# **Website Ad-Click Assignment**

R Final Project

Logistic Regression

Here, in this doc, we will focus on the approaches, steps, results found and their interpretation, significance of the variables and their business meaning etc.

1. Install required package and calling the same library at the beginning
2. Set the directory where my data is kept
3. Read the data for this project and store it in a variable i.e. DataFrame (taken a replica of the same)
4. Did basic exploration like

Structure: 6657 obs. of 14 variables. Displayed all columns and its datatypes followed by some starting values of each one.

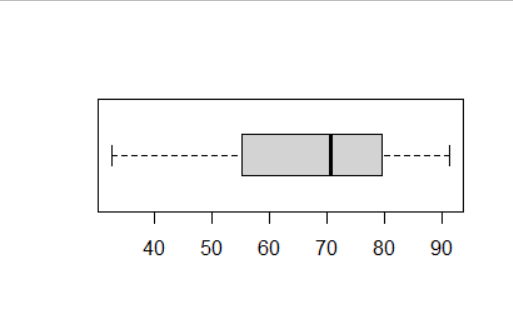
Summary: Obtained mean, median, min, max, 1st and 3rd quartile value for all continuous columns.

Dimension: 6657 14

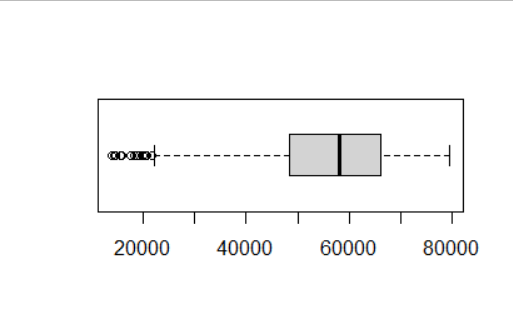
1. Changed some variables as factor variable such as City\_code, Male, Time\_Period, Weekday, Month, Year, and Clicked.
2. Moved towards missing value treatment. As per the code and data, we have not found any missing value here.
3. Then, moved towards Outlier checking for continuous variable.

Boxplot is given for Time\_Spent, Avg\_Income, and Internet\_Usage.

