**“Sales Prediction System”**

**(Using Machine Learning)**

Submitted in partial fulfillment of the requirements

of the degree of

Bachelor of Engineering

by

Ms. Archisha chandel (07)

Mst. Saurabh Dhawale (13)

Ms. Akanksha Dubey (14)

Supervisor:

**Prof. Madhuri Ghuge**



Department of Computer Engineering

Bharati Vidyapeeth College of Engineering

2018-2019

**Declaration**

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we’ve adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

----------------------- ------------------------- -------------------------

(Archisha Chandel) (Saurabh Dhawale) (Akanksha Dubey)

Date:

**Certificate**

This is to certify that the project entitled “**Sales Prediction System”** is a bonafide work of **Akanksha Dubey (14), Saurabh Dhawale (13), ArchishaChandel (07)**  submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **Undergraduate** in **Computer Engineering**.

|  |
| --- |
| **(Prof. Madhuri Ghuge)** |
| Project Guide |

|  |
| --- |
| **(Prof. Shivsagar Gondil)** |
| Project Coordinator |

|  |
| --- |
| ***(*Dr. D. R. Ingle*)*** |
| Head Of Department |

|  |
| --- |
|  |
| Principal |

**Acknowledgement**

We express esteemed gratitude and sincere thanks to our worthy project guide **Prof. Madhuri Ghuge** our vocabulary is yet to find suitable words benefiting to high standard of knowledge set by her and extreme sincerity and affection with which she has regularly encouraged us to put heart and soul in this work. We are much obliged to our honourable Head of Department **Dr. D. R. Ingle** whose support and co-operation was always helpful and encouraging. And also thank you to Project Coordinator **Prof. Shivsagar Gondil** who supported us in numerous ways.Our parents who always bear with us in every critical situation and provide the support whenever required. As we give expression to our love and appreciation our heart is fill. And we in sincere appreciation of your valuable help.

**Project Synopsis Report Approval**

This project report entitled “**Sales Prediction System”** by **Akanksha Dubey (14), Saurabh Dhawale (13), ArchishaChandel (07)** approved for the degree of **Computer Engineering.**

|  |  |
| --- | --- |
|  | Examiners |
| 1. | -------------------------------- |
| 2. | -------------------------------- |

Date:

Place:

**Abstract**

Predicting sales of a company needs time series data of that company and based on that data the model can predict the future sales of that company or product. So, in this research project we will analyze the time series sales data of a company and will predict the sales of the company for the coming quarter and for a specific product. For this kind of project of sales predict, we will apply the machine learning algorithms and evaluate the result based on the training, testing and validation set of the data. After collecting and integrating the dataset that is acquired, the nature of data needs to be understood. For doing so correlations, general trends and outliers need to be identified by calculating the mathematical statistics (mean, median, range, standard deviation etc.) for each feature.After understanding the nature of the data and finding correlation between different features and target variable i.e. sales. The erroneous values in the dataset needs to be replaced with values that make sense, the missing values need to be replaced with appropriate numerical or categorical value depending on the type of feature. Data cleansing gives us a wholesome error free dataset to work with, Feature Transformation is the family of algorithms used to create new features from existing features. This approach will predict much better than other single classifiers when more data is used. Because of the difference is not statistically significant between the proposed model and random forest, the proposed method can be used to forecast demand due to its accuracy with less data.

**Table of contents**

|  |  |  |
| --- | --- | --- |
| CHAPTER NO | TITLE | PAGE NO |
| 1. | Introduction   * 1. Problem Statement   2. Aim of the project | 01  02  02 |
| 2. | Literature Survey  2.1 Study of existing systems | 03  03 |
| 3. | 3.1 System Analysis and System Design  3.1.1 System Requirements  3.2 System design  3.2.1 System architecture design | 04  04  05  06 |
| 4. | Coding and Implementation  4.1 Code  4.2 Implementation | 07  07  09 |
| 5. | Conclusion and Future Scope   * 1. Conclusion   2. Future Scope | 11 |
| 6. | References | 12 |

**Chapter 1: Introduction**

Sales is the lifeblood of every company. The advantages of forecasting any company's sales lie mainly in giving you a firm idea of what to expect in the coming months. A standard sales forecast looks at conditions present in your business during previous months, and then applies assumptions regarding customer acquisition, the economy and your product and service offerings. Forecasting sales identifies weaknesses and strengths before you set your budget and marketing plans for the next year, allowing you to optimize your purchasing and expansion plans.

