B

Nikhil Aggarwal

Ridhima Puri

bqom 2578: data mining

Project Report: Application Behavior Analysis

# Motivation

"You've got to keep reinventing. You'll have new competitors. You'll have new customers all around you." -- Ginni Rometty, IBM CEO

Ginni Rometty has addressed the all the aspects of todays industry in one saying. The market today is very competitive for any kind of businesses. With the advent of technology, every operation inside any organization is becoming more and more effective. Today’s firms are not only making competitive products but are also craving for the attention of their customers. Almost every technology service provider has its mobile application. These mobile applications are a way to attract the users and track user activities in the application. Often, these companies provide free products/services in their mobile apps to transition their customers to a paid membership. Some of such applications are YouTube, Instacart, cam scanner and Pandora Premium to name a few.

This requires heavy marketing and a huge amount of budget in the form of offers and promotions is spent for the same. Therefore, it is very important for a company to know their target market and direct their major marketing efforts towards the users who will not subscribe to the paid memberships. We are trying to study the user app behavior and predict such users.

# Literature Review

Analyzing user behavior is a trend. There are many studies conducted on the same. The study on An Empirical Analysis of User Content Generation and Usage Behavior on the Mobile Internet (Ghose & Pil Han, 2011) also studies how the content generated in the mobile phone internet relates to the user behavior. (S.Tseng & W.Lin, 2006) have also studied mining and prediction of user behavior patterns in mobile web systems. The industry also, understands the importance of studying the user activities and tuning the products according to the consumer demands. Google analytics is used in most of the firms for studying the user activities, the clicks and the screens visited.

We are trying to apply the same idea to predict which users are not willing to subscribe to an application by studying what screens and features they use. How many users are using the premium features and the top screens and then getting enrolled in the subscription programs? We are also trying to understand the factors that may influence such activities like day of the week, some promotional games provided by the application provider, the first time the user started using the application.

# Introduction to Data Set

We are using data posted at Kaggle under the name “Fintech\_App\_Behaviour\_Analysis”. The data has 2 files. First is the app data consisting of 50K rows and 12 columns. Second is the top screens data consisting of 58 rows and 2 columns. The data gives an overview of time and date of app installation, the features used by the users and the usage of the app. Given below is the link to the data: https://www.kaggle.com/nanda1331/fintech-app-behaviour-analysis

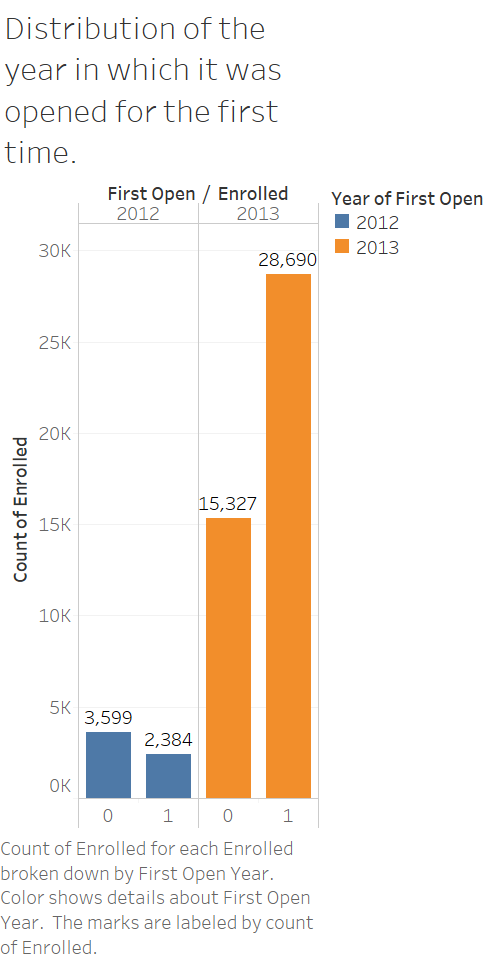
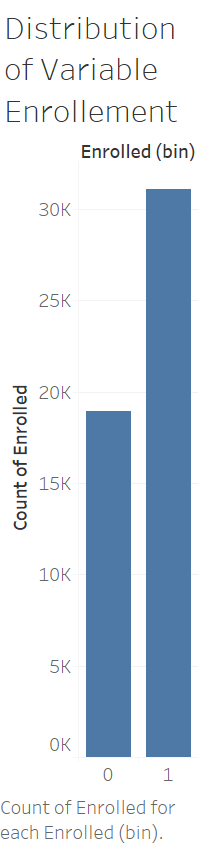
# Data Cleaning

