Report on

MODEL BUILDING FOR PREDICTION OF AD-CLICKS:

AN ML APPROACH

Submitted By:

Shaukeen

Roll no: 2K18/MBA/923

Vishakha Anand

Roll no: 2K18/MBA/929

Under the Guidance of:

Dr. Deepti Aggrawal

Assistant Professor

UNIVERSITY SCHOOL OF MANAGEMENT

& ENTREPRENEURSHIP

Delhi Technological University

MAY 2020

CERTIFICATE

This is to certify that this dissertation report titled "MODEL BUILDING FOR PREDICTION OF AD-CLICKS: AN ML APPROACH" is a bonafide work carried out by Ms. Vishakha Anand and Mr. Shaukeen of MBA- Business Analytics, 2018-20 batch, submitted to University School of Management and Entreprenuership (USME), Delhi Technological University(DTU), Vivek Vihar Phase-2, New Delhi in partial fulfillment of the requirement for the award of the degree of Masters of Business Administration in Business Analytics.

Signature of the Head

Seal of the Head

Place:

Date:"

DECLARATION

I hereby declare that the work presented in this report entitled "**MODEL BUILDING FOR PREDICTION OF AD-CLICKS: AN ML APPROACH"** in partial fulfillment of the requirements for the award of the degree of **MBA** in **Business Analytics** is submitted in the department of University School Of Management and Entrepreneurship, Delhi Technological University.

It is an authentic record of my work carried out over a period from June 2019 to July 2019 under the supervision of **Dr. Deepti Aggrawal**(Asst. Professor), Delhi Technological University(DTU).

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

(Student's Signature)

ACKNOWLEDGEMENT

We, Shaukeen and Vishakha Anand (MBA-Business Analytics) would like to convey our gratitude to the Management, Director, Dean and Career Development Center of Delhi Technological University (Formerly DCE), Delhi.

We have completed our projects based on predictive analytics. We would like to express our sincere gratitude to my Dr. Deepti Aggrawal, Asst Professor for providing her invaluable guidance, comments and suggestions throughout the course of the project. We would specially thank her for constantly motivating us to work harder.

We would like to thank Mr. Rajesh Rohilla, Head of Training & Placements (DTU), for giving us the opportunity to work on this Project namely, MODEL BUILDING FOR PREDICTION OF AD-CLICKS: AN ML APPROACH.

We extend my warm gratitude and regards to everyone who helped us during my project.

Executive Summary

Machine Learning has now variety of applications in real world whether it is for predictive analysis or automating things or decision making for business purpose.

We have done a project which is a Machine Learning approach. This project is about predicting the ad click rate of a customer, based on certain features. We found this problem interesting and therefore made a Machine Learning model to predict the Ad click.

This model we have made can be made on different platforms such as SPSS, R, Excel, python etc. We have chosen SPSS and Excel.

In this project, we worked on an advertising dataset, indicating whether or not a particular internet user has clicked on an Advertisement.

The goal is to predict if a user would click on an advertisement based on the features of the user. We are using a training dataset to find the logic that will be applied on the Test dataset to find the result.

We are using two models namely Logistic Regression and Random forest model to predict whether the user will click on the Ad or not. At then we will produce a confusion matrix to analyze how accurate was the prediction that was made on the basis of these two models.

TABLE OF CONTENTS

S.No.	Topic	Page No.
1.	Introduction	7
2.	Objective Of The Study	8-10
3.	Literature Review	11-12
4	Research Methodology	13-17
5.	Results	18-42
6.	Findings and Recommendations	43-45
7.	Limitations	46
8.	Bibliography/References	47-50

INTRODUCTION

Since the advent of technology in several sectors, the internet has been developing rapidly. With internet came applications, and with it new companies. Now because so many companies have come into existence there's a never ending competition between them. Advertising for their products online is the step that they choose to publicize themselves. Advertisement is an important factor, by analyzing the click-through rate companies can figure out the popularity of their ads. Thus, clicks can make companies or break them.

In this project what we have done is, we have made a machine learning model to help us predict whether a person will click on the intended ad or not. We have a training dataset that contains the output as 0 or 1, which means if the value is 0: the person will not click on the ad, and if 1: person will click on the ad. We will use this training dataset containing outputs to define logic.

This logic will then be applied to the test data to find out the output for the data as it was for the training data. It will help predict whether the person will click on the ad or not, after being trained on a particular data.

Two Machine Learning Models are to be used in this prediction of clicks on ads, Logistic Regression(For logical outputs, 0 and 1) and Random Forest model to make one decision tree for every independent variable available and find the maximum occurring output and considering it as the final output.

We have selected these two prediction models after analyzing the visualization of the data since these are the most appropriate models for the available data in the dataset.

Why Ad click is important?

If a company knows the Click Through Rate(CTR) of digital advertising they can easily identify whether spending their money is worth or not.

If a CTR is higher, specific ad or company is more relevant to the customer whereas lower CTR represents the ad may is not relevant enough for the customer. To find out the relevance of an ad, analyses of data is important. Analyzing whether the customer is clicking on the ad or website shows the relevance of it to the customer.

Finding out CTR of a particular website can help the website to distinguish the best features. Once the company knows its best features they can provide maximum value to the customer. If customer finds value in the content of the website, He will come back to it again and again. Eventually profiting the company itself, which is essentially the main goal of a company.

Objective

Overview:

SPSS-

The whole project is divided into 6 steps :

1. Loading Data set
2. Defining Variables
3. Exploratory Analysis
4. Data Cleaning
5. Descriptive Analysis
6. Training the Machine Learning Model
7. Testing Model Accuracy

Step 1: Loading Data set

To apply predictive analytics we first need to load the data into the software, so that we can apply Data Visualization techniques and Machine Learning models to that data.

Step 2: Defining Variables

Defining variables used in prediction is an important factor in SPSS. There are certain categories such as Categorical data, Nominal data, Ordinal data, scaled data, interval data etc.
Decimal Value setting is done to define how many decimal values will be counted in the input data and output data.

Labeling Data is done to represent the logical definition of the data i.e**. 1**- clicked on ad and **0**- not clicked on Ad

Type of data is defined as Numeric or Non-Numeric data. We have mainly considered numeric data in the prediction.

Step 3: Exploratory Analysis

Through exploratory data analysis we prepare certain graphs and plots from the data that help us with the understanding of the data and thus choose suitable Machine Learning Model.

Advantage of using such graphs and plots is that they help us find out the relation between different variables. We have made scatter plots, histograms and graphs based on the data.

Step 4: Data Cleaning

To get more accurate results we need to remove the unwanted data from the dataset.

Removing unwanted variables from the dataset and replacing the missing values with average values to help us get more accurate results and escape erroneous outputs.

Step 5: Descriptive Analysis

In this analysis we have used certain techniques to find out mean, median, mode, standard deviation, variance, ANOVA, skewness, kurtosis, regression, and correlation.

Step 6: Training the Machine Learning Model

There are two categories in output data:

1. User will on the ad (1) or
2. User will not click on the ad (0)

It means that it is a classification problem. Visualization of the data gave us an insight that there are some boundaries that can be used as the basis of selecting the Machine Learning model. Here we have decided on the basis of the visualization and other data analysis techniques that we can use Logistic regression and Random Forest model. These will be explained later at length.

Step 7: Checking Model accuracy

Final step is to check the accuracy of the Machine Learning model that we have created for ad click prediction. This will help us analyze out of all the methods we have used which one will give maximum accuracy, since prediction is based on accuracy itself. Since we have used logistic regression and random forest model, we have found out the results and compared them both.

For checking overall accuracy of the model we have prepared a confusion matrix for comparing training and test output accuracy. This will further help us decide whether the prediction model used is providing us appropriate results or not.

EXCEL-

The whole project is divided into 6 steps :

1. Load Data set
2. Data Cleaning
3. Train Test Split
4. Training the Model
5. Testing the model accuracy

Step 1: Loading Data set

To apply predictive analytics we first need to load the data into the software, so that we can apply Data Visualization techniques and Machine Learning models to that data.

Step 2: Data Cleaning

To get more accurate results we need to remove the unwanted data from the dataset.

Removing unwanted variables from the dataset and replacing the missing values with average values to help us get more accurate results and escape erroneous outputs.

We are considering only 4 independent variables and 1 dependent variable for this analysis. Representation of data should be in the way that the 1 dependent variable is placed at the rightmost column of the dataset, only independent variables are to be placed on the left columns. This is because we want the output to be represented in the similar way, on the rightmost column.

Step 4: Training and Test Dataset

Training and Test are used to train the Machine Learning model while Training is used as input for making predictions which will be then validated with the Test values.

Step 5: Training the Machine Learning Model

There are two categories in output data:

1. User will on the ad (1) or
2. User will not click on the ad (0)

We have decided on the basis of the visualization and other data analysis techniques that we can use Logistic regression and Random Forest model. We also set a threshold value for logistic regression as 0.5.

Also we do not need to assume outputs while working on excel just training the machine would work well in predicting the outputs.

Step 7: Checking Model accuracy

Final step is to check the accuracy of the Machine Learning model that we have created for ad click prediction. This will help us analyze out of all the methods we have used which one will give maximum accuracy, since prediction is based on accuracy itself. We have used logistic regression and random forest model, we have found out the results and compared them both.

For checking overall accuracy of the model we have prepared a confusion matrix for comparing training and test output accuracy. This will further help us decide whether the prediction model used is providing us appropriate results or not.

Literature Review

In this project, we have worked on an advertising dataset, showing whether a particular internet user has clicked on an Advertisement or not.

The goal is to predict if a user would click on an advertisement based on the data given in the dataset. Few assumptions made as a part of this project are:

1. Users taken into consideration are between the age group of 19 to 61.

2. There is almost equal ratio of male and female internet users.

3. The ad topic is limited to what is given in the dataset.

Challenges Faced:

A few challenges that we faced while working on this project:

1. There is very less publicly available data set for ad click.

2. New online ads that are coming up are not targeted to a particular set of users, using our prediction model will best work with a particular set of data.

Data Set

The variables consisted in the dataset:

- Daily Time Spent on Site
- Age
- Area Income
- City
- Male
- Ad Topic Line
- Daily Internet Usage
- Country

Cleaning and Approaches

Some factors do not really influence the output, so we do not consider those factors while applying the algorithms. These factors are City, Ad Topic line and Country. Gender is also not a considerable variable in the regression approach since there are far more defining factors to be considered.

.

Logistic Regression

Why have we used Logistic Regression?

Logistic Regression is a fairly easy algorithm as well as easy to train. So we have started with this algorithm. Also the main factor is that the output that is already there in the data is in logical form i.e. 0 and 1. Therefore using logistic regression will give output for the data in the form of 0 and 1 itself.

We have also set the threshold value as 0.5. In SPSS we have used Chaid growing method for making decision trees.

Random Forest

Why have we used Random Forest?

Random Forest are very flexible, easy to understand, and easy to debug. Random forest works on variable screening, where based on every independent variable we make a decision tree to predict the output through every variable considered.

For example, we predict the Click on Ad with factors such as age, time spent online, daily internet usage and income of the person.

Now, when the decision trees are prepared using all the variables we consider the output value for all those decision trees and we select the maximum occurring output as the final output of the prediction. Therefore we figure out that nonlinear relationships between variables do not affect the performance of the decision tree and further the Random Forest algorithm.

RESEARCH METHODOLOGIES AND TECHNIQUES

In this project what we have done is, we have made a machine learning model to help us predict whether a person will click on the intended ad or not. We have a training dataset that contains the output as 0 or 1, which means if the value is 0: the person will not click on the ad, and if 1: person will click on the ad. We will use this training dataset containing outputs to define logic.

This logic will then be applied to the test data to find out the output for the data as it was for the training data. It will help predict whether the person will click on the ad or not, after being trained on a particular data.

The results are influenced by the logic that has been created so if the training data trains the machine to call black red, red white and white black, then it will also do the same with the test data. Meaning- the machine will act according to the training data even if it is a wrong logic. That is an important factor to be kept in mind.

Now, we are using SPSS platform to carry out our project. First step is logistic regression. We use logistic regression when we want outputs in the form of 0 and 1(logical outputs).

Another algorithm used is Random Forest, this is used when there are multiple independent variables and a decision tree is made on the basis of all those variables. This algorithm takes the maximum time occurring output as the final output. Therefore the erroneous variables and data can be exempted.

We have used two algorithms for analysis and prediction

i) Logistic Regression,

ii) Random Forest,

Logistic Regression

Logistic Regression is used for regression analysis when the classification is binary. To predict whether a person clicked on an ad or not, that is either 1 or 0, logistic regression was considered.

We have used Logistic regression because we had the output results as 0(not clicked on ad) and 1(clicked on ad).

Logistic Regression finds out the relationship between dependent(to be predicted) and multiple independent variables(given features) by estimating probabilities in the logistic function.

These probabilities are converted to binary values for easy prediction, which is performed by logistic function. Since our data required logistic prediction (clicked-1, or not- 0), this is the most appropriate method to be used for it. This Logistic Function is an S shaped curve

that can take any real valued number and map it between the range of 0 and 1. These values then are consolidated to either 0 or 1, where threshold is considered 0.5.

Random Forest

Random forest works on variable screening, where based on every independent variable we make a decision tree to predict the output through every variable considered. Therefore it consists of a very large number of individual decision trees. Every decision tree in the random forest gives an output prediction and the class with maximum votes is considered as the model's prediction output.

If there are multiple uncorrelated decision trees working together then the accuracy will be more and the errors would be less since it's not just dependent on one single data, but a bunch of data. So the individual errors will not affect the consolidated data. Since some decision trees might give wrong outputs and some would give correct prediction output so the tree would move in the correct direction.

The prerequisites for a well performing Random Forest Model are:

1. There needs to be some actual logic in the data so that models built using those features do better than random guessing.
2. The predictions (and therefore the errors) made by the individual trees need to have low correlations with each other.

Null Hypothesis

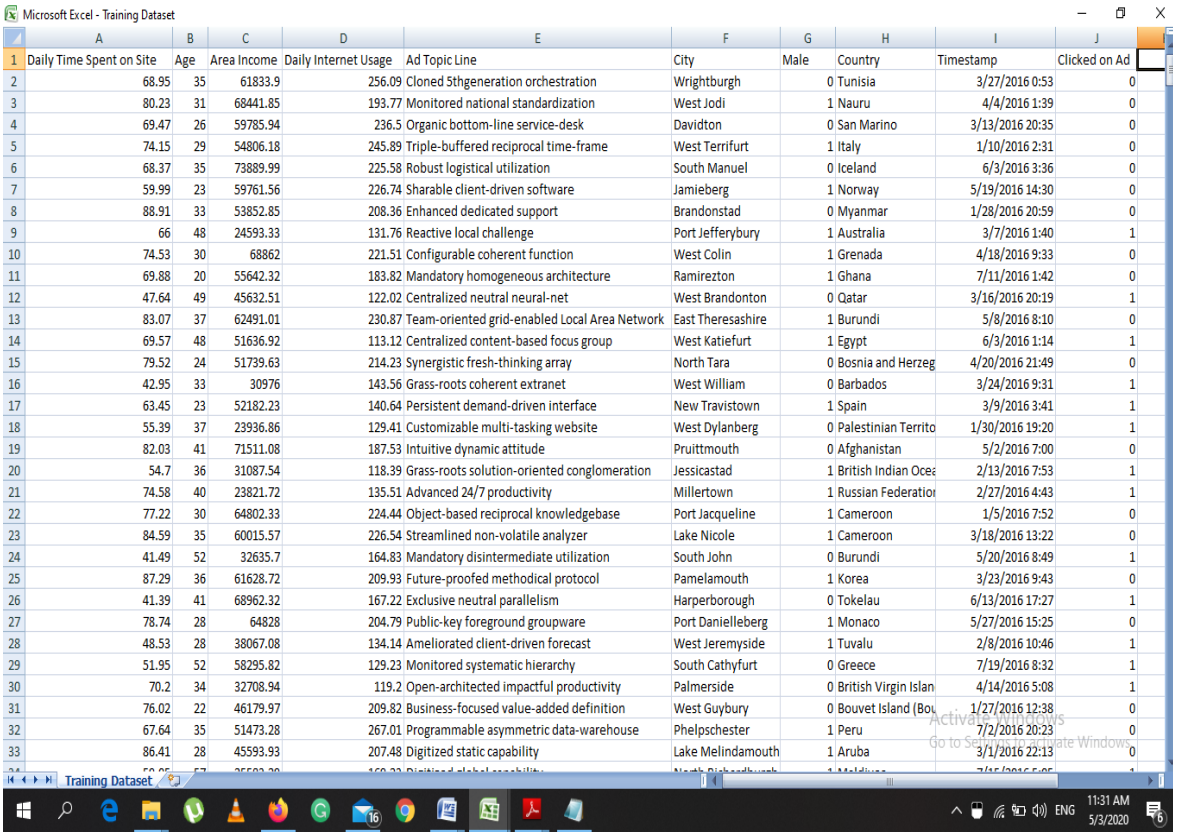
The null hypothesis (H0) generally states the currently accepted fact. It is formulated in such a way that two measured values have no relation with each other. The alternative hypothesis, H1, states that there is in fact a relation between the two values. Rejecting or disproving the null hypothesis gives support to the belief that there is a relation between the two values.

