Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data

Ching-Seh (Mike) Wu
Department of Computer Science
San Jose State University
San Jose, CA 95192
ching-seh.wu@sjsu.edu

Pratik Patil
Department of Computer Science
San Jose State University
San Jose, CA 95192
pratic, patil@sjsu.edu

Saravana Gunaseelan
Department of Computer Science
San Jose State University
San Jose, CA 95192
saravana.gunaseelan@sjsu.edu

Abstract-During the Black Friday sale, all the retail shops are crowded. Most products are marked down with discounts and customers rush in to buy the products. It is difficult for customers to buy the products even with a solid plan. But, the shop owners face even more difficulty on controlling the crowd with limited staff and in targeting prospective customers. Several techniques have been employed to tackle this problem, but they are not that successful. A prediction model is a technique that has proved promising in solving the problem. This study focuses on the field of prediction models to develop an accurate and efficient algorithm to analyze the customer spending in the past and output the future spending of the customers with same features. In this study, different machine learning techniques such as regression and neural network to develop a prediction model are implemented and a comparison is done based on their performance and accuracy of prediction. These techniques are implemented using different algorithms and on different platforms to find the best predication. We implemented seven different machine learning algorithms. Further, this study discusses the data pre-processing and visualization techniques employed to attain the optimal results.

Keywords: Black Friday Sales, Prediction model, Regression, Neural network, Machine Learning

I. INTRODUCTION

In the past, there were no supermarkets or departmental stores, only small businesses. The store owners knew their customers and their spending patterns, their likes and their dislikes. But as the small business grew into large franchises with hundreds of stores across the country, it became near to impossible to know the customers and their personal preferences. Some examples of such franchises are Costco, Walmart, and Wholefoods. These stores without any proper knowledge of their customer base are struggling to satisfy the customer needs. Thus, prediction models are needed to better understand customer preferences.

Building a prediction model depends on various features such as the location and the time. Black Friday is the largest shopping day of the year in United States 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. Finally, better visualization techniques are required to portray the findings and help the store owners understand their customers. Some questions involved in this field are: Can an accurate prediction model be developed? Which algorithm is better and efficient for such model? Will data pre-processing and visualization technique increase the accuracy?

This study is organized as follows: Section II presents a data pre-processing technique. Section III demonstrates the development of prediction model using regression and types of regression. Section IV lists the system specifications used in this study. Section V demonstrates the different machine learning algorithms used along with its accuracy and performance. Section VI concludes the study by summarizing the findings of the implementations.

II. DATAI'RE-PROCESSING

A. Dataset

The dataset used in this study is a sales transaction data. It is publicly available on the following URL. https ://datahackanalyticsvidhya.com/contest/black-friday/#data_dictionary

It has 550K sales transaction records. Each record has 12 features as listed in the Figure 1. A retail company wants to understand the customer purchase behavior (specifically,

978-1-5386-6565-7118/$31.00 ©2018 IEEE

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, marital status, city_type, stay_in_current_city), 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 offer for customers against different products.

Data

Variable	Definition
	User ID
Product_ID	Product ID
Gender	Sex of User
Age	Age in t uns
Occupation	Occupation (Masked)
City_Category	Category of the City (A,B,C)
Stay_In_Current_City_Years	Number of years stay in current city
Marital_Status	Marital Status
Product_Category_1	Product Category (Masked)
Product_Category_2	Product may belongs to other category also (Masked)
Product_Category_3	Product may belongs to other category also (Masked)
Purchase	Purchase Amount (Target Variable)

Figure 1: Feature description from the dataset.

B. *Data distribution study*

In machine learning algorithms, the dataset used must be balanced. All the classes should contain 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 9 following parameters.

Table 1: Data distribution study

1. Age group Vs Total Spending
2. Age group Vs Count
3. Age group Vs Per head spending
4. Gender Vs Total Spending
5. Gender Vs Count
6. Gender Vs Per head spending
7. Marital Status Vs Total Spending
8. Marital Status Vs Count
9. Marital Status Vs Per head spending

We noticed that there were 200% more males in the dataset than the females, 100% more people in the age group 26-35 as compared to the second highest group and 50% more unmarried people than married people. However, the per-head spending across every category is constant.

