Problem 3

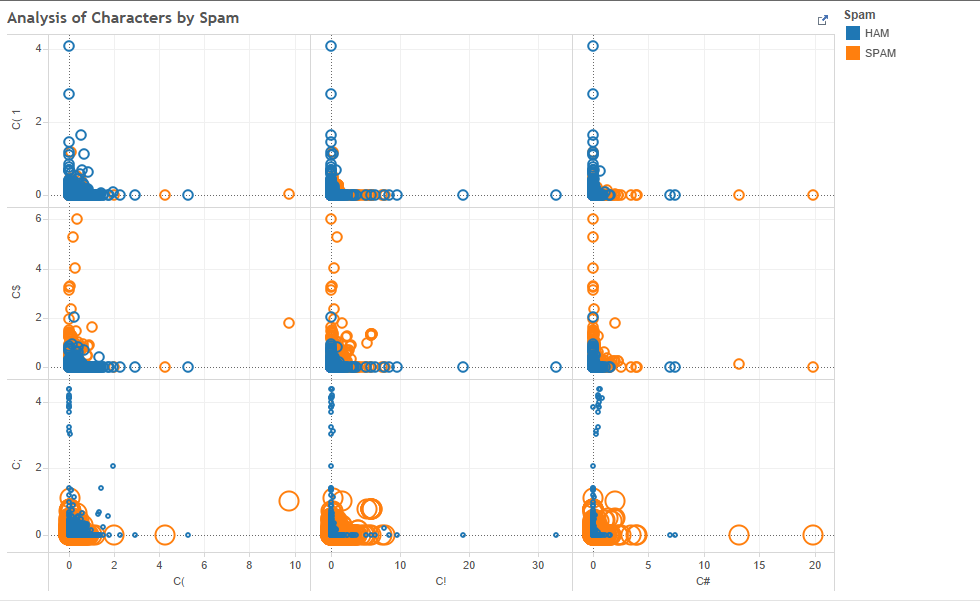
Analyzing the dataset

* There are total 48 attributes of type words
* Total 6 attributes of type characters like C
* 1 real attribute of type average of CAPITAL LETTERS
* 2 integer attributes of type total and longest of the length of CAPITAL LETTERS

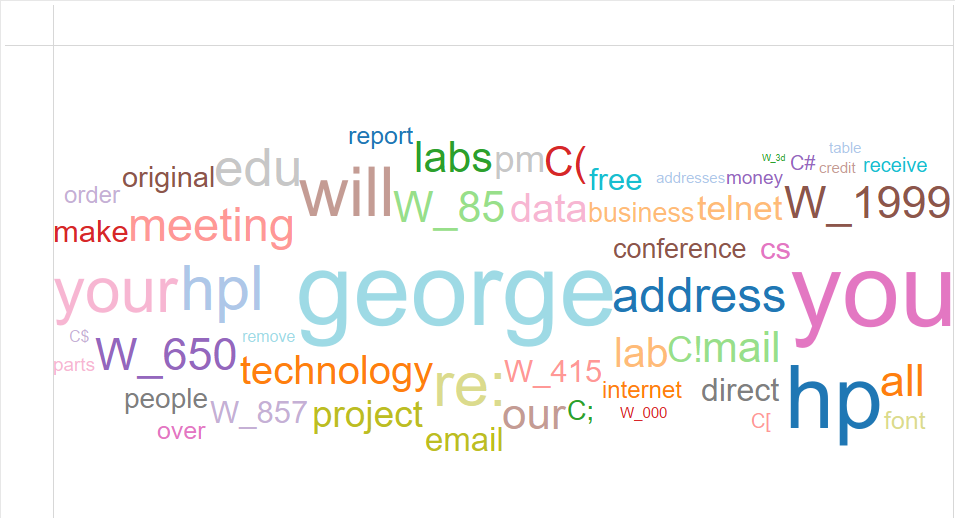
3.1 **Exploratory Analysis**:

Tool: Tableau

* Scatter Plot of Characters in the list by Spam



* Word Clouds showing Spam and Ham words by average of the percentage of the words





3.2 **Building Models**

Tool: XLMiner

Steps:

1. Data Screening and Pre-processing
2. Performed Feature Selection: Feature selection is performed to select a subset of relevant features for use in model construction. Performing feature selection gives you the chi-squared P-Value which helps in determining whether a predictor accepts or reject ‘Null Hypothesis’. There is no significant difference between observed and expected frequencies

2. Data Partition

1. Data partition has been performed through standard partition option with a 60% training and 40% testing. Random seeds has been set as 12345. Same training and testing partition will be used in all the classification models.