Here CTR is predicted using the assumption that the human features such as Daily Time Spent on Website, Income, Age, Daily Internet Usage, and Gender rely on one and another.

Data Sets Used-

We have taken two datasets-

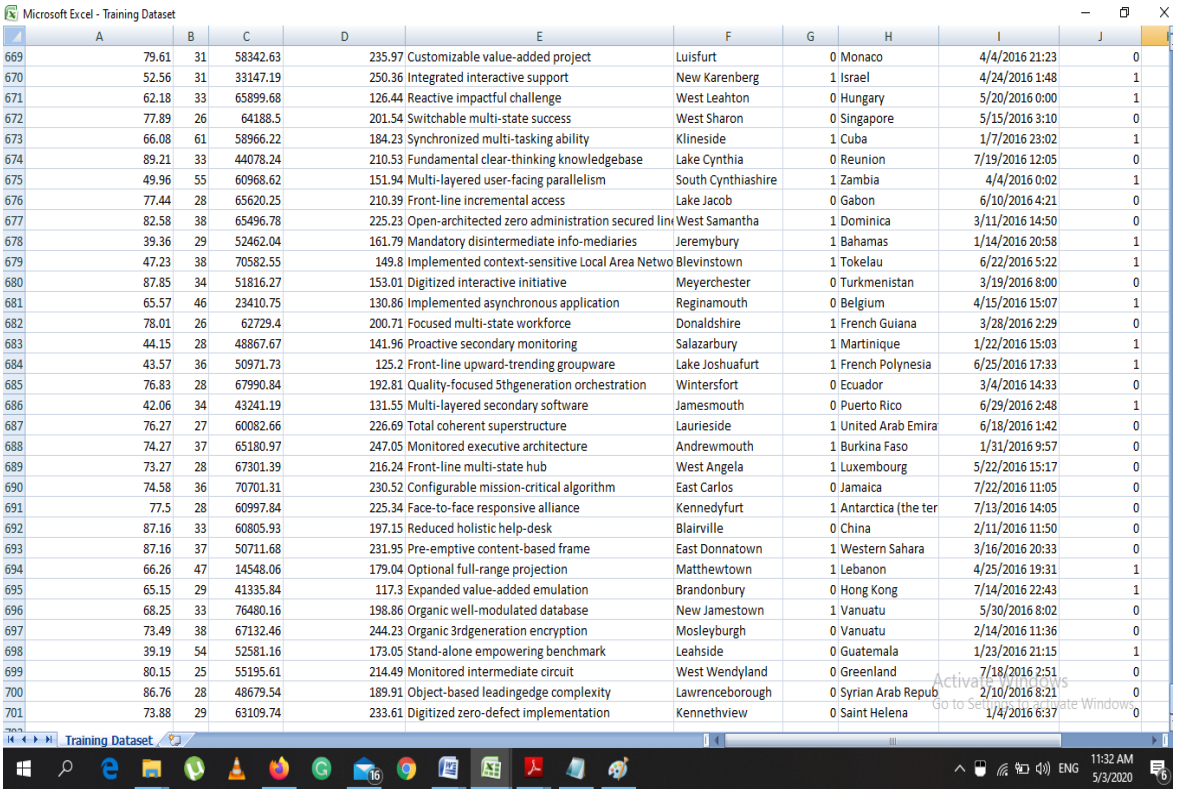
1. **Training dataset:** The dataset in which the output has been mentioned therefore we can deduce logic from that dataset. We apply logistic regression considering a null hypothesis to find out the logic which will be further applied is using the prediction for test dataset.



Microsoft Excel - Training Dataset

	A	B	C	D	E	F	G	H	I	J
1	Daily Time Spent on Site	Age	Area Income	Daily Internet Usage	Ad Topic Line	City	Male	Country	Timestamp	Clicked on Ad
2	68.95	35	61833.9	256.09	Cloned 5thgeneration orchestration	Wrightburgh	0	Tunisia	3/27/2016 0:53	0
3	80.23	31	68441.85	193.77	Monitored national standardization	West Jodi	1	Nauru	4/4/2016 1:39	0
4	69.47	26	59785.94	236.5	Organic bottom-line service-desk	Davidton	0	San Marino	3/13/2016 20:35	0
5	74.15	29	54806.18	245.89	Triple-buffered reciprocal time-frame	West Terrifurt	1	Italy	1/10/2016 2:31	0
6	68.37	35	73889.99	225.58	Robust logistical utilization	South Manuel	0	Iceland	6/3/2016 3:36	0
7	59.99	23	59761.56	226.74	Sharable client-driven software	Jamieberg	1	Norway	5/19/2016 14:30	0
8	88.91	33	53852.85	208.36	Enhanced dedicated support	Brandonstad	0	Myanmar	1/28/2016 20:59	0
9	66	48	24593.33	131.76	Reactive local challenge	Port Jefferybury	1	Australia	3/7/2016 1:40	1
10	74.53	30	68862	221.51	Configurable coherent function	West Colin	1	Grenada	4/18/2016 9:33	0
11	69.88	20	55642.32	183.82	Mandatory homogeneous architecture	Ramirezton	1	Ghana	7/11/2016 1:42	0
12	47.64	49	45632.51	122.02	Centralized neutral neural-net	West Brandonton	0	Qatar	3/16/2016 20:19	1
13	83.07	37	62491.01	230.87	Team-oriented grid-enabled Local Area Network	East Theresashire	1	Burundi	5/8/2016 8:10	0
14	69.57	48	51636.92	113.12	Centralized content-based focus group	West Katiefurt	1	Egypt	6/3/2016 1:14	1
15	79.52	24	51739.63	214.23	Synergistic fresh-thinking array	North Tara	0	Bosnia and Herzeg	4/20/2016 21:49	0
16	42.95	33	30976	143.56	Grass-roots coherent extranet	West William	0	Barbados	3/24/2016 9:31	1
17	63.45	23	52182.23	140.64	Persistent demand-driven interface	New Travistown	1	Spain	3/9/2016 3:41	1
18	55.39	37	23936.86	129.41	Customizable multi-tasking website	West Dylanberg	0	Palestinian Territo	1/30/2016 19:20	1
19	82.03	41	71511.08	187.53	Intuitive dynamic attitude	Pruittmouth	0	Afghanistan	5/2/2016 7:00	0
20	54.7	36	31087.54	118.39	Grass-roots solution-oriented conglomeration	Jessicastad	1	British Indian Ocea	2/13/2016 7:53	1
21	74.58	40	23821.72	135.51	Advanced 24/7 productivity	Millertown	1	Russian Federatior	2/27/2016 4:43	1
22	77.22	30	64802.33	224.44	Object-based reciprocal knowledgebase	Port Jacqueline	1	Cameroon	1/5/2016 7:52	0
23	84.59	35	60015.57	226.54	Streamlined non-volatile analyzer	Lake Nicole	1	Cameroon	3/18/2016 13:22	0
24	41.49	52	32635.7	164.83	Mandatory disintermediate utilization	South John	0	Burundi	5/20/2016 8:49	1
25	87.29	36	61628.72	209.93	Future-proofed methodical protocol	Pamelamouth	1	Korea	3/23/2016 9:43	0
26	41.39	41	68962.32	167.22	Exclusive neutral parallelism	Harperborough	0	Tokelau	6/13/2016 17:27	1
27	78.74	28	64828	204.79	Public-key foreground groupware	Port Danielleberg	1	Monaco	5/27/2016 15:25	0
28	48.53	28	38067.08	134.14	Ameliorated client-driven forecast	West Jeremyside	1	Tuvalu	2/8/2016 10:46	1
29	51.95	52	58295.82	129.23	Monitored systematic hierarchy	South Cathyfurt	0	Greece	7/19/2016 8:32	1
30	70.2	34	32708.94	119.2	Open-architected impactful productivity	Palmerside	0	British Virgin Islan	4/14/2016 5:08	1
31	76.02	22	46179.97	209.82	Business-focused value-added definition	West Guybury	0	Bouvet Island (Bou	1/27/2016 12:38	0
32	67.64	35	51473.28	267.01	Programmable asymmetric data-warehouse	Phelpschester	1	Peru	7/2/2016 20:23	0
33	86.41	28	45593.93	207.48	Digitized static capability	Lake Melindamouth	1	Aruba	3/1/2016 22:13	0

Training Dataset

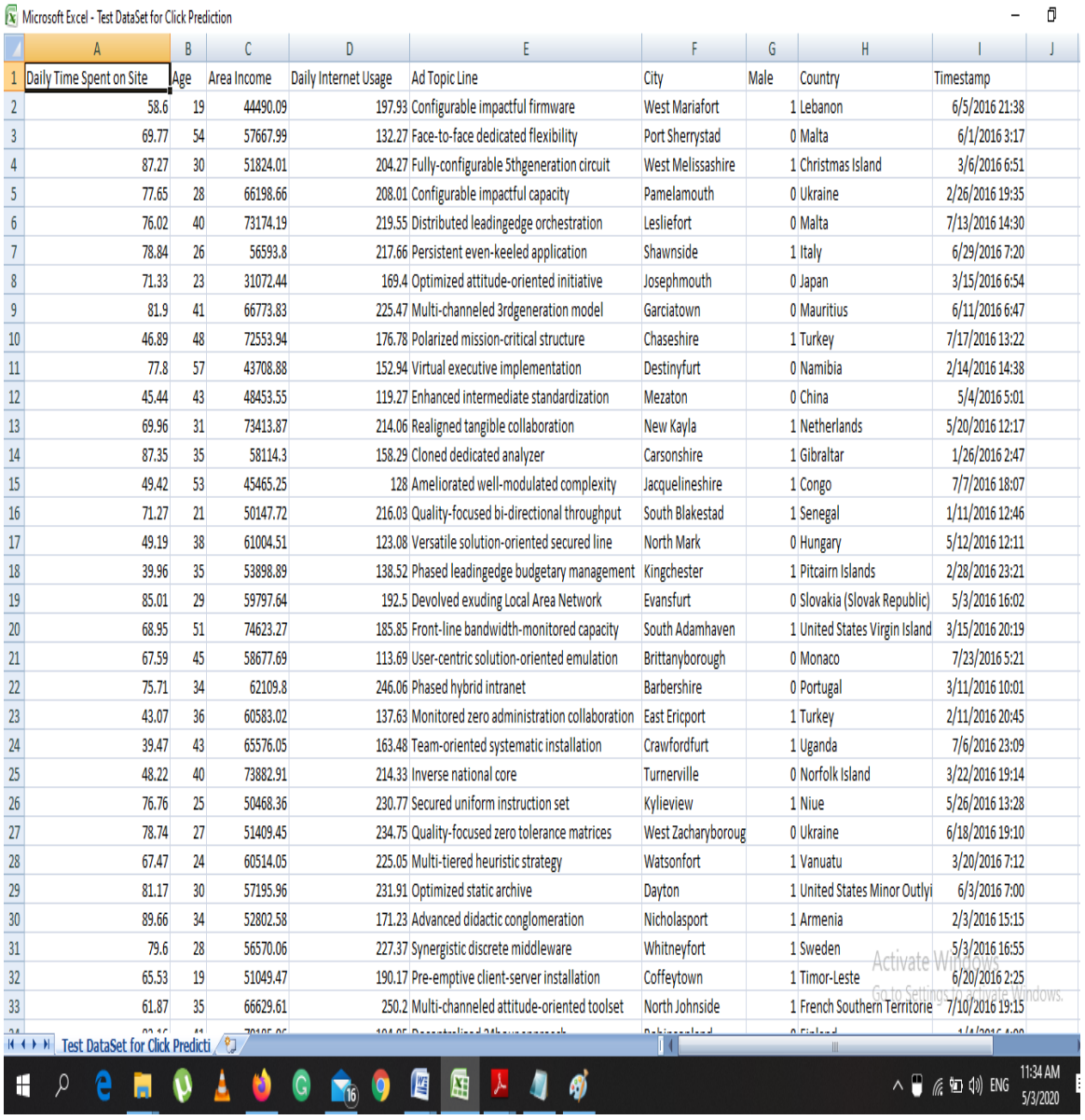


Microsoft Excel - Training Dataset

	A	B	C	D	E	F	G	H	I	J
669	79.61	31	58342.63	235.97	Customizable value-added project	Luisfurt	0	Monaco	4/4/2016 21:23	0
670	52.56	31	33147.19	250.36	Integrated interactive support	New Karenberg	1	Israel	4/24/2016 1:48	1
671	62.18	33	65899.68	126.44	Reactive impactful challenge	West Leahton	0	Hungary	5/20/2016 0:00	1
672	77.89	26	64188.5	201.54	Switchable multi-state success	West Sharon	0	Singapore	5/15/2016 3:10	0
673	66.08	61	58966.22	184.23	Synchronized multi-tasking ability	Klineside	1	Cuba	1/7/2016 23:02	1
674	89.21	33	44078.24	210.53	Fundamental clear-thinking knowledgebase	Lake Cynthia	0	Reunion	7/19/2016 12:05	0
675	49.96	55	60968.62	151.94	Multi-layered user-facing parallelism	South Cynthiashire	1	Zambia	4/4/2016 0:02	1
676	77.44	28	65620.25	210.39	Front-line incremental access	Lake Jacob	0	Gabon	6/10/2016 4:21	0
677	82.58	38	65496.78	225.23	Open-architected zero administration secured lin	West Samantha	1	Dominica	3/11/2016 14:50	0
678	39.36	29	52462.04	161.79	Mandatory disintermediate info-mediaries	Jeremybury	1	Bahamas	1/14/2016 20:58	1
679	47.23	38	70582.55	149.8	Implemented context-sensitive Local Area Netwo	Blevinstown	1	Tokelau	6/22/2016 5:22	1
680	87.85	34	51816.27	153.01	Digitized interactive initiative	Meyerchester	0	Turkmenistan	3/19/2016 8:00	0
681	65.57	46	23410.75	130.86	Implemented asynchronous application	Reginamouth	0	Belgium	4/15/2016 15:07	1
682	78.01	26	62729.4	200.71	Focused multi-state workforce	Donaldshire	1	French Guiana	3/28/2016 2:29	0
683	44.15	28	48867.67	141.96	Proactive secondary monitoring	Salazarbury	1	Martinique	1/22/2016 15:03	1
684	43.57	36	50971.73	125.2	Front-line upward-trending groupware	Lake Joshuafurt	1	French Polynesia	6/25/2016 17:33	1
685	76.83	28	67990.84	192.81	Quality-focused 5thgeneration orchestration	Wintersfort	0	Ecuador	3/4/2016 14:33	0
686	42.06	34	43241.19	131.55	Multi-layered secondary software	Jamesmouth	0	Puerto Rico	6/29/2016 2:48	1
687	76.27	27	60082.66	226.69	Total coherent superstructure	Laurieside	1	United Arab Emira	6/18/2016 1:42	0
688	74.27	37	65180.97	247.05	Monitored executive architecture	Andrewmouth	1	Burkina Faso	1/31/2016 9:57	0
689	73.27	28	67301.39	216.24	Front-line multi-state hub	West Angela	1	Luxembourg	5/22/2016 15:17	0
690	74.58	36	70701.31	230.52	Configurable mission-critical algorithm	East Carlos	0	Jamaica	7/22/2016 11:05	0
691	77.5	28	60997.84	225.34	Face-to-face responsive alliance	Kennedyfurt	1	Antarctica (the ter	7/13/2016 14:05	0
692	87.16	33	60805.93	197.15	Reduced holistic help-desk	Blairville	0	China	2/11/2016 11:50	0
693	87.16	37	50711.68	231.95	Pre-emptive content-based frame	East Donnatown	1	Western Sahara	3/16/2016 20:33	0
694	66.26	47	14548.06	179.04	Optional full-range projection	Matthewtown	1	Lebanon	4/25/2016 19:31	1
695	65.15	29	41335.84	117.3	Expanded value-added emulation	Brandonbury	0	Hong Kong	7/14/2016 22:43	1
696	68.25	33	76480.16	198.86	Organic well-modulated database	New Jamestown	1	Vanuatu	5/30/2016 8:02	0
697	73.49	38	67132.46	244.23	Organic 3rdgeneration encryption	Mosleyburgh	0	Vanuatu	2/14/2016 11:36	0
698	39.19	54	52581.16	173.05	Stand-alone empowering benchmark	Leahside	0	Guatemala	1/23/2016 21:15	1
699	80.15	25	55195.61	214.49	Monitored intermediate circuit	West Wendyland	0	Greenland	7/18/2016 2:51	0
700	86.76	28	48679.54	189.91	Object-based leadingedge complexity	Lawrenceborough	0	Syrian Arab Repub	2/10/2016 8:21	0
701	73.88	29	63109.74	233.61	Digitized zero-defect implementation	Kennethview	0	Saint Helena	1/4/2016 6:37	0

Training Dataset

2. **Test Data-** This is the data on which the prediction has to be performed. After deducing the logic from the training data we apply this logic to test data to find out the results and the sccuracy of the logic that is applied to the dataset.



