

INFO6105 (SEC 10)

Data Science Engineering Methods and Tools

Spring 2020

**AD-CLICK PREDICTION SYSTEM**

By

Gnana Tej Cherukuru

Pranathi Manusanipalli

Rishvita Bhumireddy

Poojitha Muppala

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| 1. INTRODUCTION | 3 |
| 1. KEYWORDS | 3 |
| 1. BACKGROUND | 3 |
| 1. OBJECTIVES | 4 |
| 1. METHODOLOGY | 4 |
| 1. PROBLEM STATEMENT | 5 |
| 1. SPECIFICATION OF DATASET | 5 |
| 1. APPROACH | 6 |
| 1. ALGORITHMS | 13 |
| 1. RESULT AND ANALYSIS | 32 |
| 1. CONCLUSION | 33 |
| 1. FUTURE SCOPE | 33 |
| 1. REFERENCES | 34 |

**INTRODUCTION:**

Digital advertising is a very huge business which is worth more than 50 million dollars. Because of targeted advertising the revenue that the advertisers earn is increasing. A lot of analysis has been done in the previous years, from ad-click prognosis to ad serving. There are a lot of advertisement formats such as search engines, video advertisements, textual and contextual advertisements etc on which ad-click prognosis has been done. As there is a huge growth in online activities, advertisers have been using different procedures and methods to display relevant advertisements which suits the viewer’s interest. With the growing advertising network, the data which should be analyzed for ad-click prediction is very large. With the rapid increase of advertising network, click prediction needs huge data analysis. The advertisement prediction is one of the most lucrative stories in the domain of machine learning. Also, the advance in ad serving machinery have brought a real time bidding answer where ads are short listed based on the attributes of the publishers and the viewers.

**KEYWORDS:**

* Ad-click
* Prediction
* CTR
* Features
* Machine Learning
* Analysis
* Algorithm

**BACKGROUND:**

Ad click prediction is a massive-scale learning problem that is vital to the multi-billion-dollar online advertising industry. The advertisement prediction is one of the most lucrative stories in the domain of machine learning. The advances in ad serving machinery have brought a real time bidding answer where ads are short listed based on the attributes of the publishers and the viewers. Advertising outrageously to random users extensively results in waste of resources, this can also leave a negative impact on users towards the products. Applying certain methodologies for CTR analysis helps the ad serving agency for learning customer behaviors and targeting the advertisements only to users who might be interested to in viewing the ads. This helps in maximizing the profit of the advertisers and preventing superfluous advertising. only receive the useful advertisements that have a relatively high possibility to attracts them, which is an efficient and pleasant experience to see the display advertising while surfing the Internet. It is a win-win procedure for both advertisers and users, and even for website platforms. The dataset used for prediction tends to be extremely sparse. There is a considerable fraction of Null values in the data.

**OBJECTIVES:**

The objective of the project is to predict if a customer will click on an ad based on certain features present in the dataset. We are going to use different algorithms to implement CTR prediction and check which method gives us the maximum model accuracy.

In short, we are trying to analyze the following points

1. Is it possible to predict if a customer will click on an ad which is displayed?
2. Understand which algorithm has the highest Accuracy score and AUC\_ROC curve
3. Analyze the important features in the dataset which helped us in achieving the highest accuracy
4. Analyze the speed performance of the algorithms and confusion matrix

**METHODOLOGY:**

The methods include data exploration, data cleansing, attribute extraction, analyzing various techniques for ad-click prognosis and finding the most suitable model.

***Data exploration:***

Exploring the data through different plots gives you deeper understanding of data and choosing the suitable Machine Learning Model. We used count plots and heat maps for obtaining visual insights.

***Data Cleansing:***

We used heatmaps for finding columns having null values. By understanding the interdependency among various fields, we replaced the null values with most approximate digits.

***Attribute extraction:***

By performing trend analysis of various fields in the dataset with respect to ad-clicking helped us gain insights about important features to be considered for the prediction. The analysis also gave a perspective of various columns that are interdependent. After analyzing all the fields, we developed a feature matrix that contains fields most suitable for the ad-click prediction.

***Analyzing Various Techniques for ad-click prognosis:***

We will study the CTR dataset using various algorithms. Compared various classification algorithms and selected some suitable algorithms that include Logistic, Gaussian Naïve Bayes, Gradient Boosting Machine, Decision Tree, Random Forest. Because the proportion of the positive and the negative samples is extremely uneven, we proposed the Synthetic Minority Oversampling Technique.

***Finding the Most suitable model:***

Evaluating various algorithms for aspects like fast learning speed, accurate estimation results with easily setting the weights, high precision. Based on these advantages, we concluded the most suitable algorithm for prediction.

**PROBLEM STATEMENT**

As online advertising is a very huge business, it is important to practice targeted advertising. Advertising to random users would result in wasting resources and can lead to a negative impact on the product. So, we should apply certain methods to analyze the user behavior and target advertisements which suits the viewer’s interest. We have to predict if a customer will click on an ad based on certain features present in the dataset.

**SPECIFICATIONS OF DATASET:**

The task is to select the potential users that are likely to click the ads by analyzing user's clicking/web browsing information and displaying the most relevant ads to them. The dataset contains the details of the ad displayed and the user.

There are 15 attributes in the dataset and ~460000 rows. It contains 1 week of data of 2017.

The CSV dataset contains details about users like user age group, gender, date the ad was displayed, city development rating. Product details like product name, product category, product subcategory. The ad details include the webpage id, session id, campaign id

**Column Names and Description:**

session\_id: The unique identifier of the session

DateTime: The Datetime when the ad was displayed

user\_id: The unique identifier of the user

product: The unique identifier of the product

campaign\_id: The campaign ID of the Ad displayed

webpage\_id: The unique identifier of the webpage/ website the Ad was displayed

product\_category\_1: The Product Category displayed in Ad

product\_category\_2: The Product Subcategory displayed in Ad

user\_group\_id: The unique identifier of the ad group which the user belongs to

gender: The gender of the user

age\_level: The age level of the user

user\_depth: The depth at which the Ad was displayed

city\_development\_index: The measure of development of the city

var\_1: anonymous column

is\_click: Indicator weather the Ad was clicked

Dataset: <https://drive.google.com/drive/folders/1Gv0jtk73SVfXBEhjV4sX9WU_hP_w5vA3>

Columns: 15

Rows: 463,291

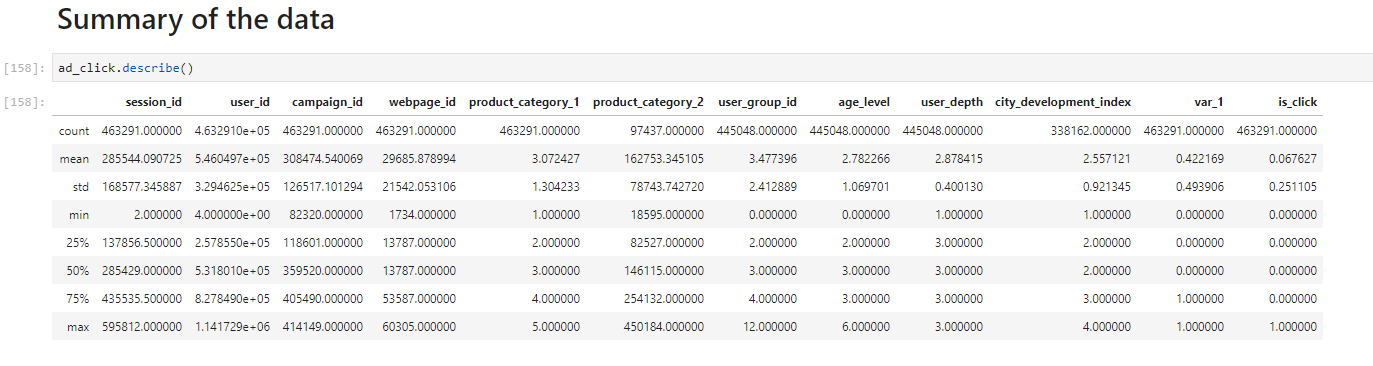
**APPROACH:**

**1. Data Exploration:**

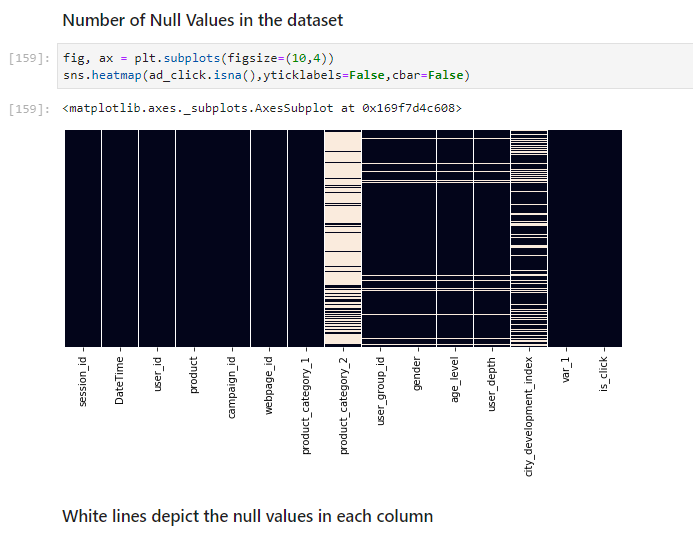
EDA refers to the critical process of performing initial investigations on data to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

We plotted different bar graphs, count plots, heat map and histograms to understand the pattern of the data and draw insights from it

