Sales forecast for Rossman Drug Store

A Capstone Project Report
Submitted to the Faculty
of the
Bennett University

By

[Rahul Singh Pundir]
[E16CSE075]

In Partial Fulfillment of the Requirements
for the Degree of
Bachelor of Technology

Major Department: Computer Science Engineering
October 2019

Greater Noida-201310, Uttar Pradesh, India

CERTIFICATE

I hereby certify that the work which is being presented in the B.Tech. Capstone Project Report entitled **"Sales forecast for Rossman Drug Store",** in partial fulfillment of the requirements for the award of the **Bachelor of Technology in Computer Science & Engineering** and submitted to the Department of Computer Science & Engineering of Bennett University Greater Noida UP is an authentic record of my own work carried out during a period from July 2019 to November 2019.

The matter presented in this thesis has not been submitted by me for the award of any other degree elsewhere.

Signature of Candidate

[Rahul Singh Pundir]

[E16CSE075]

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Head

Computer Science Engineering Department

Bennett University Greater Noida UP

ABSTRACT

In this project I tried to apply machine learning into a real-world problem of a European Store (Rossmann drug store) where I was supposed to predict the future sales of the store using the historical data provided in the form of a csv file. Given store information, and sales record I applied linear regression algorithm and then Xgboost algorithm for better accuracy and tried to predict the sales. Root Mean Square Percentage Error (RMSPE) is used to measure the accuracy. After getting the results of both the algorithms it turned out that Xgboost outshined the other model and gives a reliable forecast which helps the store managers to allocate staff and stock up accordingly.

ACKNOWLEDGEMENTS

I would like to express my special thanks of gratitude to my head of department Dr. Deepak Garg as well as our Dean Dr. Sudhir Chandra who gave me the golden opportunity to do this wonderful project on the problem of “Sales Forecast for a drug store ”, which also helped me in doing a lot of Research and I came to know about so many new things I am really thankful to them. Secondly I would also like to thank my parents and friends who helped me a lot in finalizing this project within the limited time frame.

DEDICATION

This project is dedicated to my father , who taught me that the best kind of knowledge to have is that which is learned for its own sake. It is also dedicated to my mother, who taught me that even the largest task can be accomplished If it is done one step at a time.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

LIST OF ABBREVIATIONS

[EDA] ----Exploratory Data Analysis

[Xgboost]----Extreme Gradient Boosting

[ML] ----Machine Learning

[LR] ----Linear Regression

[SKIT] ----Sci Kit

[SVM] ----Support Vector Machines

[SVR] ----Support Vector Regression

[LSTM] ----Long short-term memory

[FDR] ----Frequency Domain Regression

1. INTRODUCTION

Sales is one of the most important business domains for data science and data mining applications because of its importance it defines the growth and expansion of a business hence if we are able to predict or forecast it then we can stabilize the uncertainty of any business and control its growth and many different factors like the minimizes over and under stocking at each store thereby minimizing losses and amount of labor needed for it, further more for a new startup it can become a source of credibility so that the person can convince investors on investing on their startup on that sales forecast of a similar or somewhere close business. The major factor here plays the accuracy of prediction.

1.1. Problem Statement

Currently most of the organizations are facing the problem of managing their sales hence leaving them with shortage of products or excessive products as sales is influenced by many factors so if we predict the sales then we can maximize the profit and hence decrease the loss. Sales forecast also helps the organization to manage the work force required at the store and provides the owner with predictable outputs.

2. BACKGROUND RESEARCH

Research work based on this project has been done and several prediction models have also been made by many experts and data scientists, R or Python programming and machine learning libraries have been used for the prediction of this problem of sales but the major factor which roles is in the accuracy of my prediction model and dealing with the loss function. The supportive work of Mr. Tian Yang and Mr. Zhuyuan liu[1] on the same problem statement shows the direction and path towards the predictive analysis of sales for Rossmann Sales stores where the data provided is very difficult to analyze (as 180 stores were closed for 6 months so they were unable to fill the gap of sales for those stores.) and is in very large quantity. But as we move forward with their analysis he used Poisson function for the rest of GLM experiments and used SVM (support vector machine) Regression, for the analysis of the dataset and feature selection he moved with time series model and took help from Mr. Jianqing Fan and Mr. Qiwei Yao's "Nonlinear Time Series: Nonparametric and Parametric Methods"[2]. This book has a refined knowledge about nonlinear time series and data-analytic nonparametric methods. So, what we draw from the research about the Rossmann problem is that the analytical and predictive work has been done for this problem and as this problem was amongst one of the many Kaggle competitions and was well recognized so hence the major point still remains the working on the accuracy of prediction of my model and its betterment from other presented models. If we talk about literature or documentation present on the sales analysis, there are many articles present on internet particularly in the field of sales prediction one of them is from the page "Towards Data Science" Predicting the Sales forecasting the monthly sales using LSTM by Barış Karaman[3] and a research based post in newsletter by Giering, Michael[4] "Retail sales prediction and item recommendations using customer demographics at store level." they talk about the use of Python in a simplistic way to fuel your company's/sales

