Rossmann 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, machine learning techniques were applied to predict store sales. This kind of prediction enables store managers to benefit increased profits. Given store information and sales record we applied Multi Linear Regression, Lasso Regression and Gradient Boosted Trees to predict sales.**

***Keywords*— EDA, Multi Linear Regression, Lasso Regression, Gradient Boosted Trees, Rossmann Sales**

I. Introduction

The Rossmann Stores dataset operates over 3,000 drug stores in 7 European countries. Rossmann store managers are tasked with predicting their daily sales for up to six weeks beforehand. There are many factors that influence store sales. These factors include school and state holidays, competition, promotions, seasonality and locality.

As the dataset lacked in certain cases, like providing information about the location and weather, information about the location was inferred based on holidays, and knowing the location, a weather dataset was merged accordingly.

A brief overview of the important data fields in the dataset :

a) Store - a unique Id for each store
b) Sales - the turnover for any given day (this is what we are predicting)
c) Customers - the amount of customers on a given day
d) Open - 0 if the store is closed and 1 if its open
e) StateHoliday - indicates a state holiday. a for public holiday, b for Easter holiday, c for Christmas and 0 for none
f) SchoolHoliday - indicates if the (Store, Date) was suffering from the closure of public schools
g) StoreType - The four different store models are represented by a, b, c, d
h) Assortment - describes an assortment level:a for basic, b for extra and c for extended
i) CompetitionDistance - distance in meters to the closest competitor store
j) CompetitionOpenSince[Month/Year] - the approximate year and month of the time the closest competitor was open
k) Promo - indicates if a store is running a promo on a particular day
l) State - gives the location of the store
m) Events - indicates the weather conditions on a particular day based on the location. It is represented using keywords - ‘Fog’, ‘Wind’, ‘Rain’, ‘ThunderStorm’ and ‘Snow’.

II. Previous Work

A. *Assumptions made*

In paper [2], the data fields were discretized to work with the Hidden Markov Model and Recurrent Neural Network. For the potential ability to be close to the true value, smaller discretization works better than a larger one. However, smaller the discretization the probability estimation between states also decreases as the number of states will be more.

In paper [3], there was a difference between the scoring method of the competition and the loss function which was included in the gradient boosted tree library. This resulted in the transformation of the target variable and in writing their own loss function. So, a simple logarithmic transformation was used on the target variable. This will give similar results as that of the scoring method in the competition.

B. *Approach*

In paper [1], a simple Multiple Linear Regression (MLR) model was used. The correlation between input and output numerical variables can be realised using Linear Regression as a model.

In paper [2], three different methods for forecasting store sales of Rossmann stores have been examined:
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 helps in analysing multiple regression data which suffer from multicollinearity. It is an extension of linear regression with an improvement in the loss function to minimise the model's complexity.
Lasso Regression is an analysis method and it performs both regularization and variable selection which helps in enhancing the prediction accuracy and interpretability of the model it produces.

C. *Summary*

In paper [1], by using a linear regression machine learning 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 results in 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 maximise the accuracy for prediction of store sales using various machine learning models.

A. *Approach*

After using various research papers for understanding of concepts and as a point of reference, Multi Linear Regression, Gradient boosted Trees and Lasso Regression models have been utilised. These models have been demonstrated to show the most optimum results for prediction of Rossmann Sales. For the Rossmann Sales dataset, Gradient boosted trees [3] have shown the least prediction error.

For Linear regression and Lasso Regression, the particular attributes which show a positive correlation according to the heatmap implemented using EDA are used for analysis.

For time series analysis the data column 'Date' has been split into copious associated 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 conceptualisation of ensemble learning was used to combine multiple trivial decision tree stumps to create a gradient boosted tree.

