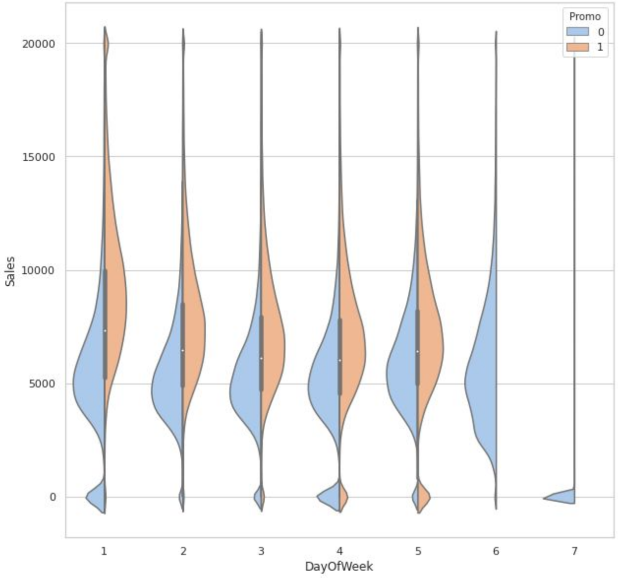


Outlier Detection using BoxenPlot



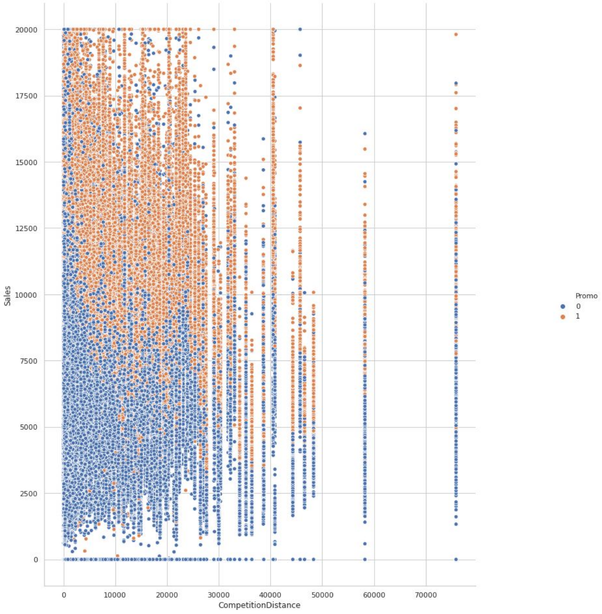
Violin Plot

The days promos were present have shown a slight improvement in Sales. This was confirmed by plotting a violin plot which shows that there was no promo offered on 6th and the 7th day of the week (Saturday and Sunday), but stores didn't suffer for doing so either, as it can be seen the no. of customers on the weekends, were more than that during the weekdays.



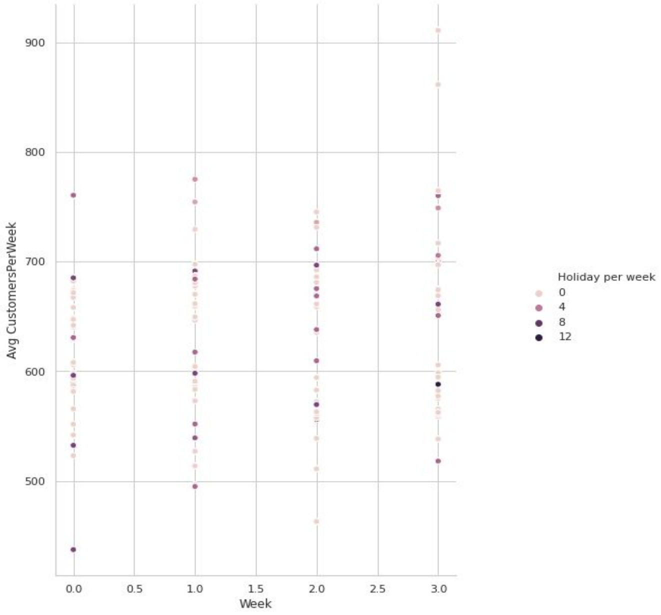
Rel Plot

Most of the stores that have very less competition Distance, but have still managed to make big Sales, by applying promo.



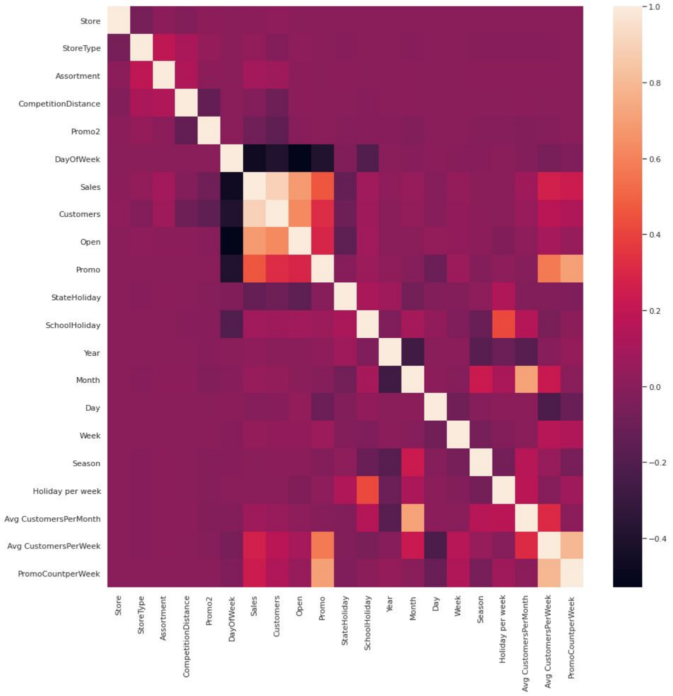
Rel Plot

Using a rel plot we found that there is no big difference in the no. of customers even if there were four School holidays a week.



Heatmap

The heatmap shows all our hypotheses were true, there is very little correlation between School Holiday, Customers and Promo, but there is a strong correlation between Promo and Sales.



References

[1] "Stock Prediction Analysis by using Linear Regression Machine Learning Algorithm", Sankranti Srinivasa Rao, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-4.

[2] "Predicting Sales for Rossmann Drug Stores", Brian Knott, Hanbin Liu, Andrew Simpson (Stanford, Harvard)..

[3] "Gradient Boosted Trees to Predict Store Sales ", Maksim Korolev, Kurt Ruegg, mkorolev@stanford.edu, kruegg@college.harvard.edu (Stanford, Harvard) Corpus ID: 16646225 .

[4] "An Effective Approach for Sales Forecasting", Mr. Faraz Hariyani, MSc IT Student and Mrs. Haripriya V, Dept of MSc IT, JAIN (Deemed to be university) International journal for research in applied science and engineering technology IJRASET Publication .