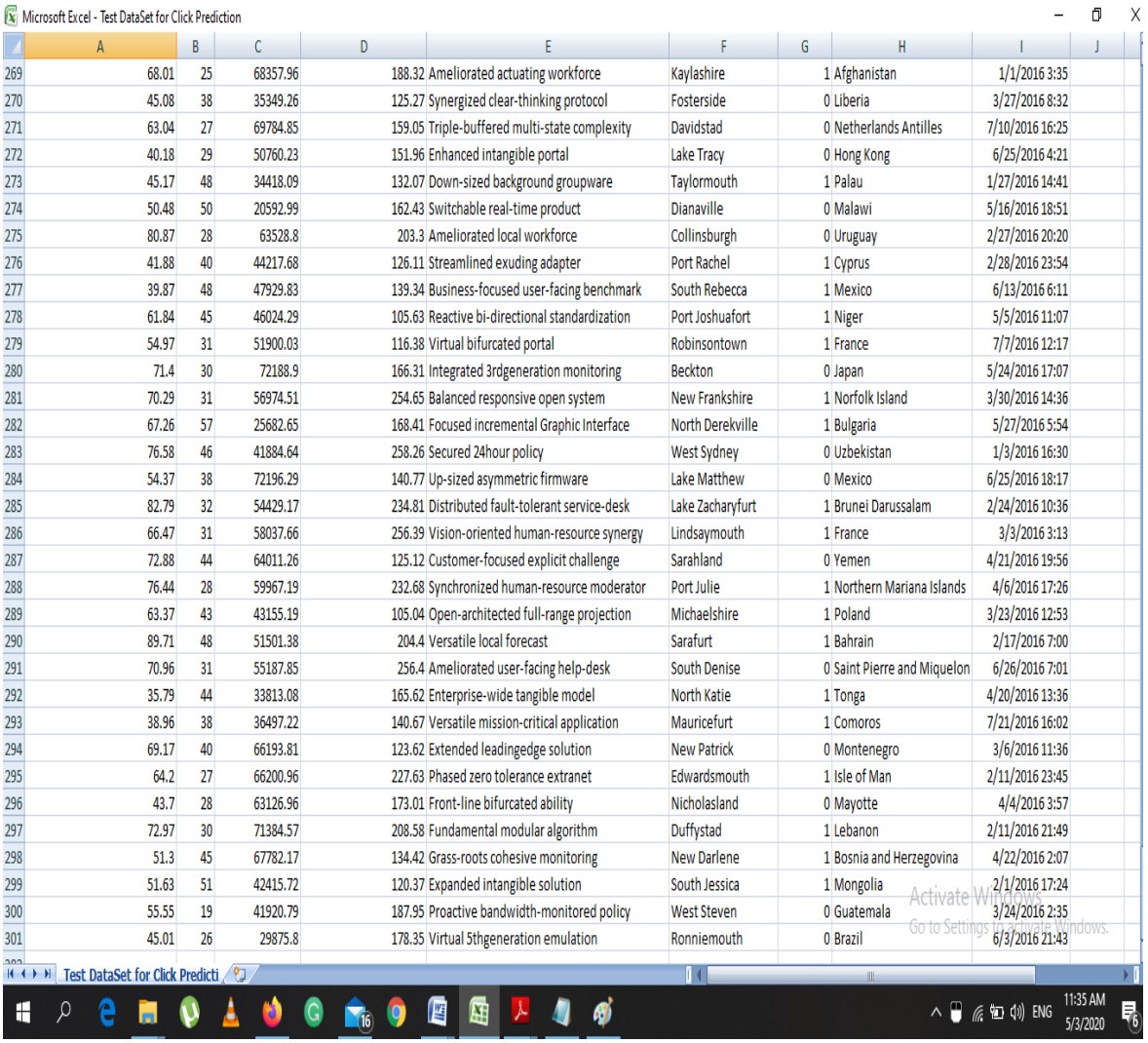
Microsoft Excel - Test DataSet for Click Prediction

	A	B	C	D	E	F	G	H	I	J
1	Daily Time Spent on Site	Age	Area Income	Daily Internet Usage	Ad Topic Line	City	Male	Country	Timestamp	
2	58.6	19	44490.09	197.93	Configurable impactful firmware	West Mariafort	1	Lebanon	6/5/2016 21:38	
3	69.77	54	57667.99	132.27	Face-to-face dedicated flexibility	Port Sherrystad	0	Malta	6/1/2016 3:17	
4	87.27	30	51824.01	204.27	Fully-configurable 5thgeneration circuit	West Melissashire	1	Christmas Island	3/6/2016 6:51	
5	77.65	28	66198.66	208.01	Configurable impactful capacity	Pamelamouth	0	Ukraine	2/26/2016 19:35	
6	76.02	40	73174.19	219.55	Distributed leadingedge orchestration	Lesliefort	0	Malta	7/13/2016 14:30	
7	78.84	26	56593.8	217.66	Persistent even-keeled application	Shawnside	1	Italy	6/29/2016 7:20	
8	71.33	23	31072.44	169.4	Optimized attitude-oriented initiative	Josephmouth	0	Japan	3/15/2016 6:54	
9	81.9	41	66773.83	225.47	Multi-channeled 3rdgeneration model	Garciatown	0	Mauritius	6/11/2016 6:47	
10	46.89	48	72553.94	176.78	Polarized mission-critical structure	Chaseshire	1	Turkey	7/17/2016 13:22	
11	77.8	57	43708.88	152.94	Virtual executive implementation	Destinyfurt	0	Namibia	2/14/2016 14:38	
12	45.44	43	48453.55	119.27	Enhanced intermediate standardization	Mezaton	0	China	5/4/2016 5:01	
13	69.96	31	73413.87	214.06	Realigned tangible collaboration	New Kayla	1	Netherlands	5/20/2016 12:17	
14	87.35	35	58114.3	158.29	Cloned dedicated analyzer	Carsonshire	1	Gibraltar	1/26/2016 2:47	
15	49.42	53	45465.25	128	Ameliorated well-modulated complexity	Jacquelineshire	1	Congo	7/7/2016 18:07	
16	71.27	21	50147.72	216.03	Quality-focused bi-directional throughput	South Blakestad	1	Senegal	1/11/2016 12:46	
17	49.19	38	61004.51	123.08	Versatile solution-oriented secured line	North Mark	0	Hungary	5/12/2016 12:11	
18	39.96	35	53898.89	138.52	Phased leadingedge budgetary management	Kingchester	1	Pitcairn Islands	2/28/2016 23:21	
19	85.01	29	59797.64	192.5	Devolved exuding Local Area Network	Evansfurt	0	Slovakia (Slovak Republic)	5/3/2016 16:02	
20	68.95	51	74623.27	185.85	Front-line bandwidth-monitored capacity	South Adamhaven	1	United States Virgin Island	3/15/2016 20:19	
21	67.59	45	58677.69	113.69	User-centric solution-oriented emulation	Brittanyborough	0	Monaco	7/23/2016 5:21	
22	75.71	34	62109.8	246.06	Phased hybrid intranet	Barbershire	0	Portugal	3/11/2016 10:01	
23	43.07	36	60583.02	137.63	Monitored zero administration collaboration	East Ericport	1	Turkey	2/11/2016 20:45	
24	39.47	43	65576.05	163.48	Team-oriented systematic installation	Crawfordfurt	1	Uganda	7/6/2016 23:09	
25	48.22	40	73882.91	214.33	Inverse national core	Turnerville	0	Norfolk Island	3/22/2016 19:14	
26	76.76	25	50468.36	230.77	Secured uniform instruction set	Kylieview	1	Niue	5/26/2016 13:28	
27	78.74	27	51409.45	234.75	Quality-focused zero tolerance matrices	West Zacharyboroug	0	Ukraine	6/18/2016 19:10	
28	67.47	24	60514.05	225.05	Multi-tiered heuristic strategy	Watsonfort	1	Vanuatu	3/20/2016 7:12	
29	81.17	30	57195.96	231.91	Optimized static archive	Dayton	1	United States Minor Outlyi	6/3/2016 7:00	
30	89.66	34	52802.58	171.23	Advanced didactic conglomeration	Nicholasport	1	Armenia	2/3/2016 15:15	
31	79.6	28	56570.06	227.37	Synergistic discrete middleware	Whitneyfort	1	Sweden	5/3/2016 16:55	
32	65.53	19	51049.47	190.17	Pre-emptive client-server installation	Coffeytown	1	Timor-Leste	6/20/2016 2:25	
33	61.87	35	66629.61	250.2	Multi-channeled attitude-oriented toolset	North Johnside	1	French Southern Territorie	7/10/2016 19:15	

Test DataSet for Click Predicti

Activate Windows
Go to Settings to activate Windows.

ENG 11:34 AM 5/3/2020



	A	B	C	D	E	F	G	H	I	J
269	68.01	25	68357.96	188.32	Ameliorated actuating workforce	Kaylashire	1	Afghanistan	1/1/2016 3:35	
270	45.08	38	35349.26	125.27	Synergized clear-thinking protocol	Fosterside	0	Liberia	3/27/2016 8:32	
271	63.04	27	69784.85	159.05	Triple-buffered multi-state complexity	Davidstad	0	Netherlands Antilles	7/10/2016 16:25	
272	40.18	29	50760.23	151.96	Enhanced intangible portal	Lake Tracy	0	Hong Kong	6/25/2016 4:21	
273	45.17	48	34418.09	132.07	Down-sized background groupware	Taylormouth	1	Palau	1/27/2016 14:41	
274	50.48	50	20592.99	162.43	Switchable real-time product	Dianaville	0	Malawi	5/16/2016 18:51	
275	80.87	28	63528.8	203.3	Ameliorated local workforce	Collinsburgh	0	Uruguay	2/27/2016 20:20	
276	41.88	40	44217.68	126.11	Streamlined exuding adapter	Port Rachel	1	Cyprus	2/28/2016 23:54	
277	39.87	48	47929.83	139.34	Business-focused user-facing benchmark	South Rebecca	1	Mexico	6/13/2016 6:11	
278	61.84	45	46024.29	105.63	Reactive bi-directional standardization	Port Joshuafort	1	Niger	5/5/2016 11:07	
279	54.97	31	51900.03	116.38	Virtual bifurcated portal	Robinsontown	1	France	7/7/2016 12:17	
280	71.4	30	72188.9	166.31	Integrated 3rdgeneration monitoring	Beckton	0	Japan	5/24/2016 17:07	
281	70.29	31	56974.51	254.65	Balanced responsive open system	New Frankshire	1	Norfolk Island	3/30/2016 14:36	
282	67.26	57	25682.65	168.41	Focused incremental Graphic Interface	North Derekville	1	Bulgaria	5/27/2016 5:54	
283	76.58	46	41884.64	258.26	Secured 24hour policy	West Sydney	0	Uzbekistan	1/3/2016 16:30	
284	54.37	38	72196.29	140.77	Up-sized asymmetric firmware	Lake Matthew	0	Mexico	6/25/2016 18:17	
285	82.79	32	54429.17	234.81	Distributed fault-tolerant service-desk	Lake Zacharyfurt	1	Brunei Darussalam	2/24/2016 10:36	
286	66.47	31	58037.66	256.39	Vision-oriented human-resource synergy	Lindsaymouth	1	France	3/3/2016 3:13	
287	72.88	44	64011.26	125.12	Customer-focused explicit challenge	Sarahland	0	Yemen	4/21/2016 19:56	
288	76.44	28	59967.19	232.68	Synchronized human-resource moderator	Port Julie	1	Northern Mariana Islands	4/6/2016 17:26	
289	63.37	43	43155.19	105.04	Open-architected full-range projection	Michaelshire	1	Poland	3/23/2016 12:53	
290	89.71	48	51501.38	204.4	Versatile local forecast	Sarafurt	1	Bahrain	2/17/2016 7:00	
291	70.96	31	55187.85	256.4	Ameliorated user-facing help-desk	South Denise	0	Saint Pierre and Miquelon	6/26/2016 7:01	
292	35.79	44	33813.08	165.62	Enterprise-wide tangible model	North Katie	1	Tonga	4/20/2016 13:36	
293	38.96	38	36497.22	140.67	Versatile mission-critical application	Mauricefurt	1	Comoros	7/21/2016 16:02	
294	69.17	40	66193.81	123.62	Extended leadingedge solution	New Patrick	0	Montenegro	3/6/2016 11:36	
295	64.2	27	66200.96	227.63	Phased zero tolerance extranet	Edwardsmouth	1	Isle of Man	2/11/2016 23:45	
296	43.7	28	63126.96	173.01	Front-line bifurcated ability	Nicholasland	0	Mayotte	4/4/2016 3:57	
297	72.97	30	71384.57	208.58	Fundamental modular algorithm	Duffystad	1	Lebanon	2/11/2016 21:49	
298	51.3	45	67782.17	134.42	Grass-roots cohesive monitoring	New Darlene	1	Bosnia and Herzegovina	4/22/2016 2:07	
299	51.63	51	42415.72	120.37	Expanded intangible solution	South Jessica	1	Mongolia	2/1/2016 17:24	
300	55.55	19	41920.79	187.95	Proactive bandwidth-monitored policy	West Steven	0	Guatemala	3/24/2016 2:35	
301	45.01	26	29875.8	178.35	Virtual 5thgeneration emulation	Ronniemouth	0	Brazil	6/3/2016 21:43	

RESULTS

Analysis of Variables:-

Frequencies:-

Statistics

		Male	Clicked
N	Valid	700	700
	Missing	0	0
Mean		.47	.48
Median		.00	.00
Mode		0	0
Std. Deviation		.499	.500
Skewness		.138	.069
Std. Error of Skewness		.092	.092
Kurtosis		-1.987	-2.001
Std. Error of Kurtosis		.185	.185
Sum		326	338