growth by applying the predictive approach to all your actions. It is a combination of programming, data analysis, and machine learning and in how many different ways and using different ML models you can approach this problem. Along with analytical study of Rossmann drug store problem I also studied similar problem statement like one presented by P. Mekala B. Srinivasan[5] which showed some light on Time series data prediction on shopping mall. At the same time I tried to have an overview of how data analytics and data mining are a big factor for business and industry for which I approached a book "A practical guide to data mining for business and industry" by Shirley Coleman Ahlemeyer Stubbe, Andrea[6], it helped me a lot knowing about some great work of data mining in the grooming businesses. As sales forecasting being the major part of the businesses these days there have been a lot of research papers published by many scholarly people using different machine learning algorithm and models to present accurate results of the retail sales forecast one of the work was done by Ankur jain, Manghat Nitish Menon, Saurabh Chandra[7] focusing on extreme gradient boosting algorithm to predict sales forecast of European retail store. After looking for different blogs and research papers I found a research paper specific to the Rossman drug store sales problem by Lin, Sen, Eric Yu, and Xiuzhen Guo. "Forecasting Rossmann Store Leading 6-month Sales."[8] it used the Frequency Domain Regression (FDR) and Support Vector Regression (SVR) for time-series prediction of Rossman Store Sales and showed how SVR clearly outperformed FDR due to extent of data variables. This research paper gave me a different approach too, but I wanted to try the one which is unique and not used much by people and still giving better results. Accordingly, I found that XgBoost was not used much by researchers and was known to give better results and as we know svm's and regression techniques were quite common and did not perform well.

2.1. Proposed System

This project aims to predict or forecast the sales using the historical data provided to us by the European company and the major factor playing the important role will be the accuracy of the prediction. Hence, by using different machine learning algorithms I found the most accurate (XGboost) serving our purpose. These insights can be very useful as they can influence future decisions for the company and can help the company grow exponentially. By optimizing our result we can assure a great help to the potential user as it makes it easy for the owner to control the goods purchased and the goods sold which reduces the chances of his loss due to the sales prediction and hence helping him to grow his business.

2.2. Goals and Objectives

Table 1: Goal and Objectives

#	Goal or Objective
1	Make the system extensible – future updates like xxx can be done easily
2	Make the system easy to support – provide good documentation, configuration/build files, administrator's manual
3	Make the system very easy to use – users would agree that minimal to no training is needed
4	Build a prototype that demonstrates the user interface by xx/xx/xx - in order to get early feedback from the customer/users
5	Have fun working on the project

3. PROJECT PLANNING

This section covers the details of the project planning. Selecting the lifecycle of the development, project stakeholders, resources required, assumptions made (if any) are detailed in the sections below.

3.1. Project Lifecycle

I used an agile approach as I was trying all the ways of data preprocessing so that I can get at most accuracy of my prediction as we know choosing different features have different impact on the accuracy of the model. The data was trained using different machine learning models as we need to check which model will be giving us the most accuracy. I was having meetings with my mentor after completing different stages of my project like doing data cleaning ,data preprocessing and then choosing the model to train my data on, so that he can always guide me through the right path for the successful completion of my project.

3.2. Project Setup

Table 2: decision description

#	Decision Description
1	Windows 10, Jupyter Notebook, Anaconda installed.
2	Coding skills in python(pandas, NumPy for data preprocessing).
3	Special access privileges needed, release to open source .
4	Choosing specific algorithms for better results.

