Literature Survey

What is the dataset about?

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 :

1. **Id** - an Id that represents a (Store, Date) duple within the test set
2. **Store** - a unique Id for each store
3. **Sales** - the turnover for any given day (this is what you are predicting)
4. **Customers** - the number of customers on a given day
5. **Open** - an indicator for whether the store was open: 0 = closed, 1 = open
6. **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
7. **SchoolHoliday** - indicates if the (Store, Date) was affected by the closure of public schools
8. **StoreType** - differentiates between 4 different store models: a, b, c, d
9. **Assortment** - describes an assortment level: a = basic, b = extra, c = extended
10. **CompetitionDistance** - distance in meters to the nearest competitor store
11. **CompetitionOpenSince**[Month/Year] - gives the approximate year and month of the time the nearest competitor was opened
12. **Promo** - indicates whether a store is running a promo on that day
13. **Promo2** - Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating
14. **Promo2Since**[Year/Week] - describes the year and calendar week when the store started participating in Promo2
15. **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.

What do we plan to do with the dataset?

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

How do we plan to go about it?

We plan to predict the “Store Sales” using different prediction models

1. Multi Linear Regression (MLR)
2. Random Forests
3. RNN, Hidden Markov Models (using Time series)
4. Gradient Boosted Trees
5. Lasso Regression

EDA:

CSV files:

- train.csv - historical data including Sales
- test.csv - historical data excluding Sales
- store.csv - supplemental information about the stores

We have used 2 of these files 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 lesser amount of data missing, the 'mode' of the very same column was used as a substitution. Later, we removed the outliers.

Are the promos effective?

The days, promos were present have indeed 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.

Does competition distance matter?

Most of the stores that have very less Competition Distance have still managed to make big Sales, by applying promo. This was done by plotting Seaborn’s Rel plot which is similar to scatter plot but lets us plot a third categorical variable.

Is there a surge of customers during SchoolHolidays?

Using a rel plot we found that there is no big difference in the no. of customers even if there were 4 School Holidays a week.

Is there an increase in promo if it is a School Holiday?

Using the rel plot we could see that the Holidays didn’t have any effect on promo and Customers.

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.

PAPER 01

Stock Prediction Analysis by using Linear Regression Machine Learning Algorithm

[https://www.ijitee.org/wp-content/uploads/papers/v9i4/D1110029420.pdf]

Source : Sankranti Srinivasa Rao, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-4

Date : February 2020

Keywords: Linear regression, Prediction, Supervised learning

The paper talks about using 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 the given paper, by using a linear regression ML algorithm, future stock values of APPLE and TSLA were predicted using past values. This paper has given me an idea on how we can apply linear regression on our chosen dataset.

PAPER 02

Predicting Sales for Rossmann Drug Stores

[http://cs229.stanford.edu/proj2015/218_report.pdf]

Source: Brian Knott, Hanbin Liu, Andrew Simpson (Stanford, Harvard)

Date : 2017

Keywords : Random Forests, Hidden Markov Models, Recurrent Neural Networks, Time Series

In this paper they have examined three different methods for forecasting store sales of Rossmann stores which we intend to apply:

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.

Whereas Hidden Markov Models and Recurrent Neural Networks predict the target variable using time-series data.

In this paper 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.

The take-away from this paper is that it shows us how to apply Random Forest Regression and RNN to our dataset.

PAPER 03

Gradient Boosted Trees to Predict Store Sales

[http://cs229.stanford.edu/proj2015/193_report.pdf]

Source : Maksim Korolev, Kurt Ruegg, mkorolev@stanford.edu, kruegg@college.harvard.edu (Stanford, Harvard) Corpus ID: 16646225

Date : February 2015

Keywords : Gradient Boosted Trees, Time Series, Bayesian Optimization

The paper talks about using Gradient Boosted Trees to predict the store sales of Rossmann stores.

In this paper, they have implemented an algorithm called Gradient Boosted trees.

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.

This paper further goes on to show us how to use SigOpt Bayesian Optimization to modify the hyperparameters of the gradient boosted trees for further reduction of prediction error.

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.

The data set also had room for improvement. There were many external factors that influence store sales that were not given. They added the ‘weatherData’ R package to get information about the weather for a given day and region. We also gathered web search rates for “Rossmann” from Google Search Trends for each region and day.

The take-away from this paper is that it shows us how to apply Gradient Boosted Trees algorithm to our dataset as it has shown to perform the best out of all decision tree learning methods.

PAPER 04

An Effective Approach for Sales Forecasting

[https://www.academia.edu/43317073/An_Effective_Approach_for_Sales_Forecasting]

Source : 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

Date : May 2020

Keywords : Regression, Lasso, Ridge, Sales Forecasting

This paper is a comparative study on sales forecasting using Supervised Machine Learning, Regression Models in particular, 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.

This paper enables us to better understand the application of regression models and how to improve the accuracy of every model used.

Amitha: Collaborated with Jigya for EDA, Literature Survey of 1 paper

Jigya: Collaborated with Amitha for EDA, Literature Survey of 1 paper

Akshaya: Collaborated with Samyuktha for Visualization, Literature Survey of 1 paper

Samyuktha: Collaborated with Akshaya for Visualization, Literature Survey of 1 paper