Figure 2: Age (Graph 1,2,3 from Table 1)

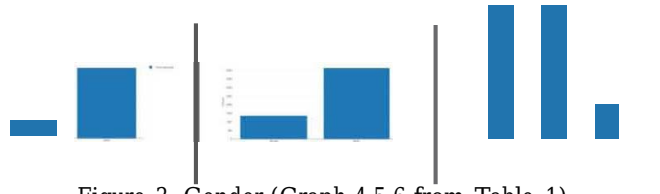

Figure 3: Gender (Graph 4,5,6 from Table 1)

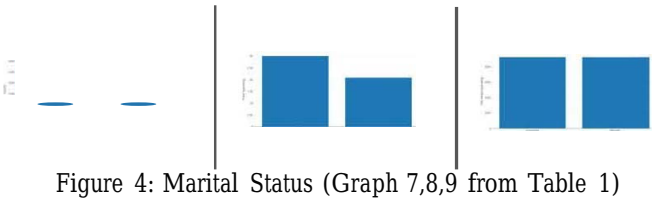

Figure 4: Marital Status (Graph 7,8,9 from Table 1)

C. *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. Barraso et al. 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 principle component analysis [3].

D. *Data Preparation*

The data was not compatible to be passed through the machine leaming 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. The following Table 2 explains the mapping and data transformation for this dataset.

Table 2: Data Mapping to Numbers

Feature	Initial Data Type	Mapping
Gender	M	1
Gender	F	2
City	A	1
City	B	2
City	C	3
Married	M	1
Unmarried	U	2
Age	0-17	17
Age	18-25	25
Age	26-35	35
Age	36-45	45
Age	46-50	50
Age	51-55	55
Age	55+	57

The product id was a semi-textual and numeric of the format P[0-9]*. We converted it to a pure number by removing the initial letter 'P'. Some of the values for the feature Product_category_2, and Product_category_3 were missing. We chose to fill it with 0 for the missing values. Google' Open Refine tool was used for the cleaning and transformation of this dataset.

III. 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 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 intercept, b is 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.

odds = p/ (1-p) _ probability of event occurrence / probability of not event occurrence

ln(odds) - ln(p/(1-p))

logit(p) _ ln(p/(1-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

4) Stepwise Regression

This form of regression is used when we deal with multiple independent variables. In this technique, the selection of independent variables is done with the help of an automatic process, which involves no human intervention.

This feat is achieved by observing statistical values like R-square, t-stats and Ale metric to discern significant variables. Stepwise regression basically fits the regression model by adding!dropping co-variates one at a time based on a specified criterion.

5) Ridge Regression

Ridge Regression is a technique used when the data suffers from multicollinearity (independent variables are highly correlated). In multicollinearity, even though the least squares estimates (OLS) are unbiased, their variances are large which deviates the observed value far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.

6) Lasso Regression

Similar to Ridge Regression, Lasso (Least Absolute Shrinkage and Selection Operator) also penalizes tbe absolute size of the regression coefficients. In addition, it is capable of reducing the variability and improving the accuracy of linear regression models. Lasso regression differs from ridge regression in a way that it uses absolute values in the penalty function, instead of squares.

7) ElasticNet Regression

ElasticNet is hybrid of Lasso and Ridge Regression techniques. It is trained with L1 and L2 prior as regularizer. Elastic-net is useful when there are multiple features which are correlated. Lasso is likely to pick one of these at random, while elastic-net is likely to pick botb.

IV. SYSTEM SPECIFICATION

The implementation was done on Intel i7 machine having Nvidia GeForce 940M with CUDA enabled. 12 GB RAM. The implementation was done with Python using python's skLearn and numpy libraries for machine learning algorithms and Kerns library for Artificial Neural Network implementation.

V. MACHINE LEARNING TECHNIQUES

To predict the purchase amount using multiple regression we implemented machine learning algorithms and compared them on accuracy and performance metric. Since it is a regression problem, the loss function used is the Root Mean Squared error (RMSE). For every machine learning algorithm, we plotted a graph of Actual value vs Predicted value for validation dataset.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

A. *Linear Regression*

The linear regression using python's skLearn library was implemented on the transformed dataset. This was the simplest of the implementations in terms of complexity of the model. The mean squared error was of the order of 4800. The below figure is the plot of actual purchase amount vs predicted amount by this model.