Predicting sales of a company needs time series data of that company and based on that data the model can predict the future sales of that company or product. So, in this research project we will analyze the time series sales data of a company and will predict the sales of the company for the coming quarter and for a specific product. For this kind of project of sales predict, we will apply the machine learning algorithms and evaluate the result based on the training, testing and validation set of the data.

A sales analysis report shows the trends that occur in a company's sales volume over time. In its most basic form, a sales analysis report shows whether sales are increasing or declining. At any time during the fiscal year, sales managers may analyze the trends in the report to determine the best course of action. Managers often use sales analysis reports to identify market opportunities and areas where they could increase volume. For instance, a customer may show a history of increased sales during certain periods. This data can be used to ask for additional business during these peak periods. A sales analysis report shows a company's actual sales for a specified period a quarter, a year, or any time frame that managers feel is significant. In larger corporations, sales analysis reports may only contain data for a subsidiary, division or region. A small-business manager may be more interested in breaking sales down by location or product. Some small, specialized businesses with a single location are compact enough to use general sales data. A sales analysis report may compare actual sales to projected sales. Machine learning models are best for this kind of problem where we can easily fit a line of high sale and low sale product, quarters and zone for a product. Also we need huge amount of data for the training of the model which we can collect from the sales data of any product or company of last 1 or 2 years for any live project. However, for this research project, the description of the dataset which we are going to use for this project is provided in the dataset portion of experimental setup section.[5]

**1.1 Problem statement**

Sales forecasting can have crucial impact on the success and performance of companies. In retail and consumer-oriented industries, accurate forecasts are essential. Sales prediction is the concept of predicting the quantity of a product that consumers will purchase during a specific time period. Predicting right demand of a product is an important phenomenon. Sales planning is essential for any company – but especially for firms who are in high growth mode or experiencing a change like adding new products or entering into new markets. The fundamental job of the prediction is to balance sales and marketing resources against supply capacity planning.

**1.2Aim of the project**

The main aim of this project is to design a model based on machine learning. The purpose of this project is to predict the sales for the coming quarter or year, for the companies. The emergence of vast amounts of data from multiple sources and platforms, generating new information every minute, has gifted companies with more consumer information than they’ve ever had before, so to help this out we are implementing this project.

1. Manual work is reduced drastically and gives a more accurate output.
2. Visual representation are used for simplifying and making the process more understandable to non-technical authority as well.
3. Its main aim is to give you a firm idea of what to expect in the coming months.
4. Proposed method identifies the weaknesses and strengths before you set your budget and marketing plans for the next year allowing you to optimize your purchasing and expansion plans.
5. On-demand sales forecasts based on most updated data.
6. Custom filters for better functionality.
7. Customizable and diverse graphs and analytics.
8. Easy access to historical sales data for reference.
9. Custom sales forecasting based on total leads and opportunities, converted leads, industry performance, etc.
10. Sales forecasting for various time ranges; day, weeks, months or years.

**Chapter 2: Literature Survey**

**2.1 Research**

To implement this concept, we have studied different research works and found following information. A sales analysis report shows the trends that occur in a company's sales volume over time. In its most basic form, a sales analysis report shows whether sales are increasing or declining. At any time during the fiscal year, sales managers may analyze the trends in the report to determine the best course of action. Managers often use sales analysis reports to identify market opportunities and areas where they could increase volume. For instance, a customer may show a history of increased sales during certain periods. This data can be used to ask for additional business during these peak periods. A sales analysis report shows a company's actual sales for a specified period a quarter, a year, or any time frame that managers feel is significant. A sales analysis report may compare actual sales to projected sales. Machine learning models are best for this kind of problem where we can easily fit a line of high sale and low sale product, quarters and zone for a product.

