A Mini Project Report

on

“Company Sales Growth Analysis”

Submitted to the

Savitribai Phule Pune University

In partial fulfillment for the award of the Degree of

Bachelor of Engineering

in

Information Technology

by

33234 - Rushikesh Landge (T150058612)

33235 - Parimal Mahindrakar (T150058615)

33244 - Prajwal Patankar (T150058638)

33246 -Tanishk Rane (T150058655)

Under the guidance of

Prof D. D. Londhe

Department Of Information Technology

Pune Institute of Computer Technology College of Engineering

Sr. No 27, Pune-Satara Road, Dhankawadi, Pune - 411 043.

2020-2021

I

CERTIFICATE

This is to certify that the project report entitled "**Company Sales Growth Analysis"** Submitted by Rushikesh Landge, Parimal Mahindrakar, Prajwal Patankar and Tanishk Rane is a bonafide work carried out by them under the supervision of Prof. D. D. Londhe and it is approved for the partial fulfillment of the requirement of Software Laboratory Course-2015 for the award of the Degree of Bachelor of Engineering (Information Technology)

Prof. D. D. Londhe
Internal Guide
Department of Information Technology

Dr. A. M.Bagade
Head of Department
Department of Information Technology

Place: Pune
Date: 03/06/2021

ACKNOWLEDGEMENT

We thank everyone who have helped and provided valuable suggestions for successfully creating a wonderful project . We are very grateful to our guide, Prof. D. D. Londhe, Head of Department Dr. A. M. Bagade and our principal Dr. R. Sreemathy. They have been very supportive and have ensured that all facilities remain available for smooth progress of the project. We would like to thank our Prof. D. D. Londhe for providing very valuable and timely suggestions and help. We would also like the entire project staff team for providing valuable reviews and suggestions from time to time. We would like to thank our entire department and college staff for the very valuable help and coordination throughout the duration of the project. We would also like to thank our families and all our friends for the valuable support they provided throughout the duration of the project.

Rushikesh Landge
Parimal Mahindrakar
Prajwal Patankar
Tanishk Rane

Abstract

Humans have had a desire to exist in a flow which remains predictable and does not bring in changes. This leads to an effort made to predict the changes yet to come in order to not be faced by surprising factors which can shock us. The predictability of the upcoming changes is a force against the conventional way of nature being unpredictable. Humans with the help of some mathematics and large memory can analyze and identify patterns hidden in a set of related data and this we may call as our ability to predict. This process of using the machine to perform such analysis is called machine learning which is a branch of artificial intelligence. Several institutions established by humans which heed to a specific purpose would require this ability to analyze and correlate the data that we produce in the interwoven fabric of existence where we only contribute to activity being carried out without a sense of integrity. The conclusion of the effects we see around us and based on these conclusions our efforts to perform some action is simply enhanced with this tool of machine learning. The data or the representation of the analysis done itself is not an indicator of the path we are supposed to follow, rather it is the interpretation of the data at hand which will shape the direction in which humanity at that point will move towards. It is to be noted that sometimes it might be the case that due to that movement of humanity in a particular direction drawn out from the previous prediction the trend itself might completely change. It is the math and the symmetry of data and its correlation with our intent of action which help us draw out a plan of action.

Making meaningful estimates of the sales of a startup and therefore predict the growth of the startup based on approximate monthly or yearly estimates. The aim is to make the predictions more accurate and meaningful by using additional methods of logistic regression & random forest regression.

