Purchase Predicting with Clickstream Data

Submitted By

Tanzim Rizwan

ID: 2013-1-60-063

Department of Computer Science and Engineering

East West University

Supervised By

Dr. Mohammad Rezwanul Huq

Assistant Professor

Department of Computer Science and Engineering

East West University

Dhaka, Bangladesh

April, 2017

Purchase Predicting with Clickstream Data

Submitted By

Tanzim Rizwan
ID: 2013-1-60-063

Supervised By

Dr. Mohammad Rezwanul Huq
Assistant Professor,
Department of CSE, EWU.

A project

Submitted in partial fulfillment of the requirements

for the degree of Bachelor of Science to

Computer Science and Engineering

Dhaka, Bangladesh

April, 2017

Declaration

This project has been submitted to the department of Computer Science and Engineering, East West University in the partial fulfillment of the requirement for the degree of Bachelor of Science in Computer Science and Engineering by us under the supervision of Dr. Mohammad Rezwanul Huq, Assistant Professor at Department of CSE at East West University under the course 'CSE 497' . I am also declaring that this project has not been submitted elsewhere for the requirement of any degree or any other purposes. This project complies with the regulations of this University and meets the accepted standards with respect to originality and quality. I hereby release this project to the public. I also authorize the University or other individuals to make copies of this project as needed for scholarly research.

_____

Tanzim Rizwan

Id: 2013-1-60-063

Department of Computer Science and Engineering

East West University.

Letter of Acceptance

The project entitled "Purchase Predicting with Clickstream Data" submitted by Tanzim Rizwan, ID 2013-1-60-063 to the department of Computer Science & Engineering, East West University, Dhaka 1212, Bangladesh is accepted as satisfactory for partial fulfillment for the degree of Bachelor of Science in Computer Science & Engineering on April, 2017.

1_____

Dr. Mohammad Rezwanul Huq

Assistant Professor Supervisor

Department of Computer Science and Engineering

East West University, Dhaka, Bangladesh

2_____

Dr. Ahmed Wasif Reza

Associate Professor and Chairperson(Acting) Chairperson(Acting)

Department of Computer Science and Engineering

East West University, Dhaka, Bangladesh

Acknowledgements

First of all, I am grateful to the Almighty God for establishing me to complete this research. I wish to express my sincere thanks and gratitude to my advisor Dr. Mohammad Rezwanul Huq, Assistant Professor at Dept. of CSE for the continuous support during my project study and related research, for his patience, motivation and immense knowledge. His guidance helped me in all the time of research and writing of the project. I will always be grateful for having the opportunity to study under him.

I am thankful to all of my teachers, Department of CSE, East West University. I also grateful to all of my primary and secondary school teachers who are my first teachers in my life and initiator of my basic knowledge.

I would like to express my thankful to my parents and siblings for supporting me spiritually throughout writing this project. And I am thankful to all my friends. And at last I again thanks to the creator Allah for everything.

Abstract

Everyday using ecommerce sites many customers purchase different types of product form online. This virtual shops are now biggest business now. One of the biggest challenge of the ecommerce sites is showing ads and offers to the correct customer. If they can do this than their sell rate will increase. For this reason customers clickstream data is rich source of customer behavior analysis. The aim of this project is to develop such a functional classifier which can predict purchase events correctly as more as possible.

In this project, I use multiple linear regression technique to predict purchase event. In this technique I build a scoring model with multiple linear regression. Then I set some thresholds manually and use it with different size of dataset. Then I pick the best acting threshold compacting precision, recall, accuracy. I also use ROC graph to verify the threshold.

Contents

List of Figures

List of Tables

Chapter 1

1 Introduction

Ecommerce has become very popular in the modern world. Nowadays people are more willing to go to the actual market than visit an ecommerce website. So ecommerce is playing a vital role to business. As ecommerce websites opens 24/7 so customers get freedom of visiting anytime and also can get a lots of option to choice. So ecommerce also plays a vital role to customers. When a customer visits a ecommerce website, a session is established for him/her. In the session he/she makes choice by a series of clicks(click no. may be 1 or many) on different items, it is called clickstream. From the clickstream many important information can be extracted from the customer's purchasing behavior. Many ecommerce websites are eager to find out the pattern of customers purchasing behavior. Because if they find out this than they can show ads or different offer to targeted customers who are predicted (more willing) to purchase. In this paper, I look at clicksteam data and build a models for classifying clicks into purchasing and not purchasing.

1.1 Clickstream reflects purchase behavior

A huge amount of information of online customers purchasing behavior including when the session started, session ended, click numbers, similar items etc. can be gotten. Clicks of a customer in a session can show the taste of the customer like which item he/she more like, when most of the customer buys etc.