Frequency Table

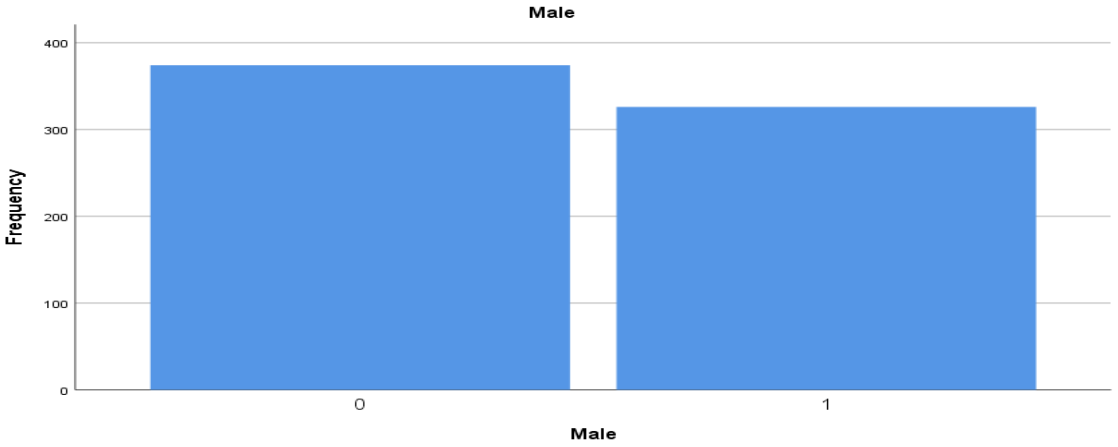
Gender

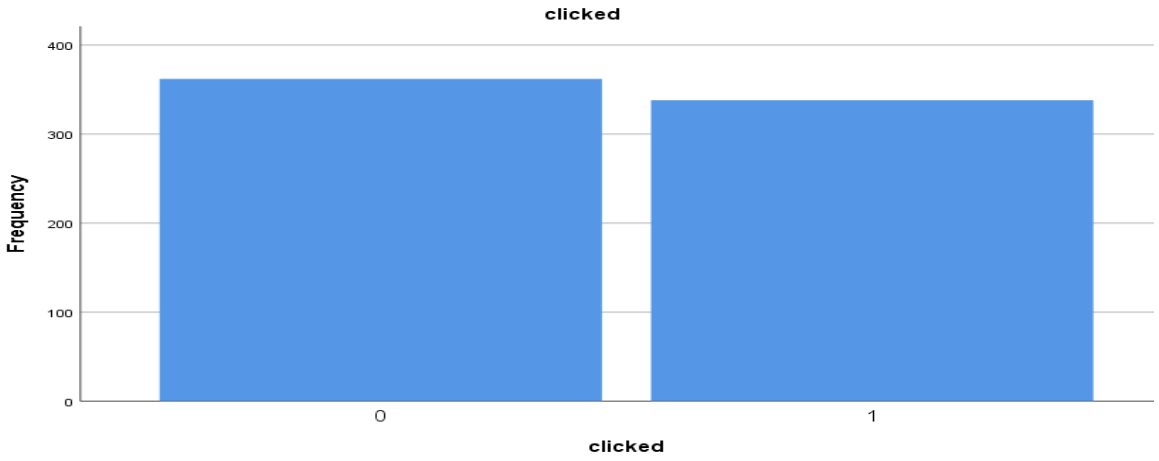
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	374	53.4	53.4	53.4
	1	326	46.6	46.6	100.0
	Total	700	100.0	100.0	

Clicked

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	362	51.7	51.7	51.7
	1	338	48.3	48.3	100.0
	Total	700	100.0	100.0	

Bar Chart


Male
400
300
200
100
0
Frequency
0
1
Male


clicked
400
300
200
100
0
Frequency
0
1
clicked

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
Clicked	.48	.500	700
Time	65.6389	15.65215	700
Age	35.93	8.754	700
Income	55363.5878	13539.53907	700
daily_usage	181.1841	43.58862	700
Gender	.47	.499	700

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Cha
1	.918	.843	.842	.199	.843	74

a. Predictors: (Constant), gender, time, income, Age, daily_usage

b. Dependent Variable: clicked

ANOVA

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	147.349	5	29.470	745.192	.000[b]
	Residual	27.445	694	.040		
	Total	174.794	699			

a. Dependent Variable: clicked

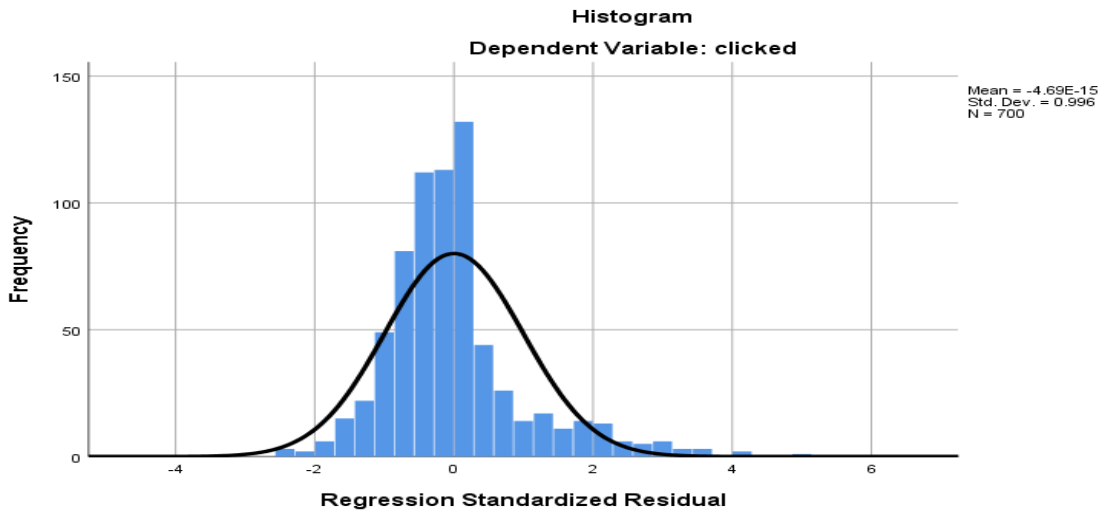
b. Predictors: (Constant), gender, time, income, Age, daily_usage

Coefficient Correlations

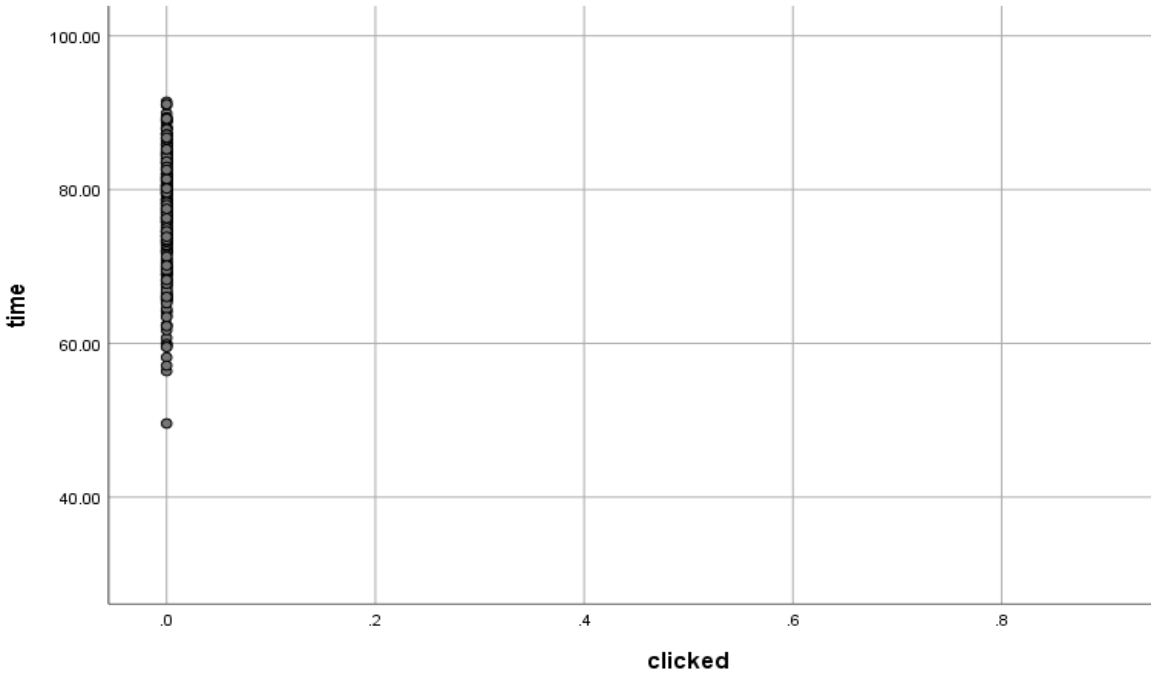
Model			gender	time	income	Age	daily_usag
1	Correlations	gender	1.000	.017	-.012	.013	-.03
		time	.017	1.000	-.144	.193	-.41
		income	-.012	-.144	1.000	.037	-.20
		Age	.013	.193	.037	1.000	.20
		daily_usage	-.032	-.412	-.209	.202	1.00
	Covariances	gender	.000	1.520E-7	-1.056E-10	1.879E-7	-1.031E-
		time	1.520E-7	3.382E-7	-5.005E-11	1.052E-7	-5.087E-
		income	-1.056E-10	-5.005E-11	3.571E-13	2.055E-11	-2.649E-1
		Age	1.879E-7	1.052E-7	2.055E-11	8.792E-7	4.016E-
		daily_usage	-1.031E-7	-5.087E-8	-2.649E-11	4.016E-8	4.498E-

a. Dependent Variable: clicked

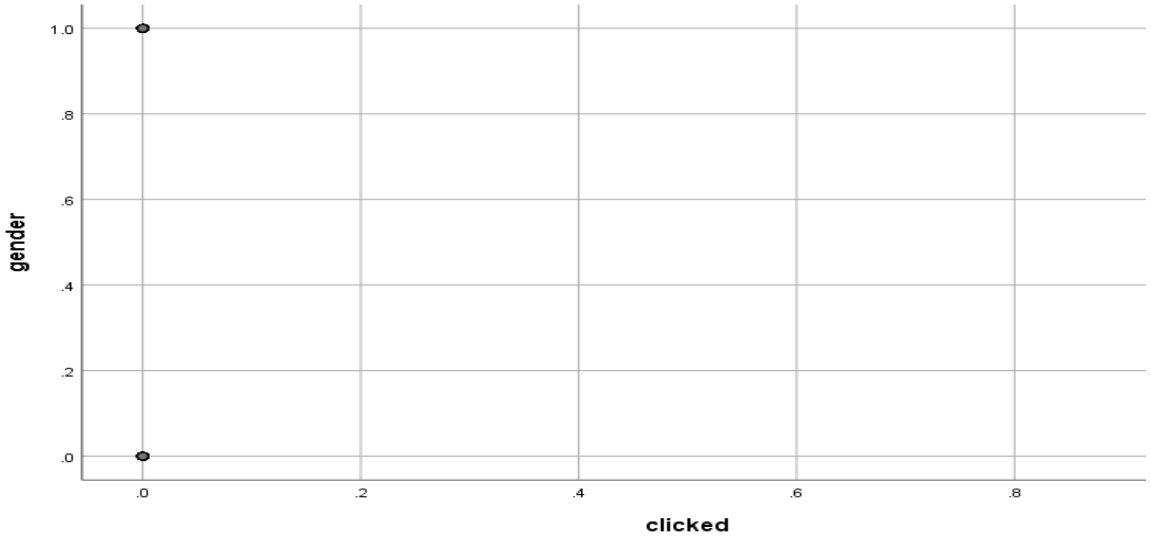
Charts

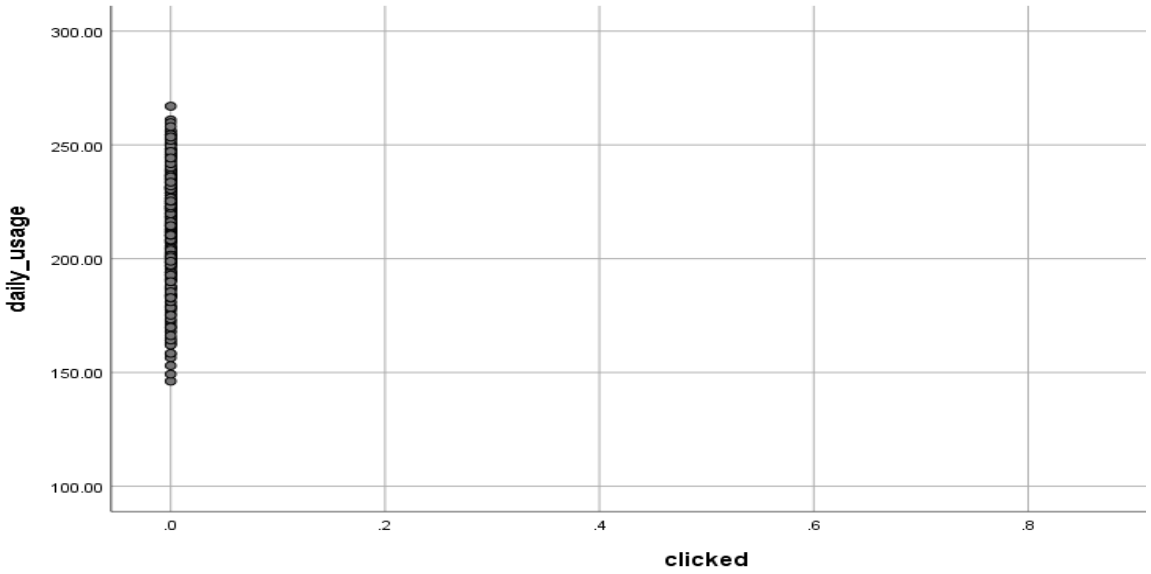


Graphs:-


100.00
80.00
60.00
40.00
time
.0
.2
.4
.6
.8
clicked


70
60
50
40
30
20
10
Age
.0
.2
.4
.6
.8
clicked


gender
1.0
.8
.6
.4
.2
.0
.0
.2
.4
.6
.8
clicked


daily_usage
300.00
250.00
200.00
150.00
100.00
.0
.2
.4
.6
.8
clicked

Outputs using SPSS:-

1. Logistic Regression-

Classification Table

			clicked		Percentage Correct
			0	1	
Step 0	clicked	0	362	0	100.0
		1	338	0	.0
	Overall Percentage				51.7

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	891.098	4	.000
	Block	891.098	4	.000
	Model	891.098	4	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	78.485	.720	.960

Classification Table

Observed			Predicted		
			clicked		Percentage Correct
			0	1	
Step 1	clicked	0	357	5	98.6
		1	8	330	97.6
	Overall Percentage				98.1

The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	time	-.280	.043	42.973	1	.000	.756
	Age	.196	.039	24.663	1	.000	1.216
	income	.000	.000	35.119	1	.000	1.000
	daily_usage	-.090	.013	46.941	1	.000	.914
	Constant	40.972	6.014	46.407	1	.000	6219151914 68059010.000

.

Case Processing Summary

Unweighted Cases		N	Percent
Selected Cases	Included in Analysis	700	100.0
	Missing Cases	0	.0
	Total	700	100.0
Unselected Cases		0	.0
Total		700	100.0

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	time	399.916	1	.000
		Age	167.823	1	.000
		income	160.605	1	.000
		daily_usage	446.673	1	.000
	Overall Statistics		589.425	4	.000

Test Data:-

Case Processing Summary

Unweighted Cases		N	Percent
Selected Cases	Included in Analysis	300	100.0
	Missing Cases	0	.0
	Total	300	100.0
Unselected Cases		0	.0
Total		300	100.0

Classification Table

Observed			Predicted		
			clicked		Percentage Correct
			0	1	
Step 0	clicked	0	0	146	.0
		1	0	154	100.0
	Overall Percentage				51.3

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.053	.116	.213	1	.644	1.05

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	time	.083	1	.773
		Age	.724	1	.395
		income	1.720	1	.190
		daily_usage	.698	1	.403
	Overall Statistics		5.050	4	.282

Omnibus Tests of Model Coefficients

		Chi-square	Df	Sig.
Step 1	Step	5.090	4	.278
	Block	5.090	4	.278
	Model	5.090	4	.278

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	410.585[a]	.017	.022

a. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.