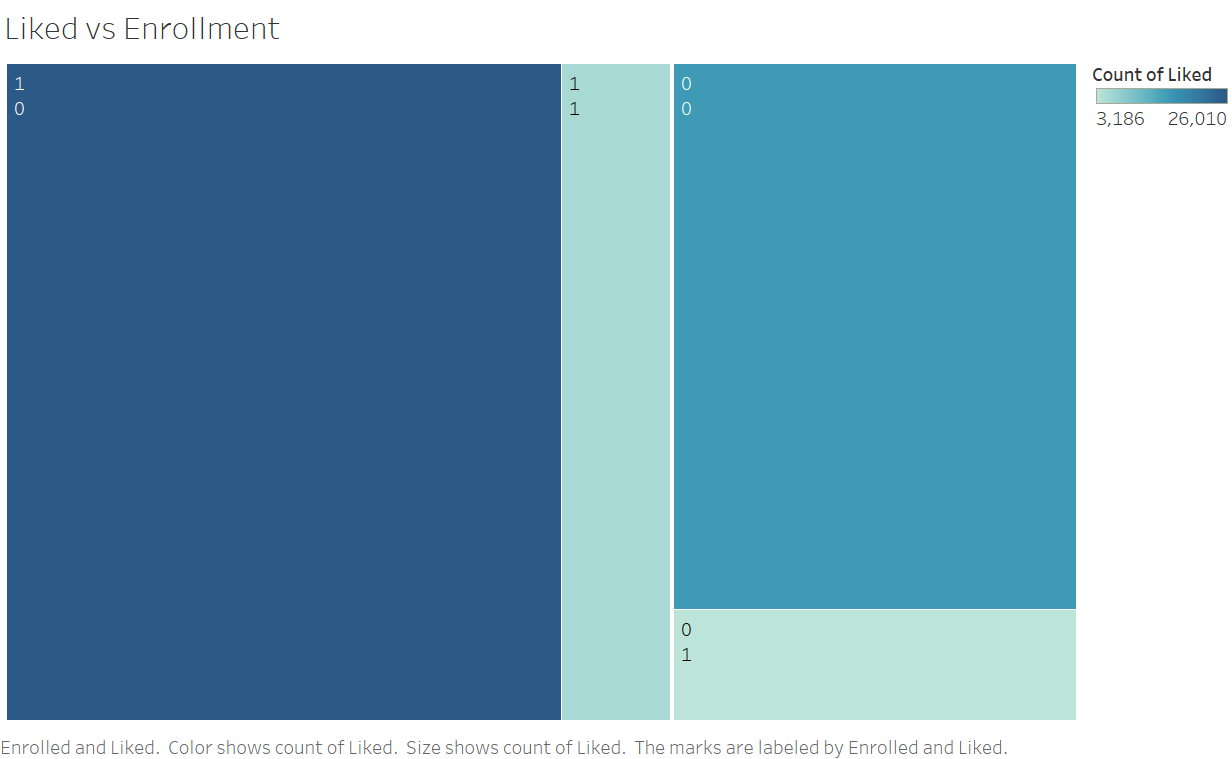
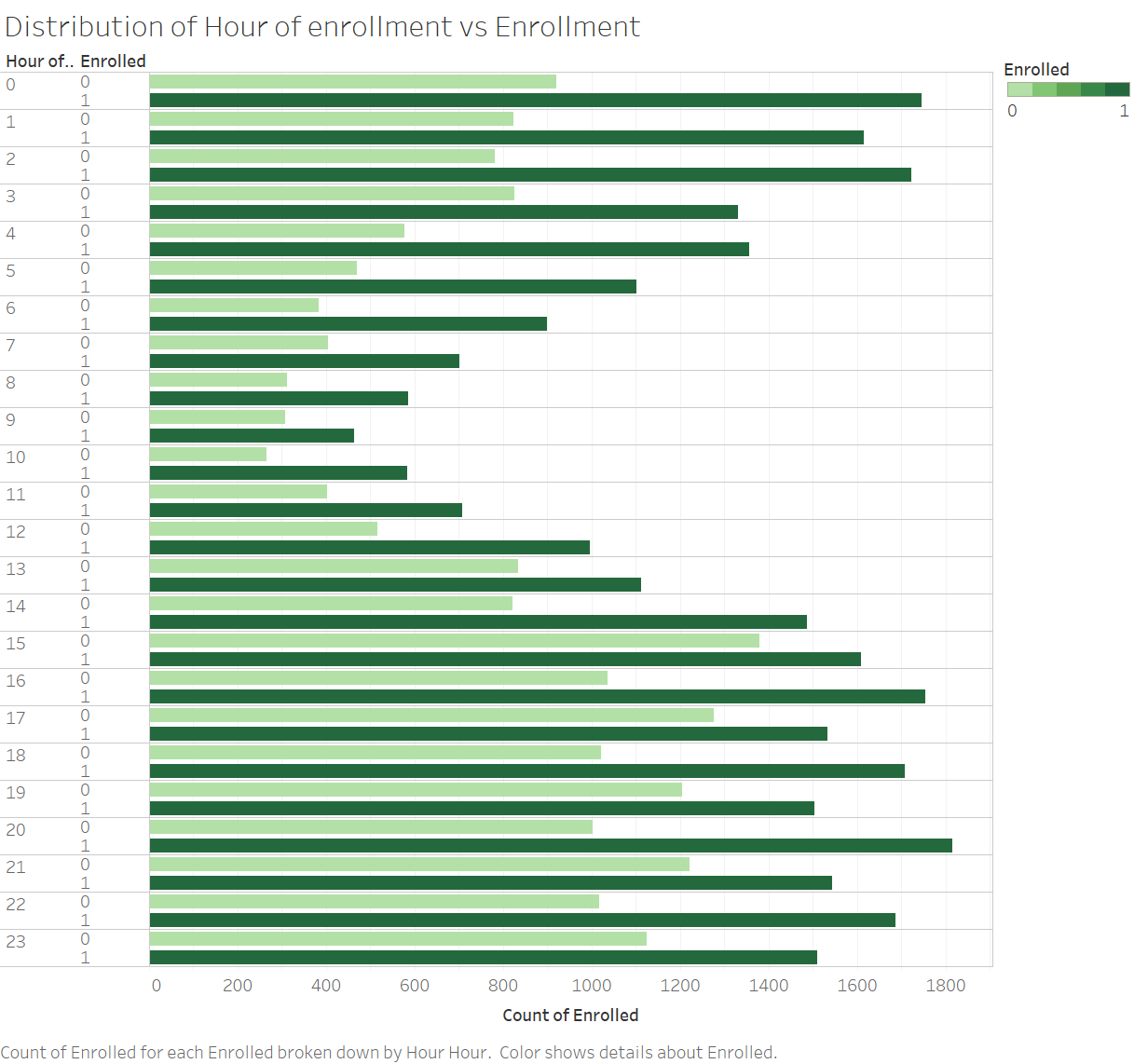
Data cleaning was an extensive process in our case. We looked for missing values in our data and we found that there all the variable of our interest has complete data. The challenge was with screen list variable. This variable contains the information about the various screens visited by a user while using the app. The number of screens visited could be different for every user and this variable consisted of all the comma separated values. We could not ignore such a variable for our analysis. We were also given the list of top screens in the app in a separate csv file. The main challenge was to use the information from both the files and make some sense out of data.