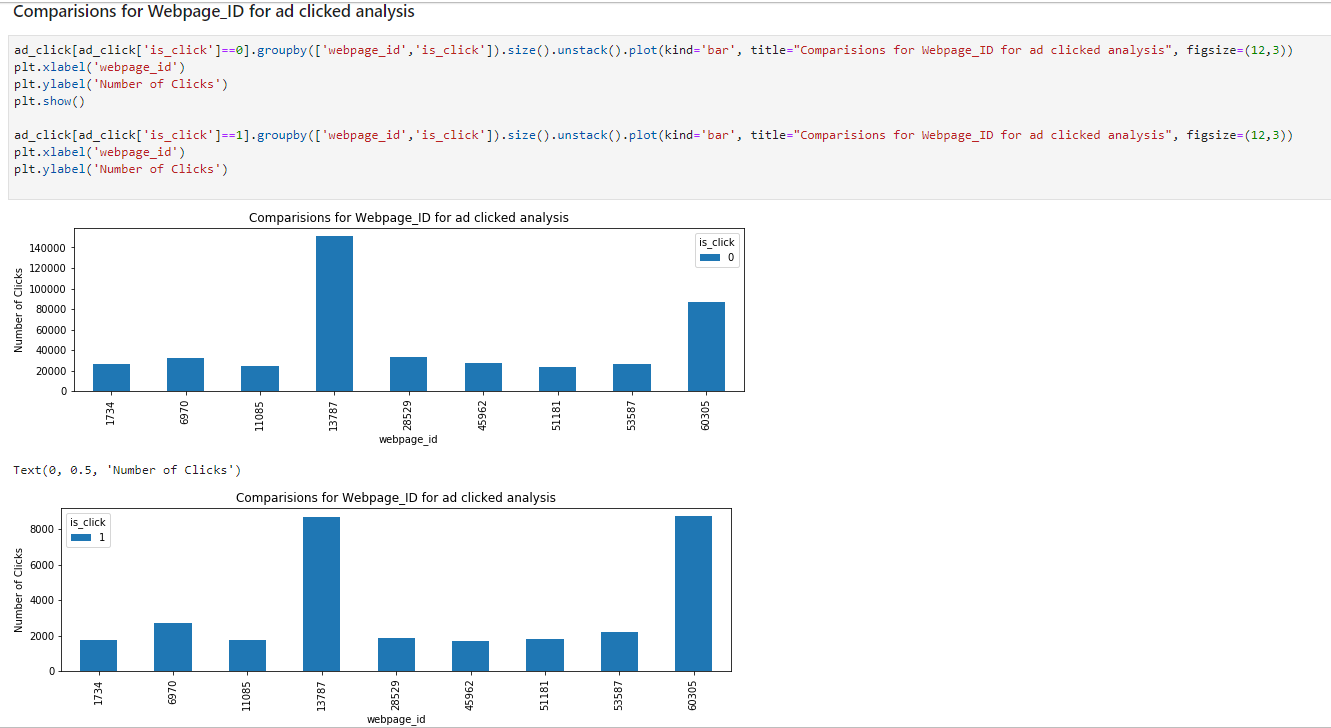
***Summary of the Data:***



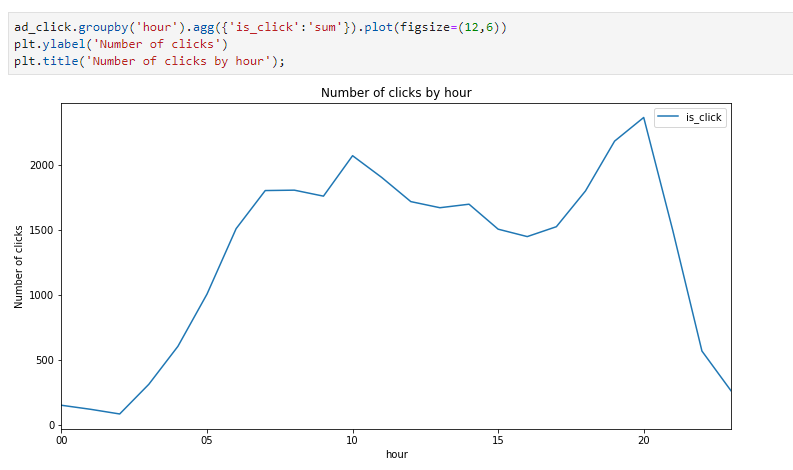
***Number of Nulls:***



***Comparison of Webpage ID and is\_click:***



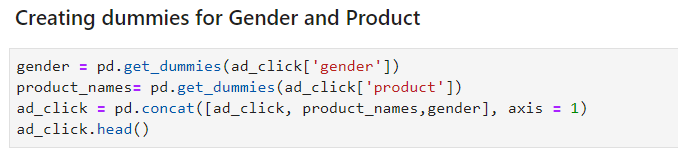
***Number of clicks by Hour:***



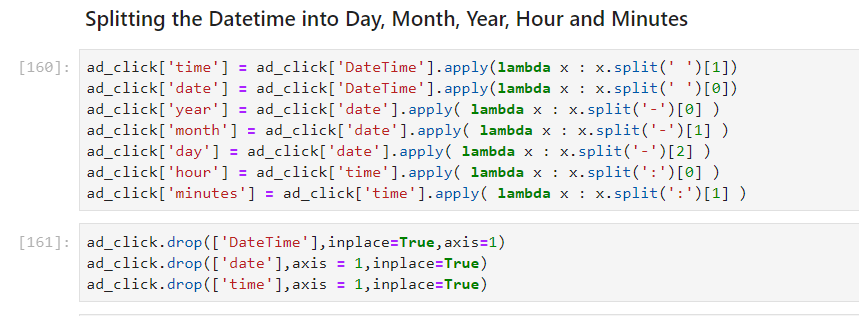
**2.Data Cleansing:**

Data cleansing is a process of preparing the data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. This data is usually not necessary or helpful when it comes to analyzing data because it may hinder the process or provide inaccurate results. It is important to do data cleaning before doing any analysis is because it improves the data quality and in doing so involves the increase of productivity.

There are categorical variables present in the data which should be converted to numerical variable to be used for regression analysis. The numerical variables created from categorical variables are called Dummy Variables. Product and Gender and converted to dummy variables.



The Datetime field present in the dataset represents the date when the ad was shown to the user. To check if the user has different behaviors at different time periods, the datetime column is split into multiple columns i.e Year, Month, Date, Hour, Minutes



***Number of Null columns present in each column of the dataset:***



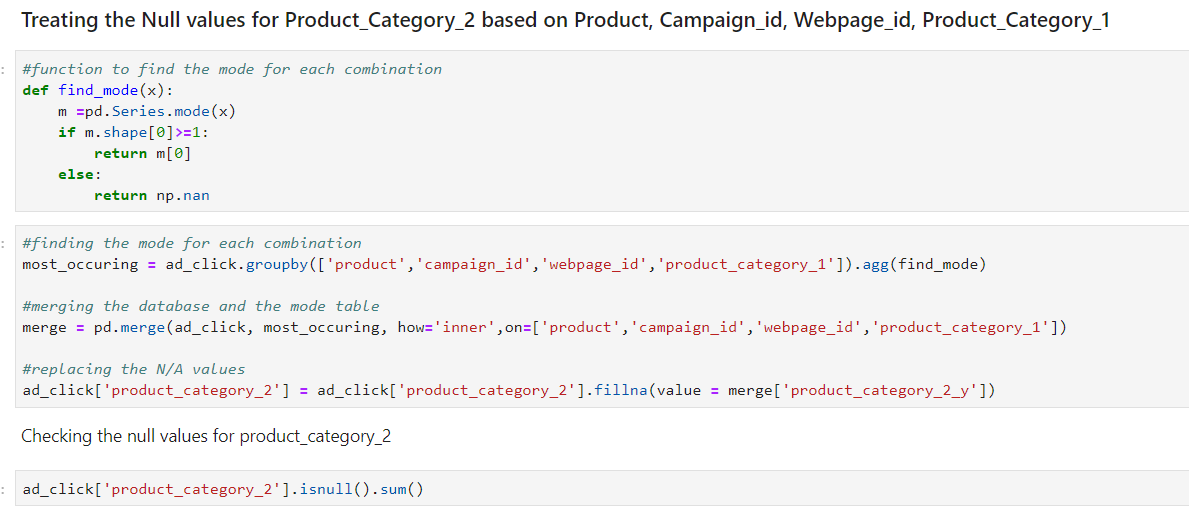
By plotting Correlation matrix we find the associations between variables:

A picture containing brick

Description automatically generated

The Null’s have been treated by

* **Product\_category\_2:** with the mode based on product, product\_category\_1, campaign\_id and webpage\_id
* **Age\_level:** with the mode based on product, product\_category\_1
* **Gender:** with the mode based on age\_level
* **User\_group\_id:** with the mode based on gender
* **User\_depth:** with the mode
* **City\_development\_index:** with the mode



**3.Attribute Extraction**

The data features that we use to train our machine learning models have a huge influence on the performance. We select those features which contribute most to the prediction variable.

We removed some irrelevant features that do not impact the prognosis. For example we removed columns like “Year” as the dataset is related to only one year (2017). We also removed columns that are unique to each row as they do not contribute for prediction (like session\_id).

**4.Analyzing Various Techniques for ad-click prognosis:**

We will use below Algorithms for prediction :