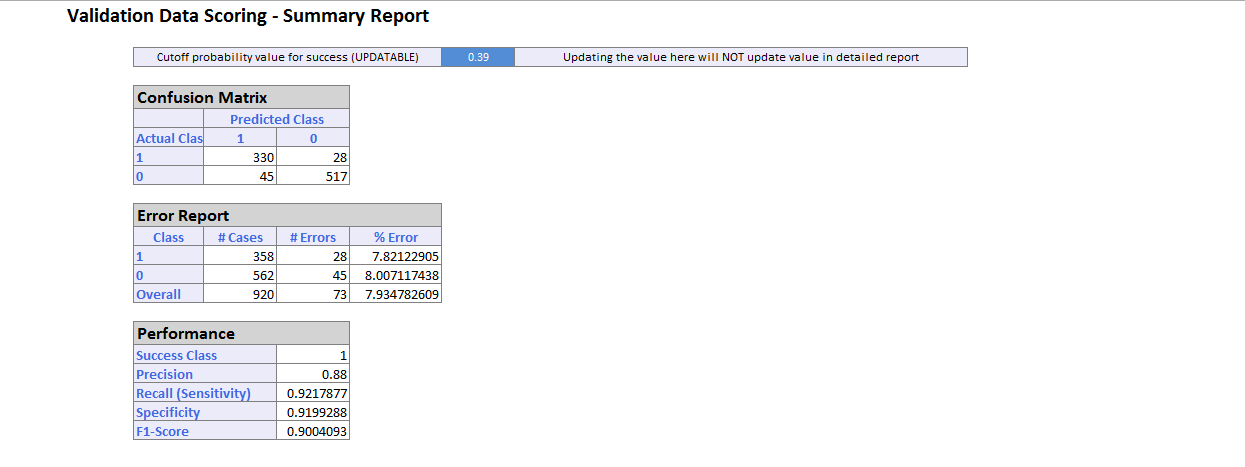
3. Build Models

1. Considering 1813 emails tagged as spam from 4601 emails. The initial cutoff probability of success is taken as 0.39. If the probability of success for an email is less than this value then the email would be a non-spam email and if it greater than this value then the email would be a spam email.
2. Based on the cut-off value: Classification has been performed using the following machine learning algorithms: Logistic regression, CART (Classification and Regression Tree) and Random forest. Default settings are applied for CART & Random Forest.

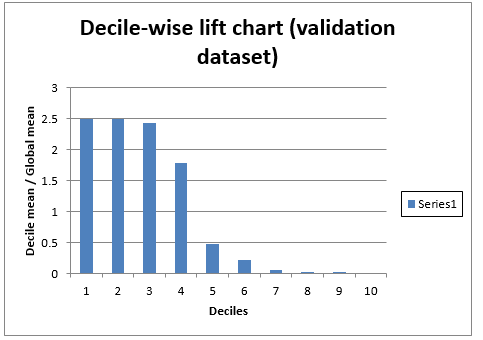
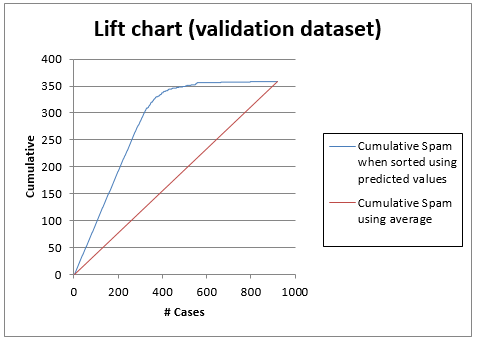
3.3 **Model Evaluation**

Measurement: Confusion Matrix, Lift Chart, Decile-wise Lift Chart

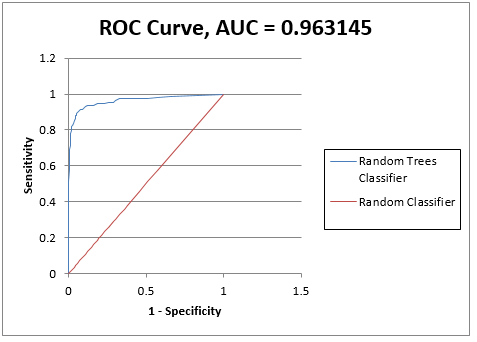
1. Logistic Regression:
2. For detecting spam messages, the error rate of spam messages from the confusion matrix of validation dataset 7.82 %



b. The curve of the lift chart for validation dataset is above the straight line which is good. The decile-wise lift chart of the validation dataset has higher heads and lower tails. The top 2 decile contains 20% of the emails most likely to be spam emails. Whereas the bottom 2 decile contains 20% of the emails least likely to be the spam emails.