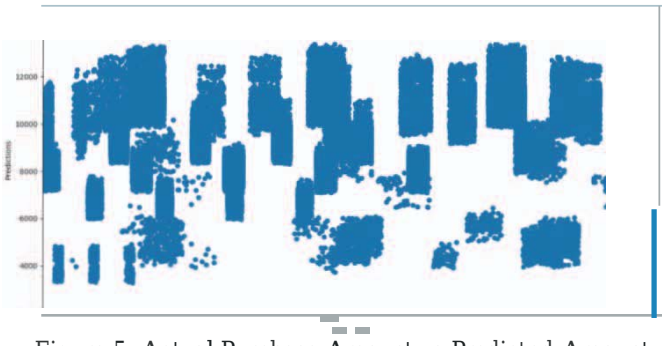

Figure 5: Actual Purchase Amount vs Predicted Amount

B. *MLK classifier*

This is skLearn's implementation of neural network. We used this implementation to build regression model which gave the RMSE of 6000 which is the worst of all the implementations. The below figure is the plot of actual purchase amount vs predicted amount by this model.

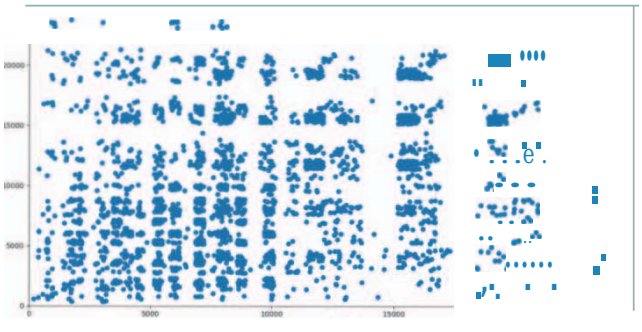

Figure 7: Actual Purchase Amount vs Predicted Amount

C. *Deep Learning Model using Keras*

For this implementation we developed a regression model using Deep neural network. We used 3 dense layers with I hidden layer having 500 nodes. For some reason the deep learning model could not predict values below $6000 as seen from the below figure. The RMSE for this model was 4200.

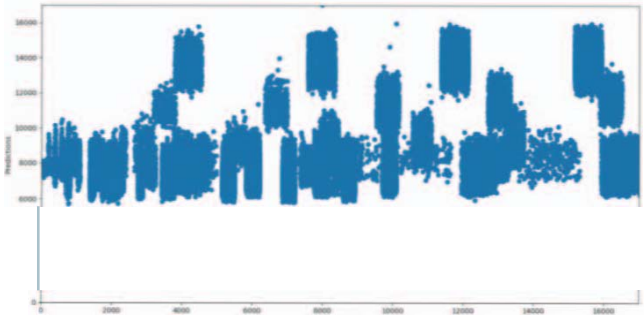

Figure 8: Actual Purchase Amount vs Predicted Amount

D. *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 patters can be used to predict the future spending. The technique involved can be classified into two groups, mathematical and statistical model, 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 RMSE for this model is 3800.

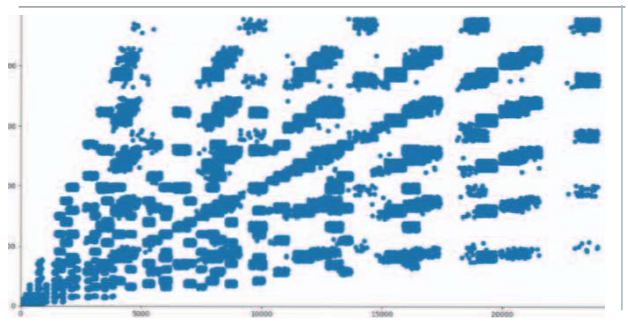

Figure 9: Actual Purchase Amount vs Predicted Amount

E. *Decision tree with Bagging.*

This is modified implementation of above version of DT. The RMSE for this implementation is 2900. Bagging is the process of developing different sets of training and testing sets for each iteration. By this method errors arising due to overfilling and bias can be rectified. The bagging method develops multiple model and chooses the best among them.

