IIMB
तेजस्वि नावधीतमस्तु
भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT
BANGALORE

INMOBI

# Predicting genuineness of consumer surveys for better product experience

INMOBI

MARCH - 2019

SRIRAM NEELAKANTA SUBRAMONEY
BDA03024

RASANA P J
BDA03019

VENKATAGIRISH PAMIDIGANTAM
BDA03027

SUKANYA KARMAKAR
BDA03025

PAVAN KUMAR KATRAPATI
BDA03016

UNDER THE GUIDANCE OF - RAHUL KUMAR

# EXECUTIVE SUMMARY

## About this report

This report is to elaborate on a forecasting method that was built to identify consumer survey responses as genuine or non-genuine. Identifying genuine survey responses would help InMobi in make their marketing efforts more effective and ultimately achieve higher consumer satisfaction. The report first explores the data to create a basic understanding and identify opportunity areas for analysis, then various predictive modelling techniques are applied and the model with the best f1 score is chosen. The report ends with a conclusion that describes the findings of the whole exercise, and next step recommendations.

## About InMobi

InMobi works in mobile advertisement services, which means dealing with a lot of consumer behavior surveys and creating strategies for better brand efforts and consumer experience. It was founded in 2007 and has offices in Bengaluru, Mumbai in India. InMobi majorly deals with brands in Asia and Africa markets and helps them by providing brand ad and digital media marketing solutions. InMobi has a robust data science team which can be credited for creating various in-house algorithms that helps their clients and the firm itself immensely.

## The demand for a forecasting model

Why is there a need for classifying survey responses as genuine or non-genuine?
In this era and age, where consumer analytics is quickly picking up, it is pivotal for consumer brands and firms working directly towards consumer catering, to understand what the customer really wants. Surveys and various methods of data collection creates opportunities for consumer behavior analytics but looking at the right kind of data in a huge pile of numbers is important. Hence, analyzing just survey responses is a futile attempt unless we first ensure that those responses could be relied upon.
This study is an attempt at creating a model that can predict the genuineness of a survey response (and associate a probability with it) so that those responses could be further utilized effectively.

## Conclusion and Recommendations

The various models which are analyzed in this report have come up with influential features to predict the genuineness of the response along with the predicted probability. We identify features that contribute highly in signifying genuineness of a survey like completeness, qualification.

The research also discoveres a number of opportunities for improvement, such as improving the quality of the data by encouraging the customers to complete the survey which is very important to make the model perform better. Another approach to model improvement could be the addition of such features that are more representative of consumer behavior traits. Some of those features could be time of the day when survey was taken, geographical profile etc.

# Table of Contents

# CHAPTER – I

# INTRODUCTION AND PROBLEM DEFINITION

## 1.1 About InMobi Private Limited

InMobi was conceptualized and born in 2007 as mKhoj and got its current name in 2009. Currently it provides mobile advertising services in Asia and Africa. They have pioneered an advertisement serving algorithm that helps in optimizing the ranking of advertisements served on mobile phones.

InMobi also provides campaign management services, such as strategy, designing, executing, managing, and optimizing campaigns, and advertisement creation, format, and targeting services for advertisers. Moreover, it offers monetization strategies to monetize site traffic by using a combination of performance and branding advertisements.

Further, InMobi provides advertisers and publishers with access to near real-time reports to monitor the performance of advertisement campaigns and the status of site monetization. Furthermore, it has inhouse algorithms for brand ad tracking, mobile campaign conversion tracking and a Lifetime Value Platform, a platform that identifies in-application user groups and provides behavior insight to publishers or application developers.

## 1.2 Project Objective:

Millions of surveys are sent out each year from organizations to understand the customer base well. Digitization has given rise to the need for e-surveys as compared to prior traditional means, since they are easier to administer and can be scaled very fast.

In this context, this study tries to help the stakeholder company InMobi to spot the non-genuine responses for the various surveys they conduct. The goal of this study is to develop a model to predict the range of non-genuine responses in the surveys undertaken. They do so by the use of bogus question indicator as a proxy for genuineness and also additionally associating a probability for genuineness. This study aims to arrive at the best F-Score on the given data. Erroneous classification of non-genuine response as being genuine will be a costly proposition for the business. And hence, the intended implementation is expected to take this aspect into consideration in its design.

## 1.3 Business Problem Description:

From millions of surveys that the customers generate, InMobi intends to understand the genuineness of those. InMobi wants to outline the genuine responses so that better decisions could be taken and, the gap between brand efforts and consumer satisfaction could be bridged.

The need for this stems from the fact that in any surveying methodology, users may not answer the questionnaire honestly due to the lack of interest, haste, sensitivity to personal information or no solid return of their time invested and, no scrutiny while responding to the survey.

To identify this, InMobi has included some bogus questions, the answers to which are less likely to be falsified or are system captured. Examples of such bogus questions are listed below:

1) Which Operating System do you have? Android/IOS/Others.
2) Please choose the option that says ‘Very True’ – False/Very True/Not True/Neutral.

The assumption is that a user who answers all bogus questions correctly is genuine. This is critical to get reliable business insights out of surveys by weeding out non - genuine responses.

## 1.4 Types of Non-Genuine responses:

In the context of defining the non-genuine responses, it is necessary to look upon the various types of non-genuine responses a researcher can come across.

- Professional competition entrants are one such type of responses who enters with an intention of winning in competitions in internet, magazines or other media. The attractive incentives is something which motivates them to participate irrespective of their subject knowledge or the relevance with it. This motive propels them to mark something random which they might think leads to a prize.
- Survey speeders are another potential group of non-genuine respondents whose whole intention is to finish the survey with possible speed no matter whether they are honest or not. Although, there might be cases where consumers start the survey genuinely but due the time taking nature of surveys makes fatigue kick in.

## 1.5 Why non-genuine responses arise?

The very first reason why non-genuinity arises in most of the cases is TUTs which technically means Tasks Unrelated Thoughts! Strong research has been shown in this area and proved that there is a mismatch between what we do and what we think about, as much as 30%. The longer the task, the greater is the mismatch.

The second is technology and pace of life. Surveys taken on laptops or desktop are less time consuming than taken in mobile devices. Observers reported that, most mobile surveys taken by respondents are mostly in the times when they are at rest or at home in the evening.

And the final problem is about the time per session. Majority of the respondents tend to give much attention to the earlier questions than later ones. The attention span is likely to be reduced as time passes.

## 1.6 What are the key challenges?

1. In the initial round of model development through this project, it is expected to rely on the bogus questions which get interleaved with real questions as part of the survey. A more long-term objective (as defined by InMobi) is to evolve the model such that the bogus questions are eliminated, and the model relies on other parameters. Based on the initial view this may or may not be feasible and the project team may have to explore additional possibilities as part of the model development
2. As this is not big data (in terms of data volume), we don't have much to play with and hence accuracy of the developed model plays a key role.

# CHAPTER – II

## Source of Data:

### 2.1 Metadata:

The initial dataset received has the data for the period 21st August 2018 to 31st August 2018.

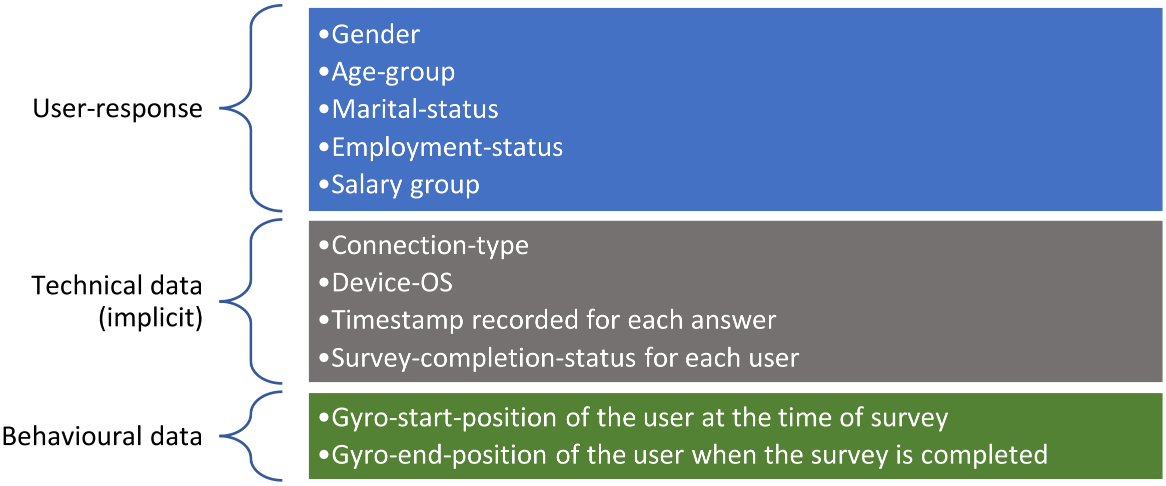
This was agreed after discussion with the client stakeholder as the data of survey was deemed as fit-for-purpose for modeling.