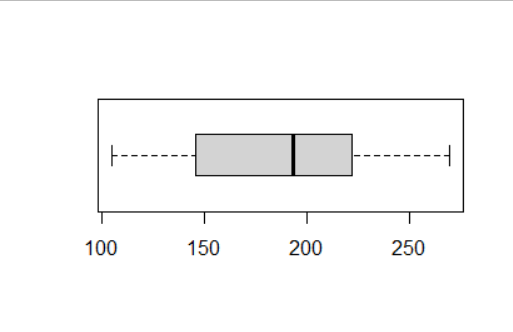
Time\_Spent



Avg\_Income



Internet\_Usage

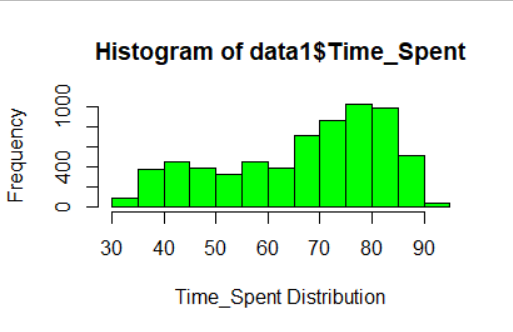


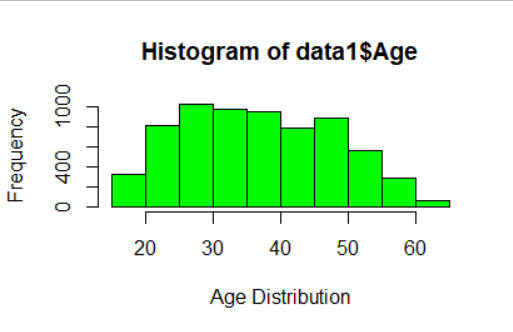
Have seen some outliers in Avg\_income

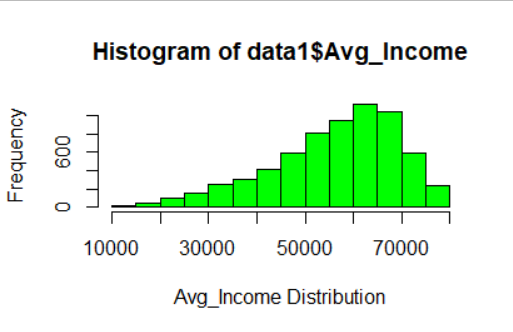
Checked with quantile method

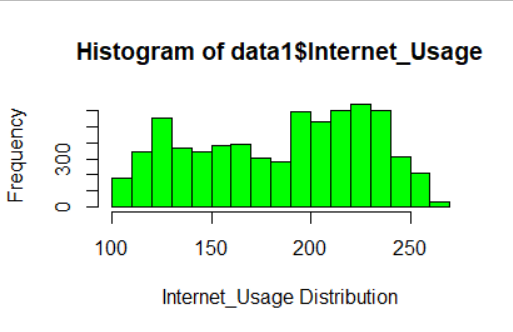
We have seen that there is not a major outlier in the Avg\_income column. Hence, ignored the outlier part (means keeping all good).

