THINK INDIA JOURNAL

ISSN:0971-1260

Vol-22-Issue-41-December-2019

Black Friday Sales Prediction and Analysis

P. Rushitha Reddy[1] and K. Sravani[2]

Student[1] and Assistant Professor[2], Department of Computer Science and Engineering, Sreyas Institute of Engineering and Technology, Hyderabad, Telengana

ABSTRACT

A company wants to understandthe customer purchase behavior (dependent variable) against different products using their demographic information as features where most of the features are self-explanatory. This dataset consists of null values, redundant and unstructured data. Machine Learning is one of the most obvious applications in the domain of retail industry. This concept helps to develop a predictor that has a clear commercial value to the store owners as it would help with their financial planning, inventory management, marketing, and advertising. This entire process of developing a model includes preprocessing, modelling, training, testing and evaluating. Therefore, frameworkshave been developed to automate some of this process and hide away its complexity. The algorithm we proposed was Random Forest regressor that performed an average accuracy of 83.6% and with minimum RMSE (Root Mean Squared Error) value of 2829 on the Black Friday sales dataset.

Keywords: Black Friday, Sales Prediction, Data Analysis, Random Forest Regressor, Testing and Training.

INTRODUCTION

Black Friday is the name given to the shopping day after Thanksgiving. It was originally called Black Friday because the volume of shoppers created traffic accidents and sometimes even violence. Police coined the phrase to describe the mayhem surrounding the congestion of pedestrian and auto traffic in downtown shopping areas.

In a retail industry, the number of sales play an important part that decide the loss or profit for the company. Predicting the sales accurately gives efficient industry management. Black Friday is the largest shopping day of the year in United States of America. 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 customers' behavior is to be analyzed in order to predict the amount of purchase to be done by him/her on a particular day. In this paper, we will predict the sales of a company on "Black Friday".

To predict the sales of different products based on their independent variables, we need to analyze the relationship between different variables and well organize the data. So that a model can perform calculations and predicts sales accurately.

OBJECTIVE

This paper emphasis two objectives. They are the following:

1. Analyzing the data of all the customers and finding relationship of independent variables with respect to the target variable.
2. Predicting the expected sales by testing and training

ALGORITHM

Regression analysis is a form of predictive modelling technique which investigates therelationship 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. Mostly dependent on 3 metrics. The shape ofregressionline, the type ofdependent variable and number ofindependentvariables.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).

Copyright © 2019Authors

ISSN:0971-1260

Vol-22-Issue-41-December-2019

In our experiments we performed data pre-processing in the first part which gives the structured data that is split into two parts called training set and testing set to check the accuracy. The below figure represents the flow of data.