List of Figures

INDEX

Introduction

1.1 Introduction to the Project:

Humans have had a desire to exist in a flow which remains predictable and does not bring in changes. This leads to an effort made to predict the changes yet to come in order to not be faced by surprising factors which can shock us. The predictability of the upcoming changes is a force against the conventional way of nature being unpredictable. Humans with the help of some mathematics and large memory can analyze and identify patterns hidden in a set of related data and this we may call as our ability to predict. This process of using the machine to perform such analysis is called machine learning which is a branch of artificial intelligence. Several institutions established by humans which heed to a specific purpose would require this ability to analyze and correlate the data that we produce in the interwoven fabric of existence where we only contribute to activity being carried out without a sense of integrity. The conclusion of the effects we see around us and based on these conclusions our efforts to perform some action is simply enhanced with this tool of machine learning. The data or the representation of the analysis done itself is not an indicator of the path we are supposed to follow, rather it is the interpretation of the data at hand which will shape the direction in which humanity at that point will move towards. It is to be noted that sometimes it might be the case that due to that movement of humanity in a particular direction drawn out from the previous prediction the trend itself might completely change. It is the math and the symmetry of data and its correlation with our intent of action which help us draw out a plan of action.

1.2 Problem Statement:

Making meaningful estimates of the sales of a startup and therefore predict the growth of the startup based on approximate monthly or yearly estimates. The aim is to make the predictions more accurate and meaningful by using additional methods of logistic regression & random forest regression.

1.3 Motivation behind the project:

We as a team of aspiring engineers and pioneers of modern explorations are very much interested in knowing how humans have moved from the fundamental basis of a subject we created to the pinnacle of its application. This endeavor requires us to explore the various aspects of the subject we explore. We have chosen this project to analyze if the understanding that has been imparted to us can be used to a dataset that provides information about a startup and help us predict its growth.

1.4 Technology Used

- **Numpy:** NumPy is a library, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
- **Pandas**: It is a python library for data manipulation and analysis.
- **Matplotlib pyplot:** Matplotlib is a plotting library for creating static, animated, and interactive visualizations in Python. Pyplot is a module in myplot to display an interface like MATLAB.

- **Scikit learn:** Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms.
- **Scipy:** SciPy is a free and open-source Python library used for scientific computing and technical computing.
- **Matplotlib pylab:** Matplotlib is the most famous library for data visualization with python. It allows to create literally every type of chart with a great level of customization.

Literature Survey

2.1 Similar works

- A paper by Francisco Ramadas da Silva, Ribeiro Bento, published in 2017, tried predicting Startup Success With Machine Learning. The paper focused on creating a model that would classify a company/startup as successful/unsuccessful based on a standard classifier, TPR and FPR. Out of the three supervised ML algorithms tested- Support Vector Machines, Random Forests and Logistic Regression - Random Forests gave the best accuracy results based on the chosen dataset.

- Another paper by Kirasich Kaitlin, Smith Trace, Sadler Bivin, published in 2018 tried comparing Random Forest vs Logistic Regression: Binary Classification for Heterogene ous Datasets. This paper compares logistic regression and random forest classification for a dataset hat has an underlying structure. It is found that when increasing the variance in the explanatory and noise variables, logistic regression consistently performed with a higher overall accuracy as compared to random forest. However, the true positive rate for random forest was higher than logistic regression and yielded a higher false positive rate for dataset with increasing noise variables.

- A paper by Shen Rong, Zhang Baowen, published in 2108, researched about the regression model in Machine Learning. The paper analyses the sale of ice products affected by variation of temperature. The final result correctly leads the company to adjust the production and sale of iced products flexibly according to the variation of temperature, which definitely provides great commercial value and offers crucial theoretical foundation for the sale of other companies who produce Iced products.

2.2 Comparison of various methods

- Linear Regression:
  - Advantages
    - Simple to implement, easier to interpret the output.
    - Best technique for analysis of continuous variables
  - Disadvantages
    - Assumes a linear relationship between the dependent and independent variables.
    - Outliers can have huge effects on the regression.
- Logistic Regression
  - Advantages
    - easier to implement, interpret, and very efficient to train.
    - No scaling of data required.
    - Low variance – less prone to overfitting
  - Disadvantages
    - Susceptible to overfitting in case number of observations is lesser than the number of features.
    - Non-linear problems cannot be solved.

- Random Forest
  - Advantages
    - Works well on data with missing values and outliers
    - Very stable to addition of new points Works well when here are both categorical and numerical features
  - Disadvantages
    - Complex, require high computation power.
    - Longer training period
- Support Vector Machine
  - Advantages
    - works well with a clear margin of separation between classes.
    - Relatively memory efficient
    - Robust to outliers
  - Disadvantages
    - Not suitable for large datasets as it takes longer training time.
    - Does not perform well with overlapping target classes.