1).Logistic Regression

2). Decision Trees Classifier

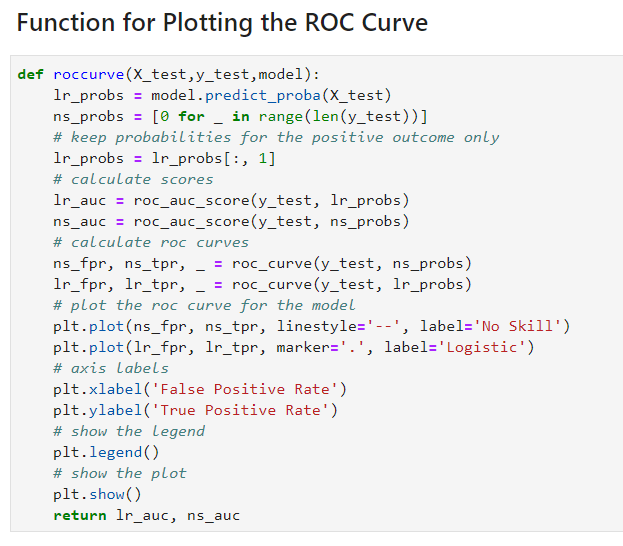
3).Gradient Boosting Machines

4).Random Forest Classifier

We used two function to plot graphs for ROC Curve and Confusion Matrix

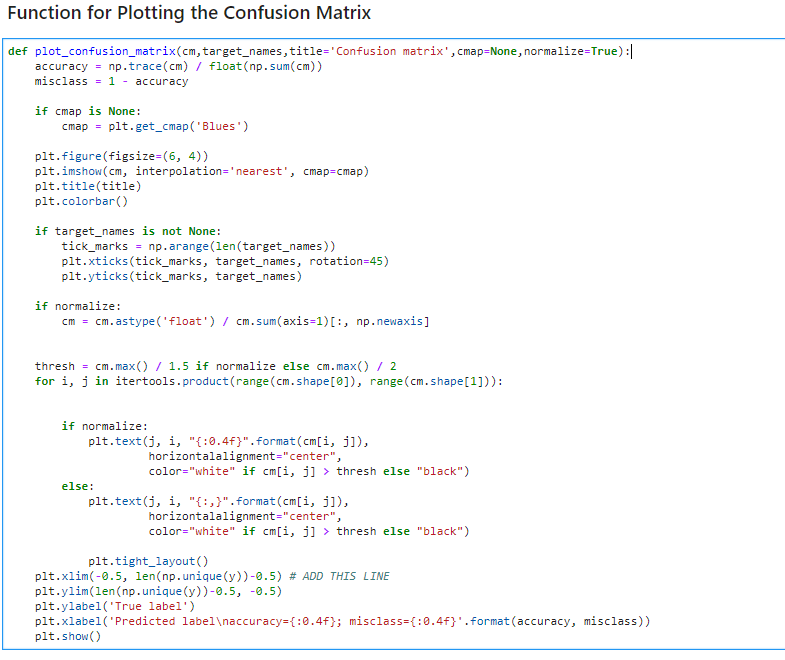
***ROC Curve:***

ROC Curves summarize the trade-off between the true positive rate and false positive rate for a predictive model using different probability thresholds.



***Confusion Matrix:***

Confusion Matrix is a table that is used to describe the performance of the model.



**ALGORITHMS:**

**LOGISTIC REGRESSION:**

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analysis, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

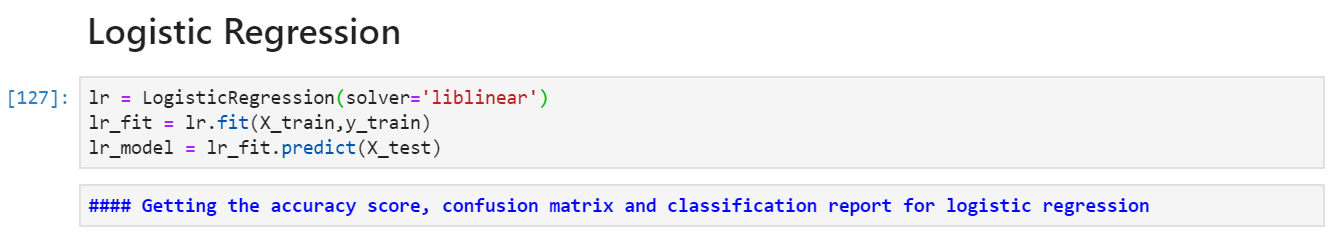
In our case since the aim is to analyses the features and determine whether the Ad is clicked or not, i.e either a ‘1’ or ‘0’, logistic regression was picked.

**Pros of Logistic Regression:**

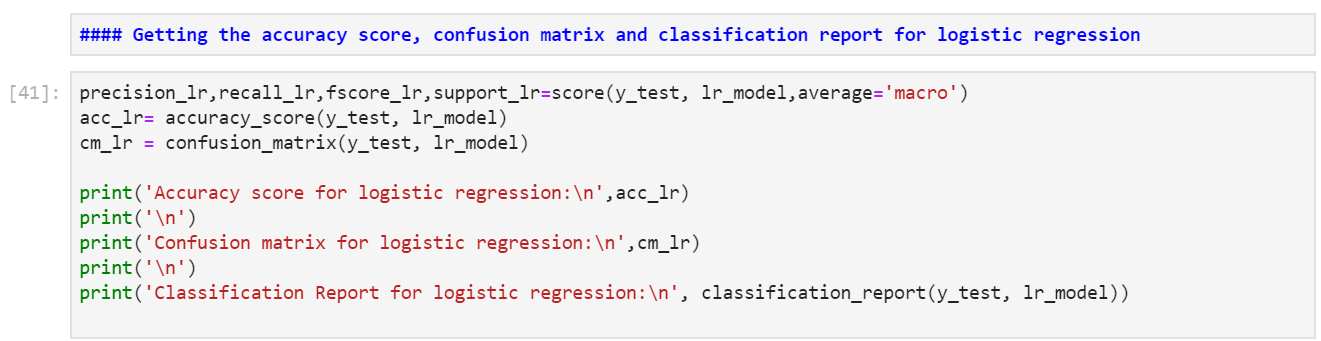
* Low Variance
* It provides Probability scores for Observations
* Easy to implement, interpret and very easy to train

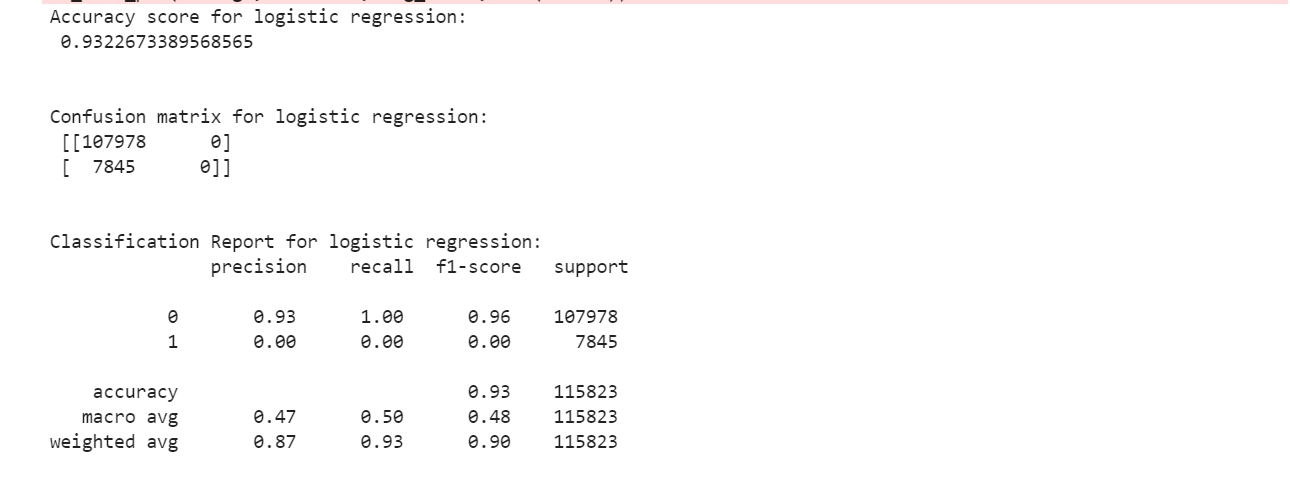
**Applying Logistic Regression:**

We trained the cleansed data using the Logistic regression and predicted the result for test data.

