**FISHERIES DATA ANALYSIS REPORT**

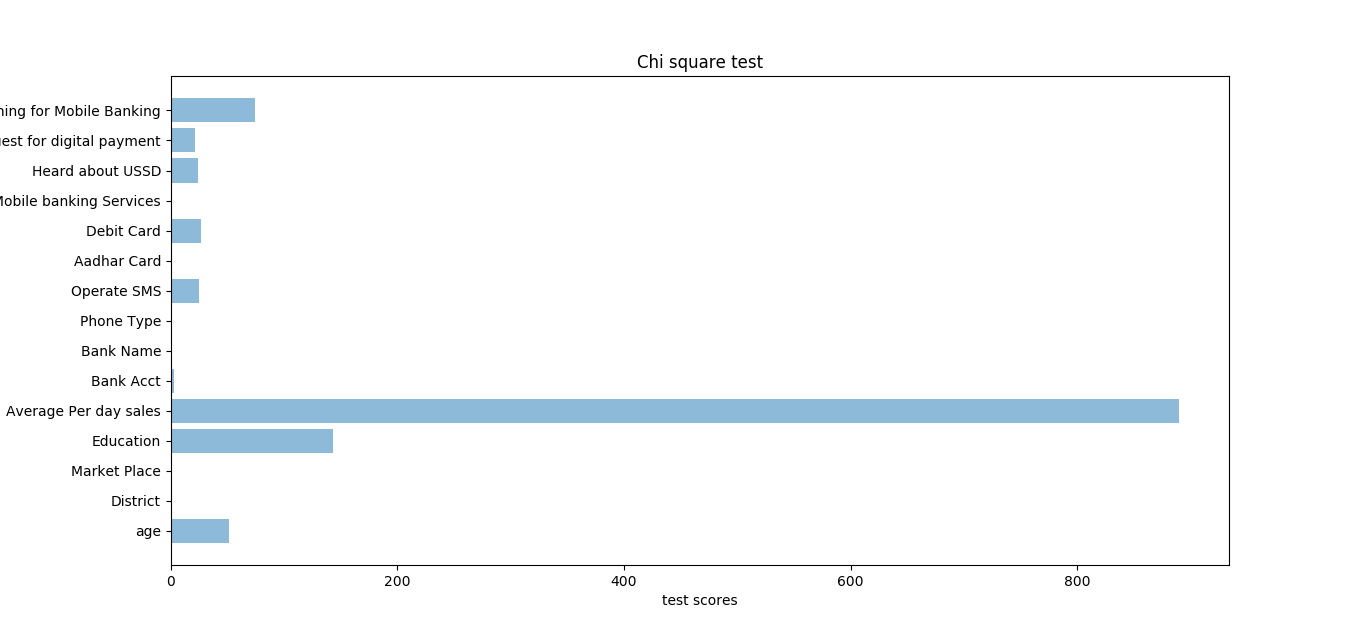
**Chi-square feature selection:**

The chi-square test is a statistical test of independence to determine the dependency of two variables. It shares similarities with coefficient of determination, R²R². However, chi-square test is only applicable to categorical or nominal data while R²R² is only applicable to numeric data.

From the definition, of chi-square we can easily deduce the application of chi-square technique in feature selection. Suppose you have a target variable (i.e., the class label) and some other features (feature variables) that describes each sample of the data.

Chi-square statistics between all of the feature variables and the target variable is calculated and check for the existence of a relationship between the variables and the target. If the target variable is independent of the feature variable, we can discard that feature variable. If they are dependent, the feature variable is very important.

Chi-square test for our data gave the following result:

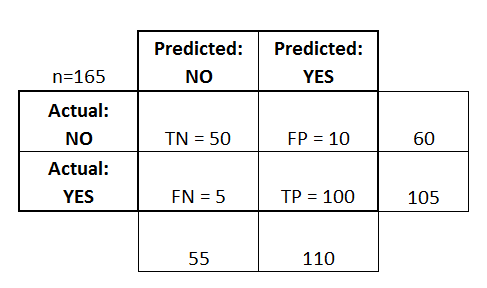


Based on the observations made in the graph we decided that age, education, Average Per day sales, Operate SMS, Debit Card, heard about USSD, Customer request for digital payment, Training for Mobile Banking as factors influencing Willingness to use Mobile Banking.

Results after applying Machine Learning models and Ensemble learning:

**Performance metrics:**

For accuracy benchmarking we use confusion matrix.