Table shown below captures key tenets of the input dataset supplied by the company.

| Key Attributes | Measure |
|---|---|
| **Survey Period** | 21st August 2018 to 31st August 2018 |
| **No. of Observations** | 1,94,965 Observations |
| **No. of Variables** | 28 Variables |

Primarily the database has been taken from inMobi to understand and to go through and explore the database that they have created or accumulated across the surveys.

Since the data set is not huge in nature, the approach is designed with the step of extracting meaningful insights out of raw data. Variable which are irrelevant for the study have been removed and emphasis has been given to relevant variables to work upon and extract insights which would eventually lead the study to arrive at a favorable F-score.



User-response

- Gender
- Age-group
- Marital-status
- Employment-status
- Salary group

Technical data (implicit)

- Connection-type
- Device-OS
- Timestamp recorded for each answer
- Survey-completion-status for each user

Behavioural data

- Gyro-start-position of the user at the time of survey
- Gyro-end-position of the user when the survey is completed

The data is contained under the following classification,

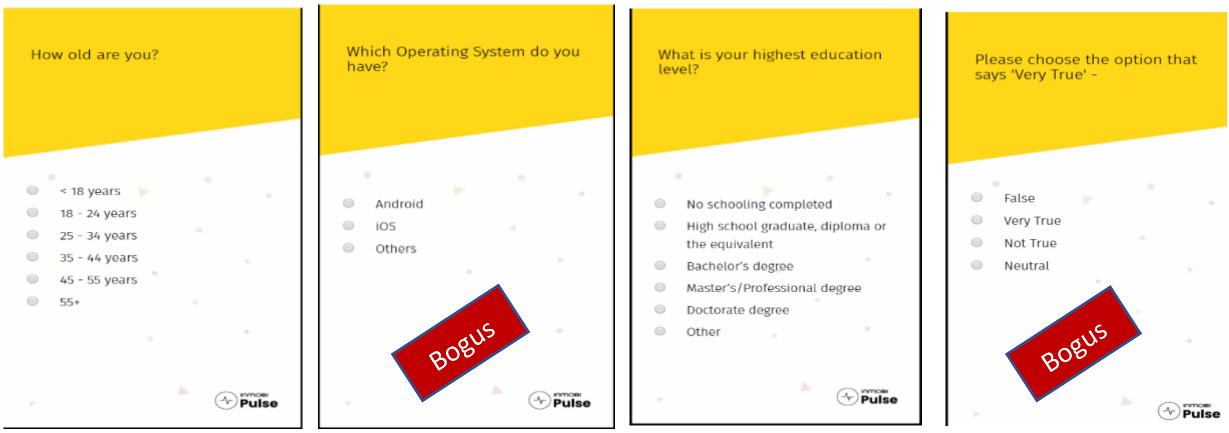
- The survey questions which were designed at the premise by InMobi for a particular study. The closed ended question which were designed with the defined choice to choose.
- The responses received for this study from the users.
- Along with this, the accumulated data grabbed by the system from the backend. And,
- Bogus questions, which are to be considered to understand the genuineness of the survey.

A Survey Response (i.e. A Dataset row) looks like:

| | |
|---|---|
| Seq | 12233103 |
| connection | Cellular |
| gyro_start | using your phone, sitting or standing |
| gyro_end | using your phone, sitting or standing |
| time | 30/08/18 1:56 |
| Q1 | Female |
| T1 | 19 |
| Q2 | 18 - 24 years |
| T2 | 22 |
| Q3 | Android |
| T3 | 27 |
| Q4 | 50-75K INR |
| T4 | 34 |
| Q5 | Master’s/Professional degree |
| T5 | 37 |
| Q6 | Unemployed |
| T6 | 43 |
| Q7 | Single |
| T7 | 45 |
| Q8 | None |
| T8 | 47 |
| Q9 | Neutral |
| T9 | 58 |
| isRewardedSlot | 0 |
| isComplete | 1 |
| isFirstBogusCorrect | 1 |
| isSecondBogusCorrect | 0 |

- Q1...Q9 correspond to survey questions in indian_demog_survey
- T1…T9 represent timestamp of respective answer in seconds from start of survey
- Seq corresponds to option position 0,1...n-1 where n is number of options of a question. Missing entry in Seq means position is not captured/not available. E.g. 12233103 means the user have chosen 2nd option (index = 1) for Q1, 3rd option (index = 2) for Q2 and so on. Index starts from 0. If Seq is shorter than # (number) of answers given, then pad them with 0. E.g. if Seq is 1212 and user has answered Q1…Q6 then actual Seq is 001212.
- **Connection** represents connection type i.e. **Wi-Fi, Cellular** etc.,
- **gyro_start, gyro_end** specify user’s position in start and end of survey respectively.
- **time** represents start time of survey in UTC
- **isRewardedSlot** is a Boolean to indicate whether the ad slot gives reward at the end of filling the survey.
- **isComplete** is a Boolean to indicate the completeness of response.

- **isFirstBogusCorrect** and **isSecondBogusCo**rrect flags indicate whether the obvious questions were answered correctly or not.



How old are you?

○ < 18 years
○ 18 - 24 years
○ 25 - 34 years
○ 35 - 44 years
○ 45 - 55 years
○ 55+

Which Operating System do you have?

○ Android
○ iOS
○ Others

What is your highest education level?

○ No schooling completed
○ High school graduate, diploma or the equivalent
○ Bachelor's degree
○ Master's/Professional degree
○ Doctorate degree
○ Other

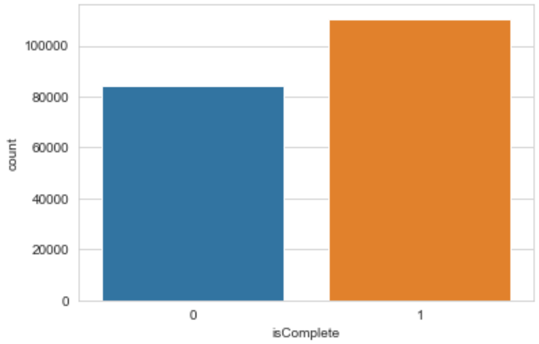
Please choose the option that says 'Very True' -

○ False
○ Very True
○ Not True
○ Neutral

How to measure these using a smartphone? There are certain hardware components in the smartphone which will automatically capture the user system information at the backend.

## 2.2 Descriptive Statistics:

Carried out a univariate analysis of the survey questions to get the initial understanding of the data set and the frequency distribution of the answers to each question.

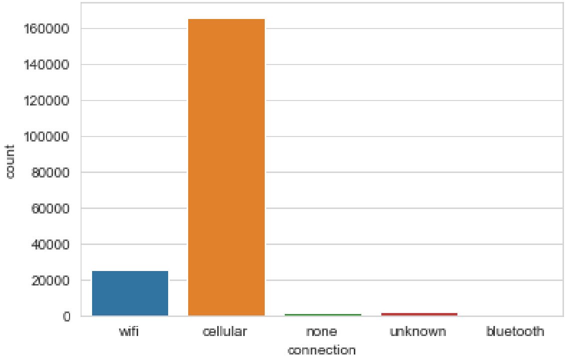


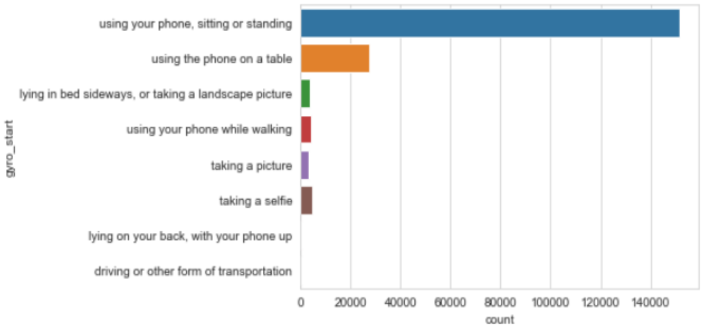
### Completeness:

Checking the completeness of the survey responses and is shown. Out of 1,94,968 user surveys, there are 110591 completed surveys and 84377 incomplete surveys. That is, 56.72% surveys are completed surveys in the given sample data set.

### Connection:

The above histogram shows that the people using Cellular connection are more compared to other type of connections, followed by Wi-Fi. None or unknown values indicate failure of JavaScript to detect the connection type since internet must be on while giving the survey.





## Gyro_start:

This variable indicates the user’s position when they start the survey.

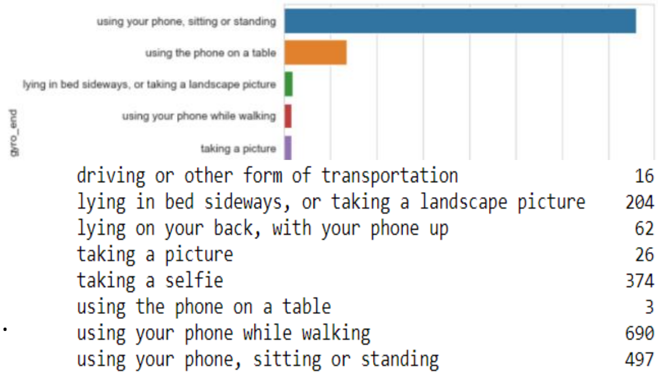
From the above histogram we can see that majority of the people are either sitting or standing, counting up to 1,51,719 responses.