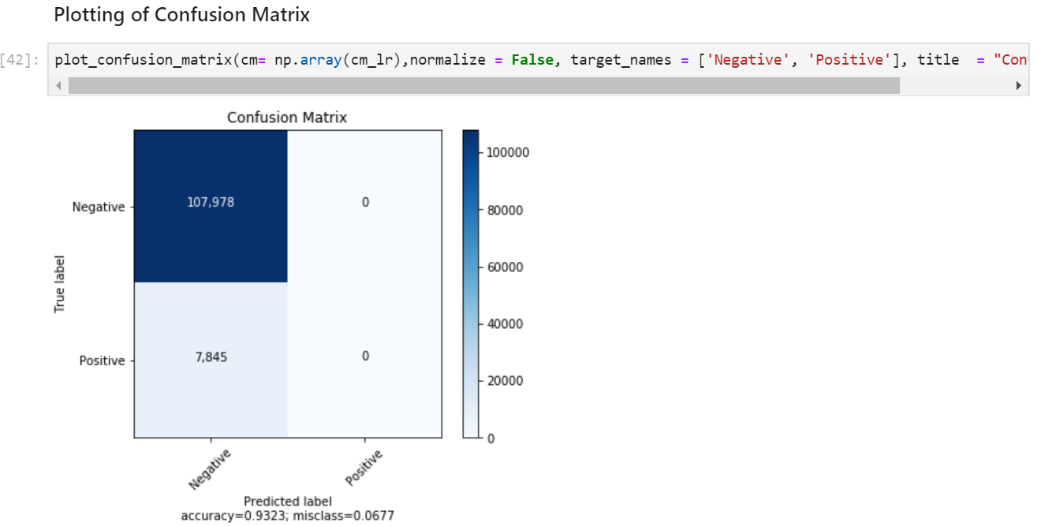


**Calculate the accuracy score, Confusion Matrix and Classification Report**





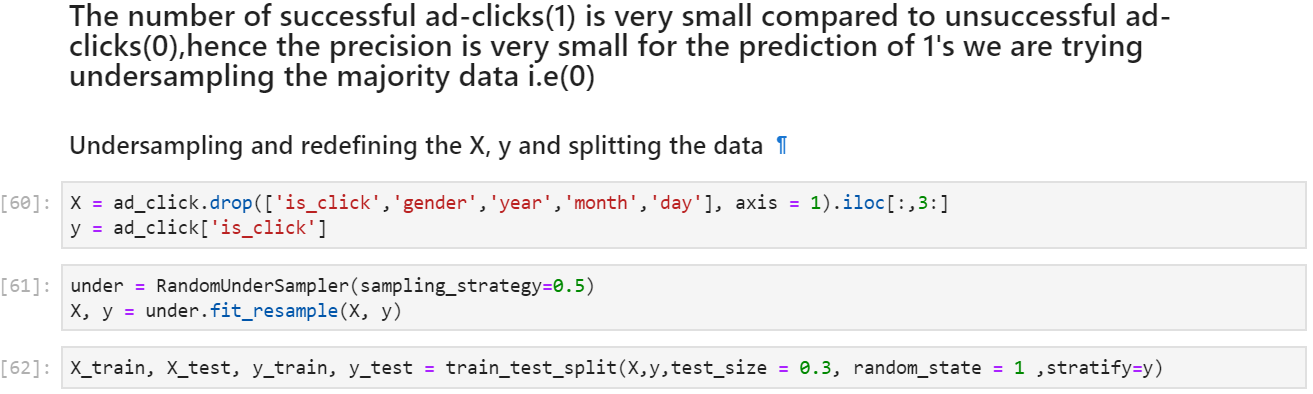
**Plotting Confusion Matrix:**



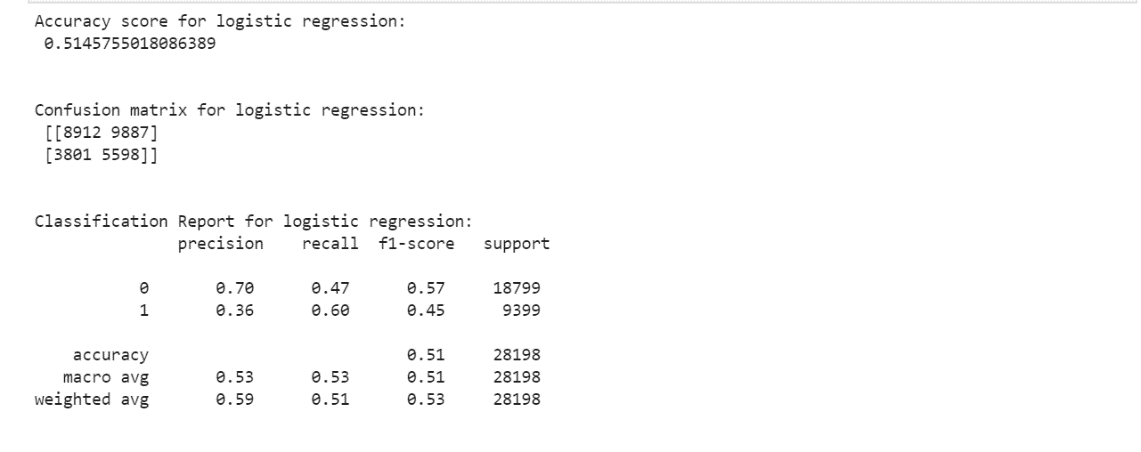
Though the accuracy is high the Recall value for ‘1’ ad-clicked is 0. Due to high imbalance in the dataset between 0’s and 1’s the model failed to predict 1’s. Hence, we proceed to do under sampling of data to achieve equality.

**Logistic Regression after Under Sampling:**

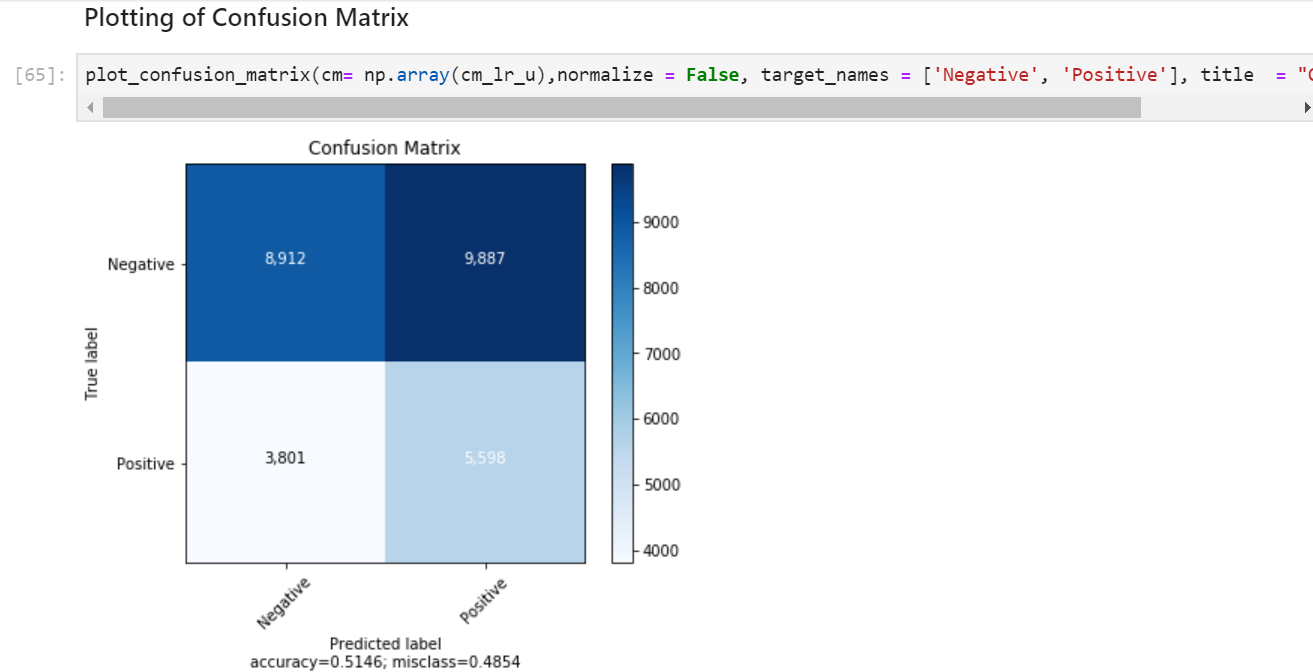
As the dataset is highly imbalanced, we performed Under Sampling on the majority set to calculate the accuracy score and plot the classification matrix.



**Calculate the accuracy score, Confusion Matrix and Classification Report**



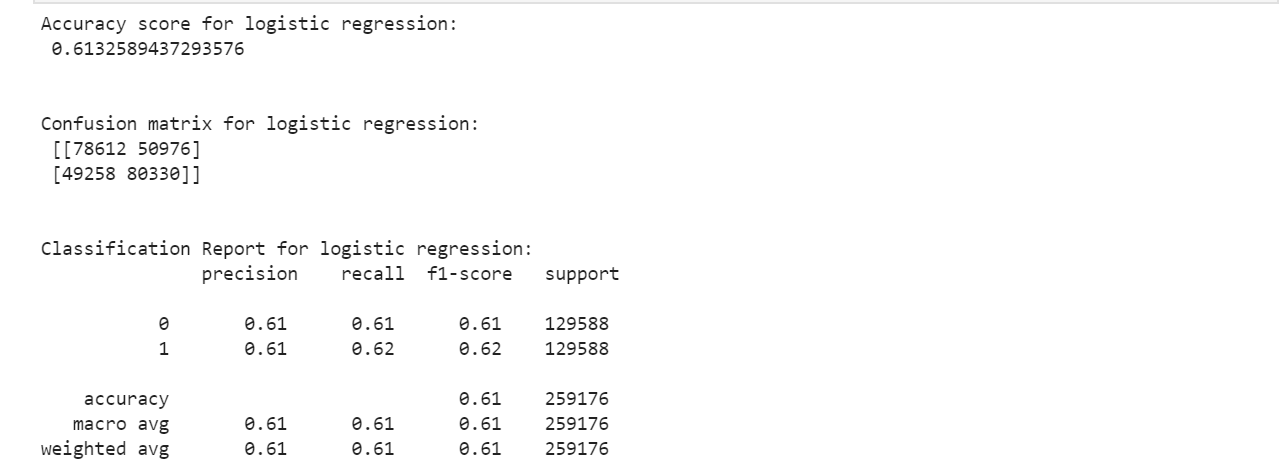
**Confusion Matrix after Under Sampling:**



As depicted the precision for predicting 1’s has increased i.e 36%. Whereas the overall accuracy stands at 51%.

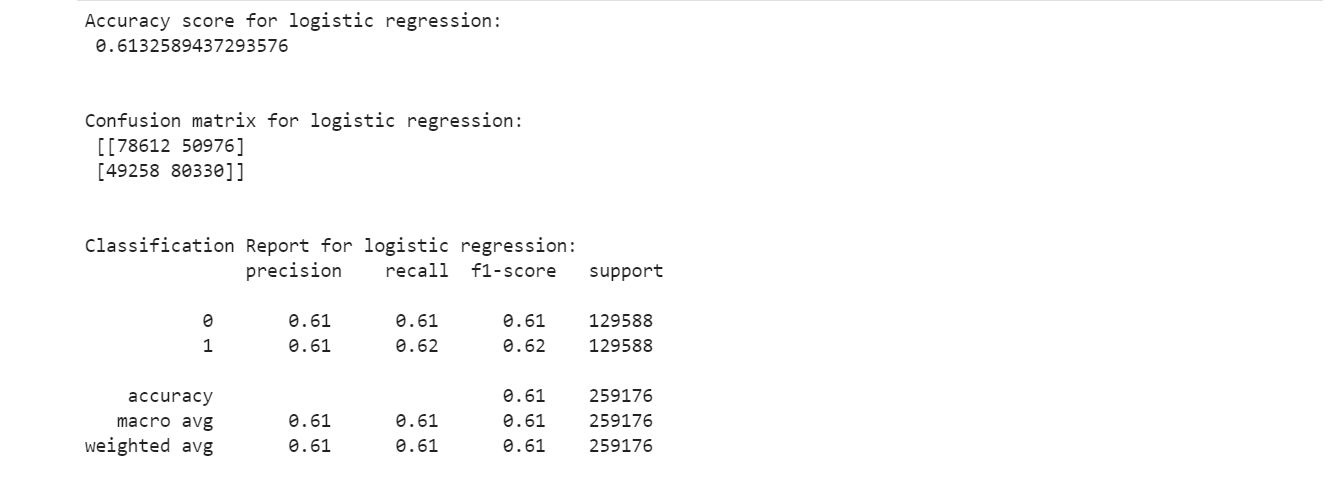
**Logistic Regression after Over Sampling:**

By performing under sampling of the data there has been an increment in the precision in predicting 1’s which is still significantly low. Hence, we performed over sampling of the minority data and applied the logistic algorithm.

