Analysis of Website Ad-Click

In this data set, we have to predict who is likely going to click on the Advertisement so it can contribute to the more revenue generation to the organisation.

- In case of a classification problem: the target variable is a categorical in nature. We will do Logistic Regression.
- Here, the target variable is "**Clicked**" which is categorical in nature.

Some basic exploratory data analysis (EDA) is done:

- Dim function returns Total dimension i.e., both the number of rows and column in a data frame.
- We can also use ncol () function to find the number of columns and nrow () to find the number of rows separately.
- Summary () function returns some basic calculations for each column like min, max, mean, median etc.

  Interpretation achieved from Summary () function: -

  1. The average age of the user is 36 years old, whilst the youngest is 19 and the oldest is 61. It is safe to conclude that the site’s target audience is adults.

  2. The percentage of males visiting the website is slightly lower than females, we can see there is a 52%:48% split in favor of women. Hence, our sample is well represented by both genders.

  3. The area income of users ranges between $13,996.50 — $79,484.80. Quite a large distribution of incomes. This tells us that site visitors hail from various social classes.

  4. The daily time spent by users on the website ranges between 32.6–91.4 minutes.
- For more deep dive statistics including standard deviation, mean absolute deviation, skew, etc.
  Describe () function is used. For this we have used Stats package called “psych”.
- Some columns are having underscores that is not appropriate for model building. To remove these, library “stringr” is used.

Nature of different columns: (Continuous / Categorical)

- Continuous- VistID, TimeSpent, Age, AvgIncome, InternetUsage, CountryName

- Categorical- Citycode, AdTopic, Male, TimePeriod, Weekday, Month, Year, Clicked
- VistID, CountryName, AdTopic are removed from the dataset as not helpful in predicition.
  6657 unique values in VistID and can't help in prediction.

Missing values interpretation:

This dataset does not contain any type of missing values.

Univariate and Bivariate Analysis:

- For Continuous Variables, use of histogram visualization is done
- For Categorical Variables, use barplot visualization is done
- For Categorical vs Continuous variables, Box Plot visualization is done
- For Categorical vs Categorical variables, Grouped Bar chart visualization is done

Statistical Tests:

1. ANOVA (Analysis of Variance)

    - For the variables TimeSpent, Age, AvgIncome, InternetUsage, the P-value is less than 0.05. This means the variables are correlated. Null hypothesis H0 is rejected. The variables are statistically significant.

2. Chi-Square Test

    - The variables (p-value < 0.05) that are found significant from Chi-square test are Male, TimePeriod, Citycode and Year. Null hypothesis H0 is rejected.
    - The variables Week and Month found with large p-value i.e., p-value > 0.05. So, we can reject these variables.

So, after doing these two statistical tests, the potential predictors are: - TimeSpent, Age, AvgIncome, InternetUsage, Male, TimePeriod, Citycode, Year.

<u>Logistic Regression model:</u>

Significant variables are achieved after building the model: -

- TimeSpent
- Age
- AvgIncome
- InternetUsage
- TimePeriod
- Citycode

Null deviance: 6882.3 -- system generated error without taking independent variable

Residual deviance: 2095.9 -- error with independent variable

AIC: 2131.9 -- adjusted r-square in logistic regression, akike information criterian

Confusion Matrix:

Confusion matrix is used to evaluate the model behaviour from a matrix. Below is how a confusion matrix looks like:

	Predicted No	Predicted Yes
Actual No	TN	FP
Actual Yes	FN	TP

TP- True Positive TN- True Negative FP- False Positive FN- False Negative

True Positive is the proportion of positives that are correctly identified. Similarly, True Negative is the proportion of negatives that are correctly identified. False Positive is the condition where we predict a result that is doesn't fulfil. Similarly, False Negative is the condition where the prediction failed when it was actually successful.

- ❖ Sensitivity / Recall = TP / (TP+FN)
- ❖ Accuracy = (TP+TN)/(TP+FP+FN+TN)
- ❖ Specificity = TN / (TN+FP)
- ❖ Precision/PPV = TP / (TP+FP)
- ❖ False Negative value = FN/(FN+TN)
- ❖ False Positive value = FP / (FP+TP)
- ❖ F1-Measures = (2*Recall*Precision)/(Recall+Precision)

Outcomes achieved from the confusion matrix are below:

- ❖ Accuracy: ------------------------ 0.9237
- ❖ Sensitivity/Recall: -------------- 0.9121
- ❖ Specificity: ------------------------0.9390
- ❖ Pos Pred Value/Precision: ---0.9514
- ❖ Neg Pred Value: ----------------0.8908
- ❖ Balanced Accuracy: ------------0.9255
- ❖ # F1-Measures/F1-score: -----0.9313356

Multicollinearity in the model:

No multicollinearity found in the logistic model as VIF value is less than 5.