## Gyro_End:

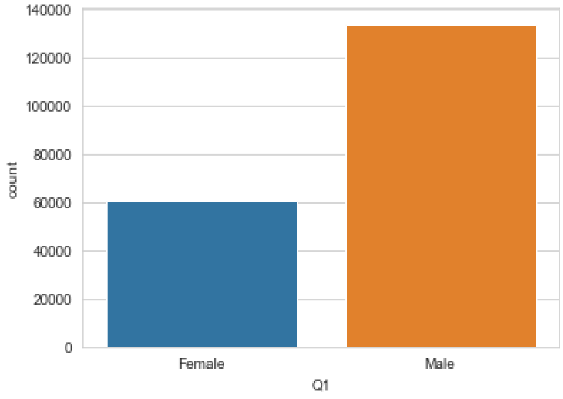
This maps the user’s position when they are finishing up the survey.

Again, either sitting or standing is the most prevalent pattern with 1,52,216 responses.

Some people have also changed their position while doing the survey and the count is as shown.



```
driving or other form of transportation             16
lying in bed sideways, or taking a landscape picture    204
lying on your back, with your phone up              62
taking a picture                                    26
taking a selfie                                    374
using the phone on a table                           3
using your phone while walking                     690
using your phone, sitting or standing              497
```



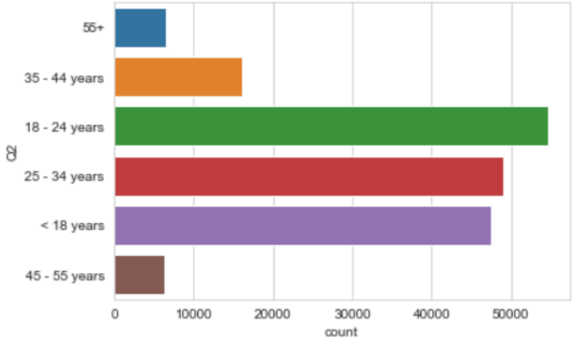
## Gender:

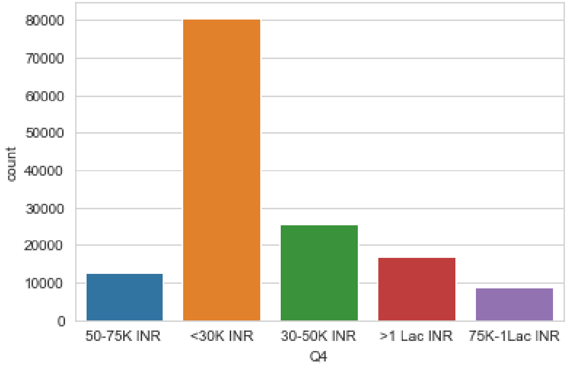
The number of Males and Females participated in the survey is as follows:

There are 1,34,012 Males and 60,956 Participated in the survey. Majority of survey participants are Men.

## Age:

People who are in the age group of 18-24 have the highest contribution to the survey, with approximately 55,000 responses.





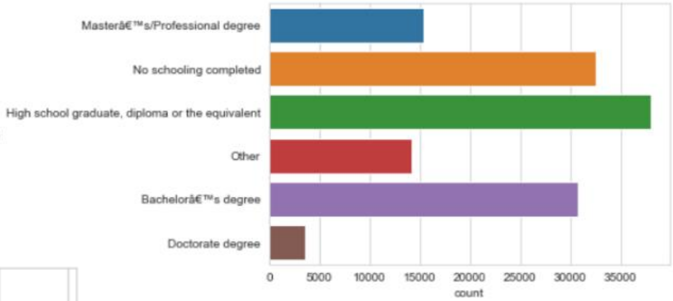
**Salary Range:**

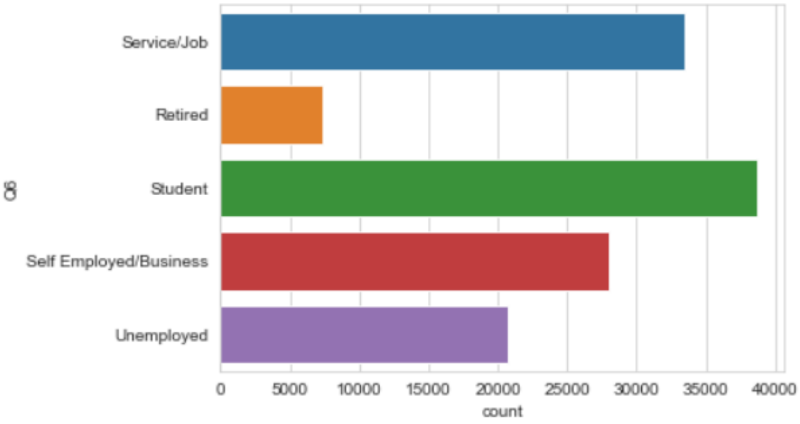
It can be observed that people who are earning less than 30K are mostly responding to the survey.

But the big question from the graph was how genuine their responses are?

**Educational Qualification:**

It can be observed that the majority of survey responses are coming from those who are high school graduates or diplomas





**Profession:**

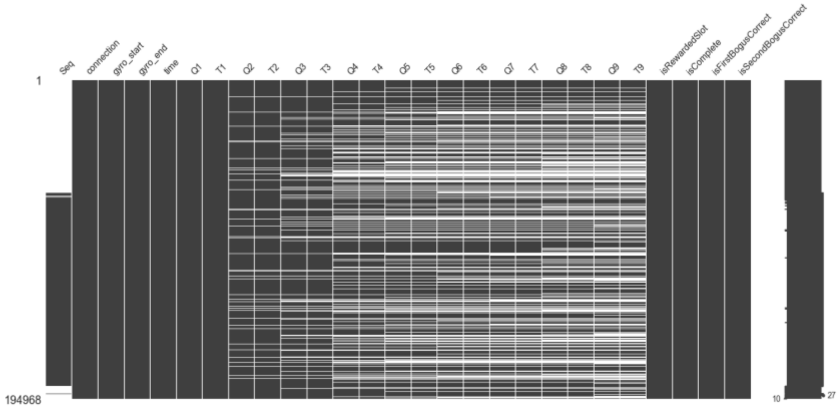
Student take the time out for responding to such surveys more than professionals. This makes sense as the average time spent online is more for students.

## 2.3 Data Cleansing and Imputation:

A data quality analysis was conducted, and a number of data cleansing and imputation steps were carried out that are discussed below.

**Missing data:**

As we can see the sequence column is not significant and we have dropped this column.



Another interesting find here, is that we see the proportion of white spaces increasing in the subsequent questions which corroborates our initial hypothesis that people drop out of surveys because of the time it takes, but the responses before that might be genuine.
Hence, we have decided to treat the incomplete responses.

***Imputation***:

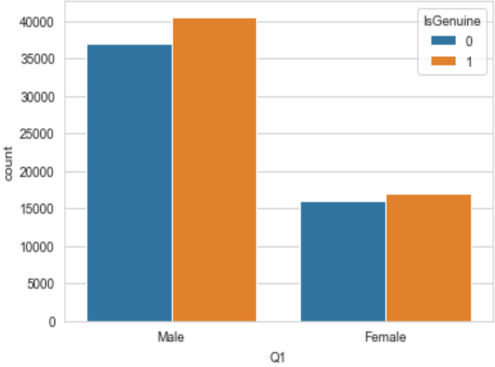
**The following data imputation steps were carried out:**

- NA values for the questions are replaced with value 'Unknown'.
- The time variables that indicate the time taken to answer each question is filled with the average time taken for the corresponding question, in case of incompleteness.

### 2.4 Labelling:

Created the target variable 'IsGenuine' using the values of bogus questions. If both the bogus questions are correct then, labelled that record as '1' which represents Genuine; otherwise '0'.

### 2.5 Bivariate Analysis:

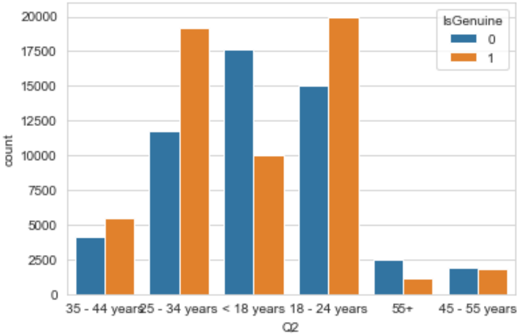


Overall, Genuine responses this way are 57,667 and non-genuine are 52,924.