IV. Proposed Solution

A. *Exploratory Data Analysis*

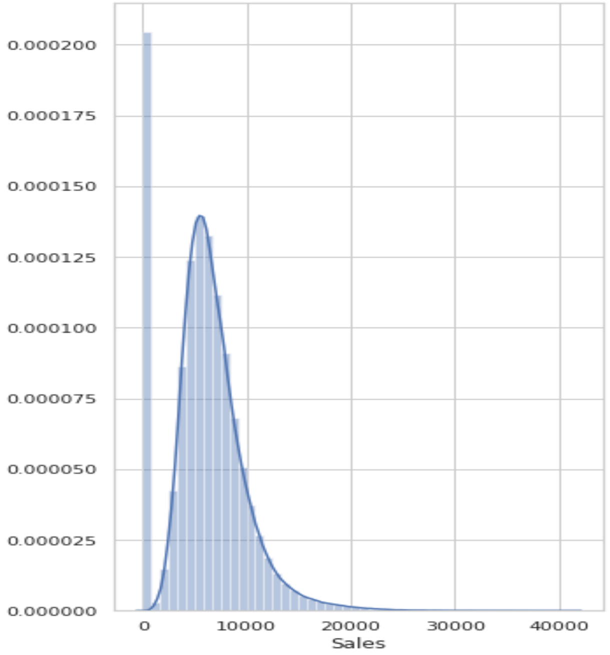
1. The dataset was cleaned by getting rid of columns ('Promo2', 'PromoSince', 'PromoInterval') that have more than 40% missing values.
2. Whereas for the columns which had a lesser amount of data missing, the previous values or the next values or the 'mode' of the very same column (depending on the situation / data column) was used as a substitution.
3. The outliers were efficiently weeded out, after visualizing it, through a boxplot.
4. The columns containing nominal categorical values were integer encoded, for the better performance of the ML models.
5. The column 'CompetitionDistance' was log transformed, as this column had huge values (10^10) whereas the rest of our dataset didn't, and this led to highly skewed visualizations.
6. Multiple additional columns were created ('Avg. number of Customers per Week', 'Avg. number of Customers per Month', 'No. of holidays per Week', 'No. of Promo per Week') for the betterment of the ML model predictions.

B. Using external data

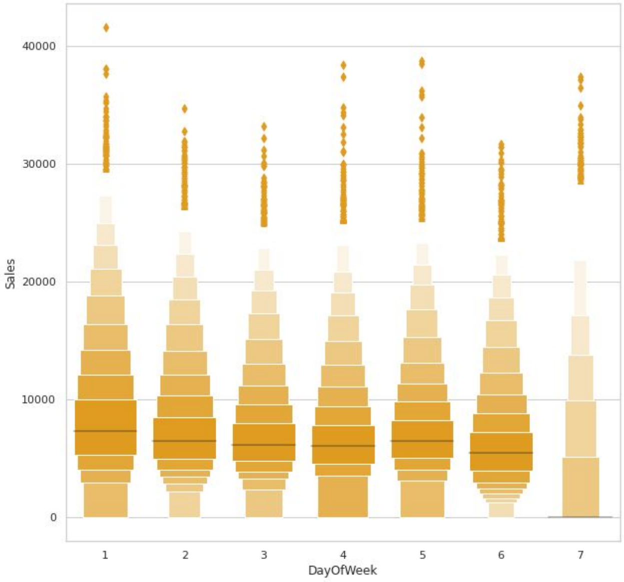
The dataset lacked in certain cases - it didn’t have any information on the store’s location, the type of items sold, the type of items the customers bought, what kind of promotional offer was running, or the weather, traffic, trends, etc. This vastly impaired the ML models and the business dashboard visuals. However, a few of the limitations were removed, by making use of external data to find the states in which the store was based, and the weather history of Germany.

1. The primary .csv files (store.csv, train.csv) were merged together using the Store ID
2. Found the state in which the store is based, by checking on which days the store celebrated a ‘State Holiday’ or a ‘School Holiday’.
3. Merged the weather dataset with the primary dataset, based on store location and dates.