Everyday thousands of people visit ecommerce website, so ecommerce websites can collect huge information about the customers. And those informations can be used to classify them between buyer and visitor. By using those information with perfect algorithm its possible to predict purchase behavior.

Many ecommerce websites like amazon, alibaba etc. are using their own algorithm to predict purchase behavior. They are still trying to develop their algorithm to get higher true prediction rate. But with clickstream data their is a problem in prediction. Let's say there are three customer A,B,C visiting an ecommerce website. A purchase a item on first click and went off the website. B click some time than purchase and after purchase visit some more time and finally went off. C click on different items and spend huge amount of time in the ecommerce website and finally went off. For A and C it quite impossible to predict their purchase behavior with clicksteam data. Because they are like random event. But for B customer it quite possible to predict purchase behavior and those type of customer are the mainly targeted for prediction.

1.2 Motivation

I have chosen to work with clickstream data since I feel it is a better to predict purchase behavior of customer. It includes rich structured information of about individuals involve in purchases in ecommerce sites. For example, it maintains information of when a session started, when a session ended, how

many click event happened, duration of surfing in the site, checking similar item to purchase etc. Another reason huge amount of data can be gotten from here which is reliable. Those data can lead to correct prediction as it is reliable. Predicting purchase behavior of customer is very important for any ecommerce website. Because if an ecommerce website can predict its customers purchasing behavior than it can generate ads and offers more efficiently for its customers. If there is no prediction system in the ecommerce website than generated ads and offers have to show to all of its customers. This a very inefficient way of business. Because if the ecommerce site have to show ads and offers to all of its customers than its server has to take a huge load to serve all customer. For this the ecommerce site has to pay much for the server. This make a lose in the business. Because the ecommerce site has to show ads and offers to all of its customers even those who are not interested in purchasing, they are only visiting the site for information collecting. This conclude that an prediction system can increase purchase rate and also reduce server cost rate.

1.3 Research Questions

I propose one technique for purchase prediction. One of the technique is Linear regression. The technique will work with selected dataset and its features.

1.3.1 Why do we need Purchase Prediction?

Everyday many people visits ecommerce websites and draw their footprint with huge amount of data in the website. This huge amount of data contains important and related to customers purchase behavior which can be used in benefit for businesses. Manually separating buyer and visitor by seeing the data is impossible. But by predicting the probability of purchase of a customer seeing his/her previous data it can be possible. With the help of prediction system it is very much possible to separate the customers into buyers and visitors.

1.3.2 Why do we use Linear Regression?

Linear Regression is easy to understand and also it is easy to implement. I think it will be a very powerful tool for my proposed technique of purchase prediction. In my technique at first I extract important features from the training data and than I will apply linear regression. This will give me two things intercept and coefficients. Than I will apply it with test data and setting a threshold for the scores I will classify them between buyer and visitor.

1.4 Overview

My main goal is correctly predict purchases as more as possible from clickstream data. My project has three parts; first one is to detect the important features from the training data, second one is to use machine learning algorithm linear regression and finally third one is use result (intercept and coefficients) on test data. After completing this classifying step we measure the accuracy by using confusion matrix.

I first am using multiple independent variables. So it it actually multiple linear regression. After getting result from regression I will use it with test dataset which will give me score for every session. Then I will set a threshold to detect the buyers.

Later on I will use confusion matrix to measure accuracy and roc graph to verify threshold.

Chapter 2

2 Background

Clickstream data has an impressive predictive power. Using this power ecommerce sites can predict which item will sell more, when will most of customer visit the site, how many quantity product customer may buy etc. Clickstream is such a dataset that reflects the taste and behavior of customer in online. As a result it is helpful for ecommerce sites. In this paper I analysis the purchase prediction over clickstream data.

For this reason I use some machine learning algorithm. They are Linear regression, Confusion Matrix and receiver operating characteristic (ROC) graph. Using such algorithms we calculate the accuracy of our experiment. The basic information of such topics is given below:

2.1 Linear Regression

Linear Regression is a supervised machine learning algorithm. It can be used for both classification and regression purposes. The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.

The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient.

For example, in a simple regression problem (a single x and a single y), the form of the model would be:

$$y = B0 + B1*x$$

Multiple Linear Regression

Multiple linear regression is the most common form of linear regression analysis. As a predictive analysis, the multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables.

The independent variables can be continuous or categorical. For more than one independent variables or input (x) the representation is:

$$y = B0 + B1*x1 + B2*x2 + B3*x3..... + Bn*xn$$

2.2 Confusion matrix

Confusion matrix is a table which describes the performance of a classification model(or "classifier") on a set of test data for which the true values are known [3]. For evaluating the performance of such systems I have to use the data in the matrix. The following table shows the confusion matrix for a two class classifier.