The graph Q1 depicts share of genuine or non genuine responses based on Gender as a metric.

We see males have more responses as genuine, although the reason behind it must be because the survey sample contains more males than females

***Age Group with Genuineness***:

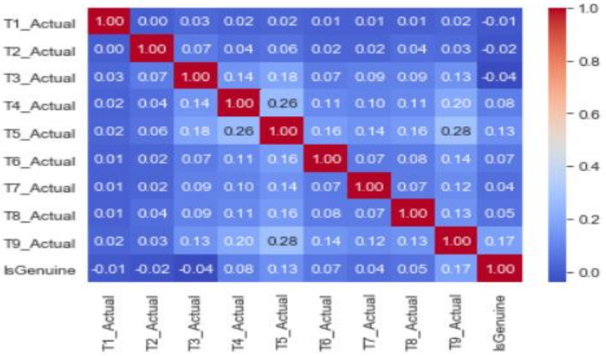


Likewise, 18-24-year olds have the most numbers of genuine responses.

People below 18years of age have a higher ratio of non genuine responses and similarly, 25-34-year olds have a higher ratio of genuine responses.
This again seems to corroborate the hypothesis that young teenagers might not take these surveys seriously, but older people are more probable to respond genuinely.

***Correlation Analysis***:

Checked the correlation between numerical variables with IsGenuine flag and the corresponding Heatmap is given as below:



T5_Actual and T9_Actual seems to have a significative correlation with the genuineness probability.

T1_Actual, T2_Actual and T3_Actual are negatively correlated with genuineness.

## CHAPTER – III

### Modelling Methodology:

Before going to start the model building, we need to check whether the model has any unwanted parameters which are not going to help us.

### 3.1 Data pre-processing:

1) Irrelevant features:

So, here the bogus questions Q9 is an option to select a given value from a list of options which doesn't contribute any information to the outcome. Similarly, Sequence column doesn't carry any information. Hence, dropped those columns.

2) Categorical features

Created dummy columns for the categorical data because most of the data in our dataset is categorical and hence, we cannot conclude with that and check are there any zero's in our dataset because it is not going to carry any information in out model building.

3) Standardization

Standardized the numeric data to avoid scaling issues.

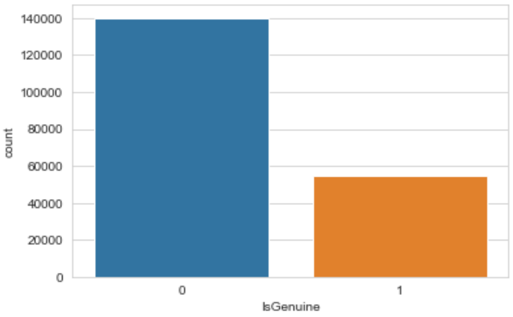
4) Data set balance:

After this step, we'll have to see whether the dataset is imbalanced or balanced.

If the number of positive samples is similar to the negative samples, the dataset is balanced. Otherwise, it is unbalanced.

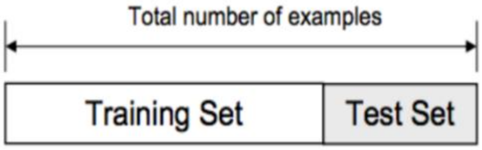
The reason why it is better to have balanced dataset is that the evaluation is easier to do since there is no bias. For example, if the percentage of positive samples in the dataset is 5%, the accuracy of the classifier which predicts all negative would be 95%. The accuracy is very high but misleading. In this case, we need to use other metrics such as precision and recall.

Using the same logic since our data is imbalanced – where it is skewed towards non genuine, we need to do sampling in order to get a balanced data.



### 3.2 Splitting the data set

It is necessary to avoid overfitting or underfitting the model as it may affect the predictability of our model — we might be using a model that has lower accuracy and/or is ungeneralized (meaning you can't generalize your predictions on other data).



So, the data set has been split into train and test set (Hold out data). The training set has been further split into train and validation set.

### 3.3 Sampling Unbalanced using SMOTE – Synthetic Minority Oversampling

This technique is followed to avoid imbalance by introducing exact replicas of minority instances to the main dataset. A subset of data is taken from the minority class as an example and then new synthetic similar instances are created. These synthetic instances are then added to the original dataset. The new dataset is used as a sample to train the classification models.

So, by using SMOTE, we got the samples as follows:

| |
|---|
| **Number transactions X_train dataset: (125736, 67)** |
| **Number transactions y_train dataset: (125736)** |
| **Number transactions X_test dataset: (43108, 67)** |
| **Number transactions y_test dataset: (43108)** |

### 3.4 Feature Selection

Feature Selection methods helps by reducing the dimensions without much loss of the total information. It also helps to make sense of the features and its importance. In other words, we choose the best predictors for the target variable.

The classes in the **sklearn.feature_selection** module can be used for feature selection/dimensionality reduction on sample sets, either to improve estimators' accuracy scores or to boost their performance on very high-dimensional datasets.

**SelectKBest Feature Selection:**

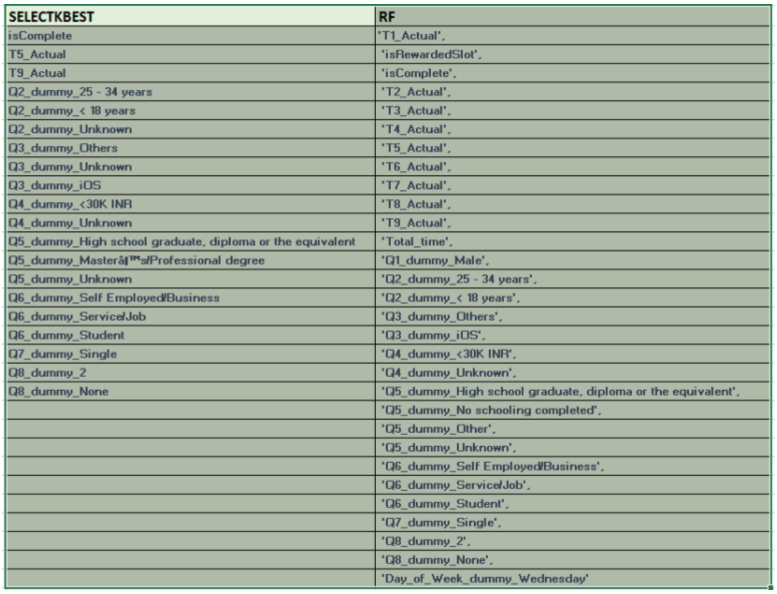
Select features according to the k highest scores. It uses the score function f_classif to score the features using ANOVA F-value between label/feature for classification tasks.

Used SelectKBest Feature Selection to our dataset in our model, to identify 20 features which are the best predictors for the target variable.

**Random Forest Feature Selection:**

Random forest consists of a number of decision trees. Every node in the decision trees is a condition on a single feature, designed to split the dataset into two so that similar response values end up in the same set. The measure based on which the (locally) optimal condition is chosen is called impurity. For classification, it is typically either Gini impurity or information gain/entropy and for regression trees it is variance. Thus, when training a tree, it can be computed how much each feature decreases the weighted impurity in a tree. For a forest, the impurity decrease from each feature can be averaged and the features are ranked according to this measure.

Random Forest Feature Selection identified 23 variables which are very near to the target variable.



| SELECTKBEST | RF |
|---|---|
| isComplete | 'T1_Actual', |
| T5_Actual | 'isRewardedSlot', |
| T9_Actual | 'isComplete', |
| Q2_dummy_25 - 34 years | 'T2_Actual', |
| Q2_dummy_< 18 years | 'T3_Actual', |
| Q2_dummy_Unknown | 'T4_Actual', |
| Q3_dummy_Others | 'T5_Actual', |
| Q3_dummy_Unknown | 'T6_Actual', |
| Q3_dummy_iOS | 'T7_Actual', |
| Q4_dummy_<30K INR | 'T8_Actual', |
| Q4_dummy_Unknown | 'T9_Actual', |
| Q5_dummy_High school graduate, diploma or the equivalent | 'Total_time', |
| Q5_dummy_Masterâ€™s/Professional degree | 'Q1_dummy_Male', |
| Q5_dummy_Unknown | 'Q2_dummy_25 - 34 years', |
| Q6_dummy_Self Employed/Business | 'Q2_dummy_< 18 years', |
| Q6_dummy_Service/Job | 'Q3_dummy_Others', |
| Q6_dummy_Student | 'Q3_dummy_iOS', |
| Q7_dummy_Single | 'Q4_dummy_<30K INR', |
| Q8_dummy_2 | 'Q4_dummy_Unknown', |
| Q8_dummy_None | 'Q5_dummy_High school graduate, diploma or the equivalent', |
|  | 'Q5_dummy_No schooling completed', |
|  | 'Q5_dummy_Other', |
|  | 'Q5_dummy_Unknown', |
|  | 'Q6_dummy_Self Employed/Business', |
|  | 'Q6_dummy_Service/Job', |
|  | 'Q6_dummy_Student', |
|  | 'Q7_dummy_Single', |
|  | 'Q8_dummy_2', |
|  | 'Q8_dummy_None', |
|  | 'Day_of_Week_dummy_Wednesday' |

Used features selected by SelectKbest as the F1 score using SelectKbest features are good compared to Random forest features.