1. Machine-Learning Models for Sales Time Series Forecasting, Bohdan M. Pavlyshenko
2. Trends in Machine Learning Applied to Demand & Sales Forecasting: A Review, Juan Pablo UsugaCadavid, Samir Lamouri , Bernard Grabot
3. Explaining machine learning models in sales predictions Marko Bohaneca,b,∗ , Mirjana Kljaji´cBorˇstnarb , Marko Robnik-Sikonja

**Chapter 3: System analysis and system design**

**3.1 System analysis**

Analysis of the system will talk about the proposed system which is delved from the existing systems with advanced features, requirements of the proposed system and risks associated with them.

**3.1.3 System Requirements**

It includes all the software and hardware requirements for the project. Further in the 4th chapter, cost estimated for the software and hardware are also discussed and the figure is drawn.

* **SOFTWARE REQUIREMENTS**

1. Operating System : Windows 7 or higher.
2. Memory : 4GB Ram or higher.
3. Disk Space : 10GB or higher.
4. Processor : Intel i5 or higher.

* **HARDWARE REQUIREMENTS**

1. Anaconda IDLE 2.0 or higher.
2. Jupyter Notebook 5.7 or higher.

**3.2 System design**

# **3.2.1 Anaconda IDLE**

The open-source [**Anaconda Distribution**](https://docs.anaconda.com/anaconda/) is the easiest way to perform Python/R data science and machine learning on Linux, Windows, and Mac OS X. It is the industry standard for developing, testing, and training on a single machine, enablingindividual data scientists to:

1. Manage libraries, dependencies, and environments with [Conda](https://conda.io/docs/)
2. Develop and train machine learning and deep learning models with [scikit-learn](https://scikit-learn.org/stable/), [TensorFlow](https://www.tensorflow.org/), and [Theano](https://pypi.org/project/Theano/)
3. Analyze data with scalability and performance with [Dask](https://dask.org/), [NumPy](http://www.numpy.org/), [pandas](https://pandas.pydata.org/), and [Numba](http://numba.pydata.org/)
4. Visualize results with [Matplotlib](https://matplotlib.org/), [Bokeh](https://bokeh.pydata.org/en/latest/), [Datashader](http://datashader.org/), and [Holoviews](http://holoviews.org/)