Applications

We have looked at different researches and implication models of these startup predicting machine learning algorithms, which aimed to predict whether the startup will be successful or not in the future. These predictions are derived by analyzing the company's sales record over a period of time, using different regression methods. Linear regression, logistic regression and random forest have been implemented in this project, before understanding how these regression models work let's look at similar examples where linear regression, logistic regression and random forest have been used.

- Studying engine performance from test data in automobiles. Linear regression models are sed to predict and examine the performance of an automobile engine. The performance of the automobile depends on various factors such as specific fuel consumption (SFC), energy generated, break mean effective pressure (BMEF), all these factors are taken under consideration while evaluating the performance of an engine and used to relate independent variables with dependent variables in LR model.
- Linear regression can also be used to analyze the marketing effectiveness, pricing and promotions on sales of a product. If a company wants to analyze the funds that they have invested on a particular brand has given them significant profit on investment, using linear regression they will be able to determine whether they should continue inviting in that specific brand in the future. This application being very similar to our application, it's quite different.
- Linear regression analysis is used in everyday astronomical research. In extra-galactic astronomy include relations between X-ray temperatures and velocity dispersions for galaxy clusters, the color-luminosity relations for field galaxies, and many other fields.
- Logistic regression works with any binary output data eg: if a customer will buy a car or not, or if a person is fit to play a game or not.
- The random forest can be used to predict if a customer is a fraud for a banking system.
- Medicines needs a complex combination of specific chemicals. Thus, to identify the great combination in the medicines, Random Forest can be used. With the help of machine learning algorithm, it has become easier to detect and predict the drug sensitivity of a medicine. Also, it helps to identify the patient’s disease by analyzing the patient’s medical record.
- Machine learning also plays role in the stock market analysis. When you want to know the behavior of the stock market, with the help of Random Forest algorithm, the behavior of the stock market can be analyzed. Also, it can show the expected loss or profit which can be produced while purchasing a particular stock.

In the next section we will understand how each of the linear regression, logistic regression and random forest models work.

Methodologies and Techniques

4.1 Linear regression:

- In the previous sections we have discussed different applications related to different regression models, before understanding how our projects works let's have a clear idea about what exactly is regression. A statistical method used in many domains which attempts to determine a relation between two variable, independent variables and dependent variables, such that we can generate an outcome which will be a predicted value based on the initial relation between dependent and independent variables.
- The linear regression attempts to represent the relationship between the dependent and independent variable by finding a linear equation. In linear regression our aim is to from a linear line, this line will be generated based on the data set provided, such that the line fits throw all the observations present in the dataset, this like is called as the regression line. The most known method to plot this linear regression line is the least square method. In a simple linear regression model, the for of the model will be:

$\hat{y} = b0 + b1 * x$ …..(i)

Where, x is the independent variable, $\hat{y}$ is the dependent variable, the slope of the line is b1 and b0 is the intercept value of the dependent variable when independent variable is 0. Also, b1 being the slop of the line, we can derive the value of b1 using:

$b1 = \sum(x-x')(y-y') / \sum (x-x')^2$

4.2 Logistic regression:

- It Logistic regression is a modification of linear regression in an exponential form to yield a binary output which we may call a classification factor. It is paralleled with the probability distribution because it gives a result between one and a zero which can be approximated to the closest it is to wither one or zero.