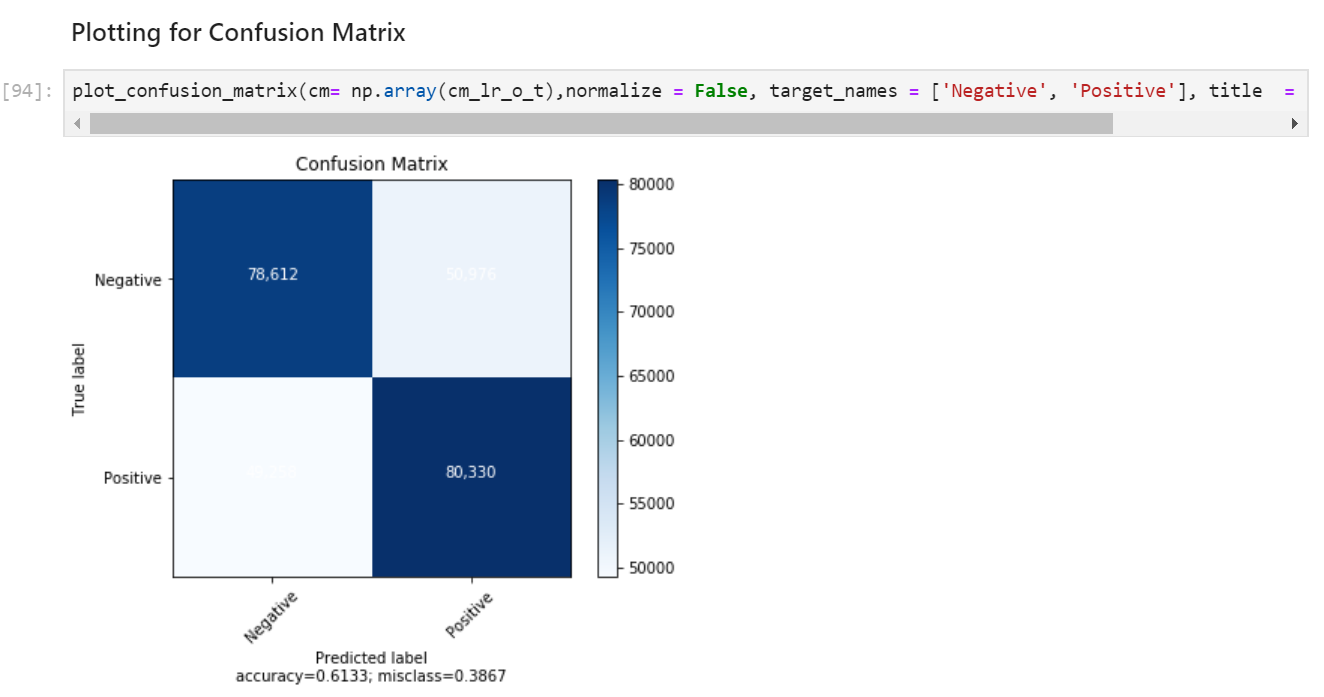


**Logistic regression after Grid Search CV:**

We further tried to enhance the model using Grid search for parameter tuning and finding the best parameters for the Logistic regression model.



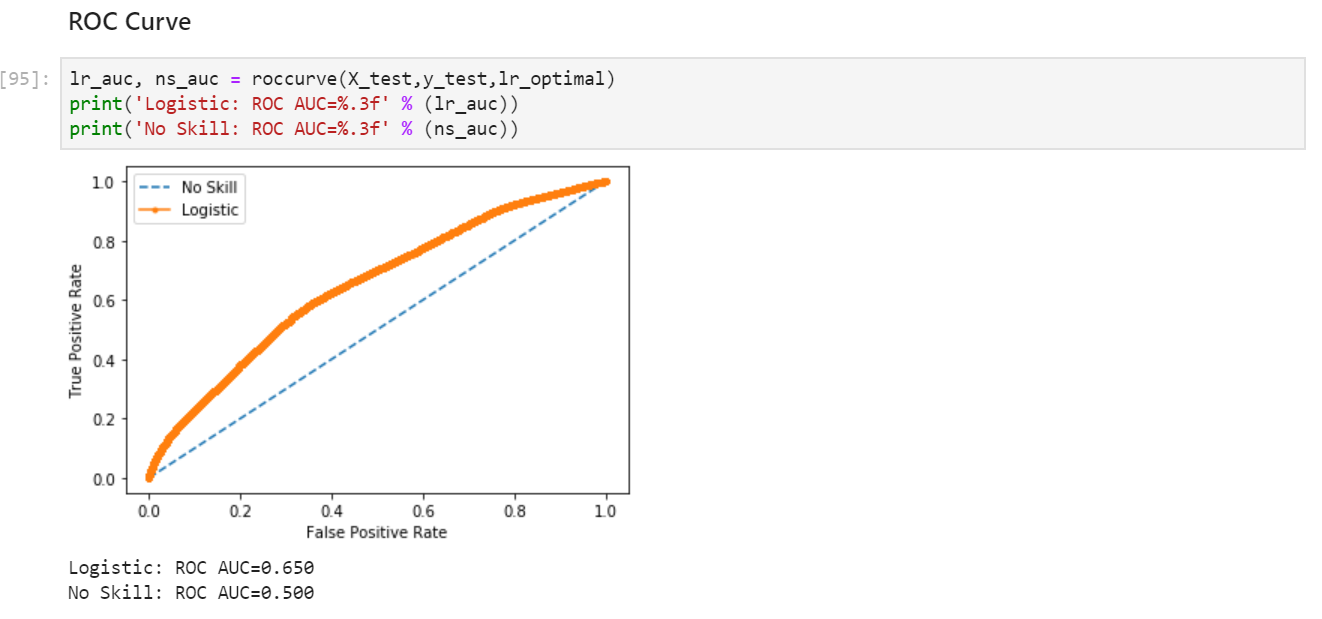
**Confusion Matric after Grid search CV:**



As the results in predicting 1’s are not satisfactory and the overall accuracy stands at 61.33%.

**ROC Curve for Logistic regression:**

ROC Curves summarize the trade-off between the true positive rate and false positive rate for a predictive model using different probability thresholds.



The area under curve value is 0.65 which shows that the logistic regression algorithm is not as efficient as the other algorithms.

**Summary and Observations of Logistic Model**

* Considered around 460k data points for running this model
* Performed SMOTE analysis to increase the precision of detecting 1’s in the data set
* After performing various techniques, we concluded that the oversampled Logistic regression with grid search CV has a better accuracy compared to the other techniques.
* AUC ROC value =0.65 and highest accuracy = 61.33

**DECISION TREE:**

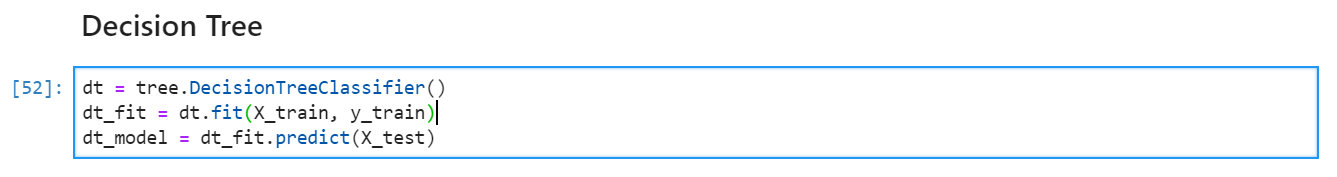
Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The Decision Tree is one of the most commonly used data mining techniques for analysis and modeling. It is used for classification, prediction, estimation, clustering, data description, and visualization.

**Pros using Decision Tree:**

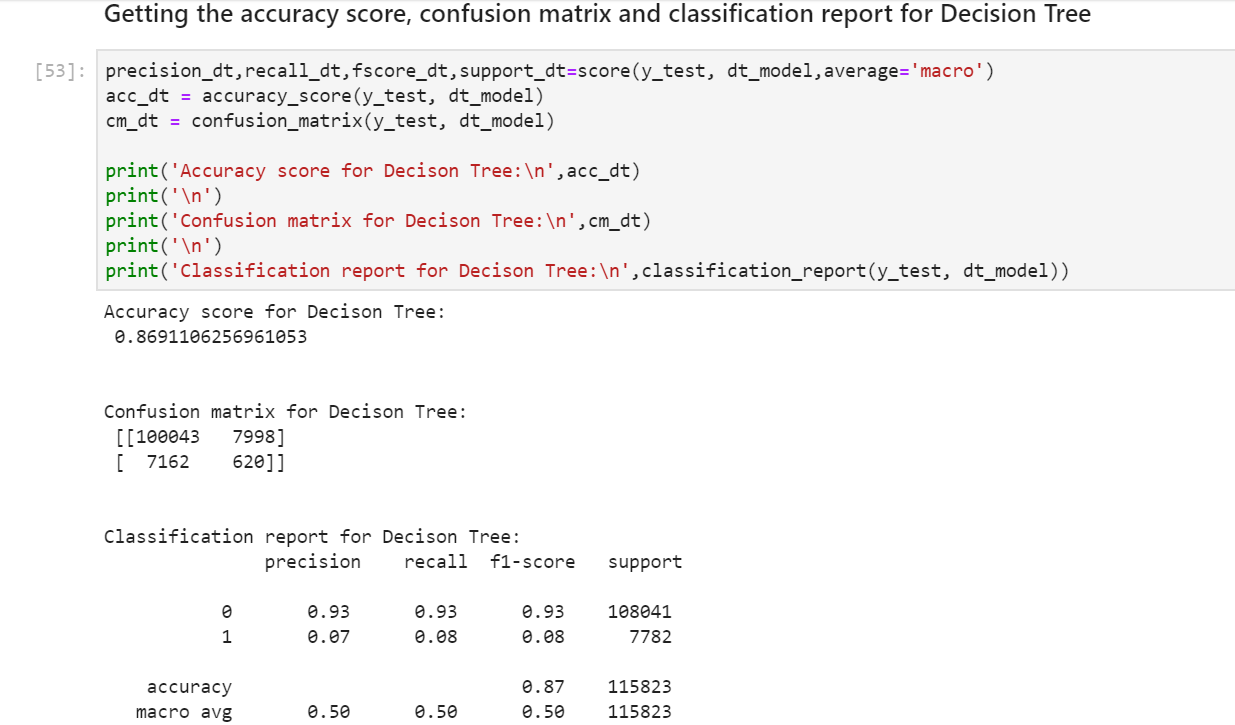
* It is very interpretable and doesn’t require normalization of data
* It deals well with noisy or incomplete data.
* It can be used for both regression and classification problems.
* Since decision trees have in-memory classification models, they do not bring in high computation costs, as they don’t need frequent database lookups.