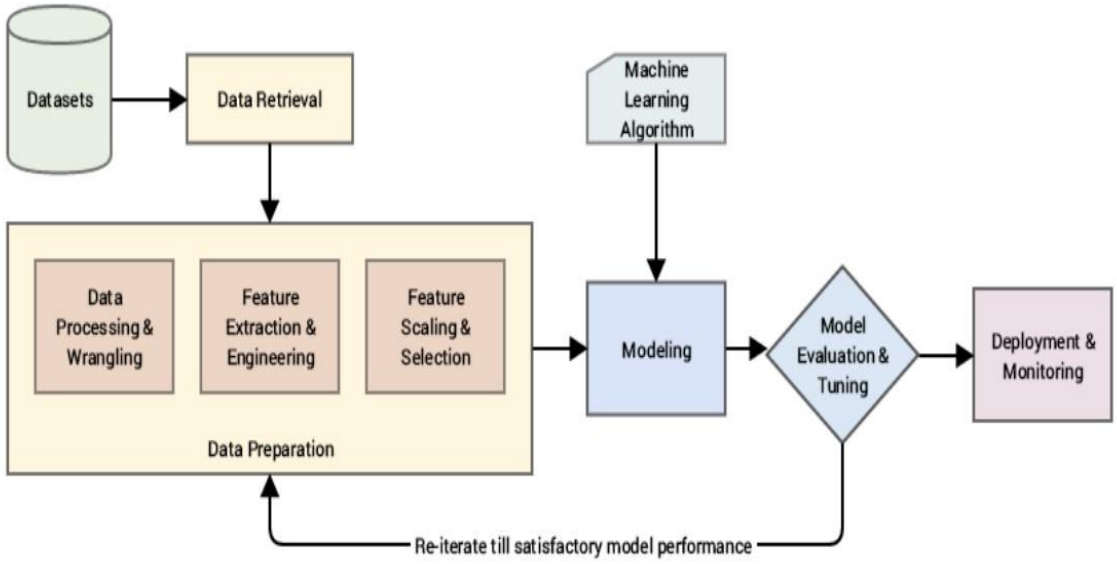

Figure 1: Data flow Architecture

FRAMEWORK

Various tools and techniques used in the work are described in this section.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 biasedtowards that categorywhere the data is skewed.

TOOLS AND ALGORITHMS

The machine learning algorithms used in the work are described in detail in this section.

RANDOM FOREST REGRESSOR

A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

The **random forest** model is a type of additive model that makes predictions by combining decisions from a sequence of base models. More formally we can write this class of models as:

g(x)=f0(x)+f1(x)+f2(x)+...

Different kinds of models have different advantages. The random forest model is good at handling tabular data with numerical features, or categorical features with fewer than hundreds of categories. Unlike linear models, random forests are able to capture non-linear interaction between the features and the target.

K-FOLD CROSS VALIDATION

Cross-validation is a statistical method used to estimate the skill of machine learning models. This technique involves randomly dividing the dataset into k groups or folds of approximately equal size. The first fold is kept for testing and the model is trained on k-1 folds on the dataset.We evaluate the model performance of testing and training based on an error metric to determine the accuracy of the model. This method however, is not very reliable as the accuracy obtained for one test set can be very different to the accuracy obtained for a different test set. *K-fold Cross Validation(CV)* provides a solution to this problem by dividing the data into folds and ensuring that each fold is used as a testing set at some point.

Copyright © 2019Authors

ISSN:0971-1260

Vol-22-Issue-41-December-2019

GRID SEARCH

It is a factor that chooses the parameters for an algorithm. It gives the hyper parameters based on the performance which results in the most accurate predictions.

Grid-searching is the process of scanning the data to configure optimal parameters for a given model. Depending on the type of model utilized, certain parameters are necessary. Grid-searching does NOT only apply to one model type. Grid-searching can be applied across machine learning to calculate the best parameters to use for any given model. It is important to note that Grid-searching can be extremely computationally expensive and may take your machine quite a long time to run. Grid-Search will build a model on each parameter combination possible. It iterates through every parameter combination and stores a model for each combination.

RMSE (ROOT MEAN SQUARED ERROR)

Root Mean Square Error (RMSE) is the standard deviation of the residuals prediction errors. Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results.

EXPERIMENTS

A set of experiments have been conducted on the dataset which were based on data mining and machine learning techniques like data preprocessing, univariate analysis, bivariate analysis and then applied for testing and training using Random forest Regressor. All these experiments performed in the model classify the data set accurately

PRE-PROCESSING

The data has to be pre-processed Before we can apply machine learning algorithms to our dataset, we need to convert it into a certain form that machine learning algorithms can operate on. The task of the learning algorithms will be to predict the value of the Purchase variable, given customer information as input.

UNIVARIATE ANALYSIS

It is the simplest form of statistical analysis. Univariate analysis can yield misleading results by studying single variable.

BIVARIATE ANALYSIS

Firstly, we individually analyzed some of the existent features, then we understand the relationship between our target variable and predictors as well as the relationship among predictors.

OUTLIER ANALYSIS

An outlier is an element of a data set that distinctly stands out from the rest of the data. We categorize the outliers into the existing class interval values and represent them using boxplot method.

DATASET

Model is trained by giving a complete data on which supervised learning can be done. It is publicly available in the following URL.

https://datahackanalyticsvidhya.com/contest/black-friday/#data_dictionary.