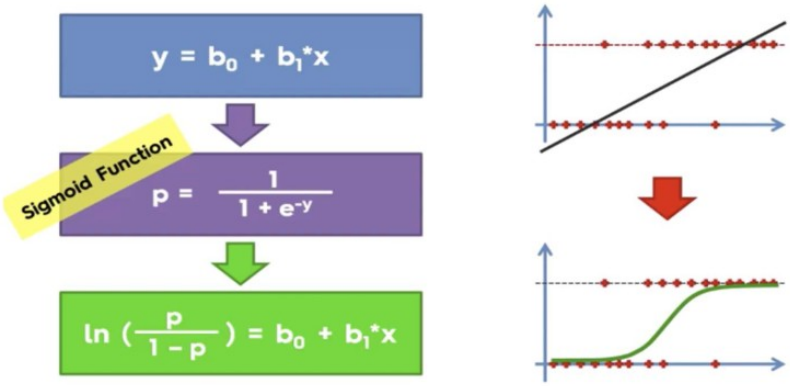

Figure 1. Activation Functions

- The above diagram shows that a linear regression curve can be converted into a logistic regression curve by strategically manipulating the data to represent a mathematical probability quantity called odds in favor or against. The natural logarithm of that odd is going to give rise to a sigmoid curve which helps us classify data logically in 1s and 0s.

4.3 Random Forest

- Random Forest is a supervised machine learning algorithm that is based on ensemble learning. Ensemble learning means joining one or more algorithms or using the same algorithm multiple times to form a powerful prediction model. In the case of Random Forest, it is latter i.e., decision tree algorithm is used multiple times which gives us a number of resultant trees. This forest of trees is what gives its name to the algorithm. Random forest can be used for both classification and regression analysis which is its biggest advantage.

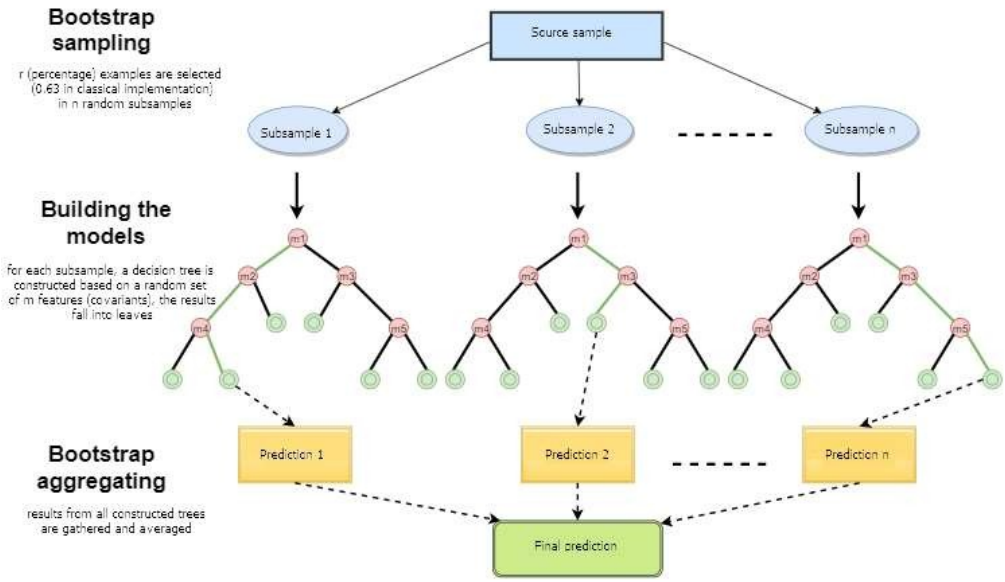

Figure 2. Random Forest

- How it works is basically by splitting the data continuously at each node based on the input features which then results in branches and more nodes called the child nodes. This process is repeated until all the data is distributed and there is no impurity or uncertainty left. At the end of the process, we have a forest of trees with an outcome each, and our goal is to combine all these outcomes and reach an average decision.