## **3.2.2 Jupyter notebook**

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

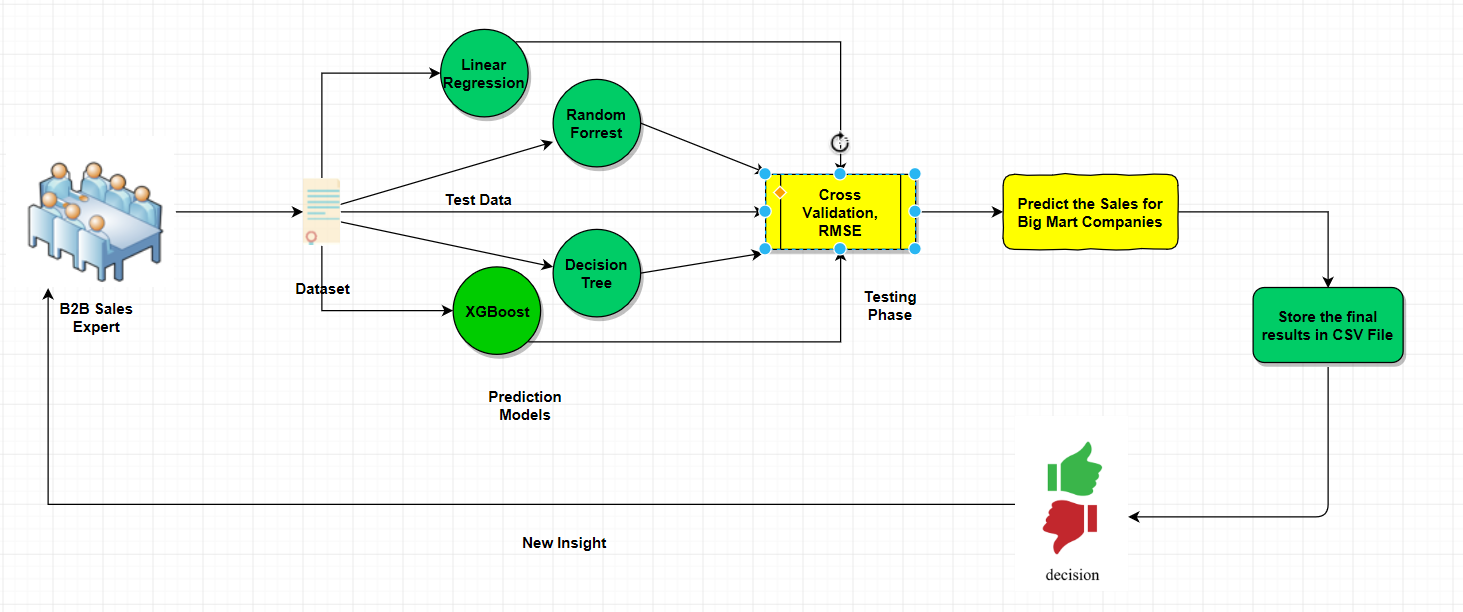
**Jupyter**is a [nonprofit organization](https://en.wikipedia.org/wiki/Nonprofit_organization" \o "Nonprofit organization) created to "develop [open-source software](https://en.wikipedia.org/wiki/Open-source_software), open-standards, and services for [interactive computing](https://en.wikipedia.org/wiki/Interactive_computing) across dozens of programming languages". Spun-off from [IPython](https://en.wikipedia.org/wiki/IPython" \o "IPython) in 2014 by [Fernando Pérez](https://en.wikipedia.org/wiki/Fernando_P%C3%A9rez_(software_developer)), Project Jupyter supports execution environments in several dozen languages. Project Jupyter's name is a reference to the three coreprogramming

languages supported by Jupyter, which are [Julia](https://en.wikipedia.org/wiki/Julia_(programming_language)), [Python](https://en.wikipedia.org/wiki/Python_(programming_language)) and [R](https://en.wikipedia.org/wiki/R_(programming_language)).

**3.2.3 Python Libraries**

Python’s standard library is very extensive, offering a wide range of facilities as indicated by the long table of contents listed below. The library contains built-in modules (written in C) that provide access to system functionality such as file I/O that would otherwise be inaccessible to Python programmers, as well as modules written in Python that provide standardized solutions for many problems that occur in everyday programming. Some of these modules are explicitly designed to encourage and enhance the portability of Python programs by abstracting away platform-specifics into platform-neutral APIs.

**3.3 System Architecture**



**Fig 3.3.1. Flowchart - Sales prediction**

**Chapter 4: Coding And Implementation**

Coding is done in Python and executed on the jupyter notebook. The program includes a no. of pre-defined libraries to work on the code, which makes the execution more faster and accurate. The proposed system works on the machine learning algorithms which takes in as input the data set of the company and then applies the algorithms like linear regression, decision tree, xgb regressor and predicts the sales for the coming quarter of the year.

**4.1 Code**

Various algorithms are used to predict highly accurate results. In the following section all the algorithms used are described:

**4.1.1 Linear Regression:**

The basic idea of this algorithm is to fit a straight line between the selected features in training dataset and a continuous target variable i.e. sales. This algorithm finds a line that best fits the data.[2]

***A. Algorithm***

Input: Dataset with proper input and output labels

Output: Predict sales value and store in csv file

begin

i. Calculate mean, variance for the list of values

Def Mean(values):

Sum(values) divided by Len(values)

Calculate Mean\_x, Mean\_y

Def Variance (values, Mean):

sum([(values-Mean) ^2])