- **TN :** Actually **negative** and predicted **negative**
- **TP :** Actually **positive** and predicted **positive**
- **FN :** Actually **positive** and predicted **negative**
- **FP :** Actually **negative** and predicted **positive**

		Predicted	
		Negative	Positive
Actual	Negative	**TN**	**FP**
	Positive	**FN**	**TP**

Table 2.1 :Confusion Matrix

Several standard terms have been defined for the 2 class matrix

$$FP\ rate = FP/(FP+TN)$$

$$TP\ rate = TP/(TP+FN)$$

$$Precision = TP/(TP+FP)$$

$$Recall = TP/(TP+FN)$$

$$Accuracy = (TP+TN)/ALL$$

- **Accuracy:** The accuracy means the proportion of the total number of predictions that is correct.
- **Precision:** The precision means the proportion of positively identified that are correct.
- **Recall :** Recall is the ratio of a number of events that correctly recall to a number of all correct events.
- **FP rate (false positive) rate:** The FP means the proportions of negatives that are incorrectly identified.
- **TP rate (true positive) rate:** The TP means the proportion of positives that are correctly identified.

2.3 ROC Graph

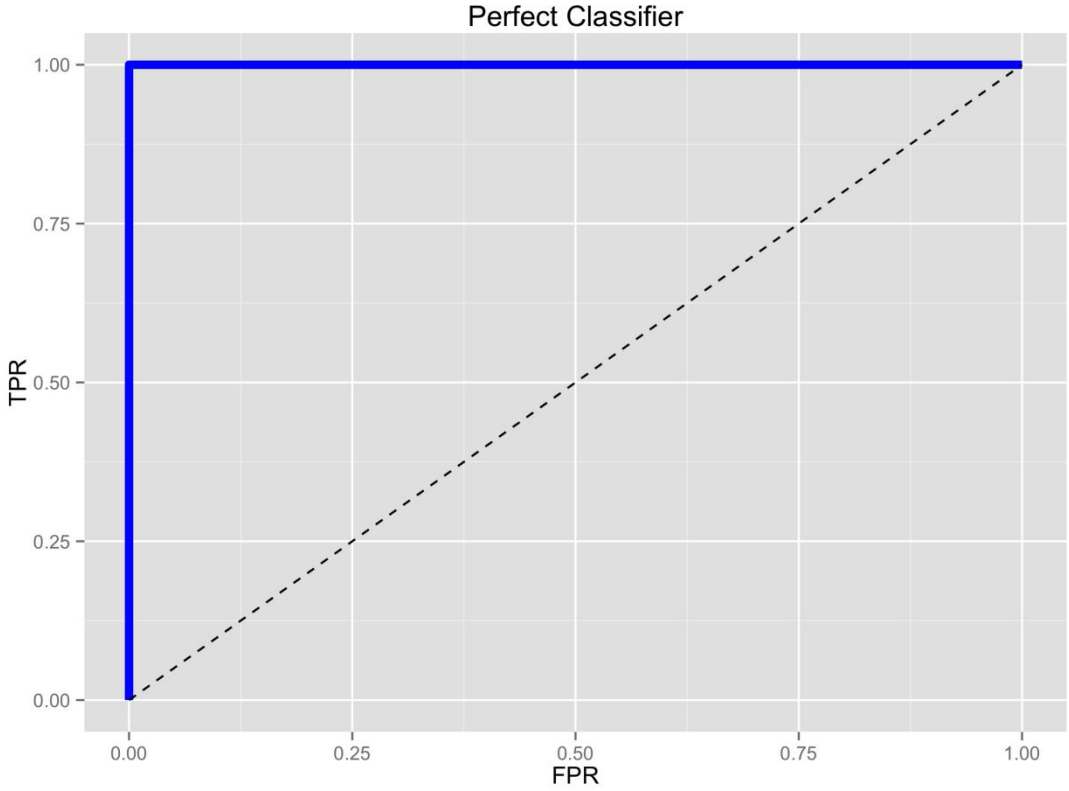
Receiving Operating Characteristic, or ROC, is a visual way for inspecting the performance of a binary classifier (0/1) [4]. In particular, it's comparing the rate at which the classifier is making correct predictions (True Positives or TP) and the rate at which the classifier is making false alarms (False Positives or FP). Definitions of True Positive Rate (TPR) or False Positive Rate (FPR) below:

TPR=TruePositives/(TruePositives+FalseNegatives)

FPR=FalsePositives/(FalsePositives+TrueNegatives)

The first there is a doted diagonal line. It indicates that the classifier is just making completely random guesses. Since the classifier is only going to be correct 50% of the time, it stands to reason that your TPR and FPR will also be equal.