Our final approach was to use the top screens data and generate additional columns in our main data set which consisted of 0 as a default value (indicating not visited). These columns were equal to the number of top screens present. For every screen visited by a user the column with that name was marked as 1 i.e. user visited this screen. This was run for all the users in our dataset.

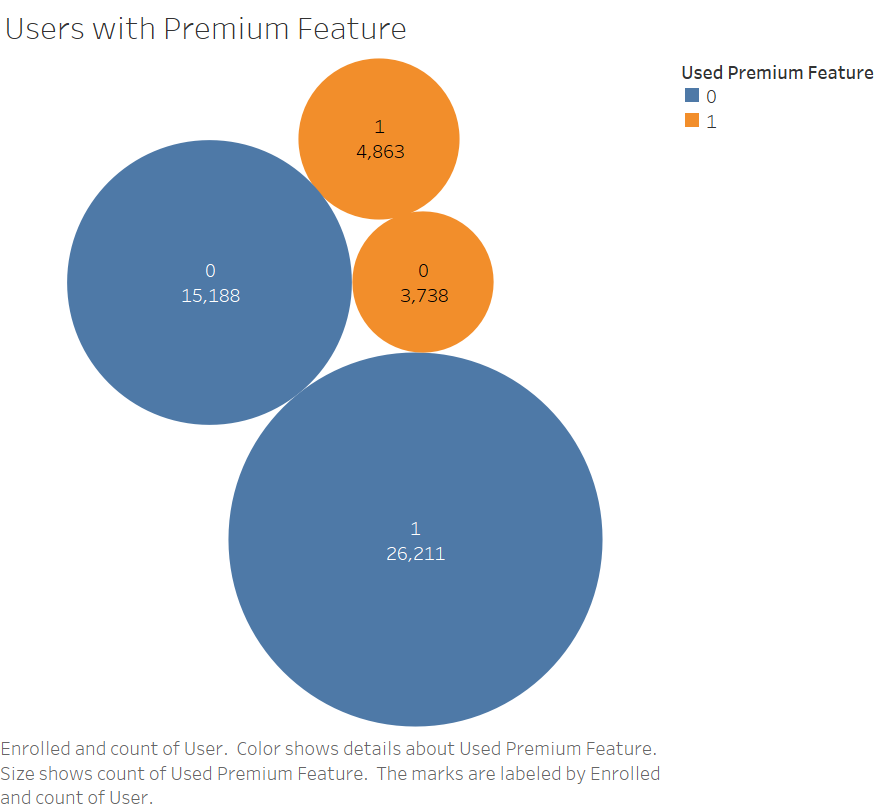
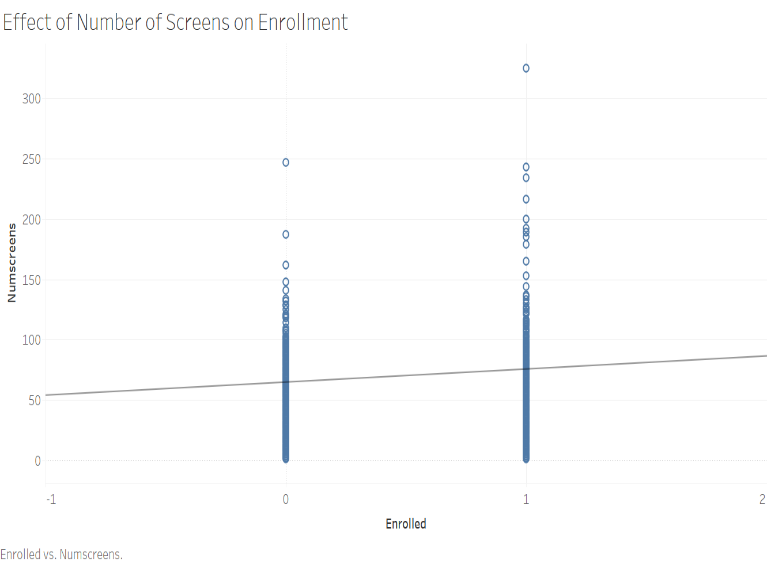
# Data Visualization

We have used Tableau and R for visualization. The main purpose of visualizing the data with the help of graphs is to see and understand trends, outliers, and patterns in data. We saw various interesting patterns while visualizing the data. We have greater number of enrolled users than not enrolled. The number of users coming and using the app and the number of users enrolling for the app are higher in 2013 as compared to 2012. Additionally, more users were active during the off-working hours and thus the late evening and the night time had greater people enrolling for the app. We also saw a strange relation between the dependent variable and liking the app. Liking the app is also negatively affecting the app.



As per our analysis of data graphically, we also saw that the age and using premium features are not increasing the number of enrollments, whereas using a greater number of screens in the app is positively affecting the enrollment. We also looked into the correlations between the variables and found no signs of multicollinearity.