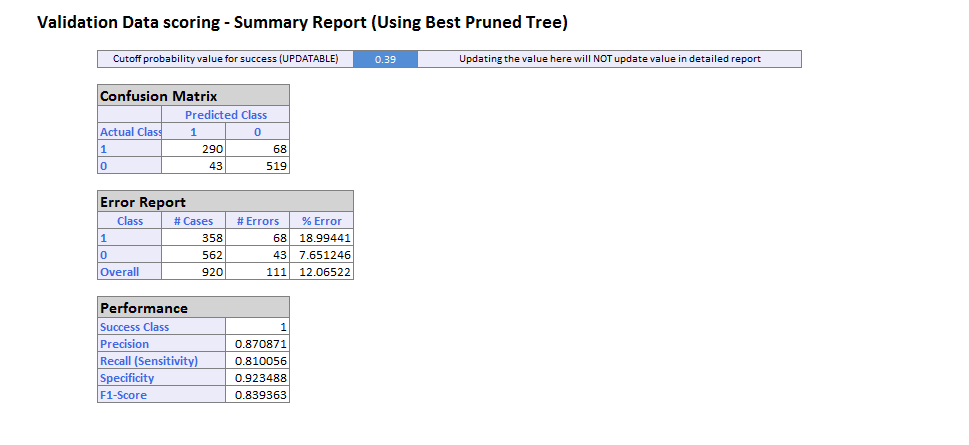


c. ROC Curve of the validation dataset is in the top left corner which indicates that the model performs better model than a random model

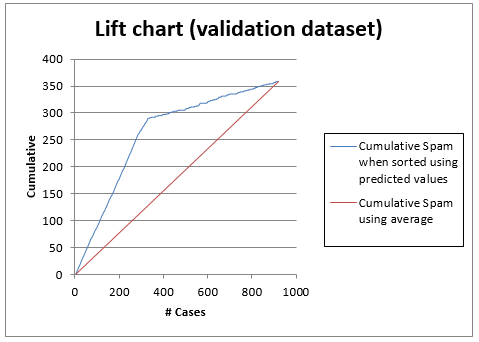
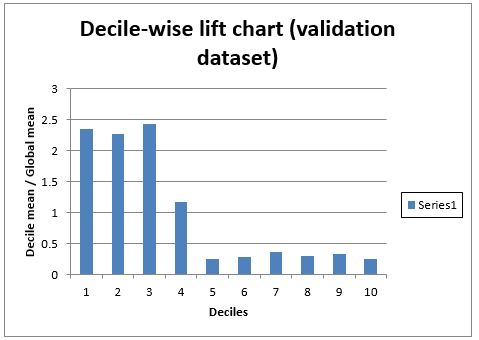


2. CART

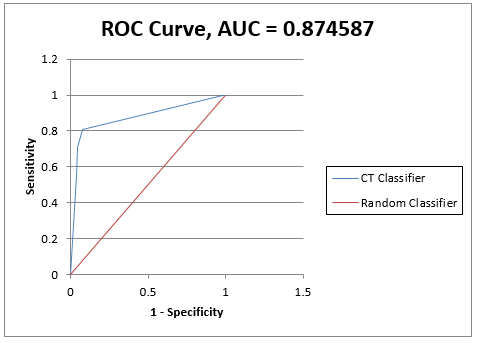
1. The error rate of spam messages from the confusion matrix is about 19% which is high compared to the error rate of Logistic regression



1. The curve of the lift chart for Validation dataset of CART is above the straight line. Whereas in the decile chart the higher and the lower heads are out of order which indicates that the model is not good compared to a random classifier model which should be in good staircase order from left to right