But for a PERFECT classifier which makes every prediction correctly. Meaning TPR of 1 and an FPR of 0. In that case, your ROC curve looks something like this.



I use this ROC graph to verify my threshold which is used to classify the score of test dataset.

Chapter 3

Related work

The contributions of this paper [5] Wendy W. Moe and Peter S. Fader developed a model of conversion behavior (i.e., converting store visits into purchases) that predicts each customer's probability of purchasing based on an observed history of visits and purchases. They offer an individual-level probability model that allows for consumer heterogeneity in a very flexible manner. They allow visits to play very different roles in the purchasing process. For example, some visits are motivated by a planned purchase while others are simply browsing visits. The Conversion Model in this paper has the flexibility to accommodate a number of visit-to-purchase relationships. Finally, consumers' shopping behavior may evolve over time as a function of past experiences. Thus, the Conversion Model also allows for non-stationarity in behavior. Specifically, their Conversion Model decomposes an individual's purchasing conversion behavior into a visit effect and a purchasing threshold effect. Each component is allowed to vary across households as well as over time. They then apply this model to the problem of "managing" visitor traffic. By predicting purchasing probabilities for a given visit, the Conversion Model can identify those visits that are likely to result in a purchase. These visits should be re-directed to a server that will provide a better shopping experience while those visitors that are less likely to result in a purchase may be identified as targets for a promotion. Understanding how conversion probabilities vary across consumers and change from visit to visit is valuable information that can allow e-commerce managers to better treat each visitor based on his/her past behavior. The patterns of

visits and purchases may reveal how best to serve a given visit. This model is a useful first step in this direction, and it encourage future researchers to build upon it.

In this paper[6] by the authors particular interest is the role of Internet clickstream data in customer relationship management by providing a rich source of behavioral information. Until recently, the bricks-and-mortar marketers have focused their measurement and modeling efforts primarily on easily available purchasing data. However, the spotlight has generally been centered on ultimate purchasing activity—when, what, and how much people buy. Customer activities such as comparison shopping and information gathering have largely been ignored for lack of data. These behaviors may be less visible but still exert strong influences on purchasing. As managers attempt to alter customer behavior using marketing tactics such as advertising and web site design, it is critical to reach and persuade customers at these earlier stages of the decision process.

Chapter 4

4 Working Procedure

I use Apache Spark for using machine learning algorithm which make the computation very fast. From the training dataset at first I extract many features for future computation. Then I use my techniques.

4.1 Data Description

There are four data files: click data file, buy data file, test data file, solution data file [7]. Click data file contains informations of session id, timestamp, item id, category. Buy data file contains informations of session id, timestamp, item id, price, quantity. Test data file contains informations of session id, timestamp, item id, category. Solution file contains informations of session id and item id. Session id is the id of the session. In one session there are one or many clicks. Timestamp is the time when the click occurred. Item id is the unique identifier of the item. Category is the category of the item. Price is the price of the item and Quantity is how many of this item were bought.

4.2 Features Extraction

In this section, I present detail about my features which I use in my techniques. First merge the click data file and buy data file, then sort them to make the training dataset. In training dataset (click and buy data file) for each buy event those features will be very much important. Before extracting the features I parse the datasets for better use.

The training dataset contains

Click Data	Buy Data	Total
500000	16376	516376

Duration

Duration is the amount of time a session established. Here I consider all duration time in seconds. This is a very important feature. Because if any session established for long time then the probability of purchasing item gets higher. In my training dataset the average duration time of a session where buy event happens is 1302.05600244s.

Number of Clicks

Number of clicks in a session. This is a important behavior of a customer. The more a customer clicks the possibility of purchasing an item gets higher. In my traning dataset the average number of clicks where buy event happens is 11.93081339.

Similar Item

Similar Item is clicking on similar type item. By clicking on similar item a customer shows his/her interest on the item, this may take him to purchase. In my traning dataset the average similar item click is 3.6971788.

Source Code Available here:

"https://github.com/BrainAxe/Purchase-Prediction"

4.3 Pseudo code of my experiment

1. **Function roc_graph**

2. **x,y ⟵ GetContent(input_file); //input_file is a text file where data is stored**

3. **Plot graph**

4. **End function**

5. **Function nsession**

6. **Open file tdata.csv**

7. **Declare c = 0**

8. **Load the csv file in reader**

```
9. Declare pre_id = ‘’
10. For row in reader
11. u_id = row[0]
12. If u_id != pre_id then
13. Increment c by 1
14. Set pre_id = u_id
15. End for loop
16. Close file
17. Return c
18. End function
19. Function check_result
20. Declare array s
21.  Open file result.csv
22. Declare reader and load the csv file
23. For row in reader
24. u_id = row[0]
25. Declare score = float(row[3])
26. If score >= threshold then
27. If u_id not in s then
28. Append u_id in s
29. End for loop
```