3.3. Stakeholders

Table 3: stakeholders and role

Stakeholder	Role
Indrajeet Sir	Mentor
Rahul Singh Pundir	Project Developer/Tester

3.4. Project Resources

Table 4: resources and resource description

Resource	Resource Description	Quantity
Data set	A dataset to train the model	1
Capstone developer team	Our team of student and mentor who will be the primary developers of the project.	2
Indrajeet Sir	The mentor who will be able to provide me with technical assistance.	1
Jupyter Notebook	A platform to develop the project, train and test our model.	1

3.5. Assumptions

Table 5: assumptions

Assumption No.	Assumption
A1	Me and mentor will be able to meet face to face once a week.
A2	Jupyter Notebook and all libraries(NumPy, pandas) will be available for successful completion of the project as they are opensource software and libraries.
A3	I will be able to familiarize myself with Jupyter Notebook and various ML models.
A4	I will have sufficient time to complete a working model to present by mid-semester.
A5	Machine Learning model will be completed in time to test on the testing dataset.
A6	The provided test dataset will be sufficient to predict the sales for the store.
A7	The ML models used for getting higher accuracy will be experimented in the given time frame

4. PROJECT TRACKING

4.1. Tracking

Table 6: details of project storage

Information	Description	Link
Code Storage	Project code will be stored on my GitHub repository	Link
Project Documents and Assignments	Milestone reports, specification and design documents and diagrams, etc. will be stored in my GitHub repository.	Link

4.2. Communication Plan

Table 7: Regularly Scheduled Meetings

Meeting Type	Frequency/Schedule	Who Attends
Team Meeting	Once in a Week	Me and Mentor
Short Meeting	Weekly	Me and mentor

Table 8: Information To Be Shared Within Our Group

Who?	What Information?	When?	How?
Mentor(Indrajeet Sir)	Project progress & General scrum information	Weekly	Team meetings, listing in Project Specification, showing updates.

Table 9: Information To Be Provided To Other Groups

Who?	What Information?	When?	How?
Sanjeet sir	Final deliverables	At completion of project	Through google form

Table 10: Information Needed from Other Groups

Who?	What Information?	When?	How?
Sanjeet sir	Information regarding project	During the capstone classes.	In person conversations.
Akshay Goel	Information regarding XgBoost algorithm.	During the capstone classes.	In person conversations.

4.3. Deliverables

Table 11: Deliverables

S.no	Deliverable
1	Study results
2	Code
3	Test and test results
4	Final report (final PowerPoint presentation, 3-minute video, and final sprint)

5. SYSTEM ANALYSIS AND DESIGN

This section describes in detail about the design part of the system.

5.1. Overall Description

This project is an attempt to apply data science and machine learning techniques to perform sales prediction which gives you a valuable insight into the inner workings of your business. Store need analysis on regular basis helps you to know the mood of customers and at the same time help your business grow rapidly. Using Jupyter notebook platform we will be performing exploratory data analysis and doing visual representation of data to understand it more clearly and by this approach we can understand our dataset more clearly. After the require analysis we will be applying certain selected machine learning models on the selected features to train them and get the required output. As the model focusses on the accuracy of the prediction we will be applying gradient boosting to the selected features(after exploratory analysis) and will be looking towards the new accuracy using Xgboost. This whole visualization and analysis of dataset will provide basic insight into the sales data for any Program Manager, or someone not experienced in data science