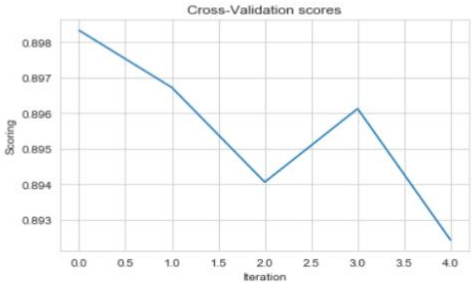
## 3.5 Model Building:

This section describes in detail the sequence of steps undertaken in the model building exercise. With the final dataset of 20 features, a number of classification techniques were attempted, and the results of those model runs are documented in the following sub-sections. The model is evaluated for a good F1-score as per the requirement by inMobi. Considering classifying non-genuine records as genuine is costlier, the model hyper parameters are tuned to get best score.

### 1. LOGISTIC REGRESSION:

Trained the logistic regression model using the training data set with 5-fold cross validation and given below the scoring of cross validation or training data.

```
Model Report:
Model score error with cv : Mean - 0.8955 | Stdf - 0.0021
```

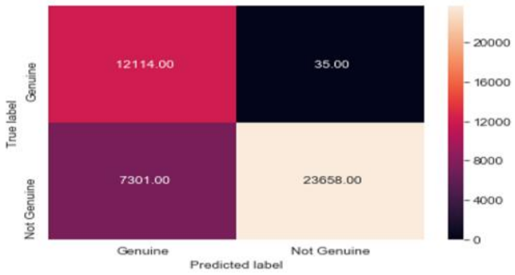


Prediction on test data got F1 score of 0.84 and the confusion matrix which is as follows:

Precision, Recall and Support:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 0.76 | 0.87 | 30959 |
| 1 | 0.62 | 1.00 | 0.77 | 12149 |
| micro avg. | 0.83 | 0.83 | 0.83 | 43108 |
| macro avg. | 0.81 | 0.88 | 0.82 | 43108 |
| weighted avg. | 0.89 | 0.83 | 0.84 | 43108 |

```
Accuracy:  0.8298227707154124
Confusion Matrix:
```

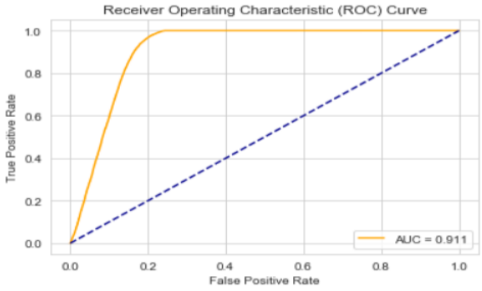


The result shows that there are 12114+23658 correct predictions and 35+7301incorrect predictions.

The model is very good at classifying genuine records, but only 76% of non-genuine records are correctly classified.

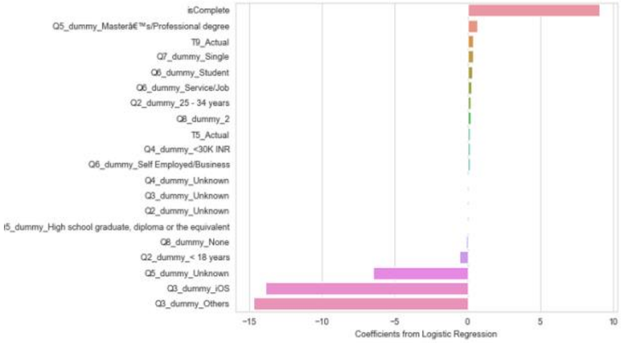
**ROC Curve:** Receiver Operating Characteristic (ROC) curve is a plot of the true positive rate against the false positive rate. It shows the tradeoff between sensitivity and specificity. AUC score in this case (test data set) is 0.91

```
AUC : = 0.911
```



So Logistic Regression gives F1-Score for non-genuine observations is 86% with specificity (True negative rate) 76% of the time.

Given below the bar plot depicting coefficient values of features in the LR model sorted in descending order.



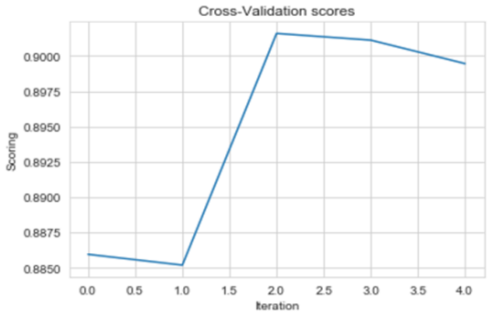
The sign of coefficient values shows the positive and negative effect of features on the response genuineness.

### 2. Decision Tree (DT):

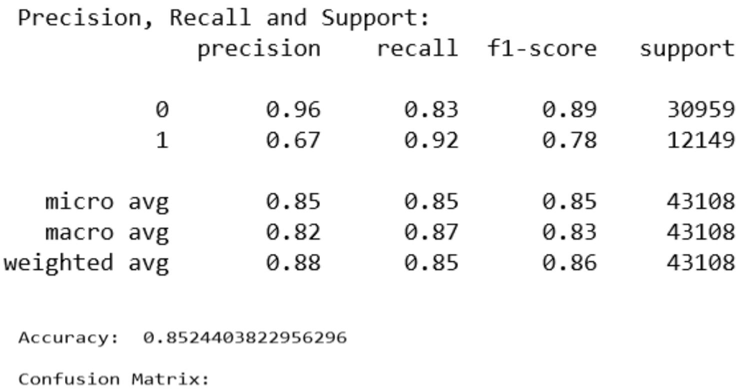
The Decision tree classifier is trained using the training data set. Built the model using “gini” and “entropy” impurity at different tree depths. The model parameters with the best score is chosen to predict the test data and the results are given below:

Cross Validation score:

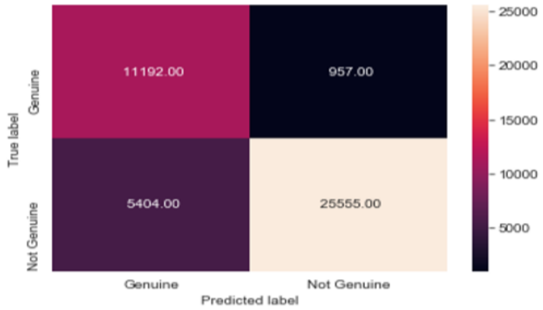
```
Model Report:
Model score error with cv : Mean - 0.8947 | Stdf - 0.0075
```



Prediction on test data got F1 score of 0.86 and the confusion matrix which is as follows:



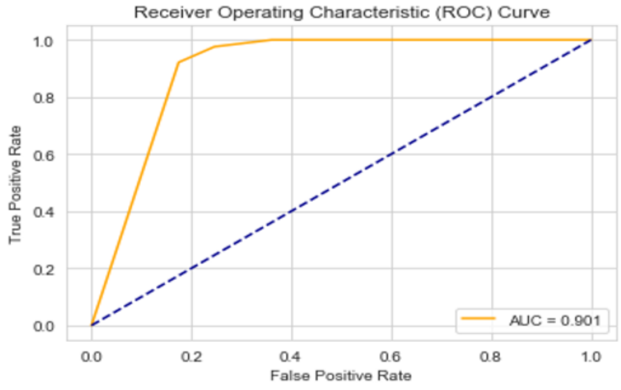
```
 Precision, Recall and Support:
              precision    recall  f1-score   support

           0       0.96      0.83      0.89     30959
           1       0.67      0.92      0.78     12149

   micro avg       0.85      0.85      0.85     43108
   macro avg       0.82      0.87      0.83     43108
weighted avg       0.88      0.85      0.86     43108

 Accuracy:  0.8524403822956296
 Confusion Matrix:
```



The result shows that there are 11192+25555 correct predictions and 957+5404 incorrect predictions.

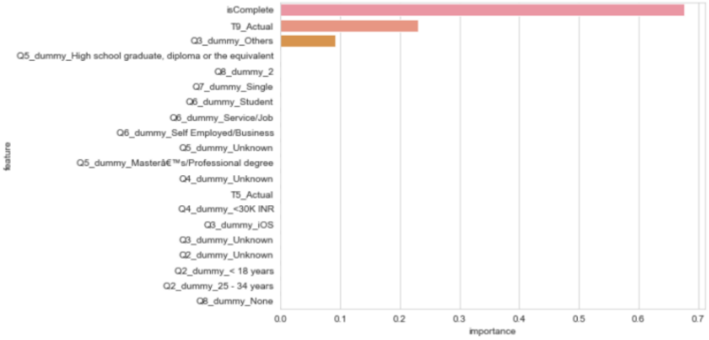
The model is good at classifying both genuine and non-genuine records.