4.4 System Architecture:

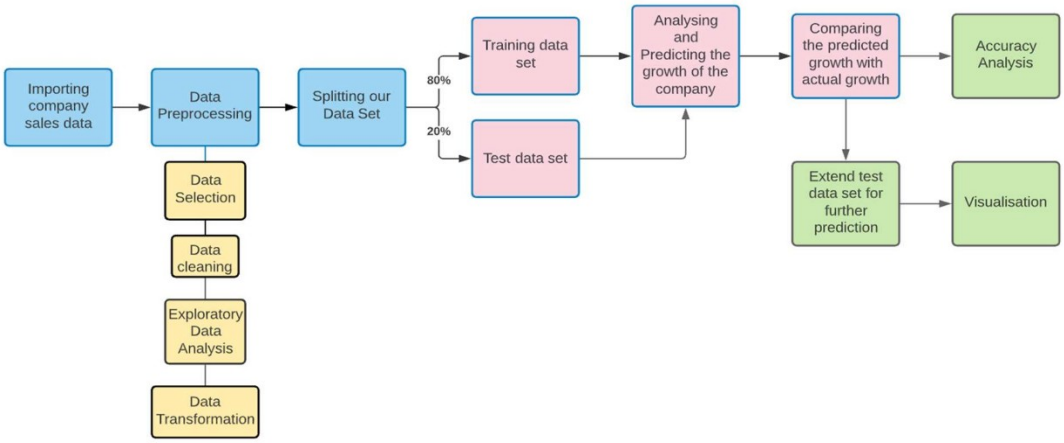

Figure 3. System architecture

- The above diagram is what happens at one particular stage, because we realized soon enough that if we had to make one diagram for the entire system that consists of 5 regression models and various manipulations and analysis, the diagram would be a couple of pages long. Thus, the diagram here shows just how each of those 5 regressions are modelled, executed, and visualized.

- Dataset Explanation: The dataset we chose was readily available on a public domain, and it contained the sales data of a new department store. This was perfect because it had the 2 essential features required for our analysis, the order dates, and the sale amount. Furthermore, it had a significant enough number of records for a reliable enough analysis, and the structure was very much like something that would be valid for any real-life company.

Conclusion

As the work on this project progressed, a number of fascinating things were discovered. Initially, we started off with the idea that we would implement a basic linear regression for quantitative prediction for sales versus time. Soon, after 3 different linear regression plots on 3 different instances of data, we realized that it was not that simple. Sales vary hugely depending on the time of the year, and if one looks at the data points individually, it is easy to predict that the sales of a company is stagnant. Later, we added the implementation of logistic regression with respect to the results found in linear regression, which was also quite intriguing. The prediction by logistic regression, although incorrect, was extremely helpful in realizing some other crucial factors, how exactly logistic regression works. To end the project, we then implemented a random forest regression algorithm, and went back from classification to numerical prediction, to predict the flow of existing data, as well as predict the progression of the later years. Here, we could notice a trend like what we found in the result of our linear regression. The sales of the upcoming year would be on average above the line of regression, which supports the claim that the sale would in fact increase. To conclude, we compared the accuracies of all the models used, namely linear, logistic, and random forest, and concluded that the random forest plot was the most accurate with the least error (approx. 87.5%) when comparing the predictions with the existing data set. With all these graphs, analysis, and prediction, we can conclude a number of things. Firstly, we can safely conclude that, according to our model, the company is expected to increase in average sales throughout the year. However, later study allowed us to realize that the predictions cannot be done solely on months, as seasons play a significant role in sales of products. Thus, we avoided any preemptive analysis of using only half a year, and thus predicted by feeding the data for entire years, even if the individual data points were in terms of dates. Thus, it is very easy to say that the sales will go down if you look at the first couple of months of any year in comparison to the previous year, and to avoid that we have used the data corresponding to each entire year. Therefore, our analysis has considered the seasonal factor and predicted the overall average sales throughout the year of 2019, the year for which no data points were present.

References

[1] A paper by Francisco Ramadas da Silva, Ribeiro Bento, published in 2017, Startup Success With Machine Learning.

[2] Paper by Kirasich Kaitlin, Smith Trace, Sadler Bivin, published in 2018 tried comparing Random Forest vs Logistic Regression: Binary Classification for Heterogene ous Datasets.

[3] A paper by Shen Rong, Zhang Baowen, published in 2108, researched about the regression model in Machine Learning.

[4] Tableau for dataset.

[5] Python Library Documentation for Regression Models