EasyChair Preprint

№ 4803

Retail Store Sale Prediction

Tanya Charan Pahadi, Palak Rani and Abhishek Verma

December 25, 2020

RETAIL STORE SALES PREDICTION

Tanya Charan Pahadi	Palak Rani	Abhishek Verma
B.Tech CSE	B.Tech CSE	B.Tech CSE
3rd Year	3rd Year	3rd Year
Galgotias University	Galgotias University	Galgotias University
tanyachapahadi@gmail.com	Palakirani018@gmail.com	jec2008.priyanka@gmail.com

ABSTRACT

In the following project, we have applied machine learning to a real world problem of predicting retail stores sales. Such predictions helps store managers in creating effective staff schedules that increase productivity. We used popular open source programming language Python and used its libraries like NumPy, scikit-learn, pandas , matplotlib for modelling, analysis and prediction and visualization. We have used different techniques like regression, ensemble and XGB regression. In view of nature of our problem, Root Mean Square Error (RMSE) is used to measure the prediction accuracy

Keywords: *NumPy, scikit-learn, machine-learning, RMSE, XGB regression, Ensemble*

1 : INTRODUCTION

Rossmann drug store has over 3,000 stores in 7 countries od Europe. Our task in this project is to predict the store sales of 1115 stores of rossmann store. Sales of stores can be influenced by a lot of factors, like promotions applied, competition distance, state holidays and school holidays, seasons, and location. Such predictions helps store managers in creating effective staff schedules that increase productivity. As feature provided to us is only store related and there is data related to the no of customer. recommendation system is a little difficult to be applied in such problem. Besides, as we have a continuous target, the problem can be a regression problem based on both categories (e.g, Store Type) and continuous (e.g, Days) features. apart from this, some feature can be assumed as categorical as well as continuous, like Day in week could be considered as continuous by , assuming there a relationship exist for adjacent days or as categorical [Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday]. The no of values of our dataset or characteristics of dataset is defined as below.

Table 1
Dataset statistic

STATISTICS	NUMBERS
Dataset size	1017209
Testing data size	41088
Total stores number	1115
Training data Time ranges	2013-01-01 to 2015-07-31
Testing data Time ranges	2015-08-01 to 2015-09-17

Table 2
Sales statistic

STATISTICS	VALUES
Global store sales average	5773.82
Max daily sales	41551
Min daily sales	46

2: LITERATURE REVIEW

Forecasting is projecting, predicting or estimating some future condition or event that is beyond an organization's power and gives a basis for efficient planning.

Forecasting is necessary for several situations of the modern business and its proper working. Organizations must make plans that will be efficient at some point in the future. And to do this they require information and data about current circumstances. It is very unfortunate that though forecasting is an important aspect yet its progress in many field or research and development has been limited.

In the past decade Machine Learning have emerged as a technology with a great promise for identifying and modeling data patterns that are not easily described by traditional statistical methods in a field as diverse as cognitive science, computer science, electrical engineering and finance. Example- studies in the "finance literature evidencing predictability of stock returns by means of linear regression can be improved by a neural network. Machine Learning have also been increasingly used in management, marketing and retailing. The types of applications include market response forecasting.

In this particular project we will give the following business insights to the owner

• What is the extend to which sales performance is influenced by factors like: promos, school and state holidays, competition distance ,competition open month. locality and seasonality,

• What model is appropriate to predict sales?

3: PROBLEM FORMULATION

Rossmann store managers had to predict the daily sales and the number of customers for up to six weeks in advance; while store sales, What is the extend to which sales performance is influenced by factors like: promos, school and state holidays, competition distance ,competition open month. locality and seasonality,

As there are so many of individual who try to forecast sales based on their unique sets of circumstances, the accuracy of such forecasts was rather varied. So our task was to make an efficient machine learning model that would predict the sales for 1,115 stores across Germany using which store managers would be able to create effective staff schedules to increase their productivity and sales turnover.