**Applying Decision Tree:**

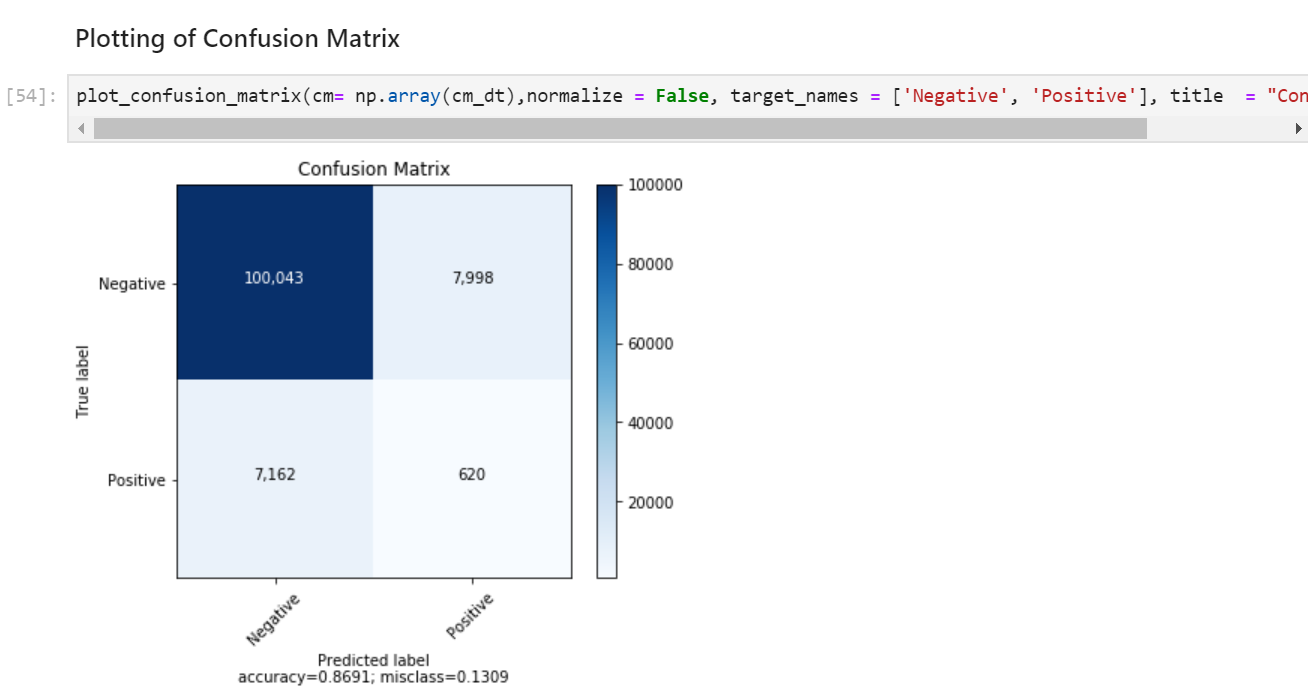
The results from the Logistic regression model were not satisfactory. Hence, we applied Decision tree algorithm as our dataset is sparsely distributed. Trained the data using the Decision tree and predicted the result for test data.



**Calculated the accuracy score, Confusion Matrix, classification report.**



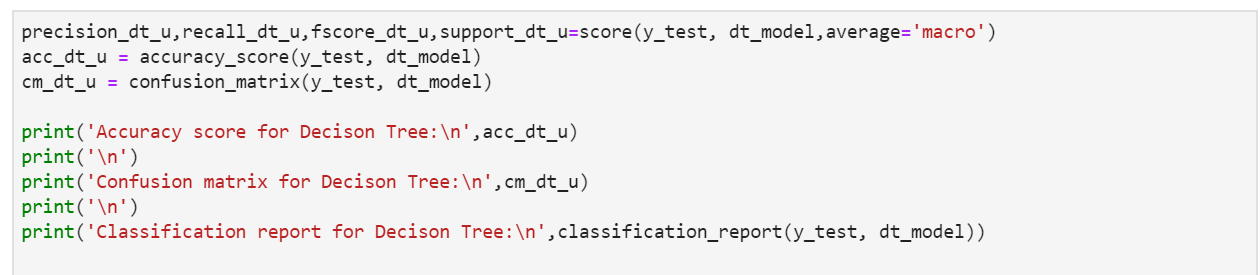
**Plotting Confusion Matrix:**

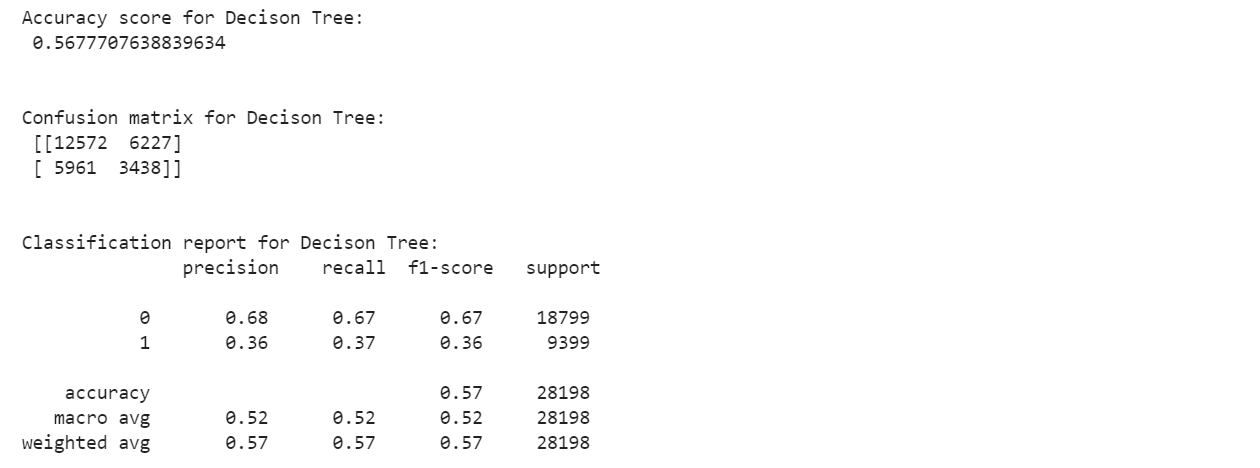


The results are not satisfactory as the precision in predicting 1’s is very as the dataset is highly imbalanced.

**Decision Tree After Under Sampling:**

As the dataset is imbalanced, we would perform SMOTE analysis by under sampling the majority. We use random under sampling technique to reduce the number of examples in the majority class to have 50 percent more than the minority class.



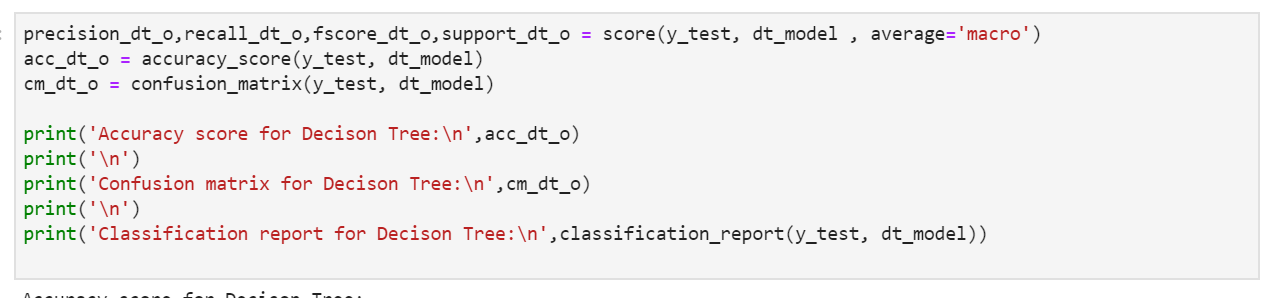


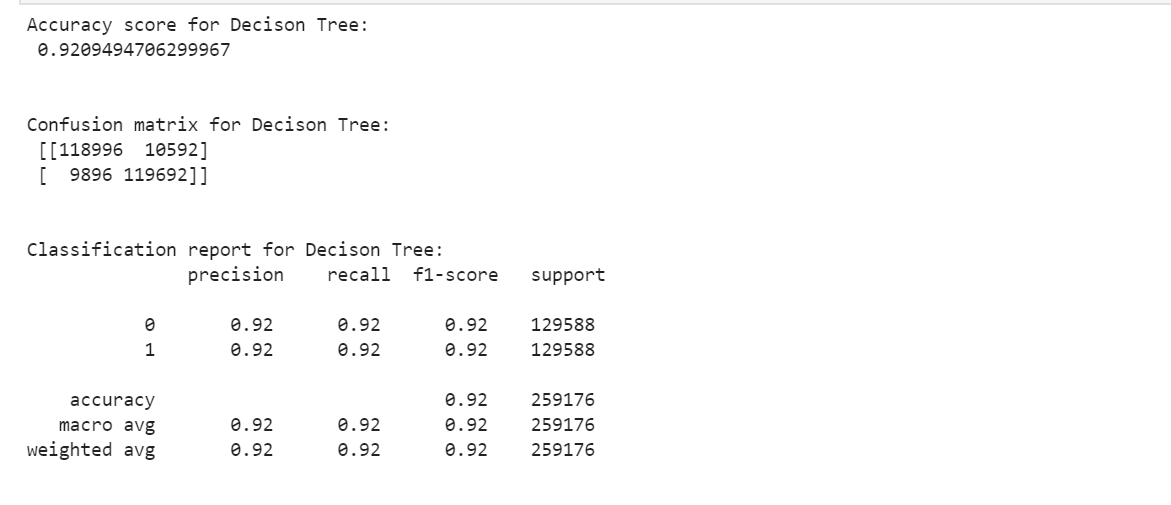
Precision for predicting 1’s has increased and stands at 36% which is significantly low. Hence, we would proceed to perform the over Sampling the minority. The overall accuracy=0.57

**Decision Tree after Over Sampling:**

By performing under sampling there has been an increment in the precision in predicting 1’s which is still significantly low. Hence, we performed over sampling of the minority data and applied the logistic algorithm

