Sales Prophesy in Business using ML

Ravi Rajan T
Computer Science Department,
Karunya Institute of Technology,
Coimbatore.

Sarbash biswas
Computer Science Department,
DIT University, Dehradu,
Uttarakhand.

Vikashini TQ
Computer Science Department,
Velammal College of Engineering,
Madurai

Sailendri Gr
Computer Science Department,
Velammal College of Engineering,
Madurai.

***Abstract*-This paper describes our research on data analytics using machine learning algorithms such as Linear regression, Logistic regression, Clustering.The project is focused on predicting the Sales in each product category for different age groups using Linear regression algorithm and visualizing them for comparison to improve concentration in the particular market .creating a GUI interface to input the data and also to predict the average sales in each category.**

I. INTRODUCTION

This document is a final report from our summer internship 2019 project as an undergraduate student. The Smartbridge organisation partnered with IBM, provided us a platform to explore the knowledge on machine learning and artificial intelligence in IBM watson studios .

The agents in the organisation provided us knowledge on the different algorithms in machine learning such as Linear Regression, Logistic Regression, Clustering, Random Forest and Decision Tree. Also working on Linux software and the Terminal.

Black Friday is the largest shopping day of the year in UnitedStates of America [1]. Black Friday is the day after Thanksgiving Day which marks the beginning of the shopping season for Christmas. A prediction model developed for Black Friday can only be used during that day because customer spending differs drastically between a normal day and a Black Friday; this is because discounts and price reductions attract more customers. A study by the National Retail Federation states that 212 million shoppers visited stores and websites over the 2010 Black Friday weekend [1]. The major problem with the existing prediction model is that the data used for development contains several irregularities such as missing values or wrong information. Also, selection of right algorithm plays a major role in developing an accurate model.

In this vast data we are taking only a part of the whole black friday dataset that is the electronics sales alone so that we can categorize them not only with money but also with an interest in buying that category of products.

II. DATASET

The dataset used in this study is a sales transaction data on a black friday. It is publicly available on the following URL https://www.kaggle.com/mehdidag/black-friday. It has 550K sales transaction records. Each record has 12 features as listed in Figure 1. A retail company wants to understand the customer purchase behavior (specifically, Figure 2: Age (Graph 1,2,3 from Table 1) purchase amount) against various products of different categories. They have shared purchase summary of various customers for selected high-volume products from last month. The data set also contains customer demographics (age, gender), product details (product_id and product category) and total purchase amount from last month. Now, this dataset can be used to train a supervised machine learning algorithm to predict the purchase amount of customer against various products which will help the retailer to create personalized offers for customers against different products and to know the level of their interest in each product .So that they could develop products meeting the user interests and requirements and they could conquer the market ...In this project we are going to consider only few features .

III.DATA DISTRIBUTION STUDY

In machine learning algorithms, the dataset used must be balanced. All the classes should contain an equal number of samples otherwise the prediction or classification will be biased towards that category where the data is skewed. To remove any presence of imbalanced data, we studied the distribution based on the product id,gender,age.

IV.SIGNIFICANCE OF DATA PRE-PROCESSING

A good machine learning algorithm is of no use without the proper data. The accuracy of the prediction model increases only if the data it is built upon is solid [5]. But the real-world data is messy and needs to be cleaned. Barrosoetal States the three ways to analyze incomplete data: elimination of the units partially observed, reweighting of units and imputation [2]. If a part of data is beyond recovery, it is best to eliminate the entire data rather than use it in its present state. Another approach is the imputation method to recover the data. A method called hot deck imputation is involved where several imputations on the same data is done without a specific purpose, so that the dataset thus obtained is not biased [2]. Even after the data is cleaned, the entire data cannot be used as a whole for developing prediction model. The features in the dataset must be ranked according to their importance. Guyon et al. explains that variable and feature selection have become the focus of much research in areas of application for which datasets with tens or hundreds of thousands of variables are available [3]. The main idea in feature selection is to remove the redundant data or data that are similar to each other. This saves a lot of processing time during the development of the model. The steps involved in achieving this include variable ranking, developing correlation factors, subset selection and dimensionality reduction through principal component analysis [3].

V. DATA PREPARATION

The data was not compatible to be passed through the machine learning algorithms since the algorithms require standardized numeric data. Some of the features from the dataset contained categorical data like age in bins. Some of the features also contained both textual data and numbers. We had to convert the feature in either number or text. We also converted categorical data to numbers as follows:

Feature	Data	Mapping
Gender	M	0
Gender	F	1
Age	0-17	0
Age	18-25	1
Age	26-35	2
Age	36-45	3
Age	46-50	4
Age	51-55	5
Age	55+	6

In this way the product Id has also been mapped. Some of the values for the feature Product~category 2, and Product~category_3 were missing. We chose to fill it with average number for the missing values.

V. REGRESSION

A. Regression

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor) This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables. There are mainly 7 different types of regression available which are mostly dependent on 3 metrics. The shape of the regression line, the type of dependent variable and number of independent variable.

1) Linear Regression Linear Regression establishes a relationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line (also known as regression line). It is represented by an equation Y=a+b*X + e, where a is the intercept, b is the slope of the line and e is error term. This equation can be used to predict the value of target variable based on given predictor variable(s).

2) Logistic Regression Logistic regression is used to find the probability of event=Success and event=Failure. We should use logistic regression when the dependent variable is binary (O! 1, True!