# Data Analysis with Modelling Techniques

After having a good understanding of data, we ran the modelling techniques for our problem. Our problem is a classification problem and the dependent variable is enrolled which indicated if a person has enrolled for the subscription or not. Since we want to help marketing team to target the people having a lower conversion chance, therefore we aim for a higher specificity of our model. Specificity is the power to correctly identify the number of people having a lower chance of enrolling in our model. In our data enrolled date is the date on which the person enrolled for our application; thus, we have not considered it in our prediction model.

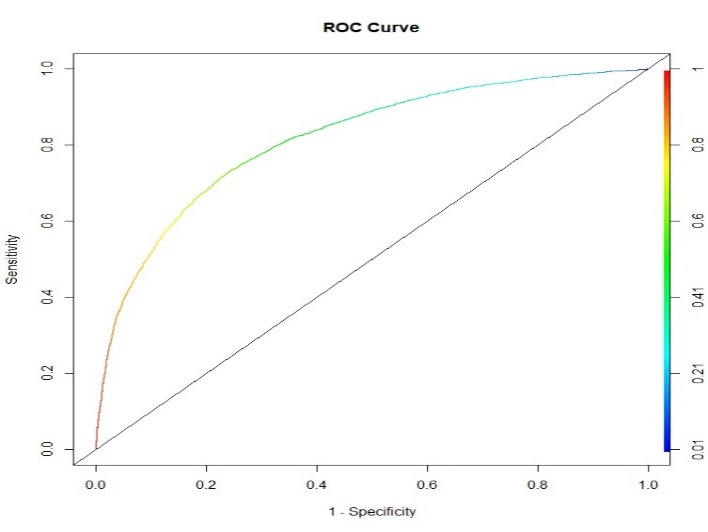
## Logistic Regression

In statistics, the logistic model (or logit model) is a widely used statistical model. In its basic form it uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model; it is a form of binomial regression.

To use this modelling technique in for our data, we have split our model into training and testing sets with 70% data in the training set and 30% in the test set. We had one variable i.e. screen list, which listed all the screen visited by the user as a comma separated string. We excluded this string variable for our first try and after building the model

The very first step was to build a logistic model with all the variables except the screen list variable. This model had a very low prediction power and we were not using one of the most significant variables for analyzing the behavior of the users. We again built a model with modified data set having all the screen list as the columns and if the values of the field are 1, it means the user visited the screen. The logistic model built with these variables could also tell us which screen has most affect on enrollment. Refer [**Appendix 1**](#_Appendix_1:) for list of all significant and Non-Significant Variables.

As we see the significant variables, we can tell what time of the day and what screens have most effect on our dependent variable. The next challenge was to choose a cut off probability for our model that supports and compliments the prediction. To tune our model, we focused on increasing the specificity of the model as we wanted to correctly identify the population for whom marketing efforts are required. Therefore, we tried various cut off values keeping in mind that we need higher specificity. After looking at various cutoff probabilities and sensitivity and specificity values we are using 0.6981603 as our cutoff probability. This is the probability where we get about 82% specificity and 64% sensitivity. The area under the roc curve is about 82%.

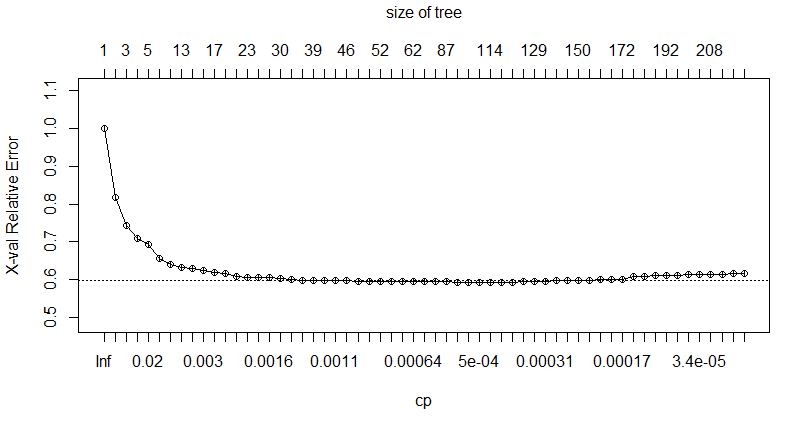


## Classification Tree

Classification and Regression Trees or CART for short is a term introduced by Leo Breiman to refer to Decision Tree algorithms that can be used for classification or regression predictive modeling problems. It provides a foundation for important algorithms like bagged decision trees, random forest and boosted decision trees. In our dataset, since our dependent variable is a categorical variable, we will be making classification tress for prediction.