5.2. Users and Roles

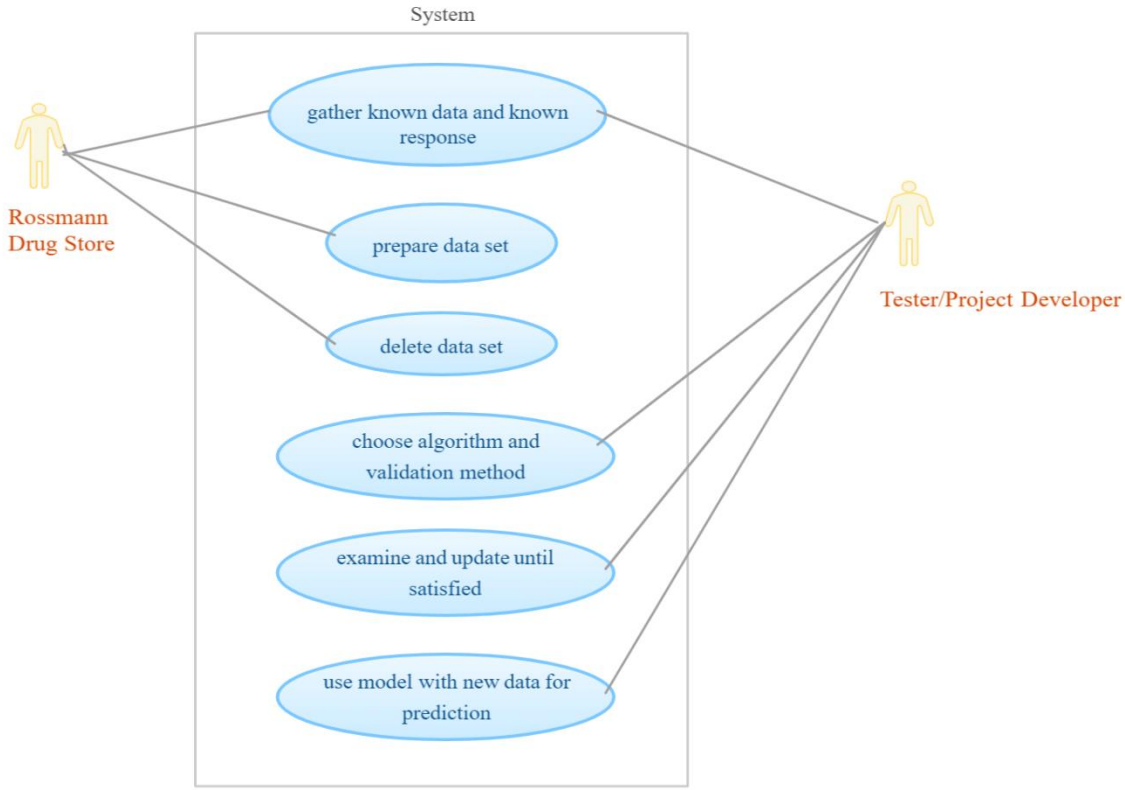
Table 12: users and their roles

User	Description
Developer (Me)	A person who will developing the project and analyzing the given dataset. He will be applying the required ML model and will be providing the end user with the highest accurate results,
Mentor	A mentor who will be guiding me throughout my project and will be providing me assistance whenever needed.
Sales manager	An end user of the prediction model who will be using the model for his/her companies future sales forecast.

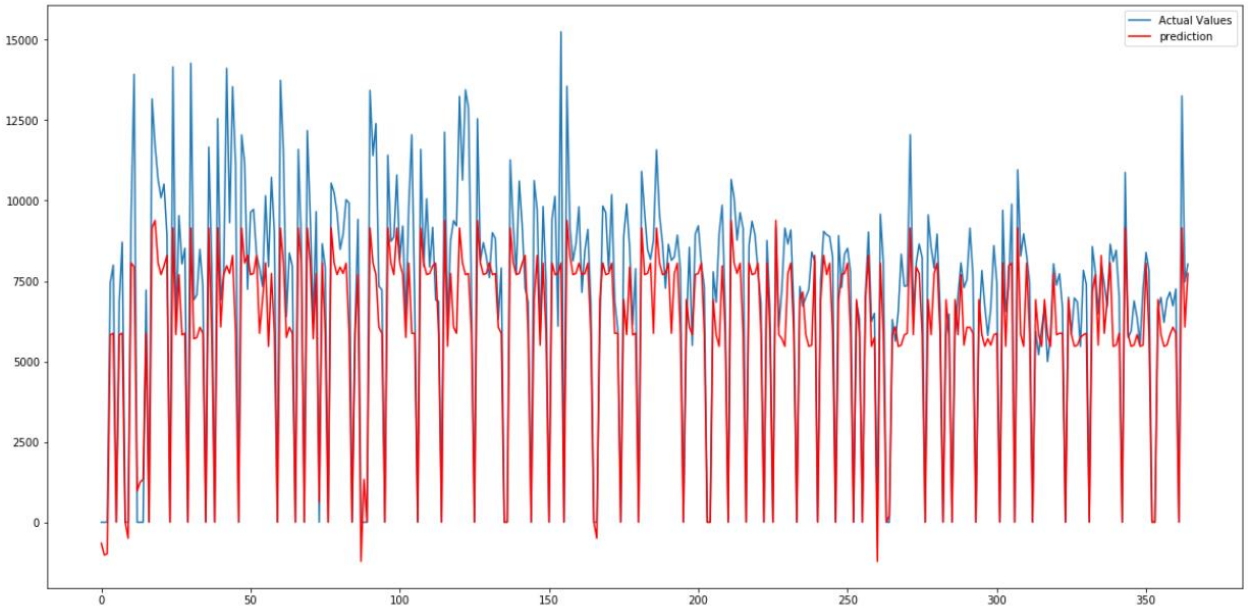
5.3. Design diagrams/ UML diagrams/ Flow Charts/ E-R diagrams