Outlier Detection using distplot

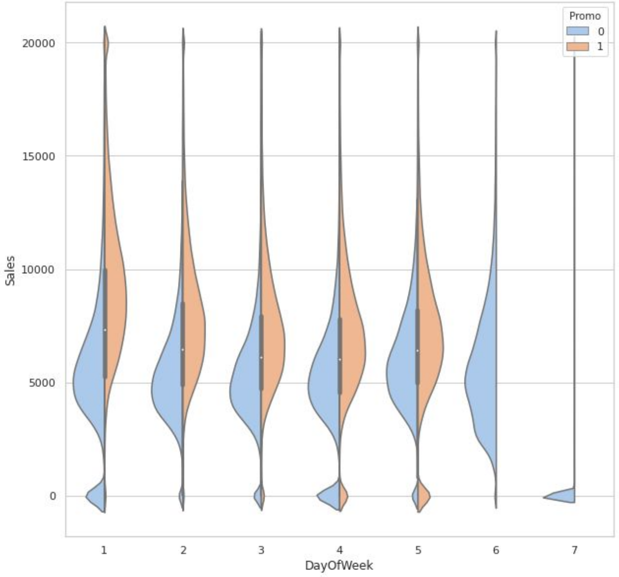


Outlier Detection using BoxenPlot



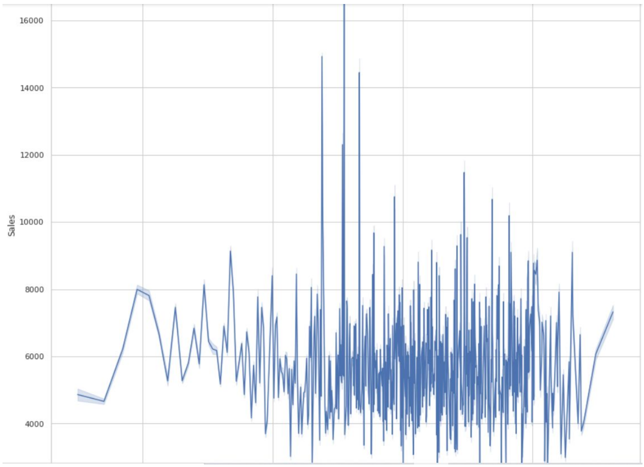
Violin Plot

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.



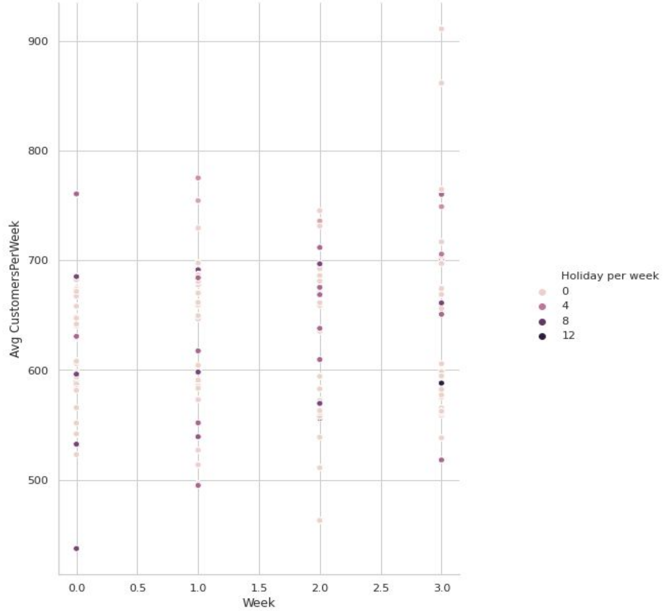
Line plot

The line plot shows us sales against the logarithm of competition distance. It shows us that the stores which have less competition distance don’t make big sales.



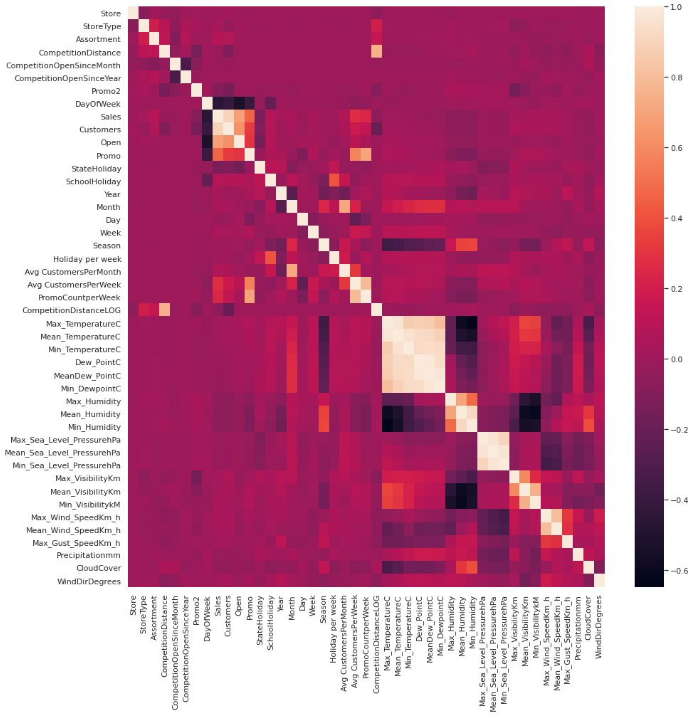
Rel Plot

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.