**ROC Curve:** Receiver Operating Characteristic (ROC) curve is shown below. It shows the tradeoff between sensitivity and specificity. AUC score in this case (test data set) is 0.90



So, Decision Tress gives F1-Score for non-genuine observations is 89% with specificity (True negative rate) 83% of the time.

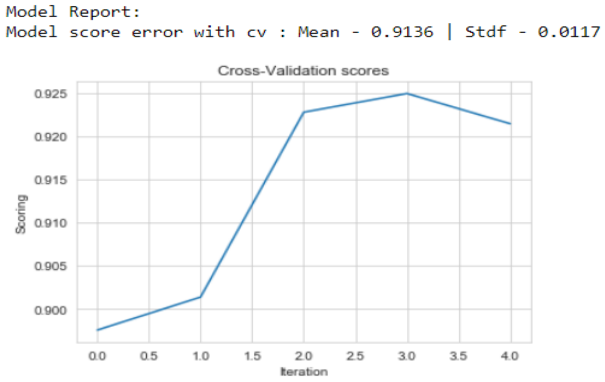
Given below the bar plot depicting importance score of features in the Decision tree classifier sorted in descending order.



3. **RANDOM FOREST (RF)**:

Random forest is one of the most popular ensemble techniques used in the industry dur its performance and scalability. As the name suggest, this algorithm creates the forest with a number of decision trees. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The hyper parameters are tuned using Grid search to increase the model accuracy. The best score when used 100 trees using entropy criteria to measure the split quality.

The cross-validation scores curve for random forest model training is shown below:

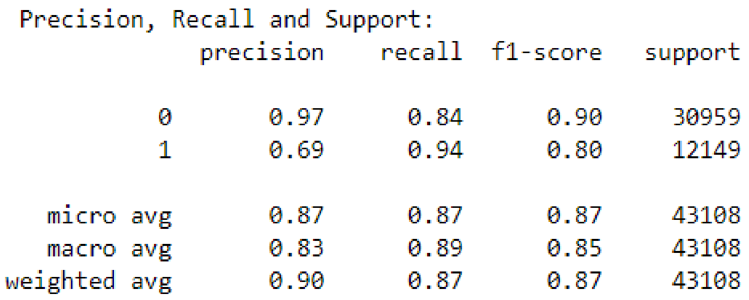


```
Model Report:
Model score error with cv : Mean - 0.9136 | Stdf - 0.0117
```

Through Random Forest, we got the F1- score of 0.87 with Specificity (True negative rate) 0.84

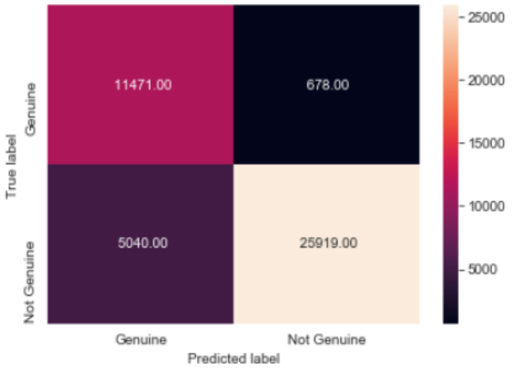
The result of the prediction of trained model on test data is given below:

**Confusion Matrix:**



```
 Precision, Recall and Support:
              precision    recall  f1-score   support

           0       0.97      0.84      0.90     30959
           1       0.69      0.94      0.80     12149

   micro avg       0.87      0.87      0.87     43108
   macro avg       0.83      0.89      0.85     43108
weighted avg       0.90      0.87      0.87     43108
```

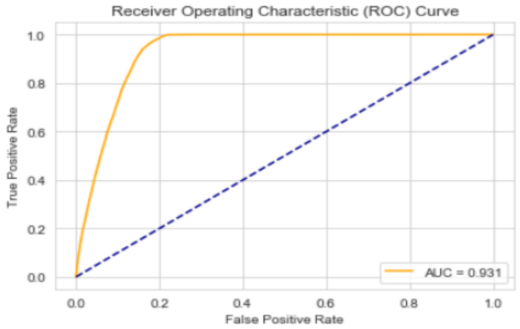
```
Confusion Matrix:
```



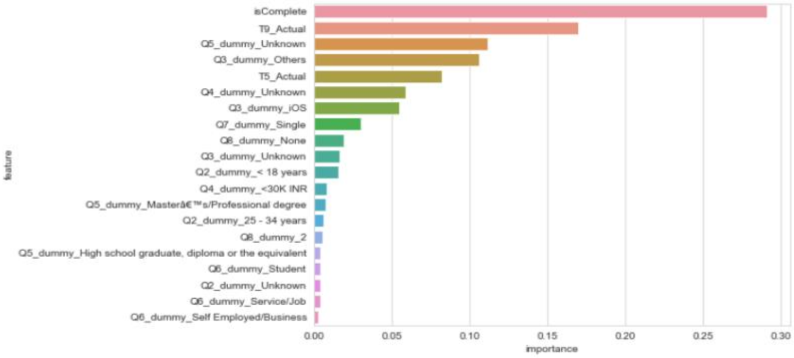
The model detects 11471 out of 12149 genuine responses and there are only 5040 false positive cases. The precision and recall are better compared to logistic regression and decision tree classifiers.

The ROC AUC curve of Random forest model is shown in the figure. The model gives an AUC value 0.93.

```
AUC : = 0.931
```



Given below the bar plot depicting importance score of features in the Random Forest classifier sorted in descending order.

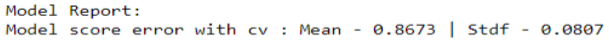


The top 5 features are isComplete flag, T9_Actual, Q5_dummy_Unknown, Q3_dummy_Others, T5_Actual.

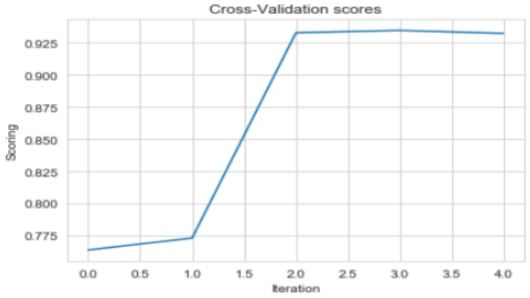
4. **Gradient Boosting (GB)**:

Gradient Boosting is another popular ensemble technique which combines multiple weak classifiers into single strong classifier. Boosting builds multiple classifiers in a sequential manner. It uses gradient descent algorithm in each stage to minimize the error and it uses decision tree as base classifier.

The cross-validation scores curve for gradient boosting model training is shown below:

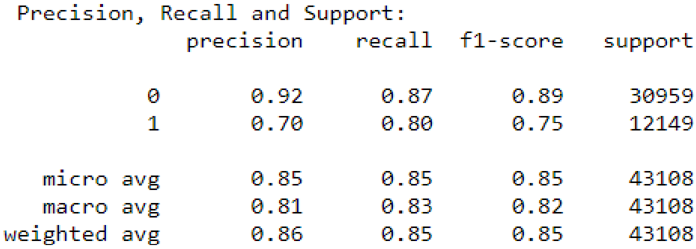


```
Model Report:
Model score error with cv : Mean - 0.8673 | Stdf - 0.0807
```

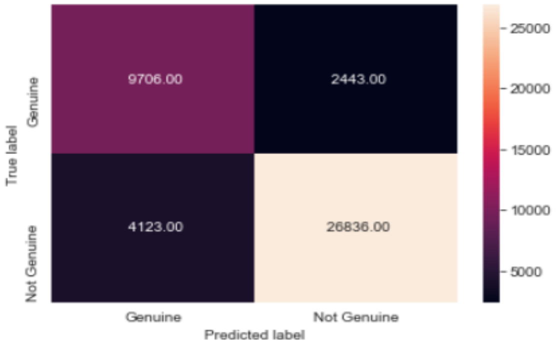


The mean model score is 0.867 for the 5-fold cross validations. The result of the prediction of trained model on test data is given below:

**Confusion Matrix:**



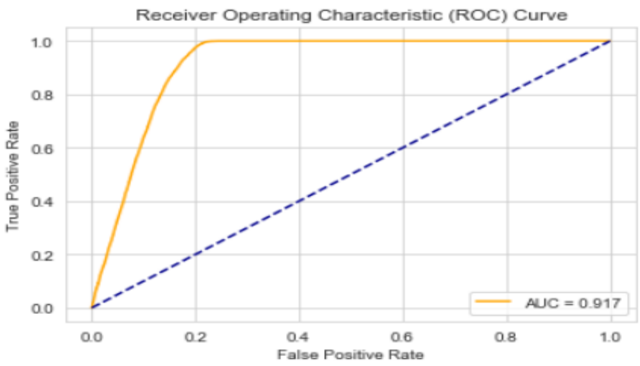
```
 Precision, Recall and Support:
              precision    recall  f1-score   support

           0       0.92      0.87      0.89     30959
           1       0.70      0.80      0.75     12149

   micro avg       0.85      0.85      0.85     43108
   macro avg       0.81      0.83      0.82     43108
weighted avg       0.86      0.85      0.85     43108
```



Through gradient boosting, we got the F1- score of 0.85 with Specificity (True negative rate) 0.87.