5.3.1. Use Case Diagrams

Figure 1: Use – Case diagram for the project



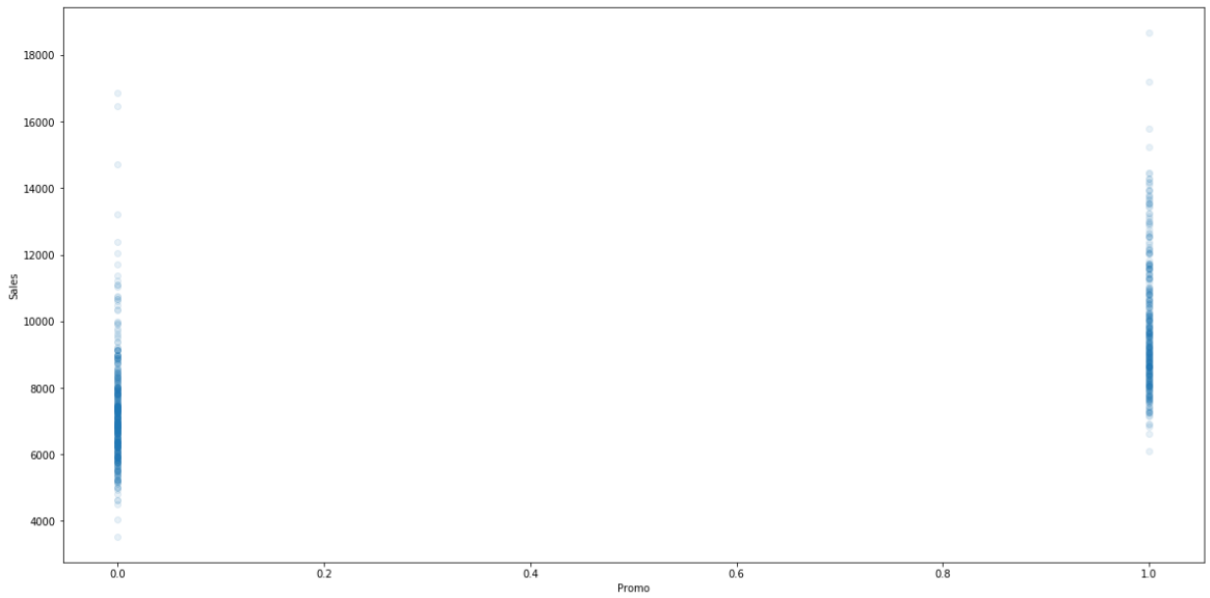
5.3.2. Difference between actual and predicted values



The blue line in the above graph shows the actual values of the sales and the red line in the graph shows the predicted values of sales when we used the linear regression model.

Figure 2: visualization of actual and predicted values

5.3.3. Boxplot of dataset



The above boxplot defines the relation between the promo applied during the purchase of goods and the sale during that particular period. This boxplot shows that how sale increased during the use of promos on the stores.

Figure 3: boxplot of dataset

5.3.4. Data Architecture

```
In [2]: data=pd.read_csv("C:/Users/Rahul Pundir/Desktop/capstone dataset/train.csv")
        data.head()
```

```
C:\Users\Rahul Pundir\AppData\Local\Continuum\anaconda3\lib\site-packages\IPython\core\interactiveshell.py:2785: DtypeWarning:
Columns (7) have mixed types. Specify dtype option on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

Out[2]:

	Store	DayOfWeek	Date	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday
0	1	5	31-07-2015	5263	555	1	1	0	1
1	2	5	31-07-2015	6064	625	1	1	0	1
2	3	5	31-07-2015	8314	821	1	1	0	1
3	4	5	31-07-2015	13995	1498	1	1	0	1
4	5	5	31-07-2015	4822	559	1	1	0	1