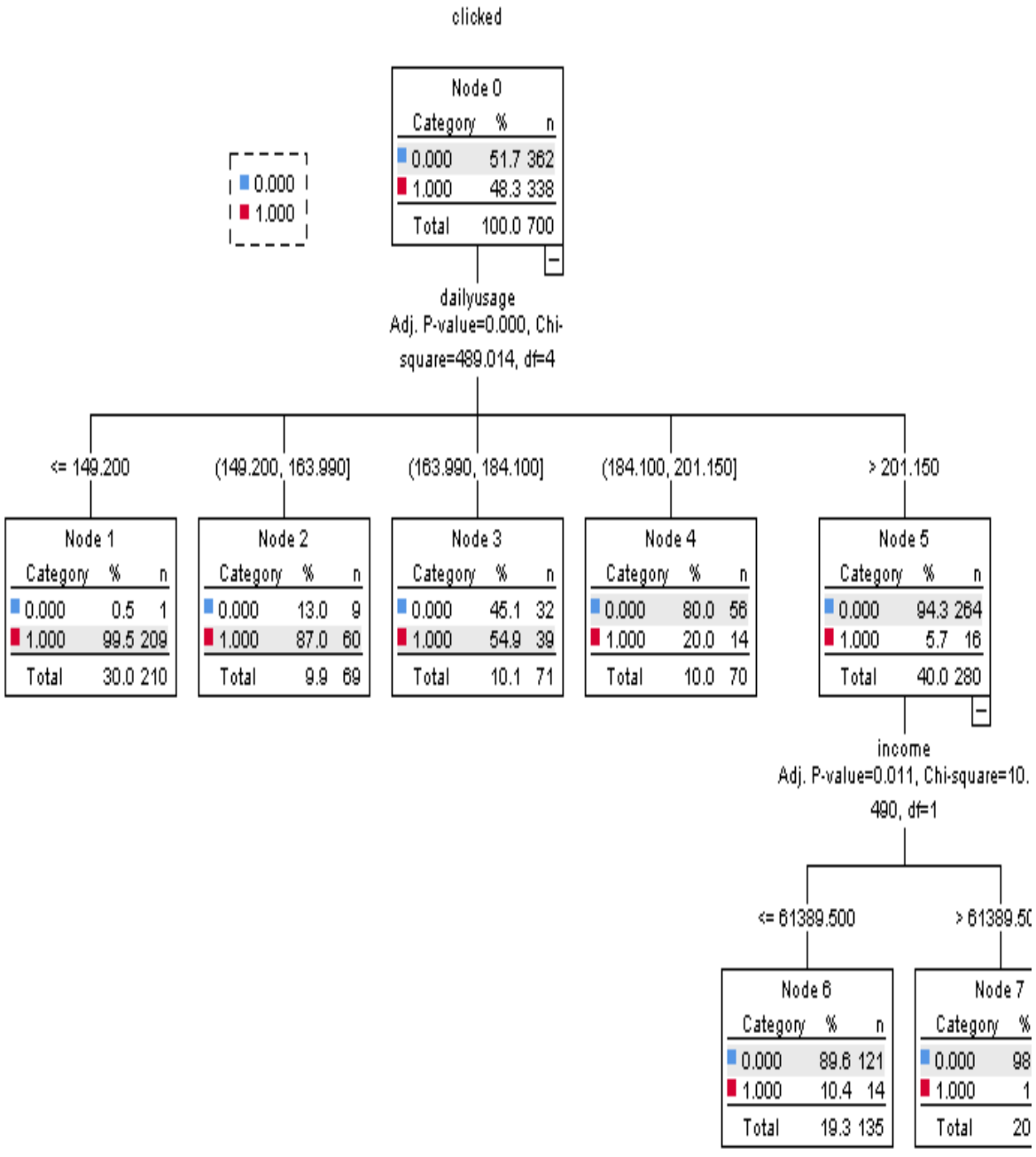
Classification Table

Observed			Predicted		
			Clicked		Percentage Correct
			0	1	
Step 1	clicked	0	69	77	47.3
		1	56	98	63.6
	Overall Percentage				55.7

a. The cut value is .500

2. Random Forest:-

Classification Tree:-



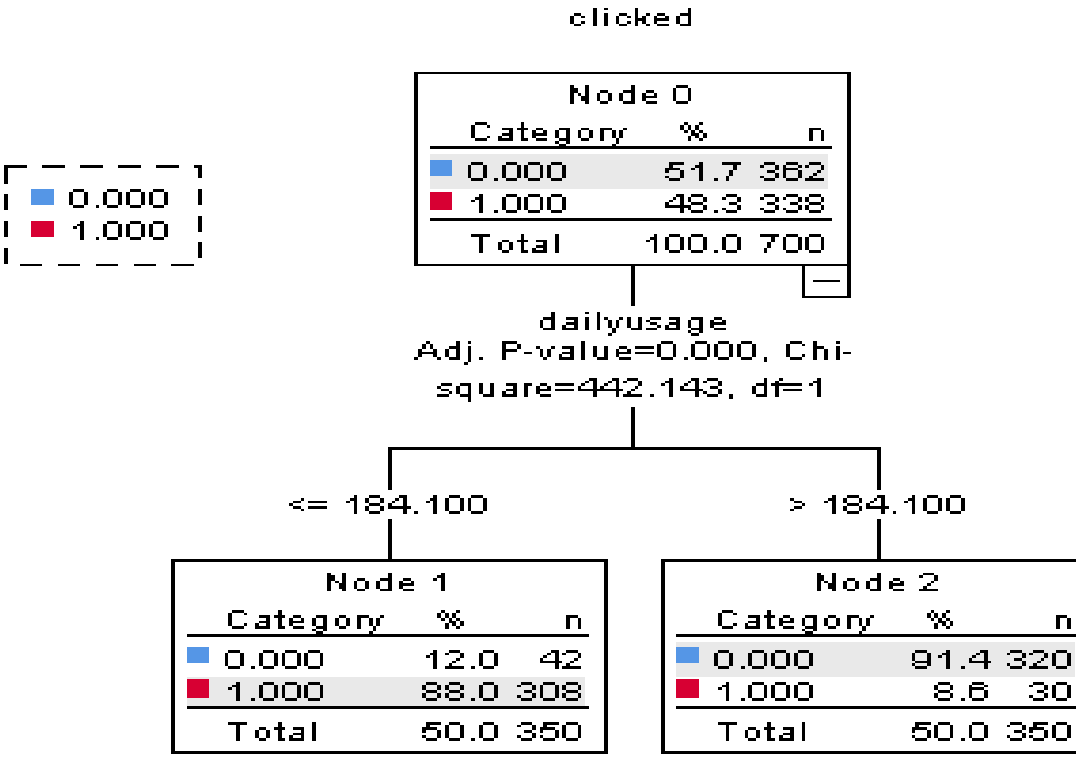
Classification

		Predicted	
Observed	0	1	Percent Correct
0	320	42	88.4%
1	30	308	91.1%
Overall Percentage	50.0%	50.0%	89.7%

Growing Method: CHAID
Dependent Variable: clicked

Model Summary

Specifications	Growing Method	CHAID
	Dependent Variable	clicked
	Independent Variables	time, Age, income, dailyusage
	Validation	None
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	400
	Minimum Cases in Child Node	200
Results	Independent Variables Included	dailyusage
	Number of Nodes	3
	Number of Terminal Nodes	2
	Depth	1



Tree Table

	0		1		Total		Predicted
Node	N	Percent	N	Percent	N	Percent	Category
0	362	51.7%	338	48.3%	700	100.0%	0
1	42	12.0%	308	88.0%	350	50.0%	1
2	320	91.4%	30	8.6%	350	50.0%	0

Growing Method: CHAID

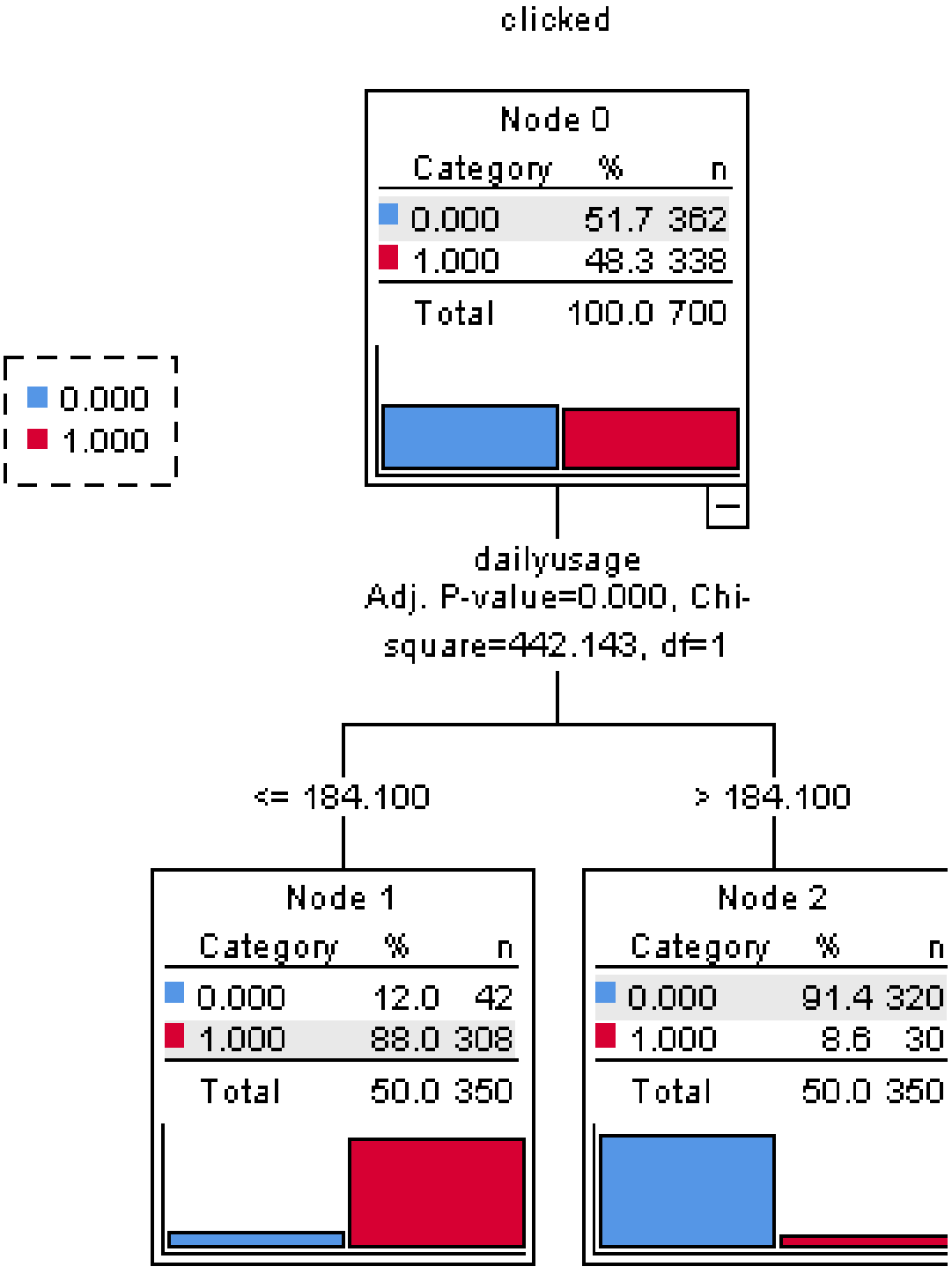
Dependent Variable: clicked

Risk

Estimate	Std. Error
.103	.011

Classification

Observed	Predicted 0	Predicted 1	Percent Correct
0	320	42	88.4%
1	30	308	91.1%
Overall Percentage	50.0%	50.0%	89.7%

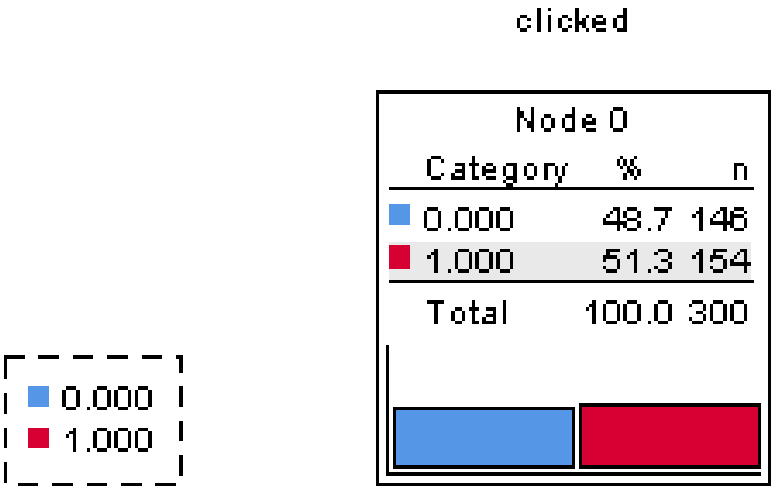

clicked
0.000
1.000
Node 0
Category % n
0.000 51.7 362
1.000 48.3 338
Total 100.0 700
dailyusage
Adj. P-value=0.000, Chi-square=442.143, df=1
<= 184.100
> 184.100
Node 1
Category % n
0.000 12.0 42
1.000 88.0 308
Total 50.0 350
Node 2
Category % n
0.000 91.4 320
1.000 8.6 30
Total 50.0 350

Test Data:-

Misclassification Costs

	Predicted	
Observed	0	1
0	.000	1.000
1	1.000	.000

Dependent Variable: clicked



Tree Table

	0		1		Total		Predicted
Node	N	Percent	N	Percent	N	Percent	Category
0	146	48.7%	154	51.3%	300	100.0%	1

Growing Method: CHAID

Dependent Variable: clicked

Risk

Estimate	Std. Error
.487	.029

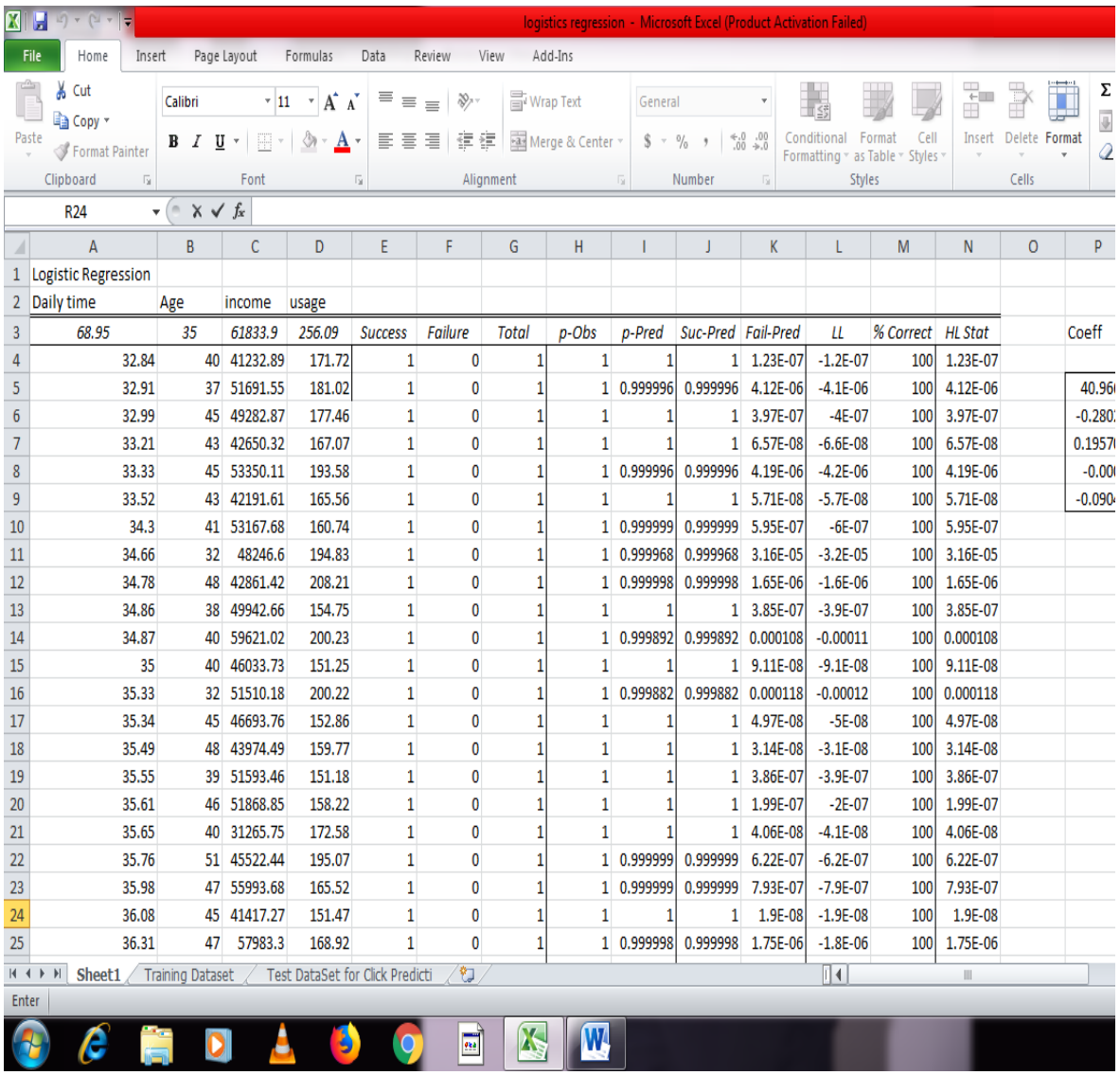
Classification

	Predicted		
Observed	0	1	Percent Correct
0	0	146	0.0%
1	0	154	100.0%
Overall Percentage	0.0%	100.0%	51.3%