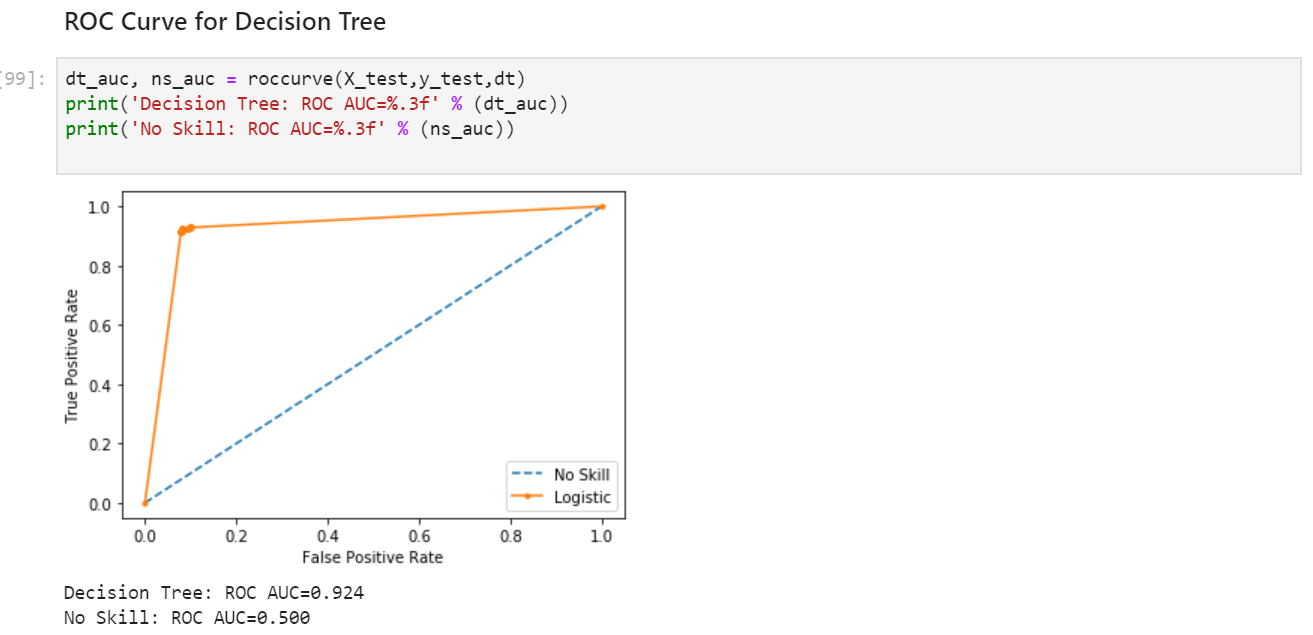
**Calculated the accuracy score, Confusion Matrix, classification report.**





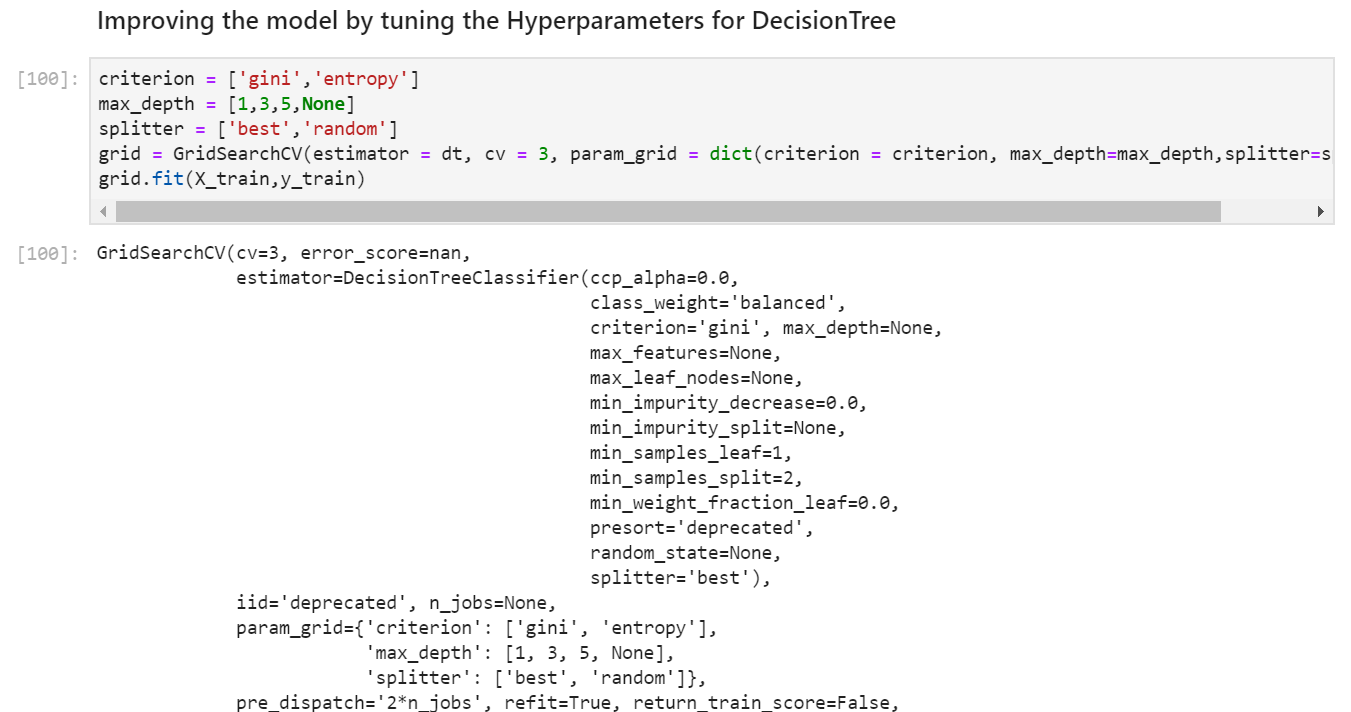
The oversampling of the minority data shows significant improvement in the precision of predicting 1’s i.e 92%. Whereas the overall accuracy= 92%, we can understand the decision tree algorithm after oversampling the data resulted in satisfactory results.

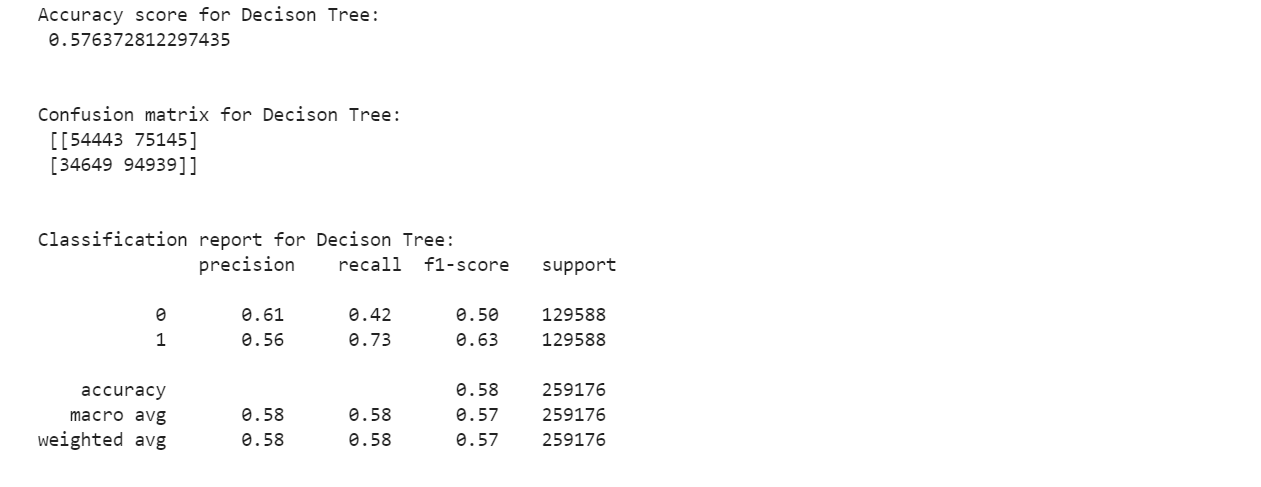
**ROC curve For Decision Tree:**



The area under curve value is AUC= 0.924 which shows that the Decision Tree algorithm is an efficient methodology in distinguishing 0’s and 1’s.

**Decision Tree Algorithm after Grid Search CV:**

We further tried to enhance the model using Grid search for parameter tuning and finding the best parameters for the Logistic regression model.



The accuracy has been reduced to 58% and the precision for predicting both 0’s and 1’s has reduced.

**Summary and Observations:**

* Performed SMOTE analysis by both under sampling and over sampling the data and Concluded that Over Sampling would be an ideal fit.
* Decision tree with oversampling the minority has the highest accuracy = 92%
* The Grid search CV has an accuracy rate less than that of the oversampled method
* AUC ROC value =0.924
* Would perform Random forest Algorithm as it is much more robust than the decision tree algorithm and compare results

**GRADIENT BOOSTING ALGORITHM**

By using Gradient Boosting we train a decision tree in which each observation is assigned an equal weight. After evaluating the first tree, we increase the weights of those observations that are difficult to classify and lower the weights for those that are easy to classify. Predictions of the final ensemble model is therefore the weighted sum of the predictions made by the previous tree models. Gradient Boosting Algorithm is implemented to overcome the shortcomings of weak learners (decision trees). The algorithm boosts its performance by using gradients in the loss function. The loss function is a measure indicating how good are model’s coefficients are at fitting the underlying data.

Initially we specify certain arguments for our function like we have

* The number of boosting stages (n\_estimators) = 100 as Gradient boosting is fairly robust to over-fitting, so a large number usually results in better performance.
* The minimum number of samples required to split (min\_samples\_split) =1 and the split point at any depth will only be considered if it leaves at least minimum training samples (min\_samples\_leaf) =1
* The number of maximum nodes (max\_depth) =3
* The number of features to consider when looking for the best split(max\_features)= sqrt(n\_features)

We then trained the model using Gradient Boosting Algorithm and predicted the output for the testing dataset

A screenshot of a social media post

Description automatically generated

In order to evaluate the performance of our model we calculated the accuracy\_score, confusion matrix, precision, recall, f1-score, support

A screenshot of a social media post

Description automatically generated

Even though the overall accuracy is 93% The confusion matrix shows that it is predicting only one class (0) i.e the majority class and is not respecting any other outcome. This suggests that the data is highly imbalanced.