30. Close file

31. Declare c = 0

32. Open file solution-sort.csv

33. Declare reader and load csvfile

34. For row in reader

35. If row[0] in s then

36. Increment c by 1

37. End for loop

38. Close file

39. Declare l with length of s

40. Declare n = functon nsession

41. Declare tp and assign c

42. Declare fp and assign l - c

43. Declare fn and assign buy_event -c

44. Declare tn and assing n - tp - fp- fn

45. Declare fpr = fp/float(fp+tn)

46. Declare tpr = tp/float(tp+fn)

47. Declare precision = tp/float(tp+fp)

48. Declare recall = tp/float(tp+fn)

49. Declare accuracy = (tp+tn)/float(n)

50. End function

```
51. Function calculate(r)
52. Open file tdata.csv
53. Declare reader and load csvfile
54. Open file result.csv as write mode
55. Declar writer and load csv file
56. Declare pre_id = ‘’
57. For row in reader
58. u_id = row[0]
59. time = float(row[1])
60. p_id = row[2]
61. click = int(row[3])
62. same = int(row[4])
63. If u_id equal pre_id then
64. t_diff = time - s_time
65. result = r[0] + float(r[1])*t_diff + float(r[2])*click +
float(r[3])*same
66. Write in the csv file
67. Else
68. s_time = time
69. t_diff = 0
```

70. **result = r[0] + float(r[1])*t_diff + float(r[2])*click + float(r[3])*same**
71. **Write in the csvfile**
72. **pre_id = u_id**
73. **End for loop**
74. **End function**
75. **Function ptest_data**
76. **Process data for use**
77. **End function**
78. **Function linear_reg**
79. **Process with apache spark**
80. **End function**
81. **Function ptraining_data**
82. **Process training data**
83. **End function**

4.4 Experiment Evaluation

My objective is to predict purchase event as more as possible with less false positive result. Before using my technique, I parse the raw data first. As I mentioned before I have a click data file and a buy data file, they are used for making training data file. I merge the click data file and buy data file and parse the data in a way so that I get some new features like duration, click no., similar. For a session data look like this

After merge click data file and buy data file:

user_id,timestamp,item_id,buy

12,2014-04-02T10:30:13.176Z,214717867,0

12,2014-04-02T10:33:12.621Z,214717867,0

12,2014-04-02T10:42:17.227Z,214717867,1

After parsing the merge file:

user_id,duration,item_id,click_no.,similar,buy

12,0,214717867,1,1,0

12,179.4449999332428,214717867,2,2,0

12,724.050999879837,214717867,3,3,1

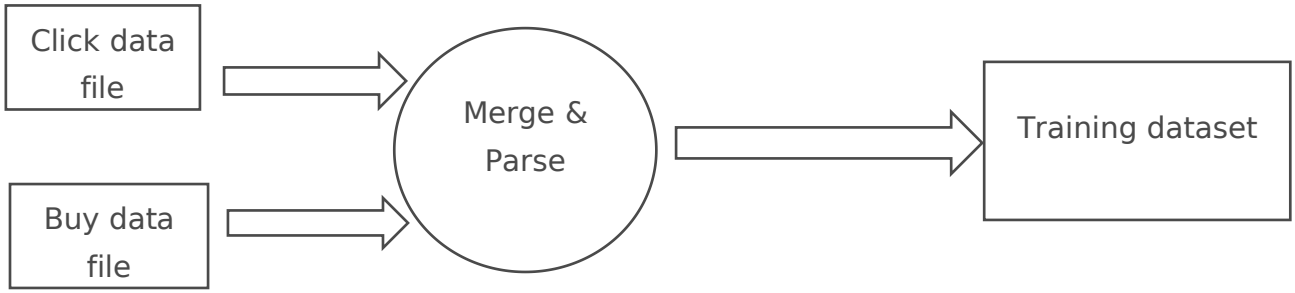


Figure 4.1: preparing training dataset

After get ready the training dataset, I use my technique which is multiple linear regression. In linear regression two types of variables are needed. They are independent variables and dependent variables. In my training dataset the independent variables are the duration, click no., similar item. For the dependent variable I use 1 and 0 for classify buy and not buy. As linear regression needs complex and huge calculation and also my training dataset is large so normal programming takes much time. So I use Apache Spark.