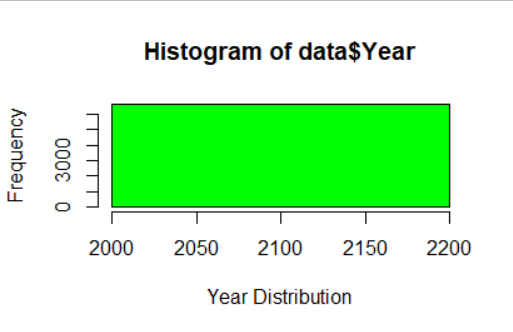
1. Univariate analysis



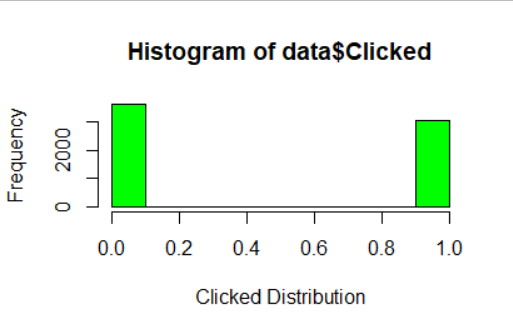




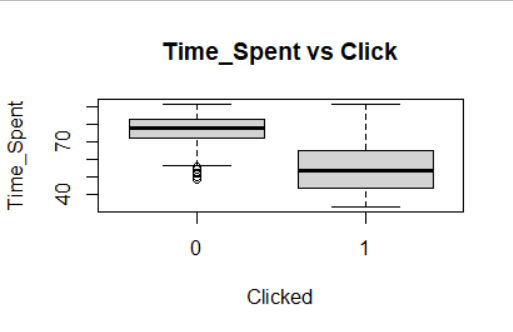


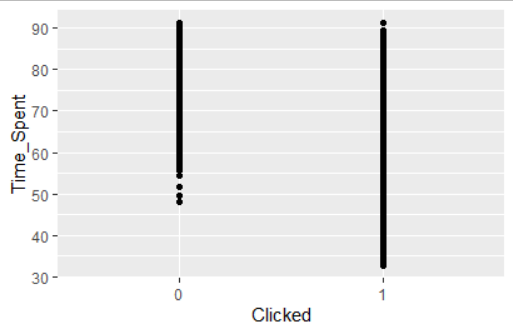


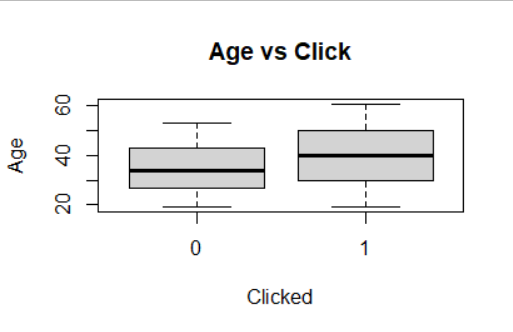
**Target variable**

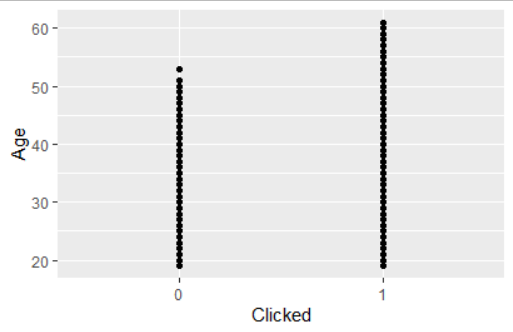


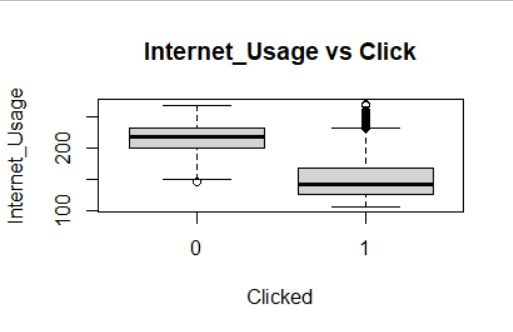
1. Bivariate analysis

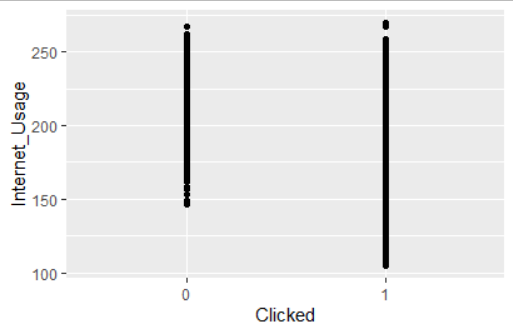


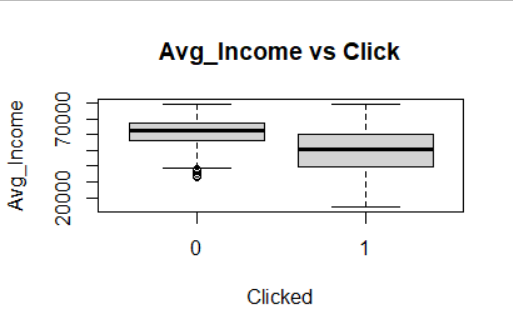


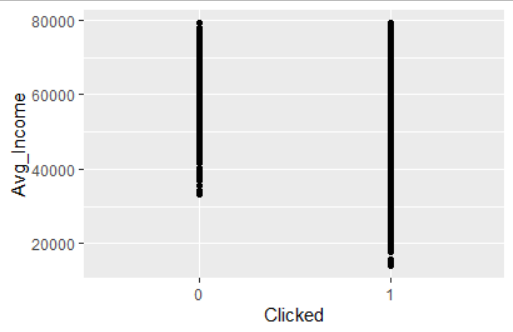












1. Feature selection:
2. Information Value (IV) Calculation (A variable reduction technique)

Creating two data sets for numeric and categorical values

Created num and cat: two DataFrame

Written two separate function for num and cat IV Calculation

Applied them on the respected variables

It is basically checking the power of independent variable to predict the dependent variable.

Result is:

|  |  |
| --- | --- |
| Variable | IV |
| Time\_Spent | 2.521434 |
| Age | 0.383866 |
| Avg\_Income | 1.198425 |
| Internet\_Usage | 2.197992 |
| Ad\_Topic | 0.13425 |
| Country\_Name | 0.260874 |
| City\_code | 0.270947 |
| Male | 0.003032 |
| Time\_Period | 0.144382 |
| Weekday | 0.002215 |
| Month | 0.003634 |
| Year | 0 |

#threshold value is >.01

Those columns that has the IV of more than 0.1, have high predicted power

1. ANOVA Test

Same like IV but it deals with probability significance to predict the target variable

It is only deals with categorical variables. Those are given below

Ad\_Topic: <2e-16 – Significant

Country\_Name: <2e-16 – Significant

City\_code: <2e-16 – Significant

Male: 0.025 – Significant

Time\_Period: <2e-16 – Significant

Weekday: 0.723 – Not Significant

Month: 0.423 – Not Significant

Year: Null – Not Significant

Result:

Based on ANOVA test and information value, final DataFrame will be created

It will be used for final modeling

Hence, removing VistID, weekday, month and year based on the test results.