4: TOOLS AND TECHNOLOGIES USED

Languages to be used: Python

Study Focus on: Data Analysis , Machine Learning

Tools to be used: Visual Studio Code, Jupyter Notebook, google collab

5: FUTURE SCOPE OF THE PROJECT

Our model will help local retailers to spike their business in the following ways:-

1.It will them to decide marketing strategies.

2. It will help them preparing the budget and for setting financial policies.

3. With effective sales forecast it is feasible to obtain an average estimate of everything in such a way that the average manpower and plant capacity is fully utilized during the entire time period. Thus the forecasting enables to overcome seasonal variations.

4. It helps in stocks organizing and prevents the risk of both the over-stocking and under stocking.

5. with the help of forecasts we can find out which product provides more profit and which product's manufactured should be stopped.

We believe, every business will at some point in the future consider forecasting their sales for the upcoming challenges. The role of Al and Deep Learning will not just be limited to technical use but will be used in every sphere of life.

6: Analysis and exploration

In the following section we try to analyze our dataset and figure out what are the most important features for our predictive model out useful features that can be used to forecast sales. At first we perform the feature extraction from our dataset and take out the derived values from the existing data given to us.. Then, to get more important features, the store information is reviewed. At the end , we will try figure out more information from store information.

Open

"Open" indicates if this store is opened or not on given specified day. Because the sales of store must be 0 if it is closed, we removed the data point with (null value removal to reduce bias)"Open = 0" and after prediction, we will replace the value of sales as 0 for the data point with "Open = 0" in testing data.

Store ID

store ID is one unique feature as every store has its own different id's. Sales may or may not change from store to store. If we use Store ID as a feature, we observe that the correlation coefficient of Store ID and Sales is 0.005.

One idea to make it better is using store daily sales average as the feature instead of store ID. The coefficient between store daily sales average and Sales is 0.53, which is pretty high in analysis. We can see the store daily sales average in Figure 1

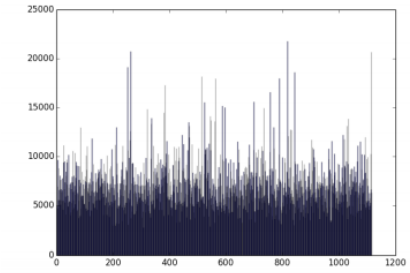


Figure 1: Store Daily Sales Average

Day of Week

On different day of week, each store will have different sales as people get used to shop in different days. This particular feature plays a significant role on the sale prediction a

State Holiday

Different People has different demands and need for drugs during holidays. We have the information of state holiday, school holidays for each store on every day. The correlation coefficient is -0.23. which indicated that there is some some correlation between state holiday and sales

Year

The sales has a relationship with years, as the brand influence, marketing and other strategies of this company may vary from year to year, which could possibly has an impact on the sales.

Month

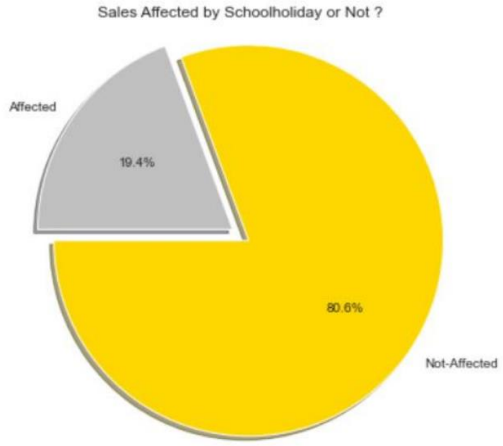
As people are more prone to having cold and other medical conditions during the winter and more sun related issue like dehydration and sunstroke during summer, people would have different demands for drugs during different month. So, the sales is possibly affected by the month of the particular year.

Day

Each single day could affect the sales. For instance, there may be people who tend to buy drugs on the first day of month or they might go to stores when get their salary. So, days must also play a role on the sale pattern.