A screenshot of a social media post

Description automatically generated

In order to solve the challenge of imbalanced dataset we used the Synthetic Minority Oversampling Technique (SMOTE).

We use random undersampling technique to reduce the number of examples in the majority class to have 50 percent more than the minority class.

A screenshot of a cell phone

Description automatically generated

After under sampling though the precision for predicting minority class(1’s) has increased to a large extent but the overall accuracy decreased fairly

A screenshot of a cell phone

Description automatically generated

As the results are not satisfactory, we tried to oversample the examples in the minority class. This can be achieved by simply duplicating examples from the minority class in the training dataset prior to fitting a model.

A screenshot of a cell phone

Description automatically generated

Using oversampled data, we predicted the testing set using the Gradient Boosting model. By learning with highly skewed data sets the prediction resulted in aided learning and decreased non-uniform misclassification costs.

A screenshot of a cell phone

Description automatically generated

**ROC Curve:**

ROC Curves summarize the trade-off between the true positive rate and false positive rate for a predictive model using different probability thresholds.

A screenshot of a map

Description automatically generated

The curve shows that we have ROC AUC value =0.92 indicating a good capability of distinguishing between 0’s and 1’s.

**Summary and Observations**:

* Considered around 460k datapoints for running this model
* SMOTE analysis helped in increasing the precision of detecting 1’s in the data set.
* Reduced the datapoints to around 94k data samples for under sampled data.
* Considered 863k data points for oversampled data.
* After applying various techniques, the oversampled data technique was most desirable and gave an accuracy of 86%
* By using this algorithm as trees are built sequentially the model took longer time.

**RANDOM FOREST ALGORITHM:**

Random Forest Algorithm combines more than one algorithm of same or different kind for classifying objects. Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object. This process helps in getting more accurate results by reducing the noise. As the data is highly unbalanced random forest may be a better choice.

Advantages of Random Forest Algorithm:

* It reduces overfitting problem observed in decision trees
* It offers feature selection for identifying most relevant features.
* As the given dataset is sparse random forest works robust for outliers

We trained the model using Random Forest Algorithm and predicted the output for the testing dataset.

A screenshot of a cell phone

Description automatically generated

In order to evaluate the performance of our model we calculated the accuracy\_score, confusion matrix, precision, recall, f1-score, support

A screenshot of a social media post

Description automatically generated

Even though the overall accuracy is around 92% The confusion matrix shows that it predicting only one class (0) i.e the majority class and is not respecting any other outcome. This suggests that the data is highly imbalanced.

A screenshot of a social media post

Description automatically generated

In order to solve the challenge of imbalanced dataset we used the Synthetic Minority Oversampling Technique (SMOTE).

We use random under sampling technique to reduce the number of examples in the majority class to have 50 percent more than the minority class. The under sampling improved the precision for predicting minority class(1’s) but the overall accuracy decreased fairly

A screenshot of a social media post

Description automatically generated

As the results are not satisfactory, we tried to oversample the examples in the minority class. This can be achieved by simply duplicating examples from the minority class in the training dataset prior to fitting a model.

By using oversampled data, we predicted the testing set using the Random Forest model. By learning with highly skewed data sets the prediction resulted in aided learning and decreased non-uniform misclassification costs.

A screenshot of a cell phone

Description automatically generated

**ROC Curve:**

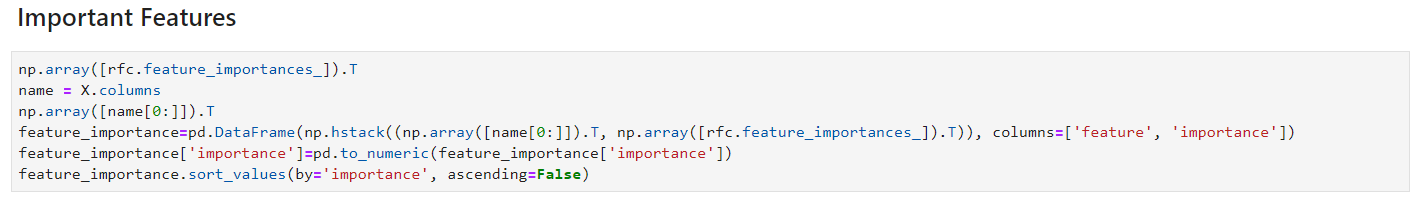
ROC Curves summarize the trade-off between the true positive rate and false positive rate for a predictive model using different probability thresholds.

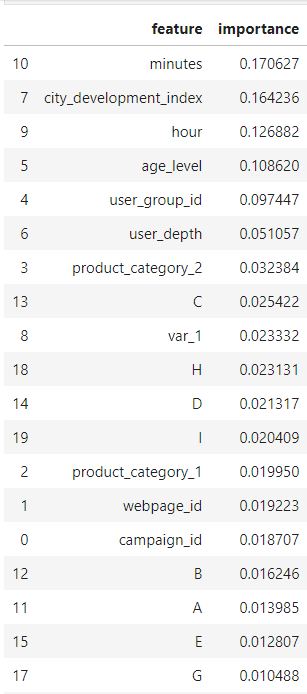
A picture containing screenshot, map

Description automatically generated

The curve shows that we have ROC AUC value = 0.974 indicating a good capability of distinguishing between 0’s and 1’s.

We analyzed the important features that are important for our model





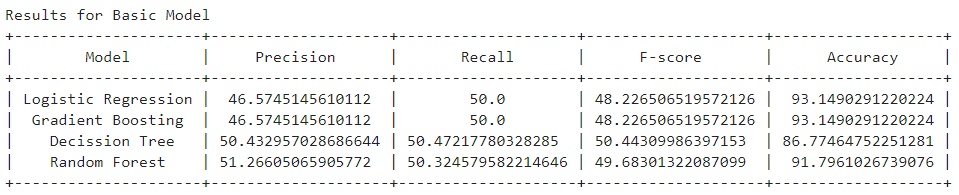
**Summary and Observations:**

* Considered around 460k datapoints for running this model
* SMOTE analysis helped in increasing the precision of detecting 1’s in the data set.
* Reduced the datapoints to around 94k data samples for under sampled data.
* Considered 863k data points for oversampled data.
* After applying various techniques, the oversampled data technique was most desirable and gave an accuracy of 94%
* The ability of classifier to find positive instances is poor in case of under sampling (Recall=0.22) while by using oversampling technique we can predict the positive instances better (Recall=0.93)
* As it is a collection of trees the model took longer time.

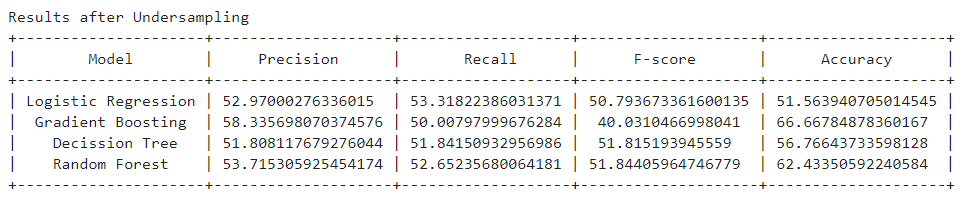
**RESULT AND ANALYSIS:**

We have applied 4 algorithms to predict the outcome.

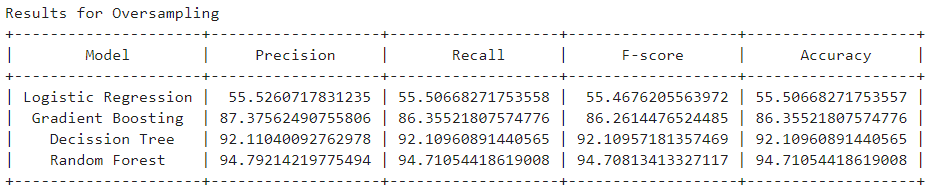
1. The dataset is very imbalanced i.e the numbers of 0’s are higher than 1’s. The model is predicting all the outcomes as 0 so the accuracy score is very good, but the precision score is not good because of the imbalance dataset.



1. As the dataset was very imbalanced, we tried undersampling the majority data and predicting the outcome. The precision score increased compared to the 1st dataset, but the accuracy is low.



1. We oversampled the data by increasing the minority data and predicting the outcome. The precision and accuracy score has increased for all the models except Logistic Regression



Finally, among the four algorithms, Random Forest has the highest accuracy

**CONCLUSION:**

Our project successfully achieved an advertisement click prediction model which has a relative high prediction and accuracy. We compared both under sampling and over sampling technique for handling imbalanced datasets and concluded that the oversampling technique gives more desirable results. Grid Search techniques allows the user to tweak the parameter values giving the user multiple ways to enhance the existing model. By using Random Forest Classifier we have achieved highest accuracy of around 95%. By using Decision Tree we have achieved an accuracy of around 92%. But the time taken for predicting the output for decision tree was less compared to random forest. By using Random Forest Classifier we achieved high accuracy(95%) and precision(93%) , we also used this model for identifying most relevant features. The system is beneficial to user, advertiser and also to the advertising platform by decreasing the advertisement pollution on the Internet.

**FUTURE SCOPE:**

* This model can be used as a recommendation system which demonstrates the potential targeting user group.
* The model can be further enhanced by considering more features that can be used for prediction. As the data keeps on growing with this system, we can design a custom model that do not compromise on accuracy while delivering fast predictions.

**REFERENCES:**

1. M. Sree Vani, Prediction of Mobile Ad Click Using Supervised Classification Algorithms, in International Journal of Computer Science and Information Technologies, Vol. 7 (2), 2016, 623-625
2. S. Saraswathi, Vallidevi Krishnamurthy, D. Venkata Vara Prasad, Tarun R K, S Abhinav and D. Rushitaa, Machine Learning Based Ad-click Prediction System, in International Journal of Engineering and Advanced Technology (IJEAT), pp.3646-3648, 2019
3. Sen Zhang, Qiang Fu and Wendong Xiao, Advertisement Click-Through Rate Prediction Based on theWeighted-ELM and Adaboost Algorithm