Copyright © 2019Authors

ISSN:0971-1260

Vol-22-Issue-41-December-2019

Data

Variable	Definition
User_ID	User ID
Product_ID	Product ID
Gender	Sex of User
Age	Age in bins
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)

RESULT AND DISCUSSION

If we compare the performance and accuracy of one algorithm with other machine learning algorithms. The algorithms applied in this work gave out high accuracy values. But among the various experiments done, the black Friday sales is best predicted by Random Forest Algorithm with an accuracy of around 81% and round mean squared error score as 2829.09. Other algorithms have similar accuracy with a variation of about 10%.

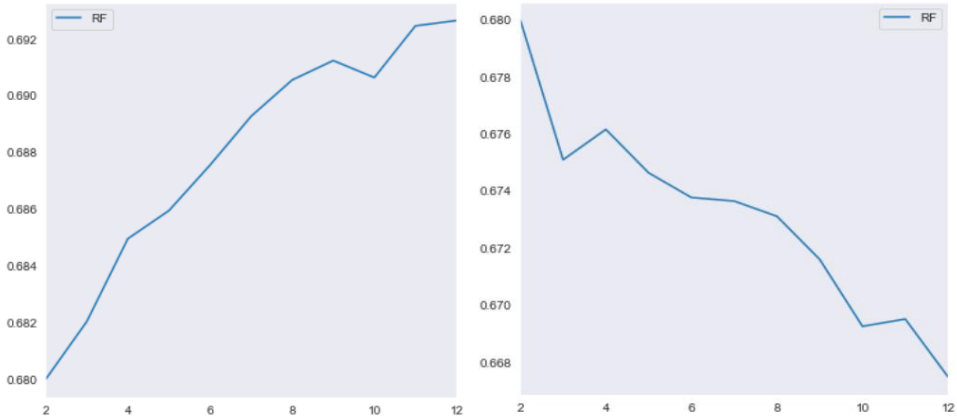

Figure 2: Graphical Representation of test and train values

CONCLUSION AND FURTHER WORK

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. It is also agreed that the data must be pre-processed to attain an effective dataset for developing the prediction model.

Data cleaning and analysis can be better done and other machine learning algorithms can be applied on the model to improve the accuracy. Increased dataset will give out more accurate predictions. 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.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.

Copyright © 2019Authors

ISSN:0971-1260

Vol-22-Issue-41-December-2019

ACKNOWLEDGEMENT

The author is extremely thankful to K. Sravani, Department of CSE, Sreyas Institute of Engineering and Technology for her guidance and valuable remarks throughout the research work. I would also like to show my gratitude towards eminent researchers whose work helped me to understand various aspects of this concept

REFERENCES

A prediction model for sales data has been developed using various methodologies. Some of the approaches are listed below.