False, Yes! No) in nature. Here the value of Y ranges from 0 to 1 and it can be represented by following equation. 18 odd s _ pi (1 -p) _ P~Obdbili ty o f event ll r e n I probability of not event occurrence I n(odd s) - In (p/ (l-p)) logit(p) _ In (p/ (l-p » _ be+blX 1+b2X2+b3X3 ...+bkXk

3) Polynomial Regression A regression equation is a polynomial regression equation if the power of independent variable is more than 1. The equation below represents a polynomial equation: y=a+b*x"2

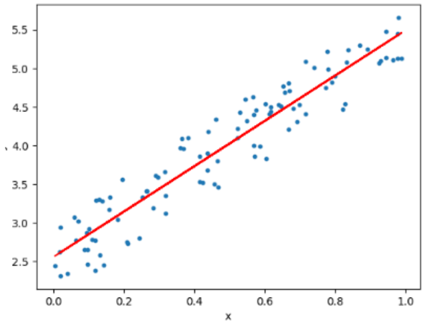
IV. SYSTEM SPECIFlCATION :

The implementation was done on AMD A6 7th generation processor with 4 GB RAM. The implementation was done with Python using python's skLearn and numpy libraries for machine learning algorithms.

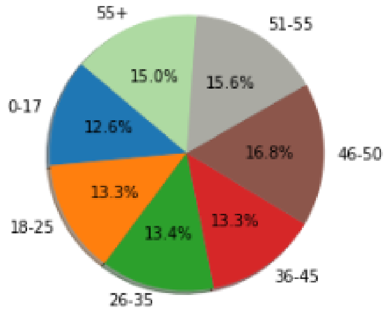
V. MACHINE LEARNING TECHNIQUES

To predict the purchase amount and which product category will be sold more, using multiple regression we implemented linear regression. For linear machine learning algorithm, we plotted a graph of Actual value vs Predicted value for validation dataset.

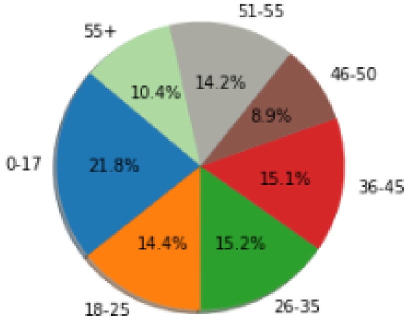
Linear Regression: The linear regression using python's skLearn library was implemented on the transformed dataset. This was the simplest of the implementation in terms of complexity of the model.The below figure is the plot of actual purchase amount vs predicted amount by this model.



BEFORE PREDICTION :
PRODUCT 1:



AFTER PREDICTION :
PRODUCT 1:



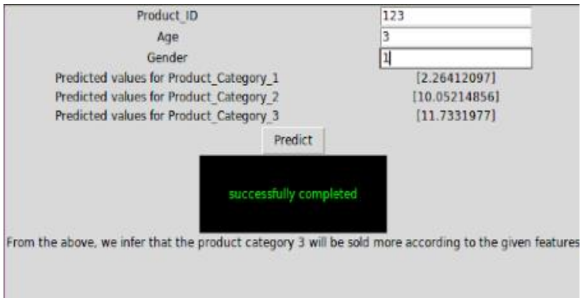
Similarly for remaining product types also,we can predict which product will have high demand and so the company can make a good profit out of that.

A) Decision Tree

Machine learning algorithms like decision tree and regression are used for developing a simple yet efficient prediction models. Guo et al. state that a time series analysis using early purchase patterns can be used to predict the future spending. The technique involved can be classified into two groups, mathematical and statistical models, and artificial intelligence model [4]. The Decision Tree technique comes under the artificial intelligence model, which develops a tree with root node containing the most important feature and subsequent nodes in the tree with less ranking features. To implement this model, skLearn was used. The RMlSE for this model is 3800. But it has less accuracy so, It's better to stick with the Linear Regression.

VI.GUI:

Any code without GUI won't be so appealing so, we made a GUI using tkinter.It takes three features like Product_ID,Age,Gender and returns predicted value which helps to determine which object will have high Demand.The GUI for this model is represented as



Product_ID	123
Age	3
Gender	1
Predicted values for Product_Category_1	[2.26412097]
Predicted values for Product_Category_2	[10.05214856]
Predicted values for Product_Category_3	[11.7331977]

successfully completed

From the above, we infer that the product category 3 will be sold more according to the given features

VII.Conclusion and Future work

We conclude that the complex models like neural network are an overkill for simple problems like regression. And simpler models along with proper data cleaning perform well for the regression.

Also, based on the current trend, the number of shoppers on the Black Friday is only going to increase. The study agrees that machine learning techniques produce better prediction models that can be used at stores and the store owners can analyze their customer base to better target the customers and increase the sales on a Black Friday.

The study also agrees that the data must be pre-processed to attain an effective dataset for developing the prediction model. Several techniques were discussed in this study to attain the best model. However, there is still no definite solution as to what the correct technique is to attain a model with high accuracy.

To improve the results, a dataset with sufficient features and increase in quantity must be obtained. Further research must be conducted in enhancing the existing machine learning techniques to work in real time and develop an efficient model. Also, the models developed must be tested on data with different volumes to test its scalability and performance.

In future work, the result of regression on balanced dataset can be studied by changing the data distribution . This can be done by selecting a sample of dataset or removing certain records to balance the type of data.