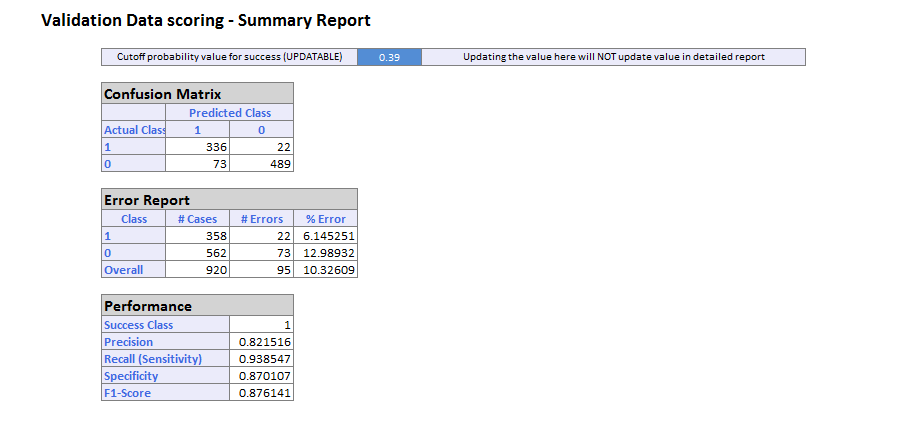


c. The ROC curve is away from top left corner which again indicates that the model is not good.

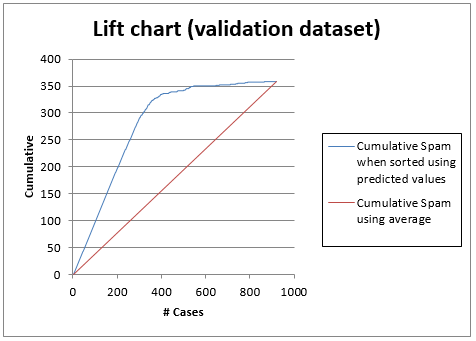
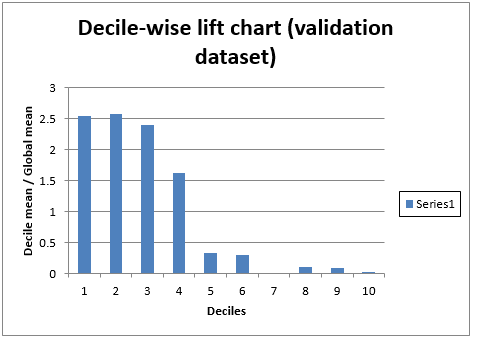


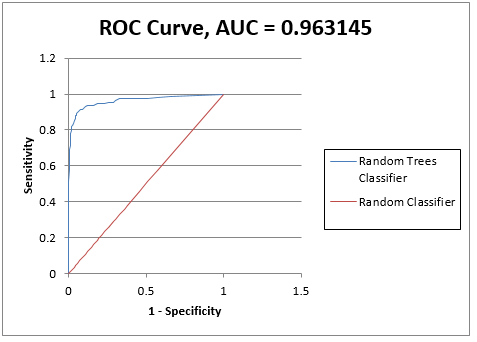
3. Random Forest

1. The error rate of the spam messages from the confusion matrix of Random Forest using Random Forest is 6.14% which is low compared to

Logistic Regression and CART

b. The Lift chart of Random Forest is above the straight line which is good. The decile-wise lift chart shows that the higher heads and the lower tails with decreasing order from which it can be determined that this is a good model compared to a random classifier.

c. The ROC curve of the Random Forest is above the straight line and near the top left corner of the chart 



3.3. Model Comparison:

1. The overall error rate of Logistic Regression is 7.9 % which is lower compared to Random Forest which is 10.32%. But for detecting the Spam emails from the validation dataset the error rate of Random Forest is lower 6.14% compared to Logistic Regression which is 7.82%.
2. CART has the highest overall error rate in the validation dataset. And, again from the decile-wise lift chart and ROC curve it can be seen that the model is not good for detecting the spam emails.
3. For detecting spam emails Random Forest provides better performance compared to Logistic Regression

3.4 **Model evaluation and Recommendation**

1. Assumption: According to the dataset, if we assume the initial cutoff probability of success to be 0.39 or 39% then Random Forest model is good taking into account the error rates which are:

Logistic Regression: 7.9%

CART: 18.99%

Random Forest: 6.14%

If we consider the initial cutoff probability of success to be 0.5 (default) or 50% for detecting spam emails then the error rates are:

Logistic Regression: 12.29%

CART: 18.99%

Random Forest: 10.89%

Hence, considering both the conditions Random Forest is a good model