Apache Spark

It is a sub-project of Hadoop. Apache Spark is a fast, in-memory data processing engine with elegant and expressive development APIs to allow data workers to efficiently execute streaming, machine learning or SQL workloads that require fast iterative access to datasets. With Spark running on Apache Hadoop YARN, developers everywhere can now create applications to exploit Spark's power, derive insights, and enrich their data science workloads within a single, shared dataset in

Hadoop. Apache spark run programs up to 100x faster than Hadoop. Apache spark is also very easy to use. Because it can be implemented with Java, Scala, Python, R. Many machine learning algorithms already implemented in apache spark. So apache spark is very helpful for doing any machine learning project.

After run through apache spark it gives intercept and coefficients. I parse the test data file. Before parsing the test data file it was like

user_id,timestamp,item_id

5,2014-04-07T17:13:46.713Z,214530776

5,2014-04-07T17:20:56.973Z,214530776

5,2014-04-07T17:21:19.602Z,214530776

After parsing the test dataset look like this

user_id,duration,item_id,click_no.,similar

5,0,214530776,1,1

5,430.25999999046326,214530776,2,2

5,452.88899993896484,214530776,3,3

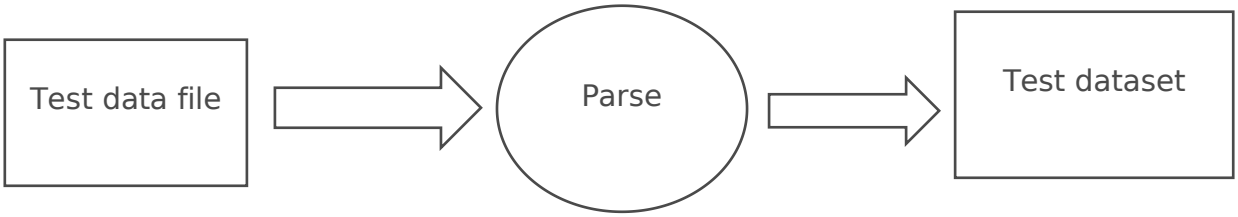


Figure 4.2: preparing test dataset

Then I use intercept and coefficients from the training dataset with the duration, click no. and similar item of test dataset to get a score.

Score = Intercept + coefficient * duration + coefficient * click_no + coefficient * similar

I classify the scores with a threshold. This threshold is set manually. I have a solution data file. Which have the result (actual buy event) of the test data file. I calculate the accuracy by using confusing matrix. Besides calculating the accuracy I also calculate the recession, recall. I apply the same process for different size of the dataset.

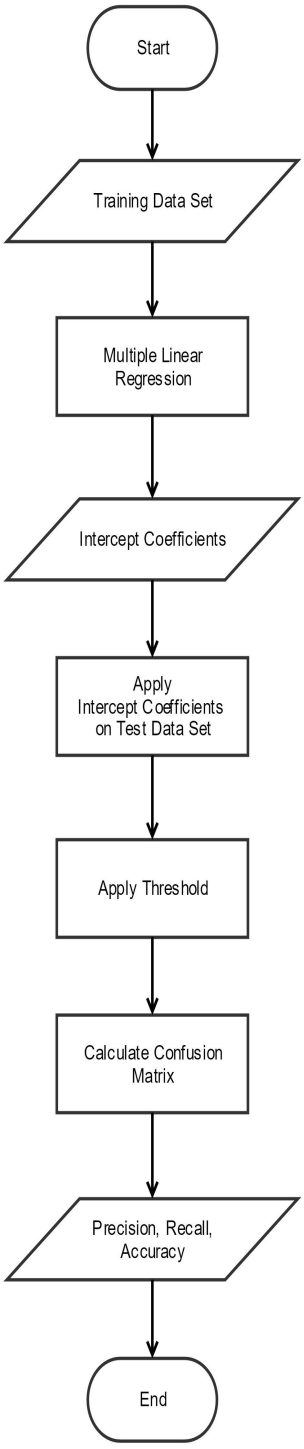


Figure 4.3: Flowchart of purchase prediction

Chapter 5

5.1 Experimental Result:

In this chapter the result of purchase prediction calculated with linear regression are represented with confusion matrix and the performance is based on Accuracy, Recall and Precision.

For 10000 clicks

Total Session = 2635

Total Buy Event = 145

Threshold	TP	TN	FP	FN	Precision	Recall	TPR	FPR	Accuracy (100%)
0.02	145	1452	1038	0	0.12	1	1	0.42	60.6
0.05	109	1861	629	36	0.15	0.75	0.75	0.25	74.7
0.08	79	2196	294	66	0.21	0.54	0.54	0.11	86.3

Table 5.1: Experimental result of 10000 click data

For 200000 clicks

Total Session = 50885

Total Buy Event = 3285

Threshold	TP	TN	FP	FN	Precision	Recall	TPR	FPR	Accuracy (100%)
0.02	3156	27792	19808	129	0.137	0.96	0.96	0.416	60.8
0.05	2299	35768	11832	986	0.163	0.7	0.7	0.25	74.8
0.08	1630	41997	5603	1655	0.22	0.49	0.49	0.12	85.7

Table 5.2: Experimental result of 200000 click data

From the table I can see that there are three attributes whose are precision, recall and Accuracy. Precision (also called positive predictive value) is the fraction of retrieved instances that are relevant while recall (also known as sensitivity) is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance.