```
In [3]: data2=pd.read_csv("C:/Users/Rahul Pundir/Desktop/capstone dataset/test.csv")
        data2.head()
```

Out[3]:

	Id	Store	DayOfWeek	Date	Open	Promo	StateHoliday	SchoolHoliday
0	1	1	4	17-09-2015	1.0	1	0	0
1	2	3	4	17-09-2015	1.0	1	0	0
2	3	7	4	17-09-2015	1.0	1	0	0
3	4	8	4	17-09-2015	1.0	1	0	0
4	5	9	4	17-09-2015	1.0	1	0	0

The above table shows the data architecture of the provided dataset for applying the machine learning model upon.

Figure 4: dataset design

6. USER INTERFACE

6.1. UI Description

I am creating a project using python programming language and ML based libraries, packages like numpy and pandas are available in python which help in exploratory data analysis this whole code will be done in Jupyter Notebook platform which will be the main means of interaction with my code. It uses a standard UI console and it is not in my scope to create my own UI on top of it. It has rows provided in which we write the commands/code and we get the output below. The end user just has to insert the dataset location into the provide code and has to run the project to get the sales forecast.

6.2. UI Mockup

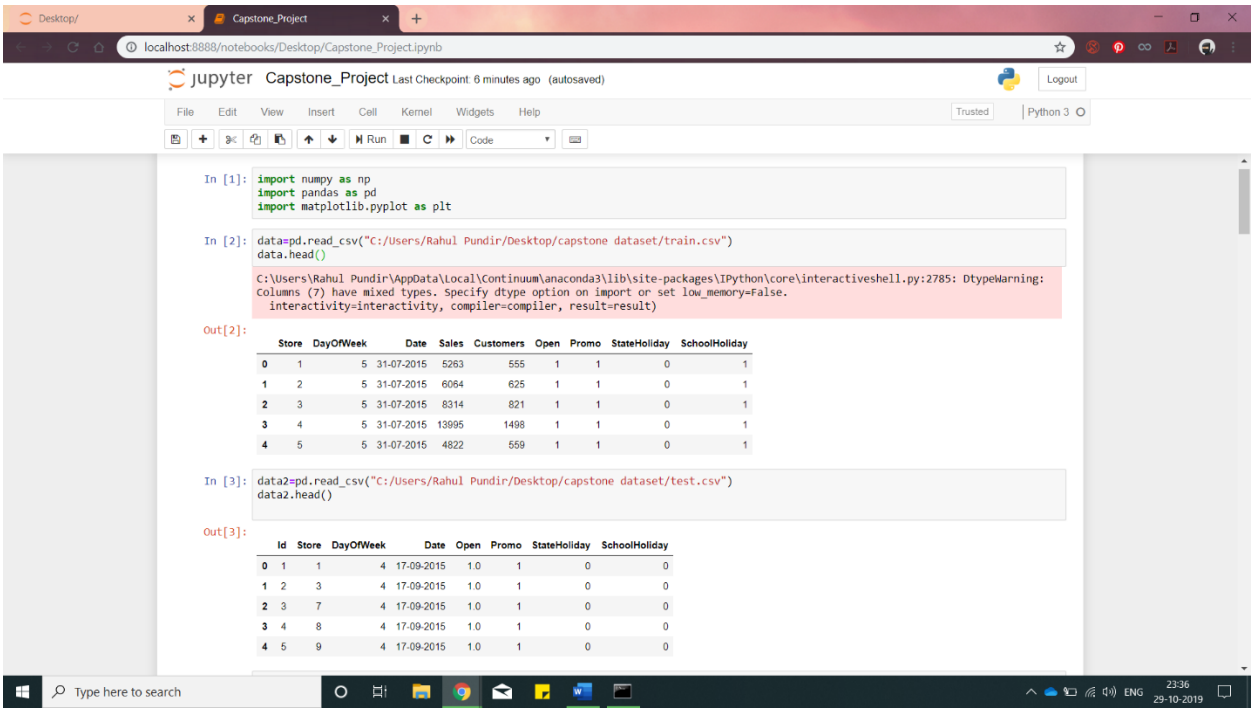


Figure5: Jupyter notebook

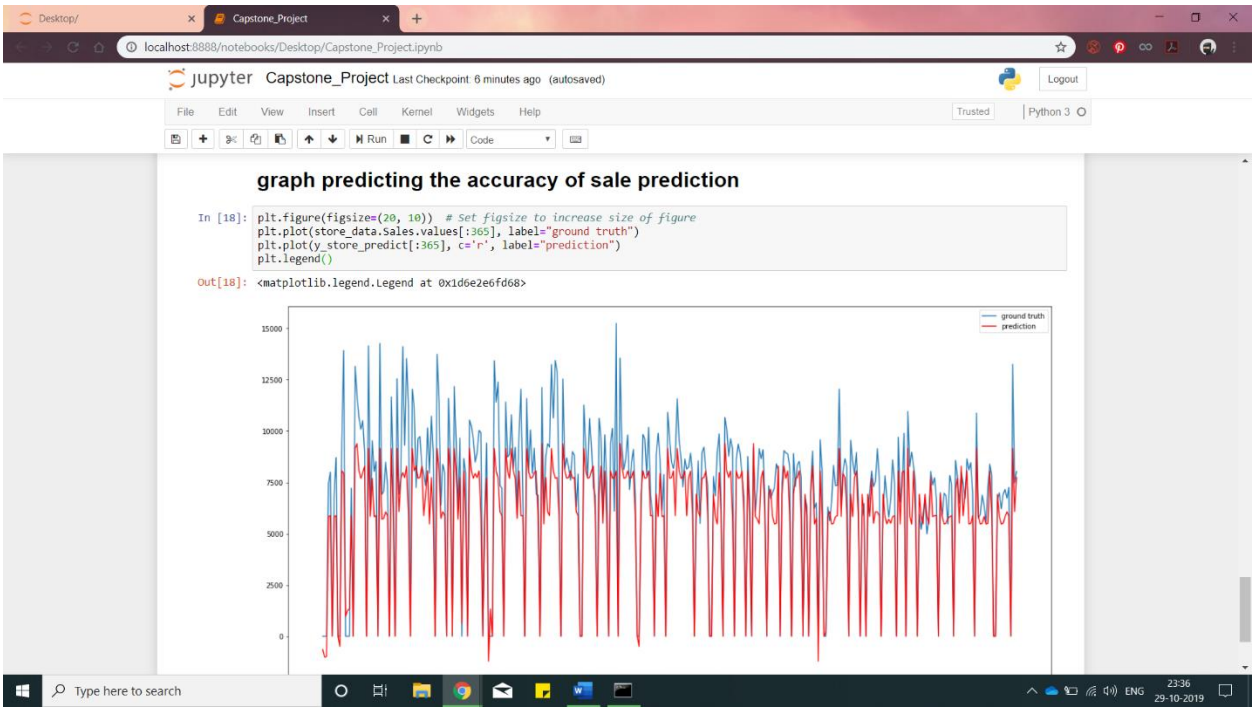


Figure 6: predicted values using model

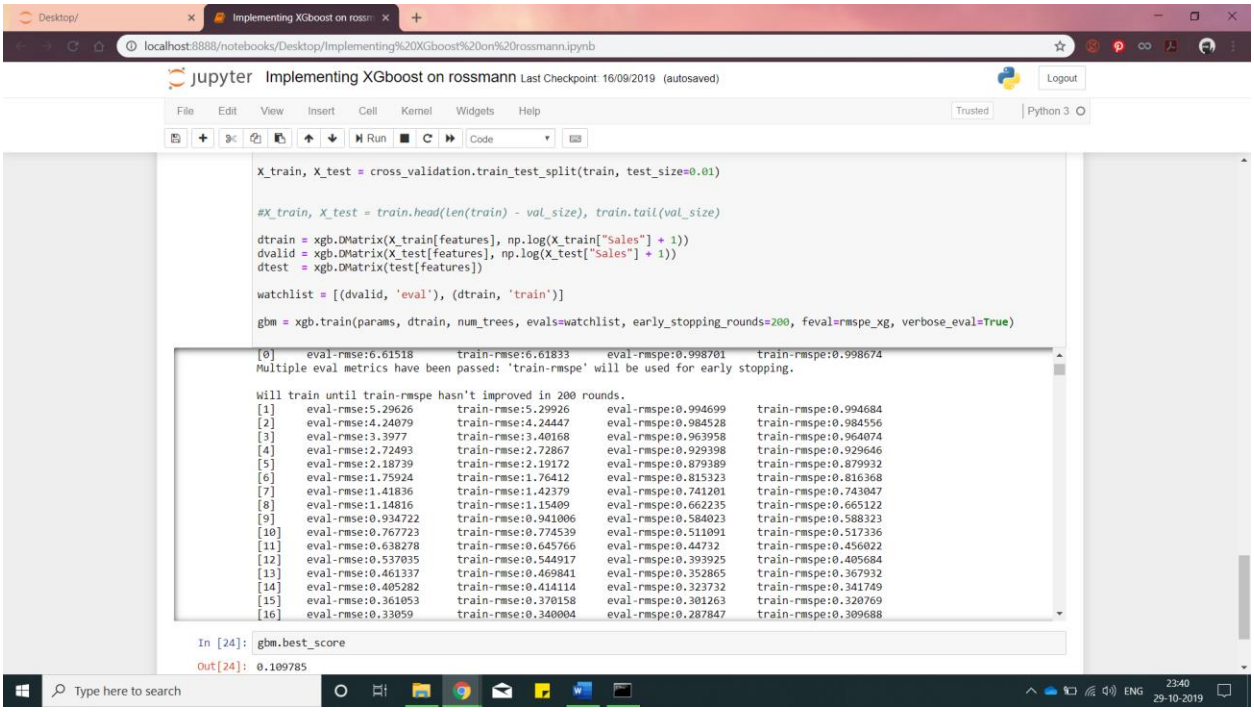


Figure 7: xgboost model

7. ALGORITHMS/PSEUDO CODE

IMPLEMENTING LINEAR REGRESSION