**1. Accuracy**: Overall, how often is the classifier correct?

Ex: (TP+TN)/total = (100+50)/165 = 0.91

**2. Precision**: When it predicts yes, how often is it correct?

Ex: TP/predicted yes = 100/110 = 0.91

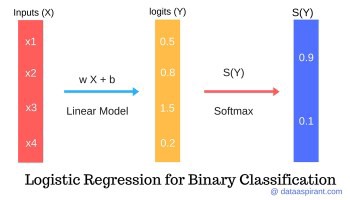
**3. Time:** Time taken to train the model by fitting the Training data and to predict values for the Test data in seconds

**4. ROC-AUC score:** An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds.

An ROC curve plots True Positive Rate (TPR) vs False Positive Rate (FPR) at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The following figure shows a typical ROC curve

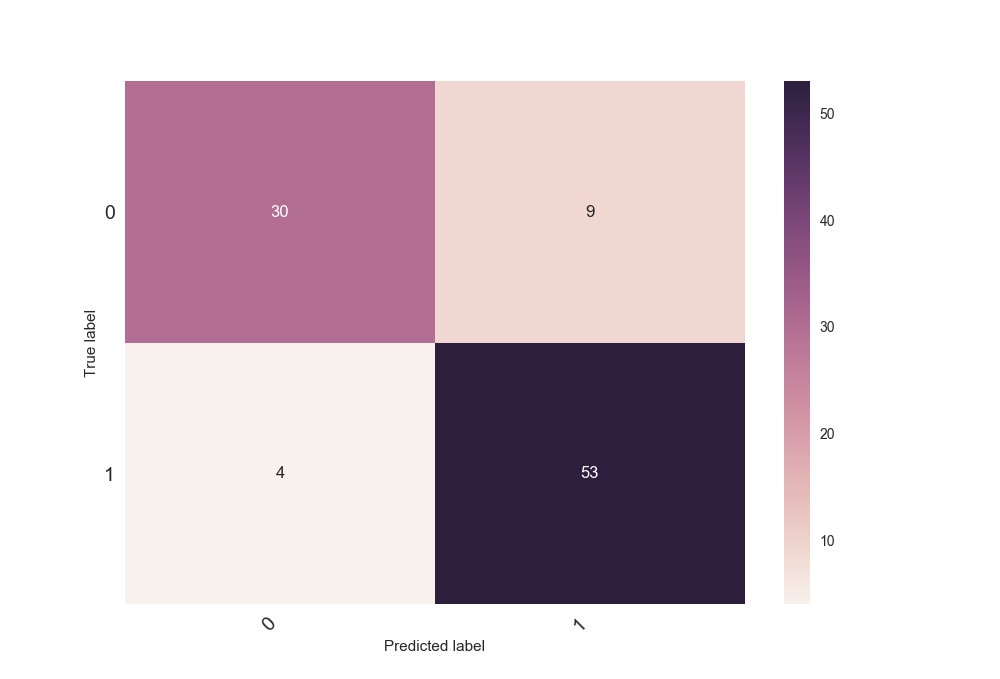
**LOGISTIC REGRESSION**

Logistic regression is another technique borrowed by machine learning from the field of statistics.



**Logistic Regression can be considered as a special case of linear regression where the outcome variable is categorical, where we use log of odds i.e. it predicts the probability of occurrence of an event by fitting data to a logit function.**

**In logistic regression there can be more than one independent variables that determines the outcome. The outcome is dichotomous which is either 0/1, true/false, yes/no.**

****

**Accuracy: 0.864583333333**

**Precision: 0.854838709677**

**Coefficients of Regression: [[-0.32450626 0.27213675 -0.68257224 0.03806945 0.21241993 0.76162507 0.16347072 1.64924434]]**

Similarly, we have used other classification models and finally Ensemble learning(bagging, boosting, stacking) to improve the results.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | ROC-AUC Score (%) | Accuracy (%) | Precision (%) |
| Decision Trees | 87.9 | 88.54 | 89.65 |
| SVM | 72.1 | 75.00 | 74.62 |
| Logistic R | 85.0 | 86.45 | 85.48 |
| Naïve Bayes | 83.6 | 84.37 | 86.20 |
| K-NN | 74.1 | 75.00 | 78.94 |
| Bagging | 86.2 | 87.45 | 88.35 |
| Stacking | 92.6 | 87.5 | 88.13 |
| Boosting | 88.24 | 89.24 | 89.67 |