5.2 Performance Evaluation

From the Experiment Result I can get an overview about the performance of the algorithms for different size of the dataset but for better analysis we should compare the results based on different parameter which is easily represented via graphs. In this graph different dataset are in horizontal line and the thresholds are in vertical line.

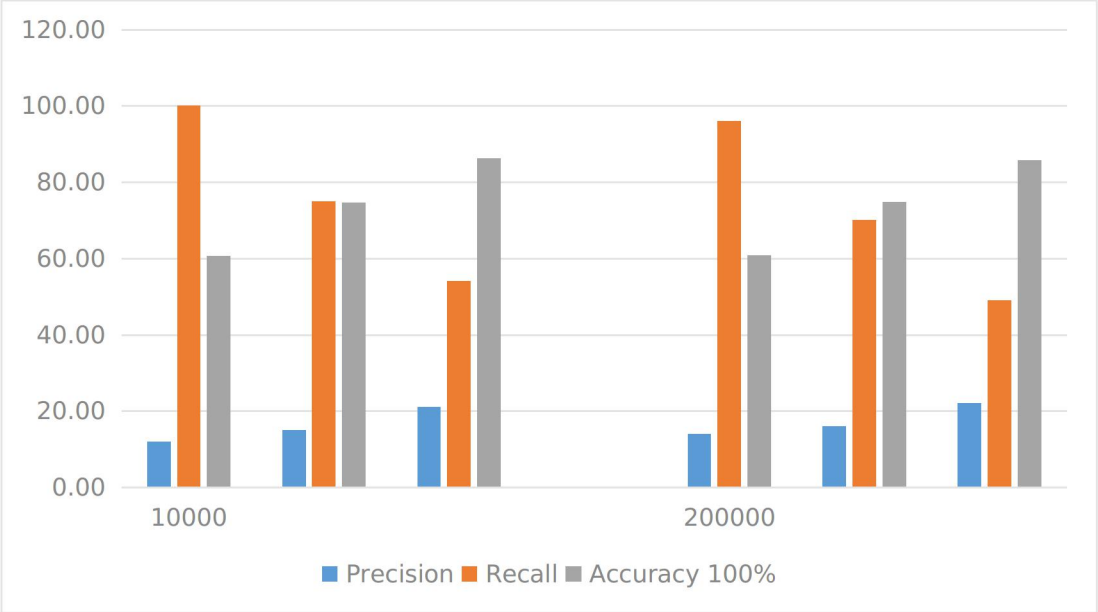


Figure 5.1: precision,recall,accuracy in different dataset

From this graph, I see that for threshold 0.02 the recall is very high but the precision is very low. Again for threshold 0.08 accuracy is high but the recall is very low. Only for threshold 0.05 recall and accuracy is in good shape. So, I think by setting 0.05 as threshold I can get better prediction result.

5.2.1 ROC Graph for Threshold Evaluation

For selecting classifiers based on their performance receiver operating characteristics (ROC) graph is a good technique. By plotting the true positive rate (TPR) against the false positive Rate (FPR) the curve is created. True positive rate indicated the sensitivity or probability of detection and false positive rate indicate the fall-out or probability of false alarm.