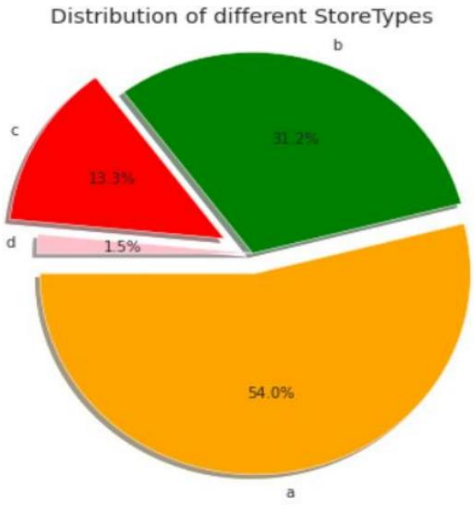
School Holiday

.On examining the effect of school holiday on the effect of sales we can see the the impact is not so significant.



Assortment

Because different stores have different assortment level, We check the percentage of sale of different types of store based on assortment we can see that the store type-a has the most powerful impact on the sale



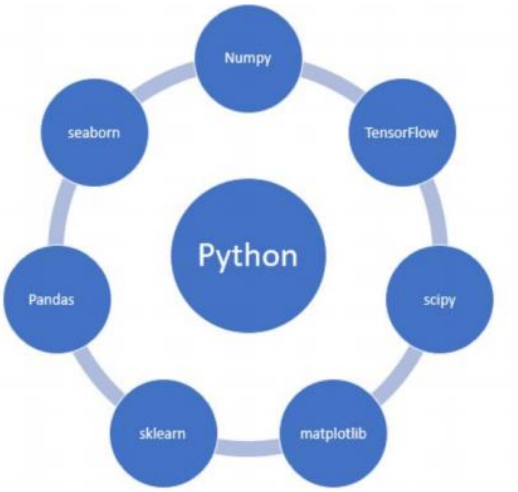
Competition Distance

People will obviously prefer going to the store which is closer to their location there competition distance can also affect the sale of a store.

7: DESIGN

Modules

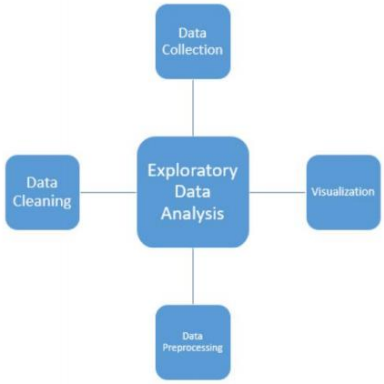
1. Loading our dataset and importing all the useful libraries



2. dealing with null values and filtering the data

3. Adding additional data field to make proper analysis

4. Exploratory Data Analysis



A lot of factors affects the sales forecasting process, which includes:

• Holiday- non holiday

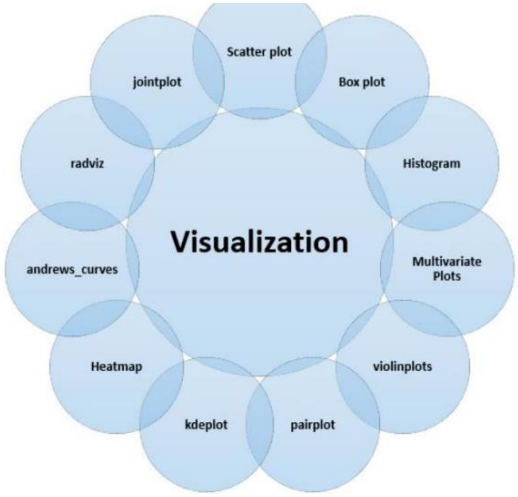
• Distance from Competition

• Seasons & Weather

• Holidays

• Supply & Demand

• Seasonality of the business We will visualize them with the help of different visualization techniques



Machine Learning Data Modeling (for our Prediction)

1. Then, we need to build a Machine Learning model that will forecast future sales.

Various methods of sales forecasting model that we will use in our project includes:

• Linear Regression

• Ridge Regression

• Lasso Regression

• Ensemble techniques

• Decision Tree Regressor

• XGB Regressor

2. A training strategy is applied to our dataset to discover the underlying relationships in the sales data.

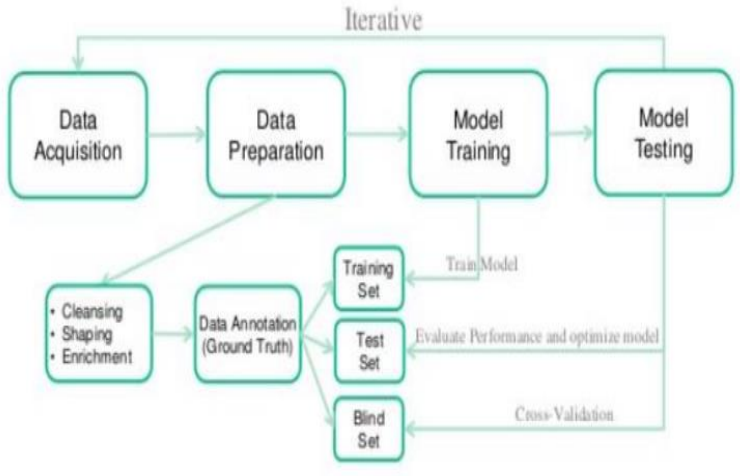
3. To improve the model's predictive capabilities, we can also apply model selection by trying combinations of variables and choosing those with more impact on sales.

4. then, the resulting model undergoes an exhaustive training analysis.

5. Finally, after model deployment and output is submitted in a csv file.

8: IMPLEMENTATION AND TESTING

Working flowchart



Testing

On our analysis using RMSE we found the the following result

Index no	Model	RMSE
1	Linear Regression	0.3731435568682413
2	Ridge Regression	0.37491234865094286
3	Lasso Regression	0.3745644170116176
4	Ensemble	0.23656802873835633
5	Decision Tree Regression	0.1816397431930095
6	XGB Regression	0.10962710044333697

It is clear from our table that XGB regression gives the least RMSE error we have therefore designed our model according to XGB regression

What is XGB?

How XGBoost optimizes standard GBM algorithm

XGBoost is a strong perspective for structuring supervised regression models. The soundness of this declaration can be tested by knowing about its base learners and objective function .

The objective function contains a regularization term and the loss function. It tells about the difference between predicted values and the actual values. Which means how close or distinct are the model real values from the model obtained. The most obvious loss functions in XGBoost for regression problems is linear, and that for binary classification is logistics.

Ensemble learning involves combining and training individual models (known as base learners) to obtain a single prediction, and XGBoost is a type of the ensemble learning methods. XGBoost anticipates to have the foundation learners which are comparatively not so good at the remainder so that when all the predictions are mixed, bad predictions calls off and better one adds up to form quality predictions.

9: CONCLUSION

As a result, we have created a model, using which Rossman store directors can predict sales for 6 weeks in .Following on from this prediction they will be able to create an effective schedule for their employees. Our next step might be the creation of a visual interface for predicting sales, so it will be possible to enter a random day e.g. the second Sunday of April 2017 and predict how many customers will attend the exact store and how much money they will spend there. In any case, at present we have a model which can be implemented into Rossman system and used for successful sales and creating schedules.

10: REFERENCES

[1] .The data used for this example can be downloaded from Kaggle.

[2]. Aburto(2009) Improved chain and supply management based on different demand forecasting.

[3]Chang, P.-C., Liu, R. K. (2008). The study of a forecasting sales model. Expert Systems with Applications, 37(12),

[4]Gering, M. (2008). Retail sales prediction and item recommendations using customer demographics at store level.

[5] HG waters.(2009). Forecasting and prediction in retail business.

WEBSITES:

[1]https://www.skyfilabs.com/project-ideas/sales-forecasting-using-walmart-dataset/i

[2] http://www.kaggle.com/

[3]https://github.com/tanyaa29/RossmannSales_Prediction (the first overview analysis code of our project on our github account)

[4]https://www.geeksforgeeks.org/xgboost-for-regression/

[5]https://colab.research.google.com/drive/1MRd93Rjw45V-XUcLl1hs4-rEyv7FVdO9?usp=sharing (ALL the source code and analysis of our project on google colab)

[6]https://www.ukessays.com/essays/commerce/literature-review-of-forecasting-and-definitions-business-essay.php