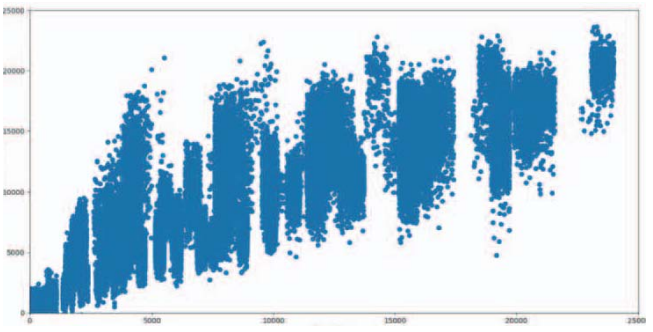

Figure 10: Actual Purchase Amount vs Predicted Amount

F. XGBoost

The XGBoost model internally implements the stepwise, ridge regression which dynamically selects the features and removes the multi-collinearity with the features. This implementation gave the bet results of this dataset. It uses ensemble model to learn from the weak predictors and eliminate the less important features to develop a strong model. We got a RMSE of 2400.

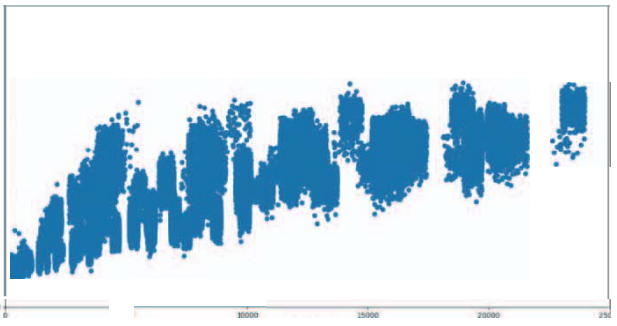

Figure 11: Actual Purchase Amount vs Predicted Amount

G. RMSE for all algorithms

The below figure depicts the plot of RMSE for all the above implementation for visual comparison.

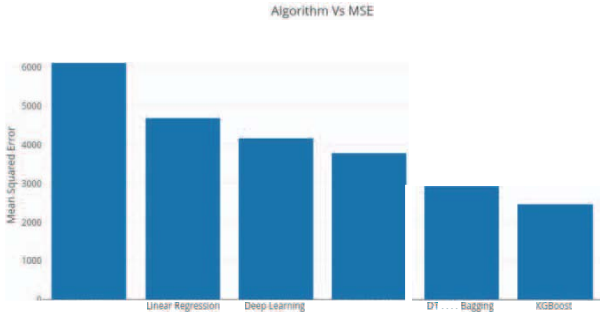

Figure 12: Algorithm Vs RMSE

VI. 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.

REFERENCES

[1] M. Petrescu and M. Murphy, "Black Friday and Cyber Monday: a case study" in International Journal of Electronic Marketing and Retailing (IJEMR), voL 5, no.3, 2013.

[2] L. P. Barroso, W. O. Bussab, and M. Knott, "Best linear unbiased predictor in the mixed model with incomplete data," Communications in Statistics Theory and Methods, vol. 27, no. 1, pp. 121-129, 1998. doi: *10.1080103610929808832654*J.

[3] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," J. Mach. Learn. Res., vol. 3, pp. 1157-1182, Mar. 2003.

[4] Z. X. Guo, W. K. Wong, and M. Li, "A multivariate intelligent decision-making model for retail sales forecasting," Decision Support Syst., voL 55, pp-247-255, Apr. 2013.

[5] A. Soroush, A. Bahreininejad, and 1. van den Berg, "A hybrid customer prediction system based on multiple forward stepwise logistic regression mode," Intell. Data Anal., vol. 16, pp. 265-278, Mar. 2012.

[6] L. Bing and S. Yuliang, "Prediction ofuser's purchase intention based on machine learning," 3rd International Conference on Soft Computing Machine Intelligence (ISCMI)., pp.99·103, Nov. 2016.

[7] Y. Qin and H. Li, "Sales forecast based on BP neural network", 2011 IEEE 3rd International Conference on Communication Software and Network., pp. 186-189, May 2011

[8] K. Singh and R. Wajgi, "Data analysis and visualization of sales data," 2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave), Coimbatore, pp. 1-6, Mar. 2016.

[9] https j/datahack.analyticsvidhya.com/contest!black·friday/#data_dictionary

[10] https .Zwww.analyticsvidhya.comlblog/2015108/comprehensive-guide-regression!

[11] https ://lmachinelearningmastery.comlregression-tutorial-keras-deep-leaming-library-python!

[12] http ://scikit-leam.org/stable/auto_examplesilinear_modellplot_ols.html