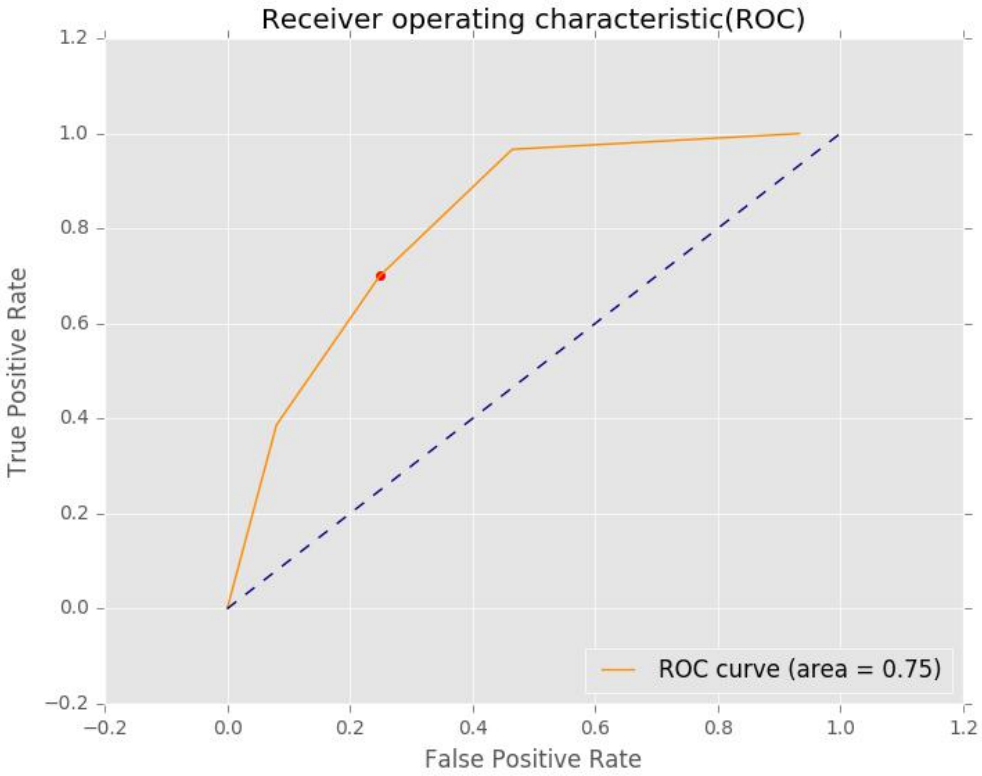


Figure 5.2: ROC graph for 5 thresholds in 200000 test dataset

According to ROC graph the performance of a threshold is better than other threshold if the true positive rate is high and false positive rate is low. Based on the concept I plot 5 thresholds in the ROC graph with their true positive rate in Y axis and false

positive rate in X axis and I get 5 points on the graph. According to the ROC graph the threshold is the best for which the data point is northwest corner compare to other threshold that threshold is best. But in my figure no data point at this can fulfill the requirement. Only one point is more closer to the northwest corner compare to other which is .05 threshold. So I can say that I can say that my threshold is pretty good enough to predict purchase events.

Chapter 6

Conclusion and Future Work

Ecommerce websites have revolutionized the way in which people purchase. So it can be beneficial to analysis user behavior and predict their purchase time. This can make a huge benefit for online business. Knowing who is going to buy ecommerce sites can show them offers and different ads.

Currently I have worked with a very simple model and in my work I design my classier with a few features and also I set a threshold manually. In this project I only focused to predict purchase event as more as possible, but my project's precision level is very low.

In future I would study further many related problem. For this I will try to improve my model by adding some extra features. In this project I only predict purchase event and I am not considered which item is going to purchase. And also the manual threshold to predict. So my next plan is I will try to make model which will set the threshold automatic and also predict which item is going to sell. From this I can say that my future work list may be contains the following things.

1. Set the threshold automatic so that no manual setting needed.

2. Predict purchase with high precision.
3. Predict Item: In future I will to predict which item is going to sell.
4. Work with others language: in my work I use python language . in future we can use different language.
5. Working with large amount of dataset: In future work with dataset which contains large amount of clicks.
6. Accuracy calculate and performance evaluation: in my work I use confusion matrix for calculate accuracy and performance evaluation. In future Apply others machine learning algorithms to calculate accuracy and performance evaluation.

In this project I am mainly focusing on purchase prediction like positive and negative event. There is potential of work in the field of prediction and I will try to use my knowledge in this field.

References

[1] J. Brownlee, "Linear Regression for Machine Learning - Machine Learning Mastery", Machine Learning Mastery, 2017. [Online]. Available: http://machinelearningmastery.com/linear-regression-for-machine-learning/. [Accessed: 08- Apr- 2017].

[2] "What is Multiple Linear Regression? - Statistics Solutions", Statistics Solutions,2017.[Online].Available:http://www.statisticssolutions.com/what-is-multiple-linear-regression/. [Accessed: 08- Apr- 2017].

[3] "Simple guide to confusion matrix terminology", Data School, 2017. [Online]. Available:http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/. [Accessed: 08- Apr- 2017].

[4]Blog.yhat.com,2017.[Online].Available:http://blog.yhat.com/posts/roc-curves.html. [Accessed: 08- Apr- 2017].

[5] W. Moe and P. Fader, "Dynamic Conversion Behavior at E-Commerce Sites", Management Science, vol. 50, no. 3, pp. 326-335, 2004.

[6]W. Moe and P. Fader, "Uncovering Patterns in Cybershopping", California Management Review, vol. 43, no. 4, pp. 106-117, 2001.

[7] 2017. [Online]. Available: http://RecSys Challenge 2015: ensemble learning with categorical features. [Accessed: 08- Apr- 2017].