Outputs Using Microsoft Excel:-

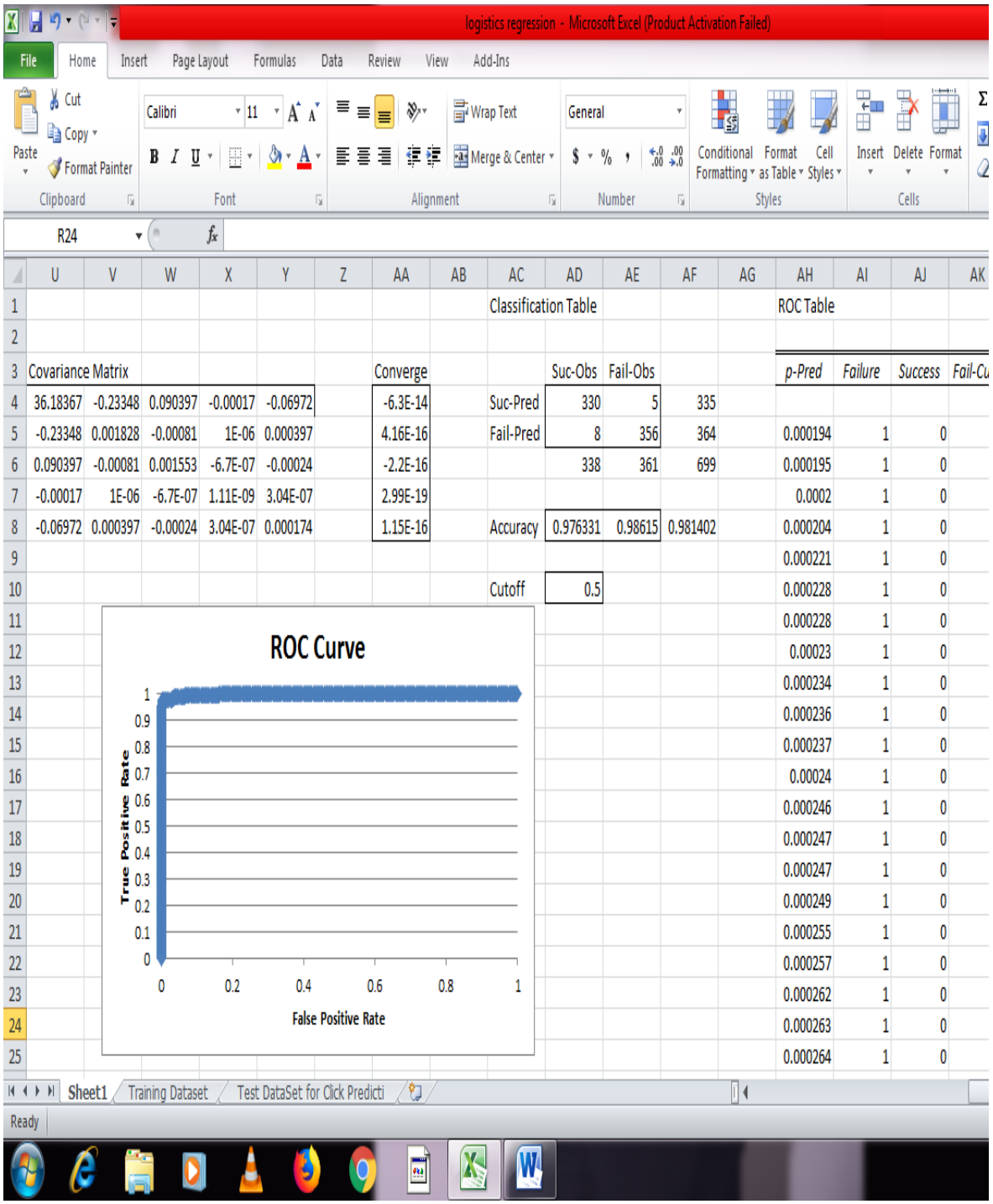
1. Logistic Regression

a. Training Dataset



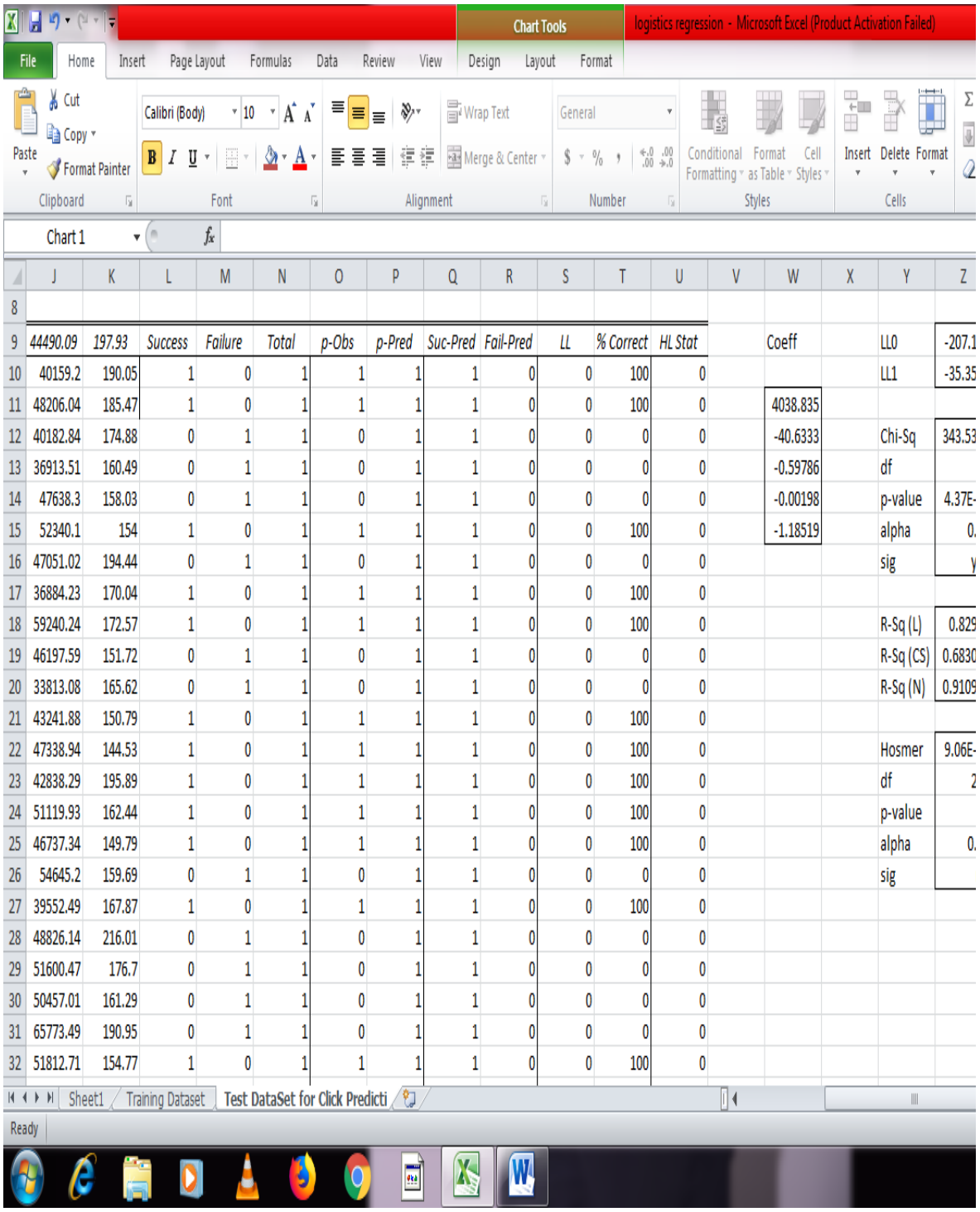
logistics regression - Microsoft Excel (Product Activation Failed)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Logistic Regression															
2	Daily time	Age	income	usage												
3	*68.95*	*35*	*61833.9*	*256.09*	*Success*	*Failure*	*Total*	*p-Obs*	*p-Pred*	*Suc-Pred*	*Fail-Pred*	*LL*	*% Correct*	*HL Stat*		Coeff
4	32.84	40	41232.89	171.72	1	0	1	1	1	1	1.23E-07	-1.2E-07	100	1.23E-07		
5	32.91	37	51691.55	181.02	1	0	1	1	0.999996	0.999996	4.12E-06	-4.1E-06	100	4.12E-06		40.96
6	32.99	45	49282.87	177.46	1	0	1	1	1	1	3.97E-07	-4E-07	100	3.97E-07		-0.280
7	33.21	43	42650.32	167.07	1	0	1	1	1	1	6.57E-08	-6.6E-08	100	6.57E-08		0.1957
8	33.33	45	53350.11	193.58	1	0	1	1	0.999996	0.999996	4.19E-06	-4.2E-06	100	4.19E-06		-0.00
9	33.52	43	42191.61	165.56	1	0	1	1	1	1	5.71E-08	-5.7E-08	100	5.71E-08		-0.090
10	34.3	41	53167.68	160.74	1	0	1	1	0.999999	0.999999	5.95E-07	-6E-07	100	5.95E-07		
11	34.66	32	48246.6	194.83	1	0	1	1	0.999968	0.999968	3.16E-05	-3.2E-05	100	3.16E-05		
12	34.78	48	42861.42	208.21	1	0	1	1	0.999998	0.999998	1.65E-06	-1.6E-06	100	1.65E-06		
13	34.86	38	49942.66	154.75	1	0	1	1	1	1	3.85E-07	-3.9E-07	100	3.85E-07		
14	34.87	40	59621.02	200.23	1	0	1	1	0.999892	0.999892	0.000108	-0.00011	100	0.000108		
15	35	40	46033.73	151.25	1	0	1	1	1	1	9.11E-08	-9.1E-08	100	9.11E-08		
16	35.33	32	51510.18	200.22	1	0	1	1	0.999882	0.999882	0.000118	-0.00012	100	0.000118		
17	35.34	45	46693.76	152.86	1	0	1	1	1	1	4.97E-08	-5E-08	100	4.97E-08		
18	35.49	48	43974.49	159.77	1	0	1	1	1	1	3.14E-08	-3.1E-08	100	3.14E-08		
19	35.55	39	51593.46	151.18	1	0	1	1	1	1	3.86E-07	-3.9E-07	100	3.86E-07		
20	35.61	46	51868.85	158.22	1	0	1	1	1	1	1.99E-07	-2E-07	100	1.99E-07		
21	35.65	40	31265.75	172.58	1	0	1	1	1	1	4.06E-08	-4.1E-08	100	4.06E-08		
22	35.76	51	45522.44	195.07	1	0	1	1	0.999999	0.999999	6.22E-07	-6.2E-07	100	6.22E-07		
23	35.98	47	55993.68	165.52	1	0	1	1	0.999999	0.999999	7.93E-07	-7.9E-07	100	7.93E-07		
24	36.08	45	41417.27	151.47	1	0	1	1	1	1	1.9E-08	-1.9E-08	100	1.9E-08		
25	36.31	47	57983.3	168.92	1	0	1	1	0.999998	0.999998	1.75E-06	-1.8E-06	100	1.75E-06		



	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK
1									Classification Table					ROC Table			
2																	
3	Covariance Matrix						Converge			Suc-Obs	Fail-Obs			*p-Pred*	*Failure*	*Success*	*Fail-Cu*
4	36.18367	-0.23348	0.090397	-0.00017	-0.06972		-6.3E-14		Suc-Pred	330	5	335					
5	-0.23348	0.001828	-0.00081	1E-06	0.000397		4.16E-16		Fail-Pred	8	356	364		0.000194	1	0	
6	0.090397	-0.00081	0.001553	-6.7E-07	-0.00024		-2.2E-16			338	361	699		0.000195	1	0	
7	-0.00017	1E-06	-6.7E-07	1.11E-09	3.04E-07		2.99E-19							0.0002	1	0	
8	-0.06972	0.000397	-0.00024	3.04E-07	0.000174		1.15E-16		Accuracy	0.976331	0.98615	0.981402		0.000204	1	0	
9														0.000221	1	0	
10									Cutoff	0.5				0.000228	1	0	
11														0.000228	1	0	
12														0.00023	1	0	
13														0.000234	1	0	
14														0.000236	1	0	
15														0.000237	1	0	
16														0.00024	1	0	
17														0.000246	1	0	
18														0.000247	1	0	
19														0.000247	1	0	
20														0.000249	1	0	
21														0.000255	1	0	
22														0.000257	1	0	
23														0.000262	1	0	
24														0.000263	1	0	
25														0.000264	1	0	

b. Test Dataset



	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
8																	
9	*44490.09*	*197.93*	*Success*	*Failure*	*Total*	*p-Obs*	*p-Pred*	*Suc-Pred*	*Fail-Pred*	*LL*	*% Correct*	*HL Stat*		Coeff		LL0	-207.1
10	40159.2	190.05	1	0	1	1	1	1	0	0	100	0				LL1	-35.35
11	48206.04	185.47	1	0	1	1	1	1	0	0	100	0		4038.835			
12	40182.84	174.88	0	1	1	0	1	1	0	0	0	0		-40.6333		Chi-Sq	343.53
13	36913.51	160.49	0	1	1	0	1	1	0	0	0	0		-0.59786		df	
14	47638.3	158.03	0	1	1	0	1	1	0	0	0	0		-0.00198		p-value	4.37E-
15	52340.1	154	1	0	1	1	1	1	0	0	100	0		-1.18519		alpha	0.
16	47051.02	194.44	0	1	1	0	1	1	0	0	0	0				sig	y
17	36884.23	170.04	1	0	1	1	1	1	0	0	100	0					
18	59240.24	172.57	1	0	1	1	1	1	0	0	100	0				R-Sq (L)	0.829
19	46197.59	151.72	0	1	1	0	1	1	0	0	0	0				R-Sq (CS)	0.6830
20	33813.08	165.62	0	1	1	0	1	1	0	0	0	0				R-Sq (N)	0.9109
21	43241.88	150.79	1	0	1	1	1	1	0	0	100	0					
22	47338.94	144.53	1	0	1	1	1	1	0	0	100	0				Hosmer	9.06E-
23	42838.29	195.89	1	0	1	1	1	1	0	0	100	0				df	2
24	51119.93	162.44	1	0	1	1	1	1	0	0	100	0				p-value	
25	46737.34	149.79	1	0	1	1	1	1	0	0	100	0				alpha	0.
26	54645.2	159.69	0	1	1	0	1	1	0	0	0	0				sig	
27	39552.49	167.87	1	0	1	1	1	1	0	0	100	0					
28	48826.14	216.01	0	1	1	0	1	1	0	0	0	0					
29	51600.47	176.7	0	1	1	0	1	1	0	0	0	0					
30	50457.01	161.29	0	1	1	0	1	1	0	0	0	0					
31	65773.49	190.95	0	1	1	0	1	1	0	0	0	0					
32	51812.71	154.77	1	0	1	1	1	1	0	0	100	0					