C. Choosing Machine Learning Models:

After referring and analysing various research papers, we found out that the models - Linear Regression, Lasso Regression, Gradient Boosted Trees, Time Series Analysis using Recurrent Neural Networks show the best results, with Gradient Boosted being the best proven model for our chosen dataset.

V. Experimental Results

As the Rossmann dataset was a time series based dataset, we sorted the entire dataset based on the Dates and then split them in a ratio of 80:20 for training and testing sets.

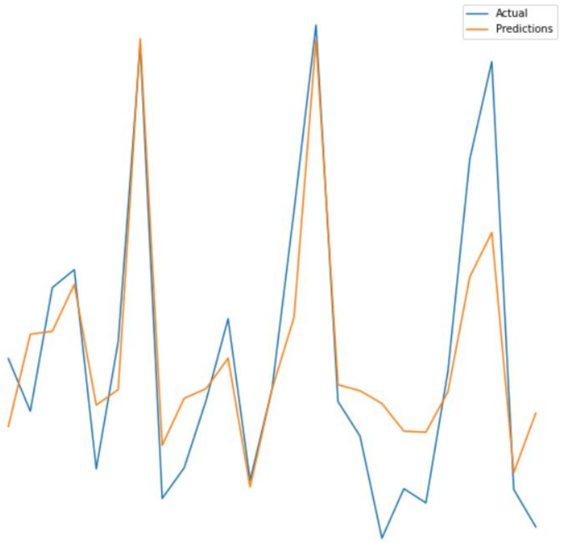
This was done in order to account for the time series based trends, which would be lost if the dataset was shuffled before splitting.

After splitting the dataset, the training set was fed into the respective models and validated using the testing set.

A. Linear Regression:

A multi linear regression model is used to find a linear relationship - best line of fit between multiple independent variables and one dependent variable which is 'sales' .

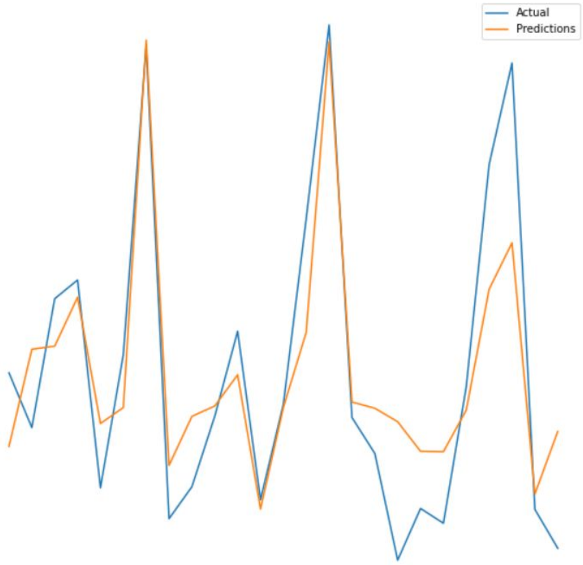
The prediction of sales was done using the Linear Regression model imported from sklearn library. The accuracy is about 86%. All the available features were used for prediction.



B. Lasso Regression:

A type of linear regression with external hyperparameters used.

Prediction of sales was done using the Lasso Regression model. The Lasso model was imported from sklearn library and obtained an accuracy of 86%. The hyperparameters used - alpha is set to 2.