Columns for final data will be Time\_Spent, Age, Avg\_Income, Internet\_Usage, Ad\_Topic, Country\_Name, City\_code, Male, Time\_Period, and Clicked.

1. Train test Split

Did splitting on the final data based on Clicked variable with 70 and 30% ratio.

The variable name for training and testing is data.train and data.test.

Dimensions are 4660 10 and 1997 10

1. Logistic Regression Model Building

Took all variables for the first time and one by one removing the insignificant variables based on p values.

Significant predictors are Time\_Spent, Age, Avg\_Income, Internet\_Usage, Ad\_Topic (product\_3, product\_8, product\_21, and product\_28), City\_code (City\_2, City\_3, City\_4, City\_5, City\_6, and City\_7), and Time\_Period (Mid-Night and Morning). All variables are significant at least 90% significance.

1. Multicollinearity check. Checked vif of the model and got values under 1.8 of each variable, which is the very good threshold of it.
2. Wald Test. Got p value 0.0. Since, p-value is less then 0.001, hence we reject Ho that the all Bi=0
3. Lagrange Multiplier or Score Test. Difference between null deviance and deviance is 4410.85. The degree of freedom for Null model and model with variables is 16. Chi-square probability is 0, which is very good.
4. Predictions. Model Accuracy is 93%. R2 value is 78%. All good.
5. Accuracy measures. Confusion matrix:

Actual

Predicted 0 1

0 1049 87

1 37 824

AUC matrix:

Metric Values

threshold 0.5365780

specificity 0.9045005

sensitivity 0.9659300

AUC 0.9653539

AccuracyRate 0.9379069

Gini 0.9307078

Other values are:

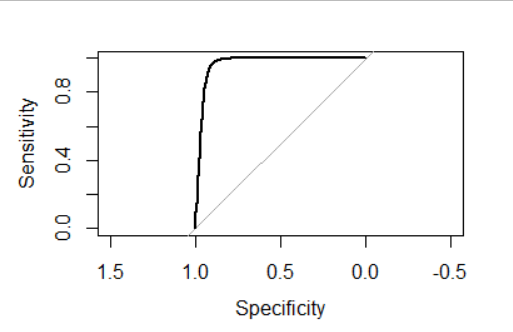
Precision value of the model: 0.95

Accuracy of the model: 0.93

Recall value of the model: 0.04

f1 score of the model: 0.08

rocCurve:



1. Important variables for the model

Among the selected variables or predictors, top three important variables are Internet\_Usage, Age, and Avg\_Income.

1. Interpretable coefficients of the final model is given below

Time\_Spent 8.887787e-01

Age 1.038061e+00

Avg\_Income 9.999411e-01

Internet\_Usage 9.580505e-01

Ad\_Topic == product\_3 6.722467e+00

Ad\_Topic == product\_8 1.956230e+00

Ad\_Topic == product\_21 2.678461e+00

Ad\_Topic == product\_28 1.749468e+00

City\_code == City\_2 6.918985e-01

City\_code == City\_3 4.568406e-01

City\_code == City\_4 4.489765e-01

City\_code == City\_5 4.490443e-01

City\_code == City\_6 2.753256e-01

City\_code == City\_7 2.538825e-01

Time\_Period == Mid-Night 6.336812e-01

Time\_Period == Morning 5.762687e-01

All coefficients have the positive impact on the model. All other interpretations are presented in PPT file.

Thanks and Regards,

Supriyo Barik