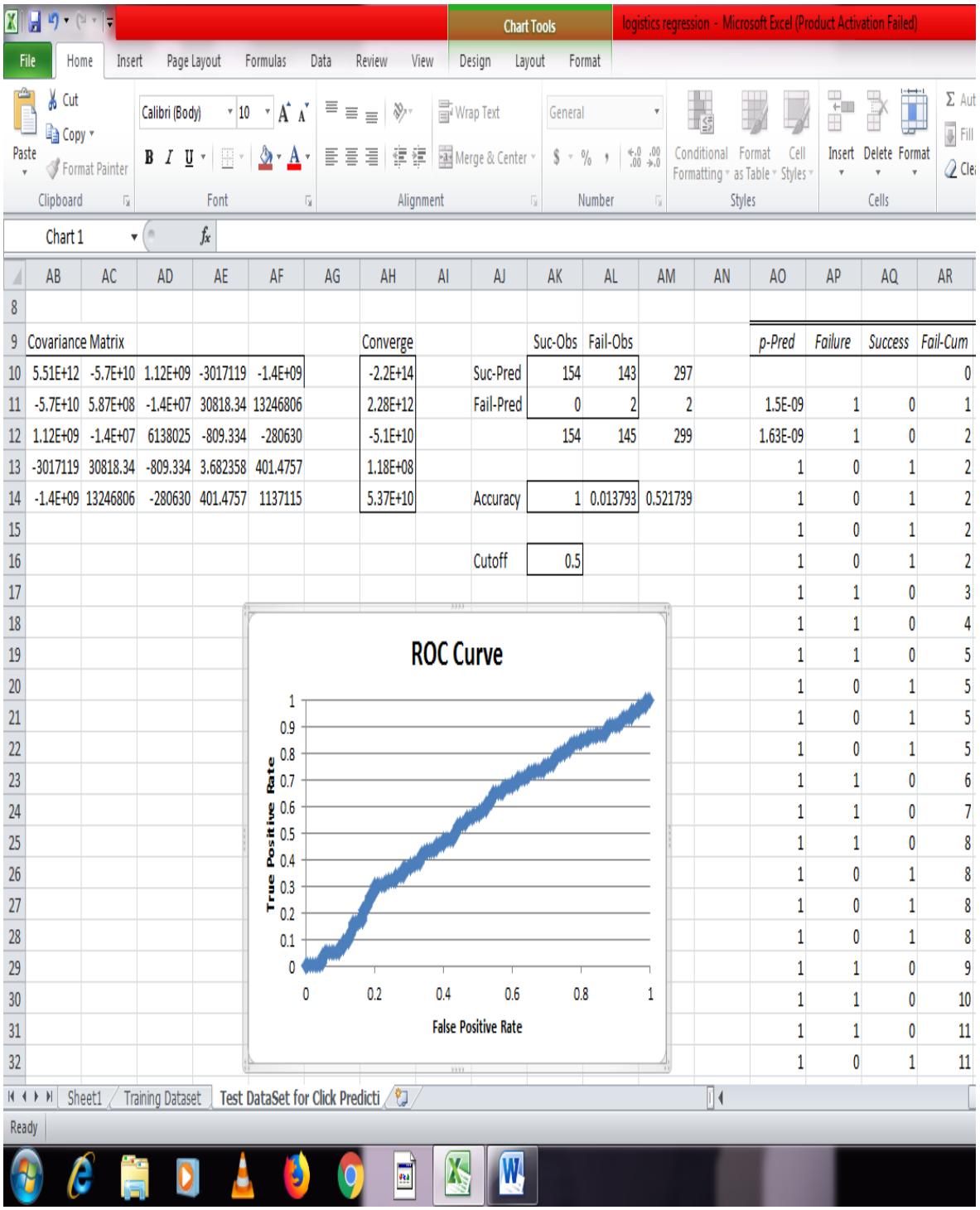
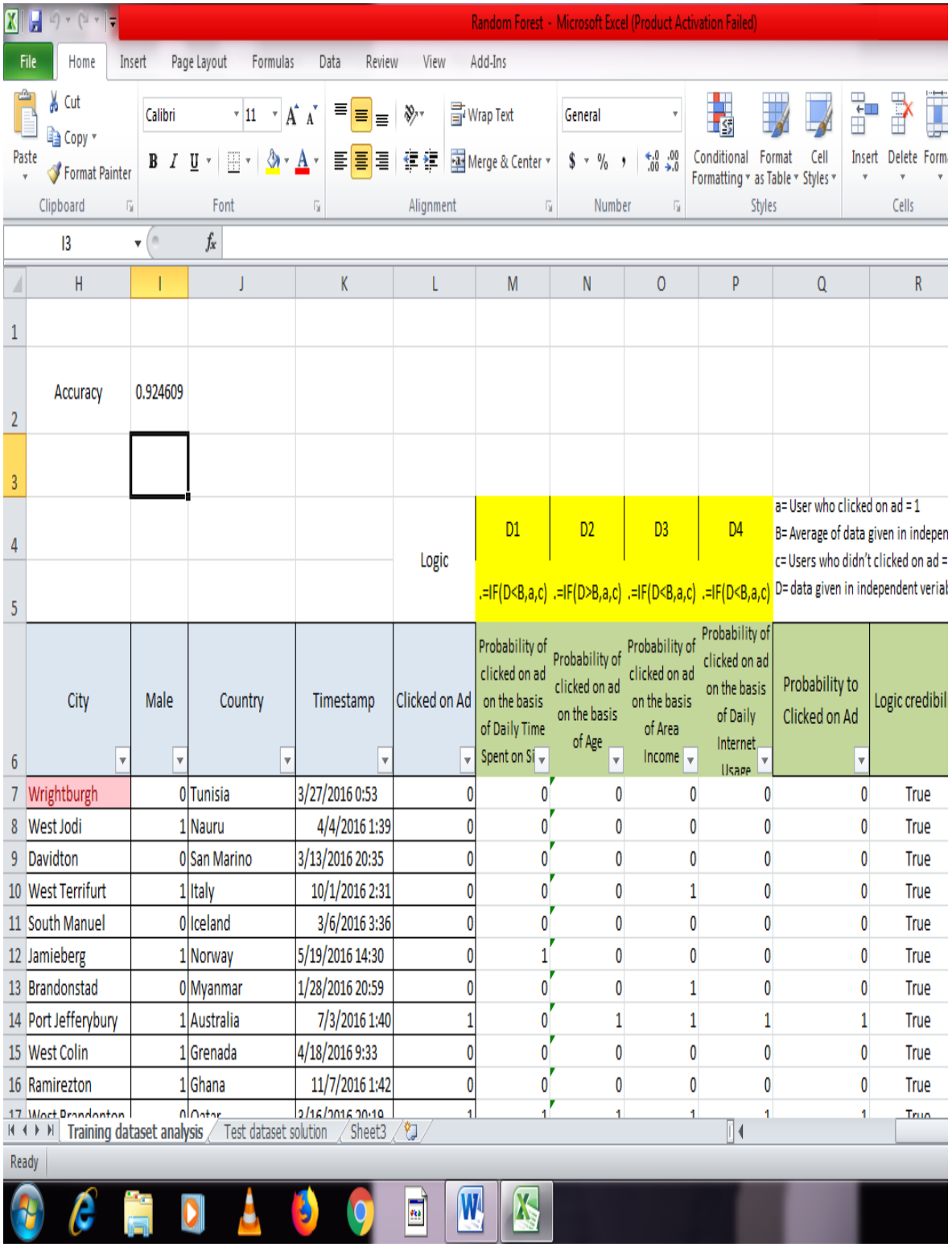


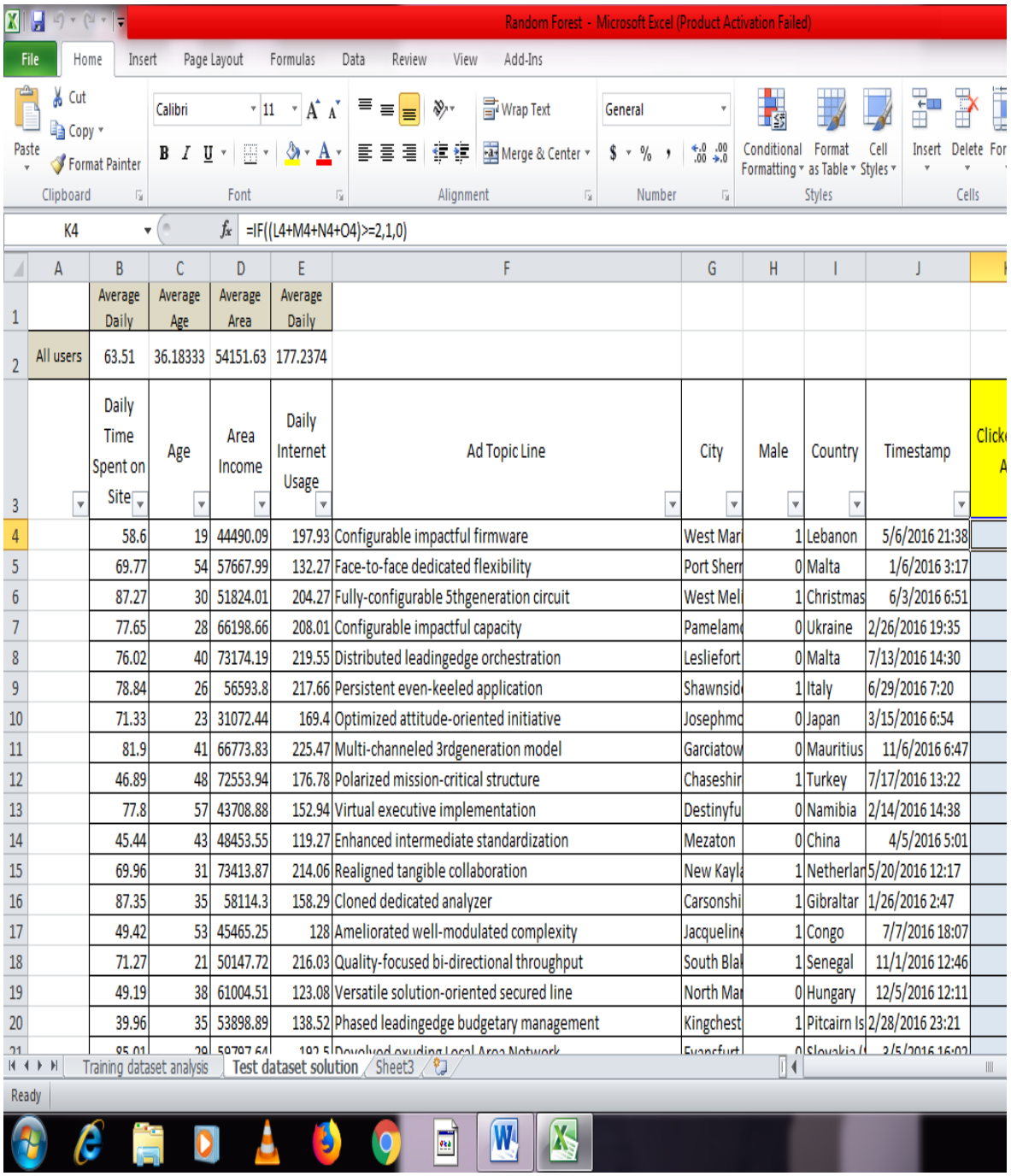
Chart Tools
logistics regression - Microsoft Excel (Product Activation Failed)
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Design
Layout
Format
Chart 1
Covariance Matrix
5.51E+12 -5.7E+10 1.12E+09 -3017119 -1.4E+09
-5.7E+10 5.87E+08 -1.4E+07 30818.34 13246806
1.12E+09 -1.4E+07 6138025 -809.334 -280630
-3017119 30818.34 -809.334 3.682358 401.4757
-1.4E+09 13246806 -280630 401.4757 1137115
Converge
-2.2E+14
2.28E+12
-5.1E+10
1.18E+08
5.37E+10
Suc-Pred
Fail-Pred
Accuracy
Cutoff
0.5
Suc-Obs
Fail-Obs
154 143 297
0 2 2
154 145 299
1 0.013793 0.521739
p-Pred
Failure
Success
Fail-Cum
ROC Curve
True Positive Rate
False Positive Rate
Sheet1
Training Dataset
Test DataSet for Click Predicti
Ready

2. Random Forest:-

a. Training Dataset-

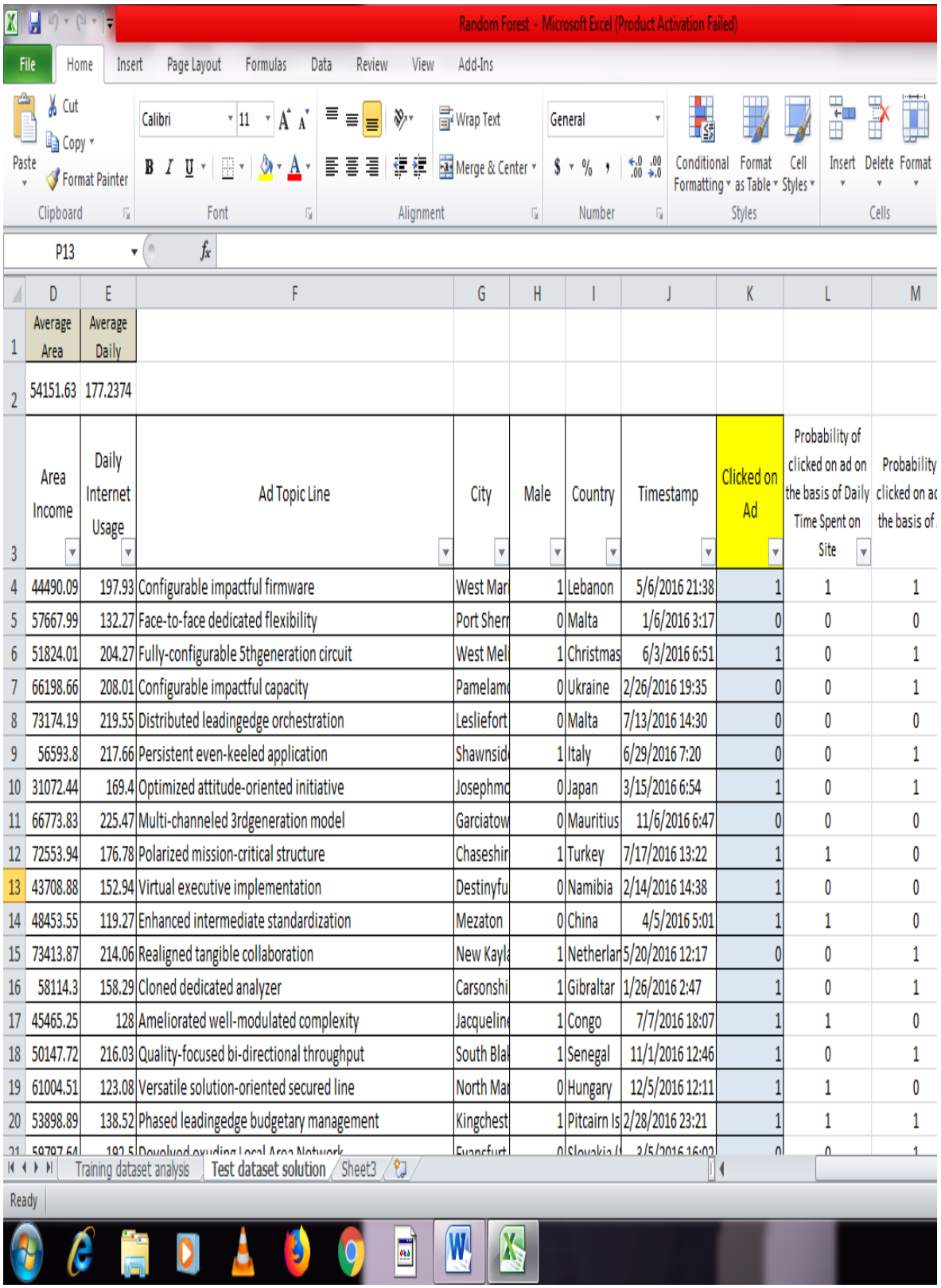


b. Test Dataset-



		Average Daily	Average Age	Average Area	Average Daily
2	All users	63.51	36.18333	54151.63	177.2374