```
from sklearn.linear_model import LinearRegression
from sklearn import cross_validation as cv

lr = LinearRegression()
kfolds = cv.KFold(X.shape[0], n_folds=4, shuffle=True, random_state=42)
scores = cv.cross_val_score(lr, X, y, cv=kfolds)

print("Accuracy: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std()))

lr = LinearRegression()
X_store = pd.get_dummies(data[data.Store!=150], columns=['DayOfWeek',
'StateHoliday']).drop(['Sales', 'Store', 'Date', 'Customers'], axis=1).values
y_store = pd.get_dummies(data[data.Store!=150], columns=['DayOfWeek',
'StateHoliday']).Sales.values
lr.fit(X_store, y_store)
y_store_predict = lr.predict(pd.get_dummies(store_data, columns=['DayOfWeek',
'StateHoliday']).drop(['Sales', 'Store', 'Date', 'Customers'], axis=1).values)
```

IMPLEMENTING XgBoost

```
import pandas as pd
import numpy as np
from sklearn import cross_validation
import xgboost as xgb

print("Train a XGBoost model")
print(train.tail(1)['Date'])
X_train, X_test = cross_validation.train_test_split(train, test_size=0.01)
dtrain = xgb.DMatrix(X_train[features], np.log(X_train["Sales"] + 1))
dvalid = xgb.DMatrix(X_test[features], np.log(X_test["Sales"] + 1))
dtest  = xgb.DMatrix(test[features])
watchlist = [(dvalid, 'eval'), (dtrain, 'train')]
gbm = xgb.train(params, dtrain, num_trees, evals=watchlist, early_stopping_rounds=200,
feval=rmspe_xg, verbose_eval=True)
```

8. PROJECT CLOSURE

This section elucidates the overall lookup at the project and some of the future works that may enhance the solution.

8.1. Goals / Vision

My original goal for this project was to take the sales data and analyze, visualize it and apply appropriate machine learning model to it in Jupyter notebook to predict accurate sales for the company based upon previous actions. Such sales forecast provides a company with decision-making insight into many key areas like customer trends and behaviors Analyze, interpret and deliver data in meaningful ways, increase business productivity, Drive effective decision-making.

8.2. Delivered Solution

My solution consists primarily of a python language machine learning based project in which data cleaning, data preprocessing, data visualization and data segregation is done then a machine learning model(Linear Regression model) is applied to check the accuracy of prediction so that we can get an idea of what a basic model provides us with the results of the model are visualized using matplot library. Then Xgboost gradient boosting is applied to the preprocessed data and further results are extracted. The provided result has RMSE (root mean square error) of 10.9785% and the features were adjusted accordingly to get the accuracy in the sales forecast.

8.3. Remaining Work

I just worked upon different Machine Learning models and did some data modeling to get the best of the accuracy and trained my model accordingly so now, I would try to embed this python model into an mobile/web based application and further convert into a handy model so that it becomes more easy for the end user. The end user or the company/business sales consultant can just enter their historic sales data into the application and get the future sales forecast. This will be very easy to handle for the end user plus will be very helpful for them and it wont require any technical assistance and would operate like any other mobile/web application.

REFERENCES

[1] Mr. Tian Yang and Mr. Zhuyuan liu "CS229 Machine Learning: Lecture Notes." (2014).

[2] FAN, JIANQING, AND QIWEI YAO. NONLINEAR TIME SERIES: NONPARAMETRIC AND PARAMETRIC METHODS. SPRINGER SCIENCE & BUSINESS MEDIA, 2008.

[3] HTTPS://TOWARDSDATASCIENCE.COM/PREDICTING-SALES-611CB5A252DE

[4]GIERING, MICHAEL. "RETAIL SALES PREDICTION AND ITEM RECOMMENDATIONS USING CUSTOMER DEMOGRAPHICS AT STORE LEVEL." ACM SIGKDD EXPLORATIONS NEWSLETTER 10.2 (2008): 84-89.

[5] P. MEKALA B. SRINIVASAN. TIME SERIES DATA PREDICTION ON SHOPPING MALL. IN INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATION AND ROBOTICS, AUG 2014.

[6] SHIRLEY COLEMAN AHLEMEYER STUBBE, ANDREA. A PRACTICAL GUIDE TO DATA MINING FOR BUSINESS AND INDUSTRY. JOHN WILEY AND SONS, 2014.

[7] JAIN, ANKUR, MANGHAT NITISH MENON, AND SAURABH CHANDRA. "SALES FORECASTING FOR RETAIL CHAINS." (2015): 1-6.

[8] LIN, SEN, ERIC YU, AND XIUZHEN GUO. "FORECASTING ROSSMANN STORE LEADING 6-MONTH SALES."