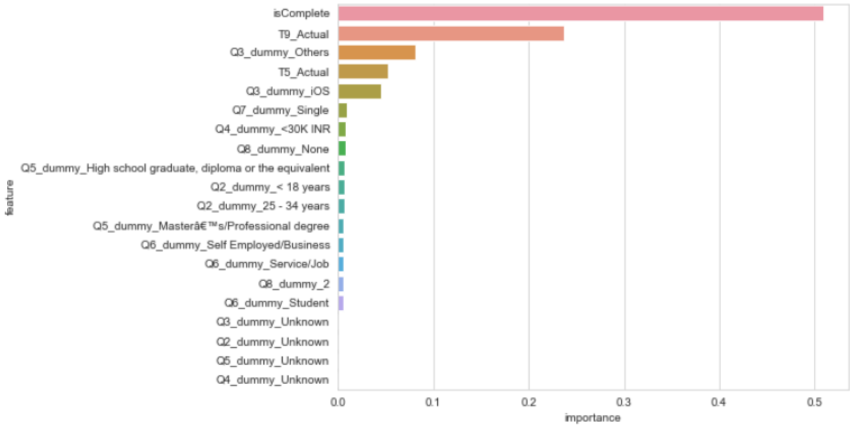
There are only 4123 false positive cases. The precision and recall are better compared to logistic regression and decision tree classifiers.

```
AUC : = 0.917
```



The ROC AUC curve of gradient boosting model is shown in the figure. The model gives an AUC value 0.917.

The feature importance scores of Gradient boosting classifier sorted in descending order is given below:

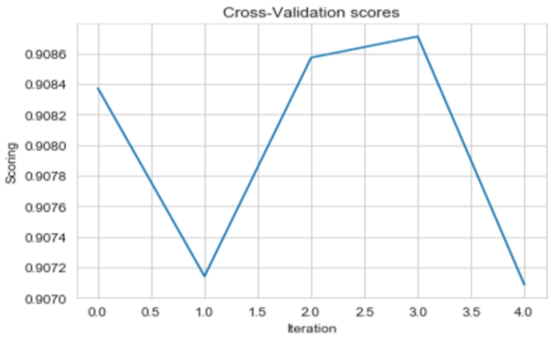


The top 5 features are isComplete flag, T9_Actual, Q3_dummy_Others, T5_Actual and Q3_dummy_iOS.

5. **Support Vector Machines (SVM)**:

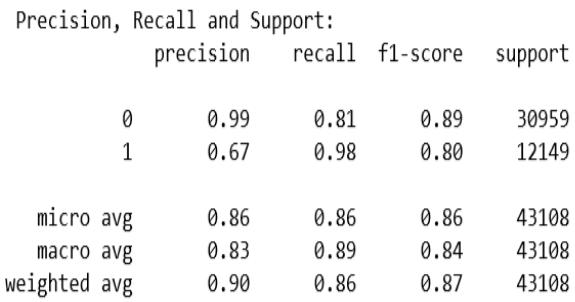
Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection.

The cross-validation scores curve for support vector classifier model training is shown below:

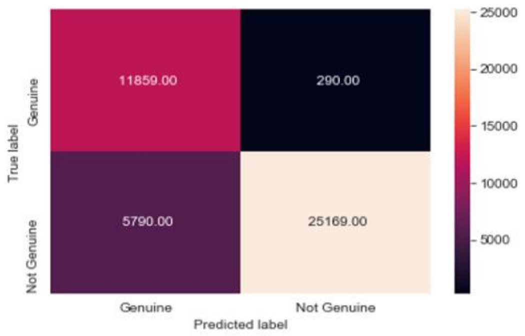


The mean model score is 0.90 for the training data. The result of the prediction of trained model on test data is given below:

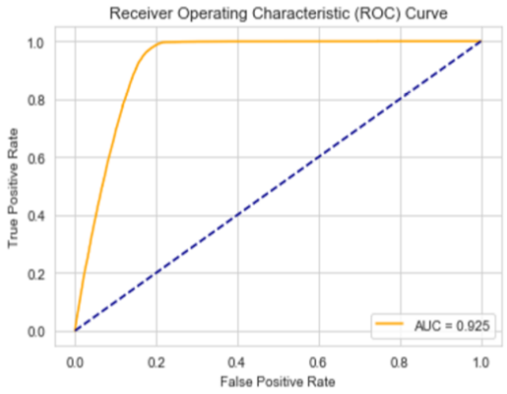
**Confusion Matrix:**



```
Precision, Recall and Support:
              precision    recall  f1-score   support

           0       0.99      0.81      0.89     30959
           1       0.67      0.98      0.80     12149

   micro avg       0.86      0.86      0.86     43108
   macro avg       0.83      0.89      0.84     43108
weighted avg       0.90      0.86      0.87     43108
```



Through SVC, we got the F1- score of 0.87 with Specificity (True negative rate) 0.81 which is less compared to random forest and gradient boosting. The model is good at classifying true positives. There are only 290 false negative cases but 5790 false positive cases.



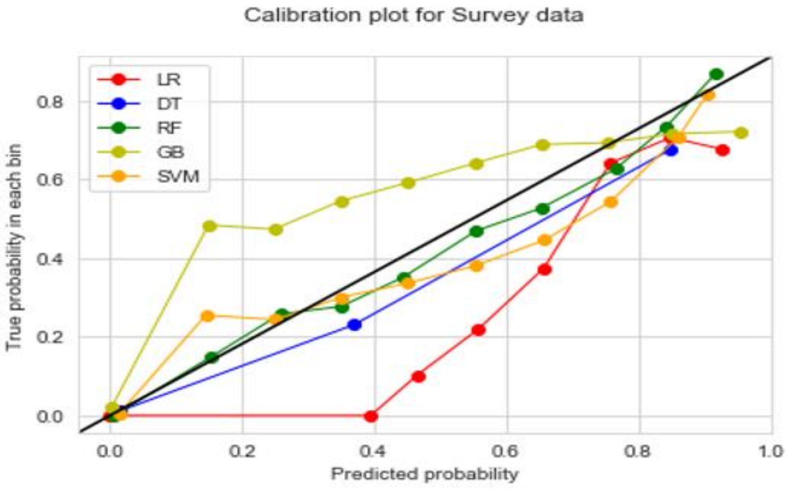
The ROC AUC curve of support vector classifier is shown in the figure. The model gives an AUC value 0.925 .

## CHAPTER – IV

**Main results:**

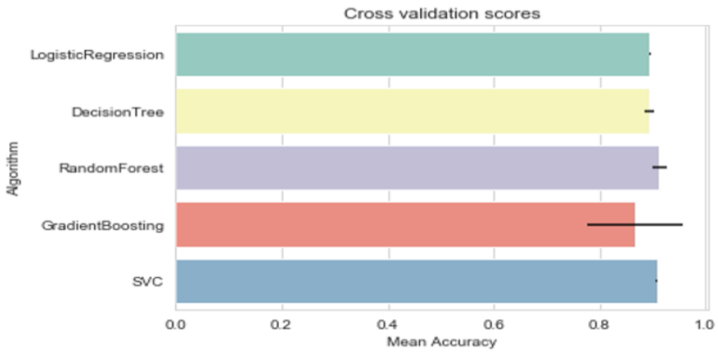
### 4.1 Probability calibration curve:

The objectives of this project is, not only to predict the class label, but also obtain a probability of the respective label. This probability gives the stakeholders some kind of confidence on the prediction. Some models can give you poor estimates of the class probabilities and some even do not support probability prediction. The calibration curve allows you to better calibrate the probabilities of a given model, or to add support for probability prediction.

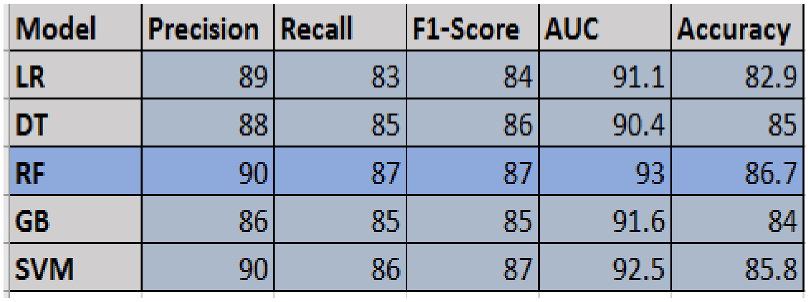


From the calibration plot we can observe that Random Forest returns a well calibrated predictions while Logistic regression and GB performs very badly. Decision Tree and SVM are also giving good performance.