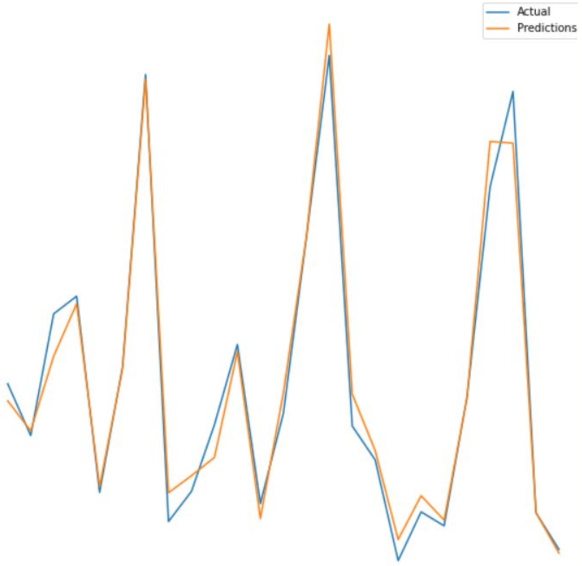


C. Gradient Boosted Decision Trees:

Light GBM is a type of gradient boosting learning model, which helps the tree grow vertically leaf-wise and not level-wise.

Light GBM model was used and an accuracy of 98% was achieved. The parameter used were -

maximum leaves in base learners is 50, n_estimators is 700, learning rate is 0.3, subsample is 1, colsample_bytree is 0.8, regularisation values are 0.1 and 1.



VI. Conclusion

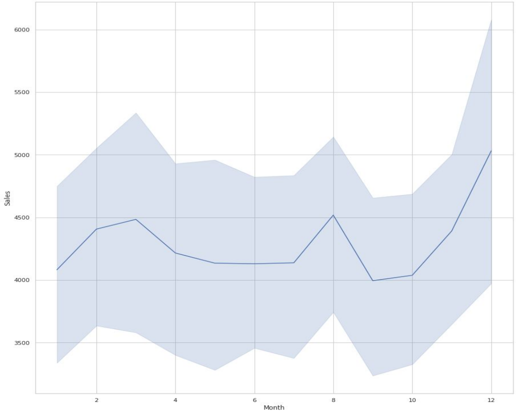
After the prediction of store sales using different models, we came to a conclusion that gradient boosted trees, Light GBM model gave the best accuracy.

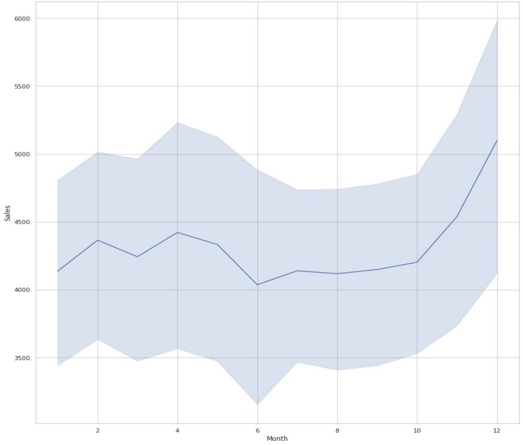
This was premeditated as Gradient boosted trees, uses ensemble learning - boosting technique. It uses multiple weak learners with high bias, uniquely used to predict Sales in certain concept spaces.

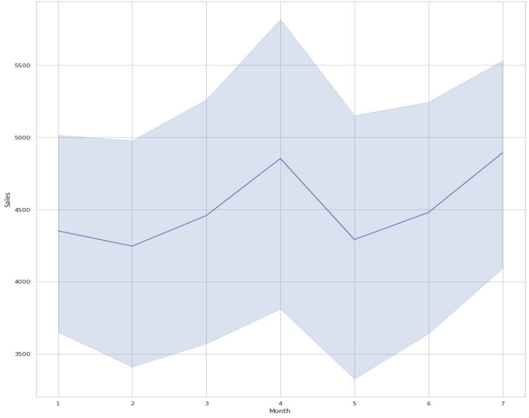
On trying to find patterns between time series and Sales, we found no regular cyclic trends.

It seems that while Recurrent Neural Networks might be useful theoretically and might find trends, it doesn't show the expected results practically for this dataset.

We didn’t find any clear trends, even after plotting sales for a random choosen store for all the years (2013, 2014 & 2015)







Contributions:

Amitha Nayak : Contributed in Exploratory Data Analysis and procured the Gradient Boosted Decision Trees model.

Akshaya J : procured the Lasso Regression model.

Jigya Shah : Contributed in Exploratory Data Analysis and procured the Gradient Boosted Decision Trees model.

Samyuktha Prakash : procured the Linear Regression model.

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 .