	Daily Time Spent on Site	Age	Area Income	Daily Internet Usage	Ad Topic Line	City	Male	Country	Timestamp	Click A
4	58.6	19	44490.09	197.93	Configurable impactful firmware	West Mari	1	Lebanon	5/6/2016 21:38	
5	69.77	54	57667.99	132.27	Face-to-face dedicated flexibility	Port Sherr	0	Malta	1/6/2016 3:17	
6	87.27	30	51824.01	204.27	Fully-configurable 5thgeneration circuit	West Meli	1	Christmas	6/3/2016 6:51	
7	77.65	28	66198.66	208.01	Configurable impactful capacity	Pamelam	0	Ukraine	2/26/2016 19:35	
8	76.02	40	73174.19	219.55	Distributed leadingedge orchestration	Lesliefort	0	Malta	7/13/2016 14:30	
9	78.84	26	56593.8	217.66	Persistent even-keeled application	Shawnsid	1	Italy	6/29/2016 7:20	
10	71.33	23	31072.44	169.4	Optimized attitude-oriented initiative	Josephmo	0	Japan	3/15/2016 6:54	
11	81.9	41	66773.83	225.47	Multi-channeled 3rdgeneration model	Garciatow	0	Mauritius	11/6/2016 6:47	
12	46.89	48	72553.94	176.78	Polarized mission-critical structure	Chaseshir	1	Turkey	7/17/2016 13:22	
13	77.8	57	43708.88	152.94	Virtual executive implementation	Destinyfu	0	Namibia	2/14/2016 14:38	
14	45.44	43	48453.55	119.27	Enhanced intermediate standardization	Mezaton	0	China	4/5/2016 5:01	
15	69.96	31	73413.87	214.06	Realigned tangible collaboration	New Kayl	1	Netherlar	5/20/2016 12:17	
16	87.35	35	58114.3	158.29	Cloned dedicated analyzer	Carsonshi	1	Gibraltar	1/26/2016 2:47	
17	49.42	53	45465.25	128	Ameliorated well-modulated complexity	Jacqueline	1	Congo	7/7/2016 18:07	
18	71.27	21	50147.72	216.03	Quality-focused bi-directional throughput	South Blak	1	Senegal	11/1/2016 12:46	
19	49.19	38	61004.51	123.08	Versatile solution-oriented secured line	North Mar	0	Hungary	12/5/2016 12:11	
20	39.96	35	53898.89	138.52	Phased leadingedge budgetary management	Kingchest	1	Pitcairn Is	2/28/2016 23:21	



	D	E	F	G	H	I	J	K	L	M
1	Average Area	Average Daily								
2	54151.63	177.2374								
3	Area Income	Daily Internet Usage	Ad Topic Line	City	Male	Country	Timestamp	Clicked on Ad	Probability of clicked on ad on the basis of Daily Time Spent on Site	Probability clicked on ac the basis of
4	44490.09	197.93	Configurable impactful firmware	West Mar	1	Lebanon	5/6/2016 21:38	1	1	1
5	57667.99	132.27	Face-to-face dedicated flexibility	Port Sherr	0	Malta	1/6/2016 3:17	0	0	0
6	51824.01	204.27	Fully-configurable 5thgeneration circuit	West Mel	1	Christmas	6/3/2016 6:51	1	0	1
7	66198.66	208.01	Configurable impactful capacity	Pamelam	0	Ukraine	2/26/2016 19:35	0	0	1
8	73174.19	219.55	Distributed leadingedge orchestration	Leslieforl	0	Malta	7/13/2016 14:30	0	0	0
9	56593.8	217.66	Persistent even-keeled application	Shawnsid	1	Italy	6/29/2016 7:20	0	0	1
10	31072.44	169.4	Optimized attitude-oriented initiative	Josephmc	0	Japan	3/15/2016 6:54	1	0	1
11	66773.83	225.47	Multi-channeled 3rdgeneration model	Garciatow	0	Mauritius	11/6/2016 6:47	0	0	0
12	72553.94	176.78	Polarized mission-critical structure	Chaseshir	1	Turkey	7/17/2016 13:22	1	1	0
13	43708.88	152.94	Virtual executive implementation	Destinyfu	0	Namibia	2/14/2016 14:38	1	0	0
14	48453.55	119.27	Enhanced intermediate standardization	Mezaton	0	China	4/5/2016 5:01	1	1	0
15	73413.87	214.06	Realigned tangible collaboration	New Kayl	1	Netherlan	5/20/2016 12:17	0	0	1
16	58114.3	158.29	Cloned dedicated analyzer	Carsonshi	1	Gibraltar	1/26/2016 2:47	1	0	1
17	45465.25	128	Ameliorated well-modulated complexity	Jacqueline	1	Congo	7/7/2016 18:07	1	1	0
18	50147.72	216.03	Quality-focused bi-directional throughput	South Blak	1	Senegal	11/1/2016 12:46	1	0	1
19	61004.51	123.08	Versatile solution-oriented secured line	North Mar	0	Hungary	12/5/2016 12:11	1	1	0
20	53898.89	138.52	Phased leadingedge budgetary management	Kingchest	1	Pitcairn Is	2/28/2016 23:21	1	1	1
21	59797.64	192.5	Developed exuding Local Area Network	Evansfurt	0	Slovakia (	3/5/2016 16:02	0	0	1

FINDINGS AND RECOMMENDATIONS

Accuracy:-

1. Using IBM SPSS-

    a. Logistic Regression

    Training Dataset - 98.1%

    Test Dataset - 55.7%

    Threshold Value - 0.5

    b. Random Forest

    Training Dataset - 89.7%

    Test Dataset - 51.3%

    Growing Method - CHAID

2. Using Excel-

    a. Logistic Regression

    Training Dataset - 98.1%

    Test Dataset - 52.1%

    Threshold Value - 0.5

b. Random Forest

Training Dataset - 92.4%

Growing Method - Decision Tree

Random Forest:-

- It can use for both Regression and classification
- Commonly used predictive modeling and M.L. techniques
- In this we select some most appropriate variables to make algorithm more intelligent.

Random Forest Algorithm:-

1. Randomly select m features from T (m<<T)
2. For node d calculate the best split point among m feature.
3. Split the node into two daughter nodes using the best split
4. Repeat first three steps until n number of nodes has been reached.
5. Build your forest by repeating steps 1-4 for D number of times.

T – Number of features

D – Number of trees to be constructed

V – Output

Features:-

1. Most Accurate – Because using of number of trees parallel.
2. Works for both classification and regression.
3. Runs efficiently on large datasets.
4. Requires almost no input preparation.
5. Can be easily grown in parallel.

Logistic Regression:-

It produces results in binary format which is used to predict outcome of a categorical dependent variable.
Outputs – 0 or 1
True or False
High or Low
We use sigmoid curve to make curve continuous.

Note – Here we are getting more accuracy by Logistic Regression. So I would like to suggest logistic regression is more useful for categorical data.

LIMITATIONS OF THE STUDY

Logistic Regression

- Main limitation of Logistic Regression is the **assumption of linearity** between the dependent variable and the independent variables. In the real world, the data is rarely linearly separable. Most of the time data would be a jumbled mess.

- If the numbers of observations are lesser than the number of features, Logistic Regression should not be used; otherwise it may lead to over fitting.

- Logistic Regression can only be **used to predict discrete functions**. Therefore, the dependent variable of Logistic Regression is restricted to the discrete number set. This restriction itself is problematic, as it is prohibitive to the prediction of continuous data.

Random Forest

- The main disadvantage of Random forests is their complexity. They are much harder and time-consuming to construct than decision trees.

- They also require more computational resources and are also less intuitive. When you have a large collection of decision trees it is hard to have an intuitive grasp of the relationship existing in the input data.

- In addition, the prediction process using random forests is time-consuming than other algorithms.

BIBLIOGRAPHY

1.Goodman S. A dirty dozen: twelve P-value misconceptions. Semin Hematol. 2008.

2. Banerjee A, Chitnis UB, Jadhav SL, Bhawalkar JS, Chaudhury S.□Hypothesis testing, type I and type II errors.

3. Kuhn M, et al. Caret: classifcation and regression training, 2016.

4. Pedregosa F, et al. Scikit-learn: machine learning in Python. J Mach Learn.

5. Anaconda distribution: The most popular Python/R data science distribution. Available: https://www.anaconda.com/distribution/.

6. R: The R Project for Statistical Computing. Available: https://www.r-project.org/.

7. RStudio: Open source and enterprise-ready professional software for R, RStudio. Available: https://www.rstudio.com/products/rstudio/.
F. J. W. M. Dankers et al.

8. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning– with applications in R. 1st ed. New York: Springer; 2013.

9. Python: The offcial home of the Python Programming Language. Available: https://www.python.org/.

10. Spyder: The Scientific PYthon Development Environment. Available: https://pythonhosted.org/spyder/index.html.

11. Jupyter: Open-source web application for live coding, data visualizations, numerical simulation, statistical modeling and more. Available: http://jupyter.org/.

12. Müller A, Guido S. Introduction to machine learning with Python. Sebastopol: O'Reilly Media; 2016.

13. Matlab: The easiest and most productive software environment for engineers and scientists. Available: https://www.mathworks.com/products/matlab.html.

14. Murphy P.Machine learning: a probabilistic perspective. Cambridge: The MIT Press; 2012.

15. SPSS: The world's leading statistical software used to solve business and research problems by means of ad-hoc analysis, hypothesis testing, geospatial analysis and predictive analytics. Available: https://www.ibm.com/analytics/spss-statistics-software.

16. George D, Mallery P.IBM SPSS statistics 23 step by step: Pearson Education; 2016.

17. SAS: SAS/STAT State-of-the-art statistical analysis software for making sound decisions. Available: https://www.sas.com/en_us/software/stat.html.

18. SAS/STAT® 13.1 User's Guide. SAS Institute Inc, 2013.

19. Orange: Open source machine learning and data visualization for novice and expert. Available: https://orange.biolab.si/.

20. Weka: Data mining software in Java. Available: https://www.cs.waikato.ac.nz/ml/index.html.

21. Witten I, Frank E, Hall M, Pal C. Data mining: practical machine learning tools and techniques. Burlington: Morgan Kaufmann; 2016.

22. RapidMiner Studio: Visual workflow designer for data scientists.Available: https://rapidminer.com/products/studio/.

23. Efron B.Logistic regression, survival analysis, and the Kaplan-Meier curve. J Am Stat Assoc.

24. Walters SJ. Analyzing time to event outcomes with a Cox regression model. Wiley Interdiscip
Rev Computer Stat.

25. van der Maaten LJP, Postma EO, van den Herik HJ.Dimensionality reduction: a comparative review. Tilburg University Technical Report TiCC TR 2009-005; 2009.

26. Vrieze SI.Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Psychol Methods. 2012.

27. Dash M, Liu H.Feature selection for classi□cation. Intell Data Anal. 1997.

28. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B Methodol.

29. Zou H, Hastie T.□Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005.

30. Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett. 2006.

31. Steyerberg EW. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology. 2010.

32. Moons KGM. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012.

33. Steyerberg EW, Vergouwe Y.Towards better clinical prediction models: seven steps for devel-
opment and an ABCD for validation. Eur Heart J. 2014.

34. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models.

35. Xu Q-S, Liang Y-Z. Monte Carlo cross validation. Chemom Intell Lab Syst. 2001

36. Adams, M.N.: Perspectives on Data Mining. International Journal of Market Research 52(1), 11–
19 (2010)CrossRefGoogle Scholar

37. Asur, S., Huberman, B.A.: Predicting the Future with Social Media. In: ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 1, pp. 492–499 (2010)Google Scholar

38. Bakshi, K.: Considerations for Big Data: Architecture and Approaches. In: Proceedings of the IEEE Aerospace Conference, pp. 1–7 (2012)Google Scholar

39. Cebr: Data equity, Unlocking the value of big data. in: SAS Reports, pp. 1–44 (2012)Google Scholar

40. Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J.M., Welton, C.: MAD Skills: New Analysis Practices for Big Data. Proceedings of the ACM VLDB Endowment 2(2), 1481–1492 (2009)CrossRefGoogle Scholar

41. Cuzzocrea, A., Song, I., Davis, K.C.: Analytics over Large-Scale Multidimensional Data: The Big Data Revolution! In: Proceedings of the ACM International Workshop on Data Warehousing and OLAP, pp. 101–104 (2011)Google Scholar

42. Economist Intelligence Unit: The Deciding Factor: Big Data & Decision Making. In: Capgemini Reports, pp. 1–24 (2012)Google Scholar

43. Elgendy, N.: Big Data Analytics in Support of the Decision Making Process. MSc Thesis, German University in Cairo, p. 164 (2013)Google Scholar

44. EMC: Data Science and Big Data Analytics. In: EMC Education Services, pp. 1–508 (2012)Google Scholar

45. He, Y., Lee, R., Huai, Y., Shao, Z., Jain, N., Zhang, X., Xu, Z.: RCFile: A Fast and Space-efficient Data Placement Structure in MapReduce-based Warehouse Systems. In: IEEE International Conference on Data Engineering (ICDE), pp. 1199–1208 (2011)Google Scholar

46. Herodotou, H., Lim, H., Luo, G., Borisov, N., Dong, L., Cetin, F.B., Babu, S.: Starfish: A

Self-tuning System for Big Data Analytics. In: Proceedings of the Conference on Innovative Data Systems Research, pp. 261–272 (2011)Google Scholar

47. Kubick, W.R.: Big Data, Information and Meaning. In: Clinical Trial Insights, pp. 26–28 (2012)Google Scholar

48. Lee, R., Luo, T., Huai, Y., Wang, F., He, Y., Zhang, X.: Ysmart: Yet Another SQL-to-MapReduce Translator. In: IEEE International Conference on Distributed Computing Systems (ICDCS), pp. 25–36 (2011)Google Scholar

49. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: Big Data: The Next Frontier for Innovation, Competition, and Productivity. In: McKinsey Global Institute Reports, pp. 1–156 (2011)Google Scholar

50. Mouthami, K., Devi, K.N., Bhaskaran, V.M.: Sentiment Analysis and Classification Based on Textual Reviews. In: International Conference on Information Communication and Embedded Systems (ICICES), pp. 271–276 (2013)Google Scholar

51. Plattner, H., Zeier, A.: In-Memory Data Management: An Inflection Point for Enterprise Applications. Springer, Heidelberg (2011)CrossRefGoogle Scholar

52. Russom, P.: Big Data Analytics. In: TDWI Best Practices Report, pp. 1–40 (2011)Google Scholar

53. Sanchez, D., Martin-Bautista, M.J., Blanco, I., Torre, C.: Text Knowledge Mining: An Alternative to Text Data Mining. In: IEEE International Conference on Data Mining Workshops, pp. 664–672 (2008)Google Scholar

54. Serrat, O.: Social Network Analysis. Knowledge Network Solutions 28, (2009)Google Scholar