1. Ching-Seh Mike Wu ; Pratik Patil ; Saravana Gunaseelan(2018). “Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data.”2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS).
2. Sunitha Cheriyan ; Shaniba Ibrahim ; Saju Mohanan ; Susan Treesa(2018). “Intelligent Sales Prediction Using Machine Learning Techniques.” 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE).
3. Gopalakrishnan T, Ritesh Choudhary, Sarada Prasad(2018). “Prediction of Sales Value in Online shopping using Linear Regression.” 2018 4th International Conference on Computing Communication and Automation (ICCCA).
4. Frank M. Thiesing, Ulrich Middelberg, Oliver Vornberger. “Short term prediction of sales in supermarkets.” Proceedings of ICNN'95 - International Conference on Neural Networks.
5. K.Sravani,K.Manohar,S Irfan.” Object Recognition with Improved Features Extracted from Deep Convolution Networks(IJET) vol 7 pp-1203-1209
6. Kumari Punam ; Rajendra Pamula ; Praphula Kumar Jain (2018). "A Two-Level Statistical Model for Big Mart Sales Prediction," 2018 International Conference on Computing, Power and Communication Technologies (GUCON).
7. 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
8. K.L.A.Nivedita,K.Sravani.”Fuzzy type ahead search in XML Data”(IJERA) Vol 3 Issue 4,Jul-Aug 2013.
9. K. Singh and R. Wajgi, "Data analysis and visualization ofsales data," 2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave), Coimbatore,pp. 1-6, Mar. 2016.
10. Rohit Raja, Tilendra Shishir Sinha, Ravi Prakash Dubey (2016), Orientation Calculation of human Face Using Symbolic techniques and ANFIS, Published in International Journal of Engineering and Future Technology, Vol. 7, Iss.7, pp. 37-50, ISSN: 2455-6432.
11. Rehmat Khan, Rohit Raja (2016) Introducing L1- Sparse Representation Classification for facial expression, Published in Imperial Journal of Interdisciplinary Research (IJIR), Vol. 2, Iss. 4, pp. 115-122, ISSN: 2454-1362.
12. Nikita Rawat, Rohit Raja (2016), A Survey on Vehicle Tracking with Various Techniques”, International Journal of Advanced Research in Computer Engineering and Technology (IJARCET), Vol. 5 Iss. 2, pp. 374-377, ISSN: 2278-1323.
13. Nikita Rawat, Rohit Raja (2016), Moving Vehicle Detection and Tracking using Modified Mean Shift Method and Kalman Filter and Research, International Journal of New Technology and Research (IJNTR), Vol. 2, Iss. 5, pp. 96-100, ISSN: 2454-4116.
14. Ramya laxmi K., Pallavi S., Ramya N. (2020) A Hybrid Approach of Wavelet Transform Using Lifting Scheme and Discrete Wavelet Transform Technique for Image Processing. In: Satapathy S., Raju K.,

Copyright © 2019Authors

ISSN:0971-1260

Vol-22-Issue-41-December-2019

Shyamala K., Krishna D., Favorskaya M. (eds) Advances in Decision Sciences, Image Processing, Security and Computer Vision. ICETE 2019. Learning and Analytics in Intelligent Systems, vol 3. Springer. (Scopus)

15. Sumati Pathak, Rohit Raja, Vaibhav Sharma, and K. Ramya Laxmi, (2019) A Framework Of ICT Implementation On Higher Educational Institution With Data Mining Approach, European Journal of Engineering Research and Science, Vol. 4, Iss. 5, pp. 34-38, Publication date 13/5/2019, ISSN (Online) : 2506-8016.
16. K. Ramya Laxmi, N Ramya, S. Pallavi, (2018) A Survey on Automatically Mining Facets for Queries from their search Results, International Journal of Management Technology and Engineering IJMTE Vol. 8, Iss. 7 July 2018. ISSN NO: 2249-7455 (UGC Approved).
17. S. Pallavi, K. Ramya Laxmi, N. Ramya, Rohit Raja (2018), Study and Analysis of Modified Mean Shift Method and Kalman Filter for Moving object Detection and Tracking, Published in 3rd International Conference on Computational Intelligence and Informatics (ICCII-2018), held during 28-29 Dec 2018.
18. K. Ramya laxmi, S. Pallavi, N. Ramya, (2019) A Hybrid Approach of Wavelet Transform using Lifting Scheme and Discrete Wavelet Transform Technique for image processing, 2nd National Conference on Cyber Security, Image Processing, Graphics, Mobility and Analytics (NCCSIGMA 2019), Organized by Department of CSE at CMR Technical Campus, Hyderabad in association with DIV – 5 Education & Research, CSI India from 24th – 25th Jan 2019.
19. K. Ramya laxmi, N. Ramya, S. Pallavi, K. Madhuravani, (2019) Study and Analysis of Apriori and K-Means Algorithms for Web Mining, 8th International Conference On "Innovations In Electronics & Communication Engineering (ICIECE-2019)" On August 02-03, 2019.
20. K. Ramya laxmi, Marri Abhinandhan Reddy, CH. Shivasai, P. SandeepReddy, 8th International Conference On "Innovations In Electronics & Communication Engineering (ICIECE-2019)" On August 02-03, 2019.

Copyright © 2019Authors