Calculate Variance\_x, Variance\_y

ii. Calculate covariance

Covar<=0

Def Covariance (x, Mean\_x, y, Mean\_y):

For length of x do:

Covar<= Covar +(x[i]-Mean\_x)\*(y[i]-Mean\_y)

End

iii. Estimate coefficients

B0 <= covariance (x, Mean\_x, y, Mean\_y) / variance (x, Mean\_x)

B1 <=Mean\_y – B0\*Mean\_x

iv. Predict

For every X in test set do:

Y <=B0+B1X

End

Store predicted values in csv file

**4.1.2 XGBoost Regressor:**

XGBoost stands for e**X**treme **G**radient **B**oosting. The implementation of the algorithm was engineered for efficiency of compute time and memory resources. [2]

***A. Algorithm***

XGBoost’s split finding greedy algorithm

Input: Dataset with proper input and output labels, I instance set of current node.

Output: Predict sales value and store in csv file

Begin

i. Gain <= 0

ii. For every i belongs to I do:

G <= sum (g i), H <= sum (h i)

iii. For k=1 to m do:

GL <=0, HL <=0

For j in sorted (I, by Xjk) do:

GL <= GL + gj, HL <= HL + hl

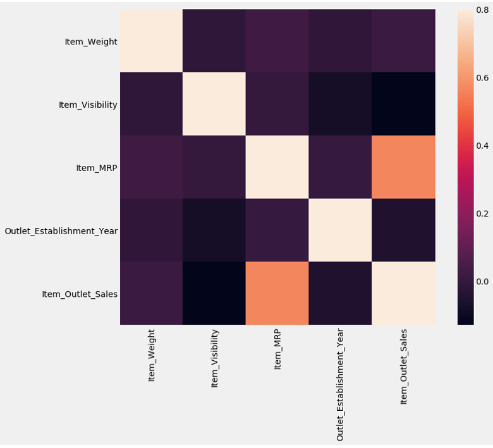
GR <= G - GL, HR <= H – HL

Score <= [ Score, GL^2/ (HL + λ) + GR^2/ (HR + λ) - G^2/ (H + λ)]