To start with the prediction model, we split the dataset into two sets i.e. train set and test set keeping 70% of the data in training set and remaining 30 percent in testing dataset. Then, we build our model using complex parameter = 0 and minimum bucket size of 30. Since, complexity parameter is a significant factor in determining the size of the tree, for the base case, we build a tree with cp as 0. Next, we build the cross-validation graph for the tree that we build and looked for the most optimal value of cp for which the error is minimum. Based on the graph, we found that optimal value of cp between 0.001 and 0.004. We build different trees for values between these ranges and calculated both sensitivity and specificity for all these trees. Finally, we selected cp = 0.0014 as for this cp, value of sensitivity (78.97%) and specificity (76.66%) is maximum. Overall accuracy of the model is 78.10%.

Refer [**Appendix 2**](#_Appendix_2:) for Confusion Matrix, [**Appendix 3**](#_Appendix_3:) for Final Decision Tree and [**Appendix 6**](#_Appendix_6:) for source code.



## Random Forest

Random forests or random decision forests are an ensemble learning method for classification and regression trees. They operate by constructing a multitude of decision trees using training data and predicting the class of the dependent variable that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Again, to build Random Forest, we split our dataset into training and test data set keeping 70% of observations in training set and remaining 39 percent in test set. To build a model, we need to find out the optimal value of factor mtry. In order to calculate that, we build two different data frames, one consisting only dependent variable and other containing all independent variables. We used tuneRF function of library Random Forest to calculate value of mtry. Based on the graph, we found mtry =12 is having lowest OOB error. Hence, we selected value of mtry to be 12.

Now, we build our model keeping node size = 30 and number of trees to be 300. Looking at the cross-validation graph of this model, we found that the value of error stops changing after 100 trees. Hence, to reduce the complexity and for a better prediction, we selected the number of trees to be 100 which is also a significant amount. Finally, we used this model to predict our dependent variable using test data and we compared it with values of test set. We build a confusion matrix between the actual and predicted values and found sensitivity to be 83.95%, specificity to be 73.88% and overall accuracy to be 80.14%.

|  |  |
| --- | --- |
|  |  |

Refer [**Appendix 4**](#_Appendix_4:) for OOB Error with 300 trees, [**Appendix 5**](#_Appendix_5:) for Confusion Matrix and its statistics and [**Appendix 7**](#_Appendix_7:) for source code.

# Managerial Implications and Summary

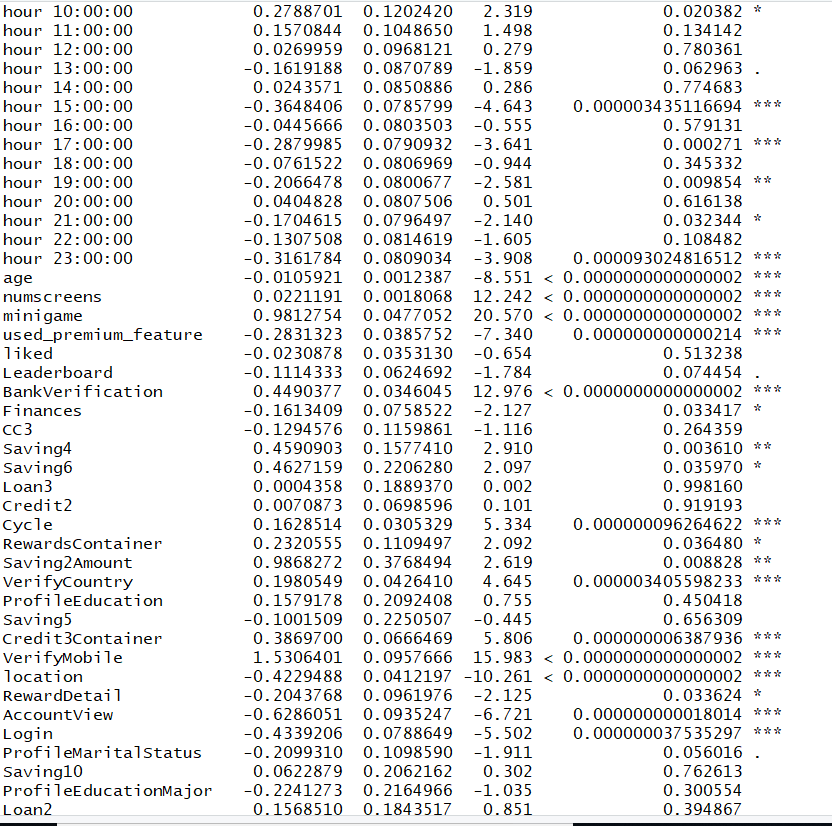
Out of the three models implemented, the logistic regression model gives us the best predictive power to classify the people who would not enroll for the app. This model has other advantages as well. This is a transparent model and gives the information of all the variables which are impacting the dependent variables along with the direction and the magnitude of the effect. Therefore, if a manager wants to decide on which day in the week, what time of the day or age group to target for promotions could be easily explained by this model. We can also increase our promotion for the screen which are used by the people the most and are positively impacting the enrollment.