## 4.2 Cross Validation scores:



Shown above the cross-validation scores of training data for each model and the corresponding model prediction scores on test data is given below.

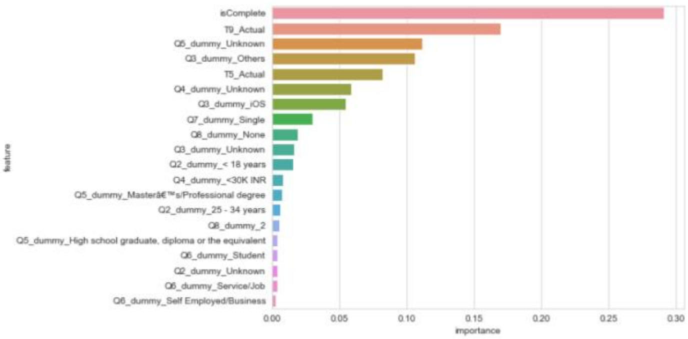


| Model | Precision | Recall | F1-Score | AUC | Accuracy |
|---|---|---|---|---|---|
| LR | 89 | 83 | 84 | 91.1 | 82.9 |
| DT | 88 | 85 | 86 | 90.4 | 85 |
| RF | 90 | 87 | 87 | 93 | 86.7 |
| GB | 86 | 85 | 85 | 91.6 | 84 |
| SVM | 90 | 86 | 87 | 92.5 | 85.8 |

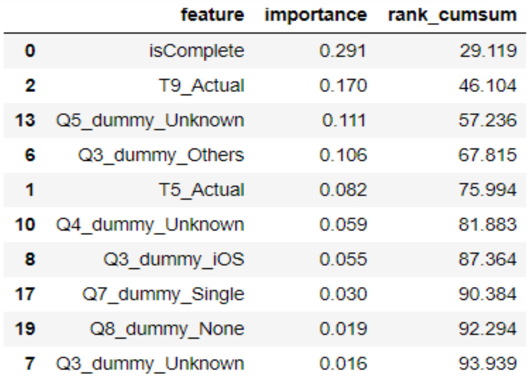
From these analyses selected Random forest classifier as a good model compared to other models.

## 4.3 Final Model:

The top 5 features are isComplete flag, T9_Actual, Q5_dummy_Unknown, Q3_dummy_Others, T5_Actual:



The importance score is normalized and shows the relative importance of features.



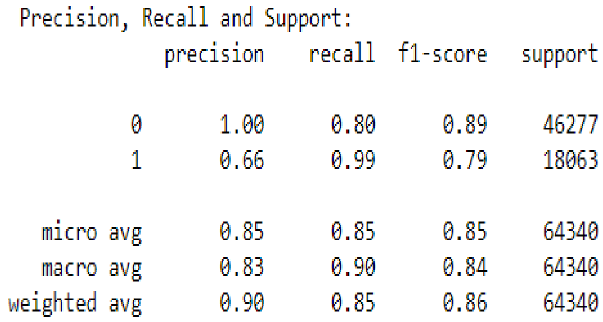
| | feature | importance | rank_cumsum |
|---|---|---|---|
| 0 | isComplete | 0.291 | 29.119 |
| 2 | T9_Actual | 0.170 | 46.104 |
| 13 | Q5_dummy_Unknown | 0.111 | 57.236 |
| 6 | Q3_dummy_Others | 0.106 | 67.815 |
| 1 | T5_Actual | 0.082 | 75.994 |
| 10 | Q4_dummy_Unknown | 0.059 | 81.883 |
| 8 | Q3_dummy_iOS | 0.055 | 87.364 |
| 17 | Q7_dummy_Single | 0.030 | 90.384 |
| 19 | Q8_dummy_None | 0.019 | 92.294 |
| 7 | Q3_dummy_Unknown | 0.016 | 93.939 |

The top five features provide 75% of the information in the data with respect to the outcome variable.

## 4.4 Evaluation of Random Forest model on Validation data:

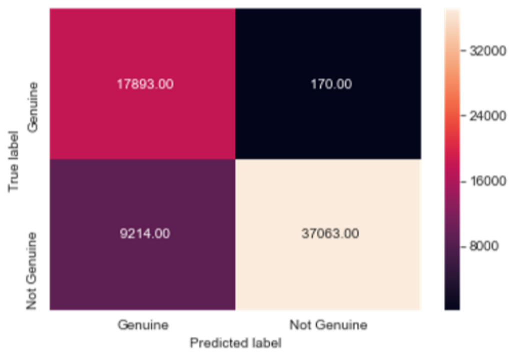
Validated the random forest model on an unseen validation data set and the results are shown below.

### Confusion Matrix:



```
Precision, Recall and Support:
              precision    recall  f1-score   support

           0       1.00      0.80      0.89     46277
           1       0.66      0.99      0.79     18063

   micro avg       0.85      0.85      0.85     64340
   macro avg       0.83      0.90      0.84     64340
weighted avg       0.90      0.85      0.86     64340
```

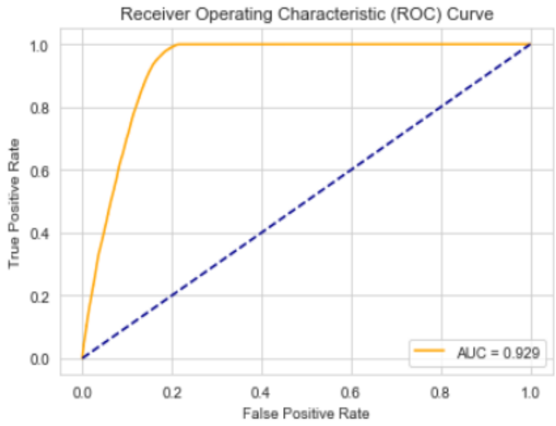
Confusion Matrix:



The random forest model is classifying the validation data with an F1 score of 0.86 and the false negatives (non-genuine classified as genuine) are 9214 out of 46277 which is a good score. The model is very good at classifying genuine records as genuine.

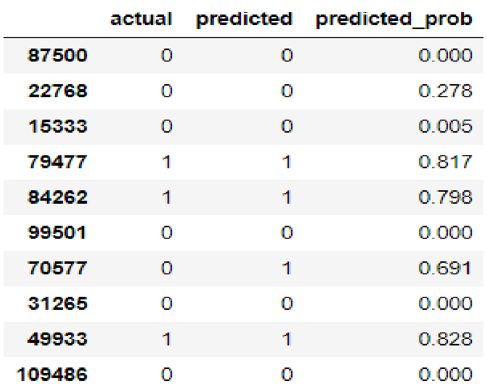
### ROC -AUC curve:

```
AUC : = 0.929
```



The model has an AUC value of 0.929 on validation data set which is approximately equal to the AUC of training data. Hence, we can trust the model as it works well on the validation data.

The figure shows a snapshot of the probability predicted by the model for the data set.



| | actual | predicted | predicted_prob |
|---|---|---|---|
| **87500** | 0 | 0 | 0.000 |
| **22768** | 0 | 0 | 0.278 |
| **15333** | 0 | 0 | 0.005 |
| **79477** | 1 | 1 | 0.817 |
| **84262** | 1 | 1 | 0.798 |
| **99501** | 0 | 0 | 0.000 |
| **70577** | 0 | 1 | 0.691 |
| **31265** | 0 | 0 | 0.000 |
| **49933** | 1 | 1 | 0.828 |
| **109486** | 0 | 0 | 0.000 |

## CHAPTER – V

### 5.1 Conclusion:

To summarize, based on the current metrics that are captured via a survey, we are able to predict if a consumer has responded to a survey truthfully or not, with an approximate accuracy of 92%.

The most important contributing factor is this prediction is the completeness of the responses from each consumer who has taken the survey. Other factors like education and gender play a very subtle role towards genuineness, although we cannot say for sure that a certain gender or education segment is more genuine that the others.

The platform for taking the survey, whether android or iOS, is not a contributing factor. Salary bracket, marital status contributes neutrally, which means they are not very helpful in differentiating a genuine response from a non genuine one.

Caveat: This model will be tested on a completely unseen dataset that is with inMobi and has not been used in the model training and validation process. The results of that test set should provide more insights.

### 5.2 Recommendations:

- As we have already seen, response completeness is a major contributor so any effort to drive the surveys in a way where consumers are encouraged to respond till the end would make the model more efficient and better.
- Including more metrics that indicate specific consumer behavior or helps create consumer cohorts with defined behavioral traits would further enable the analyst to understand the thought process of the survey taker and help the model predict better.
- Administering the survey at different time periods might also provide some insights into the consumer. This could be an iterative exercise, to collect multiple samples of responses in different time durations of the day, week month etc., and then further analysis could reveal if when the survey is taken affects the genuineness. This would also help in the overall understanding required for effective marketing and ad placement strategies.