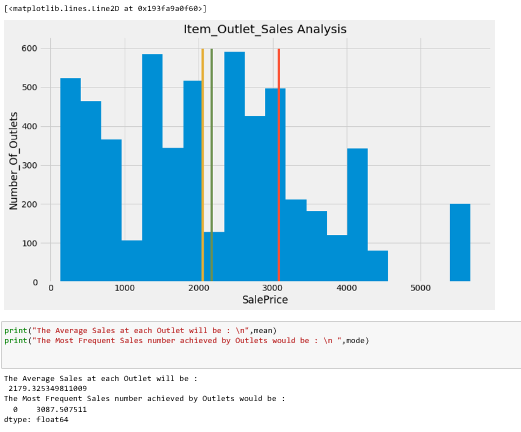
End

End

* 1. **Implementation**

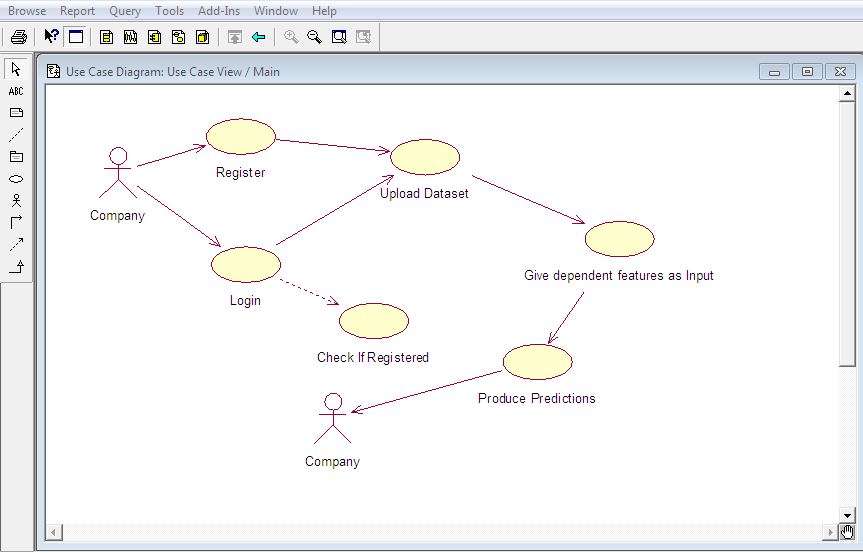


**Fig 4.2.1 Correlation Matrix**

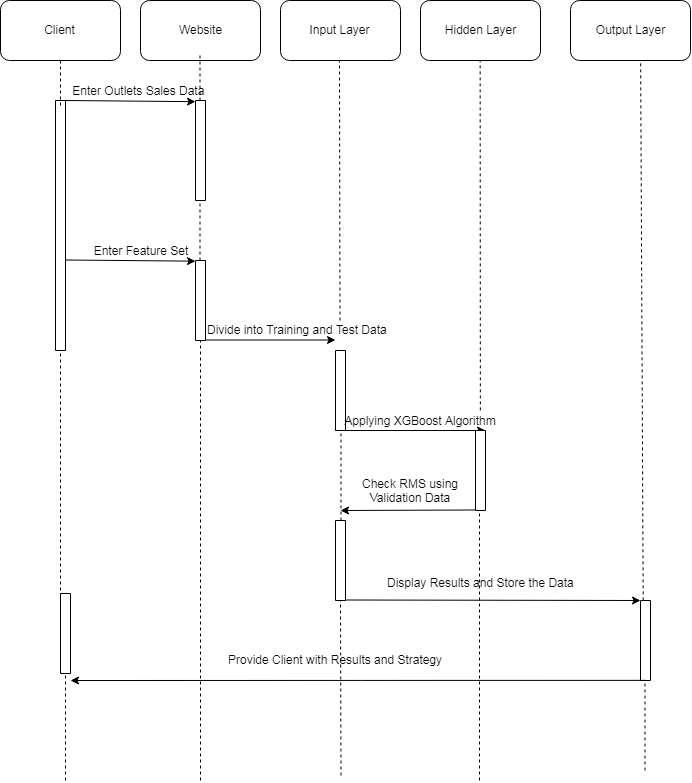


**Fig 4.2.2 Analysis Graph**

This approach will predict much better than other single classifiers when more data is used. Because of the difference is not statistically significant between the proposed model and random forest, the proposed method can be used to forecast demand due to its accuracy with less data.



**Fig 4.2.3 Use Case Diagram**



**Fig 4.2.4 Sequence diagram**

**Chapter 5: Conclusion and Future Scope**

**5.1 Conclusion**

The proposed system examines the problem of Sales forecasting on a dataset provided by the seller. Experiments have shown that this approach predicts demand up to 87% accuracy. The proposed method will predict much better when more data is used. Because the difference is not statistically significant between the proposed model and random forest, the proposed method can be used to forecast demand due to its accuracy with less data.

## **5.2Future scope**

In the future, this system will implement various technologies to enhance the prediction and use the output of this project as part of price optimization problem which we are planning to work on over a wide range of branches working on even large amount of data. Thus, making the analysis system faster for the companies and making them generate a better revenue.

**References**

1. http://www.esalestrack.com/article/sales-forecasting.html
2. [https://www.semanticscholar.org/paper/Explaining-machine-learning-models-in-sales-3. Bohanec-Borstnar/5da7d4b2f83a4e889ac4a1a03a05131da786ab4d](https://www.semanticscholar.org/paper/Explaining-machine-learning-models-in-sales-3.%20%20%20Bohanec-Borstnar/5da7d4b2f83a4e889ac4a1a03a05131da786ab4d)
3. <https://rstudio-pubs-static.s3.amazonaws.com/105869_f6e7f8d4e0434c40bd939a3d1e792af9.html>
4. <https://www.kaggle.com/c/competitive-data-science-predict-future-sales>
5. <https://ieeexplore.ieee.org/document/8073492/>