Studying user behavior in general can have many managerial usages.

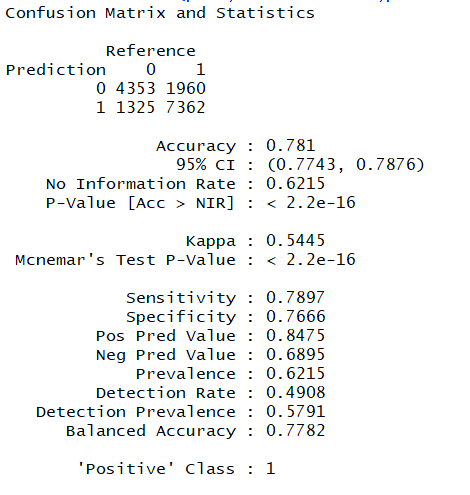
1. We can study the screens used the most by any user and use this information for advertisements.
2. We can put a greater marketing effort for the users who have lower chance of conversion.
3. By studying the features used the most, we can lure the consumers to buy subscription by enhancing such features.

# APPENDIX

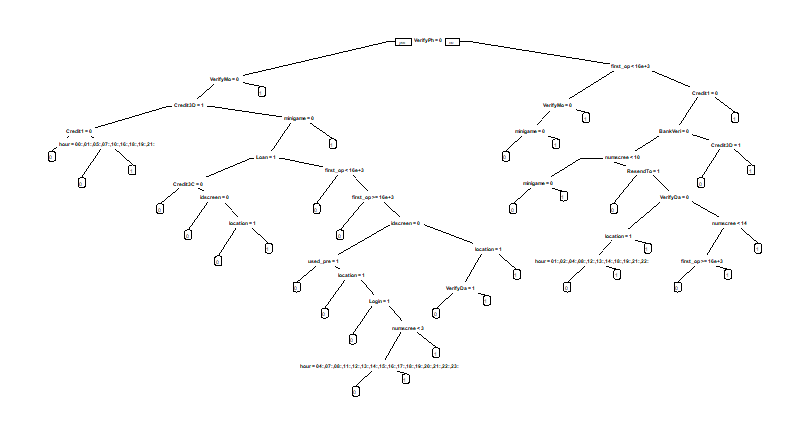
## Appendix 1:



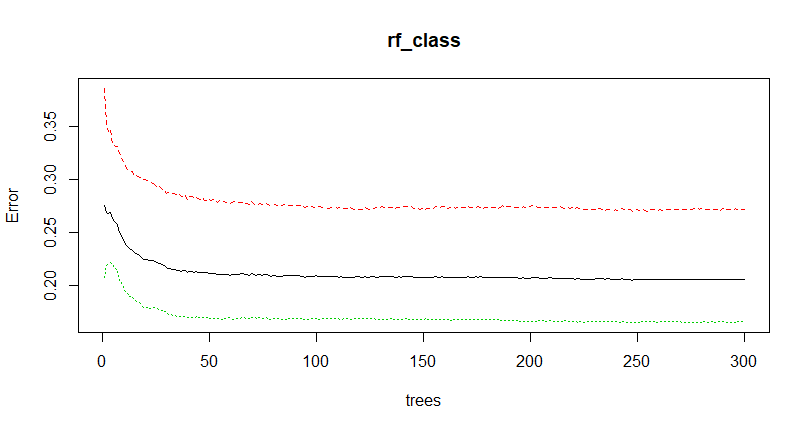
## Appendix 2:



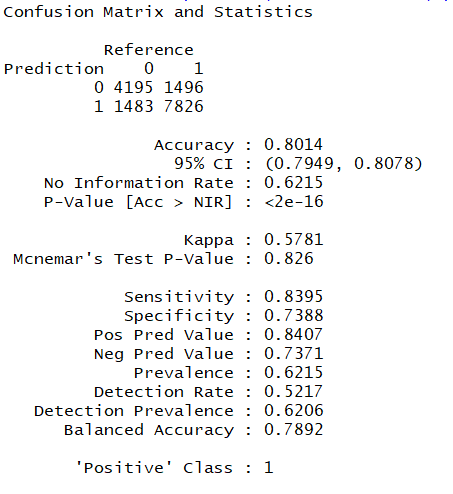
## Appendix 3:



## Appendix 4:



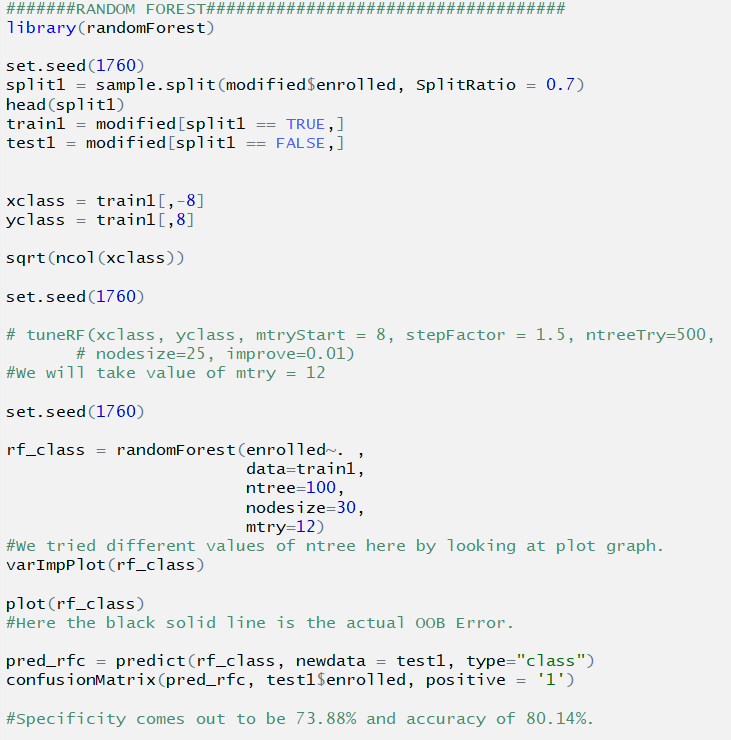
## Appendix 5:



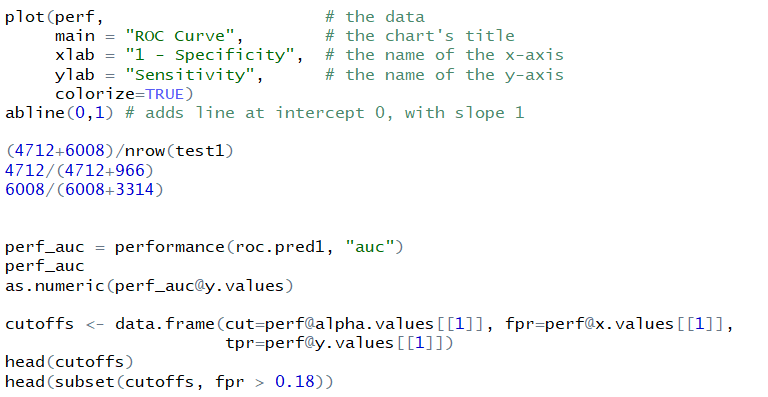
## Appendix 6:



## Appendix 7:



## Appendix 8:



## Appendix 9:

