# German Credit Data Risk Analysis

Souratosh Khan

## I. INTRODUCTION

A most optimal machine learning method is determined to predict whether customers at a bank are creditable or not. Both the training and test data are downloaded from the following site : https://onlinecourses.science.psu.edu/stat857/node/215/. The data is divided equally between the training and test sets with 500 instances, 20 input features and one target variable which determines whether a customer can be given a loan or not.

## II. EXPLORATORY ANALYSIS

In this section data is cleaned and preprocessed before fitting a model to test the accuracy. The training set has a list of 357 creditable customers and 143 non-creditable customers classified on the basis od 20 different features. These input features have both categorical and continuous data types. The 'Occupation' feature has been excluded since it contains only one category in the training set. Categories with too few observations are combined. For example, the 'Purpose' feature has four different categorical values. 1 and 2 are combined to 1. 3 is mapped to 2 and 4 is mapped to 3. Likewise, the four different categorical values of the savings and stocks variable are mapped to two values. An exhaustive description of the variables, data type and the categorical values have been provided below.

### A. Variable Description

- Creditability : Target variable . Creditable - 1 ; Non-Creditable - 0
- Account.Balance : Balance of current account ( 1 - No running account; 2 - No balance/debit; 3 - Some balance)
- Duration.ofCredit.month : Duration in months (continuous variable)
- Payment.Status.of.Previos.Credit : Payment of previous credits ( 1 - Some problems; 2 - Paid up; 3 - No problems with current credits at this bank)
- Purpose : Purpose of credit (1 - car; 2 - furniture; 3 - radio/television)
- Credit.Amount : Amuont of credit in DM (continuous variable)
- Value.Savings.Stocks : Value of savings/stocks (1 - No; 2 - Yes)
- Length.of.current.employment : Number of years of current employment (1 - unemployed or $<$ 1 year; 2 - 1 $\leq$ ... $<$ 4 years; 3 - 4 $\leq$ ...$<$ 7 years; 4 $\geq$ 7 years)
- Instalment.per.cent : Instalment in % of available income (1 - $\geq$ 35; 2 - 25 $\leq$ ...¡ 35; 3 - 20 $\leq$...$<$ 25; 4 - $<$ 20)
- Sex...Marital.Status : Marital Status/Sex (1 - separated male; 2 - single male; 3 - married/widowed male)
- Gurantors : Further debtors/Gurantors (1 - No; 2 - Yes)
- Duration.in.Current.address : Number of years at present address (1 - $<$ 1 year; 2 - 1 $\leq$...$<$ 4 years; 3 - 4 $\leq$...$<$ 7 years; 4 - $\geq$ 7 years)
- Most.valuable.available.asset : Most valuable available assets (1 - property; 2 - savings contract with a building society/life insurance; 3 - car/other; 4 - no assets)
- Age..years : Age of customers in years (continuous variable)
- Concurrent.Credits : Further running credits (1 - other banks/department stores; 2 - None)
- Type.of.apartment : Type of apartment (1 - free; 2 - rented flat; 3 - owner-occupied flat)

- No.of.Credits.at.this.Bank : Number of previous credits at this bank (inlcuding the running one) (1 - 1; 2 - > 1)
- Occupation : Occupation of customer ( dropped from analysis since only 1 category is reported )
- No.of.dependents : Number of persons entitled to maintenance (1 - $\geq$ 3; 2 - $\leq$ 2)
- Telephone : Customers who own a telephone (1 - No; 2 - Yes)
- Foreign.Worker : Customer is a foreign worker or not (1 - Yes; 2 - No)

### B. Feature Plots

In order to better understand how the input variables affect the outcome, a $K1 \times 2$ contingency table can be evaluated with the Input variable and Creditability as the two observables. For the purpose of visualization, barplots of all the features with categorical values for creditable and non-creditable customers are provided below. The chi-Squared test is performed based on the contigency tables which gives the Pearson's chi-square value ($\chi^2$) and the p-value ($p$) for each observable. A large value of chi-square (small p-value) would lead to a rejection of the null hypothesis. 'Age', 'Credit Amount' and 'Duration of Credit (months)' are continuous variables and hence histogram plots are shown.

| Account Balance | Creditability 0 | Creditability 1 | Total |
|---|---|---|---|
| 1 | 54 | 62 | 116 |
| 2 | 56 | 89 | 145 |
| 3 | 33 | 206 | 239 |
| Total | 143 | 357 | 500 |

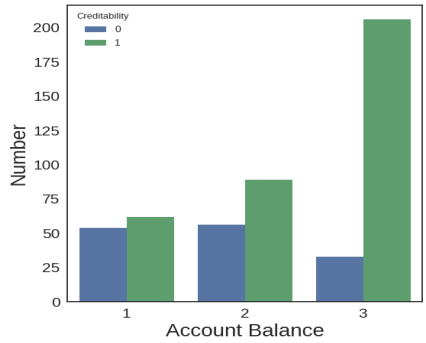


FIG. 1: Account Balance : Contingency Table and Barplot; $\chi^2 = 51.047, p = 0.000$

| Previous Credit | Creditability 0 | Creditability 1 | Total |
|---|---|---|---|
| 1 | 25 | 12 | 37 |
| 2 | 77 | 181 | 258 |
| 3 | 41 | 164 | 205 |
| Total | 143 | 357 | 500 |

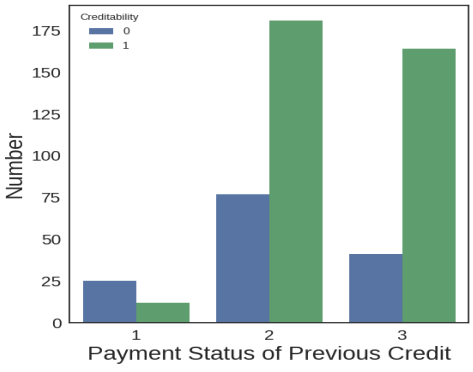


FIG. 2: Status of Previous Credits : Contingency Table and Barplot; $\chi^2 = 35.134, p = 0.000$

| Purpose | Creditability 0 | Creditability 1 | Total |
|---|---|---|---|
| 1 | 30 | 108 | 138 |
| 2 | 44 | 135 | 179 |
| 3 | 69 | 114 | 183 |
| Total | 143 | 357 | 500 |

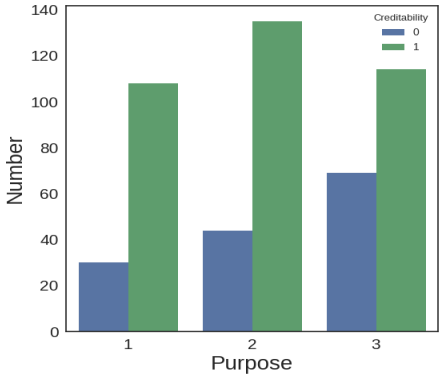


FIG. 3: Purpose : Contingency Table and Barplot; $\chi^2 = 12.026, p = 0.002$

| Most Valuable Asset | Creditability 0 | Creditability 1 | Total |
|---|---|---|---|
| 1 | 31 | 113 | 144 |
| 2 | 28 | 71 | 99 |
| 3 | 51 | 130 | 181 |
| 4 | 33 | 43 | 76 |
| Total | 143 | 357 | 500 |

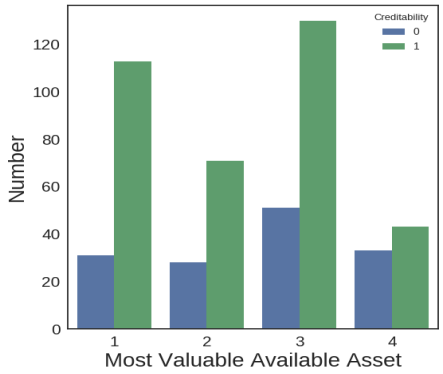


FIG. 4: Most Valuable Available Asset : Contingency Table and Barplot; $\chi^2 = 11.723, p = 0.008$

| Savings / Stocks | Creditability 0 | Creditability 1 | Total |
|---|---|---|---|
| 1 | 104 | 199 | 303 |
| 2 | 39 | 158 | 197 |
| Total | 143 | 357 | 500 |


FIG. 5: Values of Savings/Stocks : Contingency Table and Barplot; $\chi^2 = 11.635, p = 0.001$

| Concurrent Credits | Creditability 0 | Creditability 1 | Total |
|---|---|---|---|
| 1 | 36 | 46 | 82 |
| 2 | 107 | 311 | 418 |
| Total | 143 | 357 | 500 |

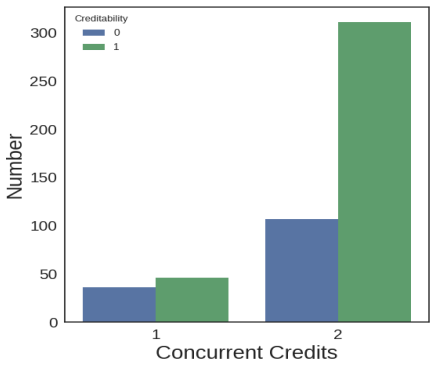


FIG. 6: Concurrent Credits : Contingency Table and Barplot; $\chi^2 = 10.369, p = 0.001$

| Length of Employment | Creditability 0 | Creditability 1 | Total |
|---|---|---|---|
| 1 | 41 | 72 | 113 |
| 2 | 52 | 117 | 169 |
| 3 | 21 | 75 | 96 |
| 4 | 29 | 93 | 122 |
| Total | 143 | 357 | 500 |

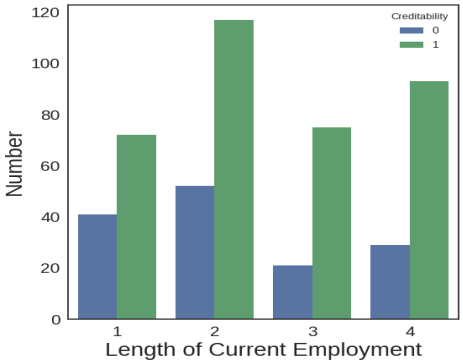


FIG. 7: Length of Current Employment : Contingency Table and Barplot; $\chi^2 = 7.176, p = 0.067$

| Type of Apartment | Creditability 0 | Creditability 1 | Total |
|---|---|---|---|
| 1 | 31 | 59 | 90 |
| 2 | 91 | 264 | 355 |
| 3 | 21 | 34 | 55 |
| Total | 143 | 357 | 500 |

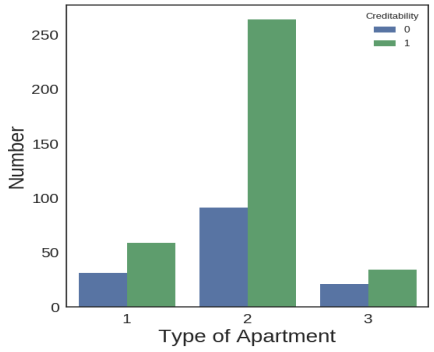


FIG. 8: Type of Apartment : Contingency Table and Barplot; $\chi^2 = 5.508, p = 0.064$

| Duration in Current Address | Creditability 0 | Creditability 1 | Total |
|---|---|---|---|
| 1 | 14 | 51 | 65 |
| 2 | 46 | 110 | 156 |
| 3 | 27 | 50 | 77 |
| 4 | 56 | 146 | 202 |
| Total | 143 | 357 | 500 |

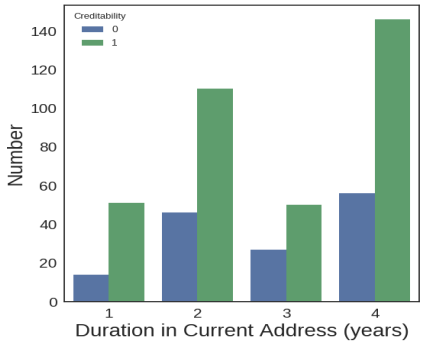


FIG. 9: Duration in Current Address : Contingency Table and Barplot; $\chi^2 = 3.299, p = 0.348$

| Instalment per cent | Creditability 0 | Creditability 1 | Total |
|---|---|---|---|
| 1 | 18 | 49 | 67 |
| 2 | 29 | 84 | 113 |
| 3 | 18 | 59 | 77 |
| 4 | 78 | 164 | 242 |
| Total | 143 | 357 | 500 |

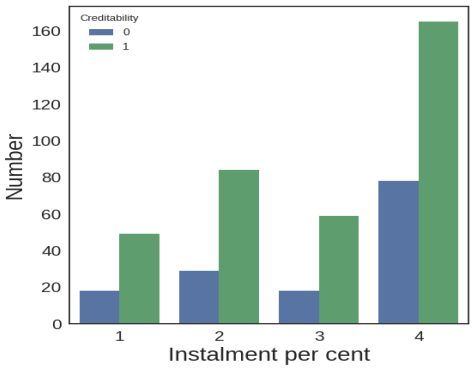


FIG. 10: Instalment : Contingency Table and Barplot; $\chi^2 = 3.061, p = 0.382$

| Sex + Marital Status | Creditability 0 | Creditability 1 | Total |
|---|---|---|---|
| 1 | 58 | 119 | 177 |
| 2 | 72 | 201 | 273 |
| 3 | 13 | 37 | 50 |
| Total | 143 | 357 | 500 |

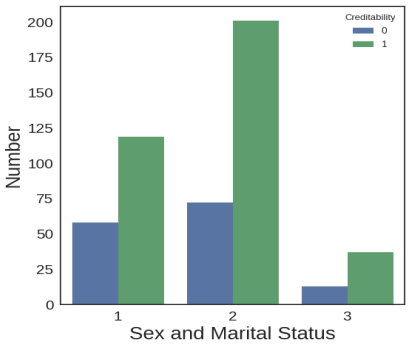


FIG. 11: Sex and Marital Status : Contingency Table and Barplot; $\chi^2 = 2.334, p = 0.311$

| Total Credits | Creditability 0 | Creditability 1 | Total |
|---|---|---|---|
| 1 | 97 | 223 | 320 |
| 2 | 46 | 134 | 180 |
| Total | 143 | 357 | 500 |

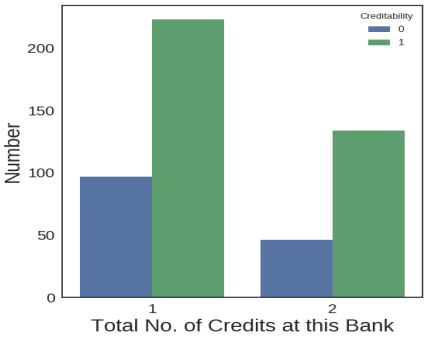


FIG. 12: Total number of Credits at this Bank : Contingency Table and Barplot; $\chi^2 = 1.054, p = 0.304$

| Guarantors | Creditability 0 | Creditability 1 | Total |
|---|---|---|---|
| 1 | 127 | 329 | 456 |
| 2 | 16 | 28 | 44 |
| Total | 143 | 357 | 500 |

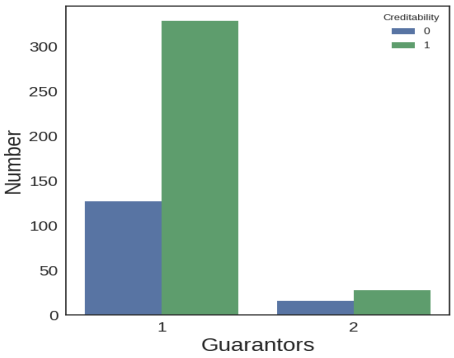


FIG. 13: Guarantors : Contingency Table and Barplot; $\chi^2 = 1.038, p = 0.308$

| Dependents | Creditability 0 | Creditability 1 | Total |
|---|---|---|---|
| 1 | 120 | 308 | 428 |
| 2 | 23 | 49 | 72 |
| Total | 143 | 357 | 500 |

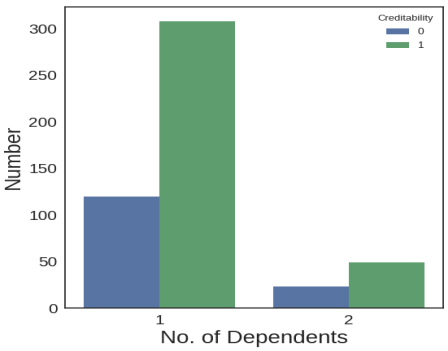


FIG. 14: Number of Dependents : Contingency Table and Barplot; $\chi^2 = 0.289, p = 0.591$

| Telephone | Creditability 0 | Creditability 1 | Total |
|---|---|---|---|
| 1 | 84 | 217 | 301 |
| 2 | 59 | 140 | 199 |
| Total | 143 | 357 | 500 |

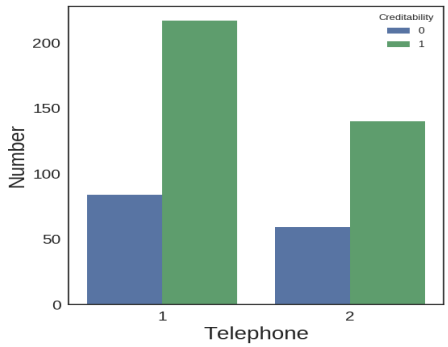


FIG. 15: Telephone : Contingency Table and Barplot; $\chi^2 = 0.103, p = 0.748$

| Foreign Worker | Creditability 0 | Creditability 1 | Total |
|---|---|---|---|
| 1 | 139 | 344 | 482 |
| 2 | 4 | 13 | 17 |
| Total | 143 | 357 | 500 |

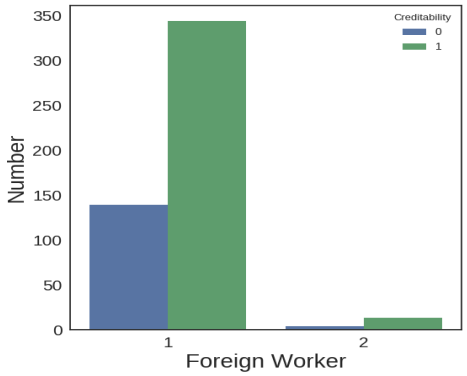


FIG. 16: Foreign Worker : Contingency Table and Barplot; $\chi^2 = 0.039, p = 0.843$

Features with $\chi^2 < 5.0$ and $p > 0.3$ are excluded from the datasets from here on.

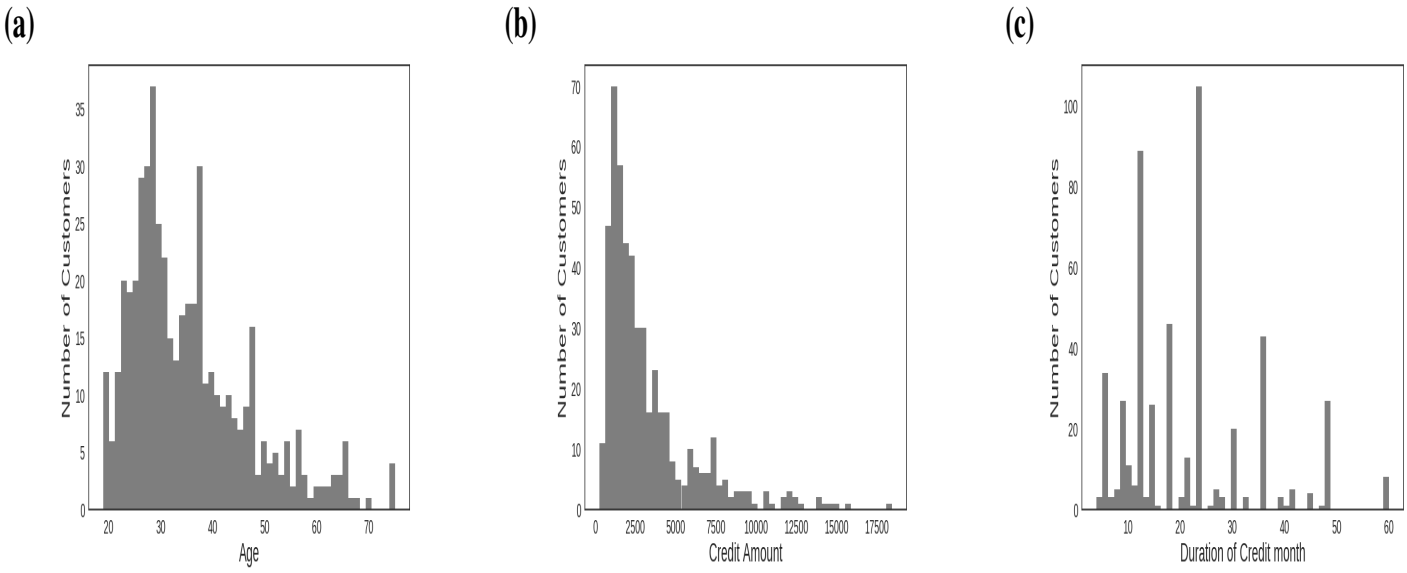


FIG. 17: Histogram plots of (a) Age, (b) Credit Amount and (c) Duration of Credit (months). All plots have positive skewness.

A summary of the statistics of the continuous variables is provided below

TABLE I: Mean, Standard Deviation, Min, Max, 1st quantile, 2nd quantile and 3rd quantile of Age, Credit Amount and Duration of Credit

| | Age | Credit Amount | Duration of Credit (months) |
|---|---|---|---|
| **mean** | 35.458000 | 3242.096000 | 21.552000 |
| **std** | 11.419152 | 2841.763537 | 12.321645 |
| **max** | 75.000000 | 18424.000000 | 60.000000 |
| **min** | 19.000000 | 276.000000 | 4.000000 |
| **25%** | 27.000000 | 1360.750000 | 12.000000 |
| **50%** | 33.000000 | 2243.500000 | 18.000000 |
| **75%** | 41.000000 | 3983.500000 | 24.000000 |

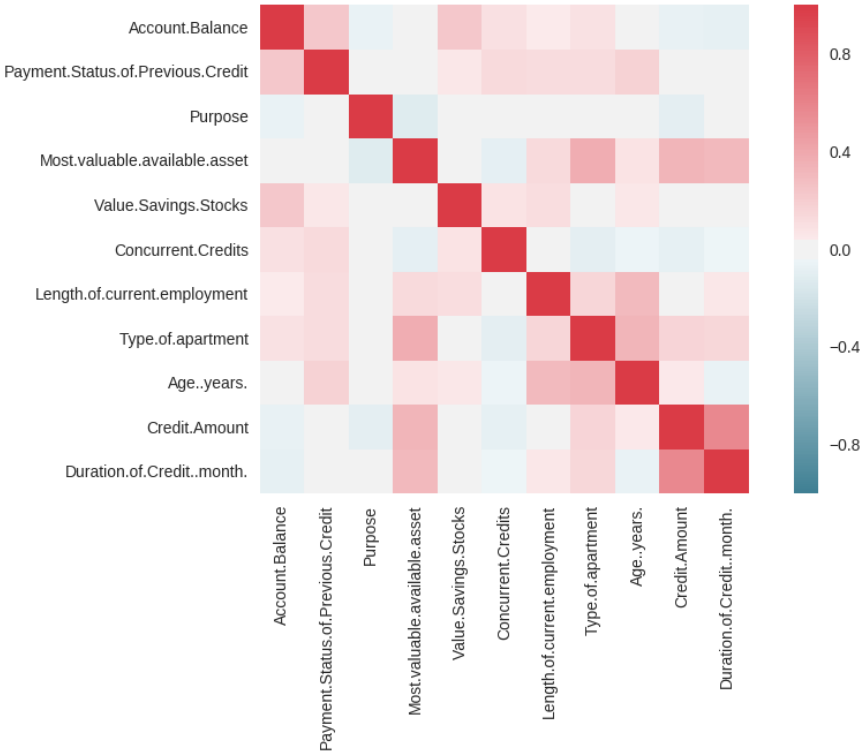


FIG. 18: Plot of the Correlation matrix to test the independence of the input features. It is obvious that the Credit Amount and the Duration of Credit in months are highly correlated. As a result, the Duration of Credit in months is dropped from the analysis since the variables are assumed to be independent. Other pairs of variables that are moderately correlated are i) (Age, Length of current employment), (Age, Type of Apartment), (Most valuable available asset, Type of Apartment) and (Most valuable available asset, Credit Amount)

## III. EVALUATION METRICS FOR SUPERVISED LEARNING METHODS

Before we fit various models to our data, we would like to familiarize ourselves with the different evaluation metrics that will be used to finally select the most effective predictive model. For such a purpose, a dummy classifier is

invoked which provides a null accuracy baseline and serves as a sanity check for the classifiers that will be used in the following sections. We use the most-frequent strategy since 70% of the customers in the training set are creditable. The most basic evaluation metric is the accuracy which is defined in the equation below.

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{1}$$

where $TN, TP$ are the true negatives and true positives respectively and $FN, FP$ are the false negatives and false positives respectively. If the selected model's accuracy is close to the accuracy of the dummy classifier then the model is ineffective. However, there are other evaluation metrics that can navigate us in selecting our model eg. Precision, Recall and F1-score.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

While Recall oriented machine learning tasks is used for tumor detection, Precision oriented tasks involve customer identification. Hence, for our purposes, Precision is going to determine the most efficient method since larger Precision means smaller false positives (FP) which prevents the bank from falsely identifying customers as creditable. This is highlighted in the ROC- curve which is a plot of the True Positive Rate versus the False Positive Rate. For an ideal situation, where the FP rate is absent, the Area under the ROC curve is 1.

## IV. DUMMY CLASSIFIER RESULTS

Accuracy on test set = 0.686

## V. LOGISTIC REGRESSION

Logistic Regression (LR) transforms several input values using a sigmoid function to a probability. If the probability is greater than 0.5, then the customer is labeled creditable else it is classified in the negative class. In order to avoid overfitting the data, LR is performed over three different values of the regularization parameter (C = 0.1, 1.0, 100.0). The ROC curves along with the area under the curve and the different evaluation metrics are provided below.

TABLE II: Logistic Regression : ROC Curves with C = 0.1, 1.0, 100.0

| | C = 0.1 | C = 1.0 | C = 100.0 |
|---|---|---|---|
| **Accuracy - Training Set** | 0.754 | 0.788 | 0.792 |
| **Accuracy - Test Set** | 0.718 | 0.742 | 0.748 |
| **Precision** | 0.737 | 0.797 | 0.809 |
| **Recall** | 0.915 | 0.837 | 0.828 |
| **F1-score** | 0.817 | 0.817 | 0.818 |

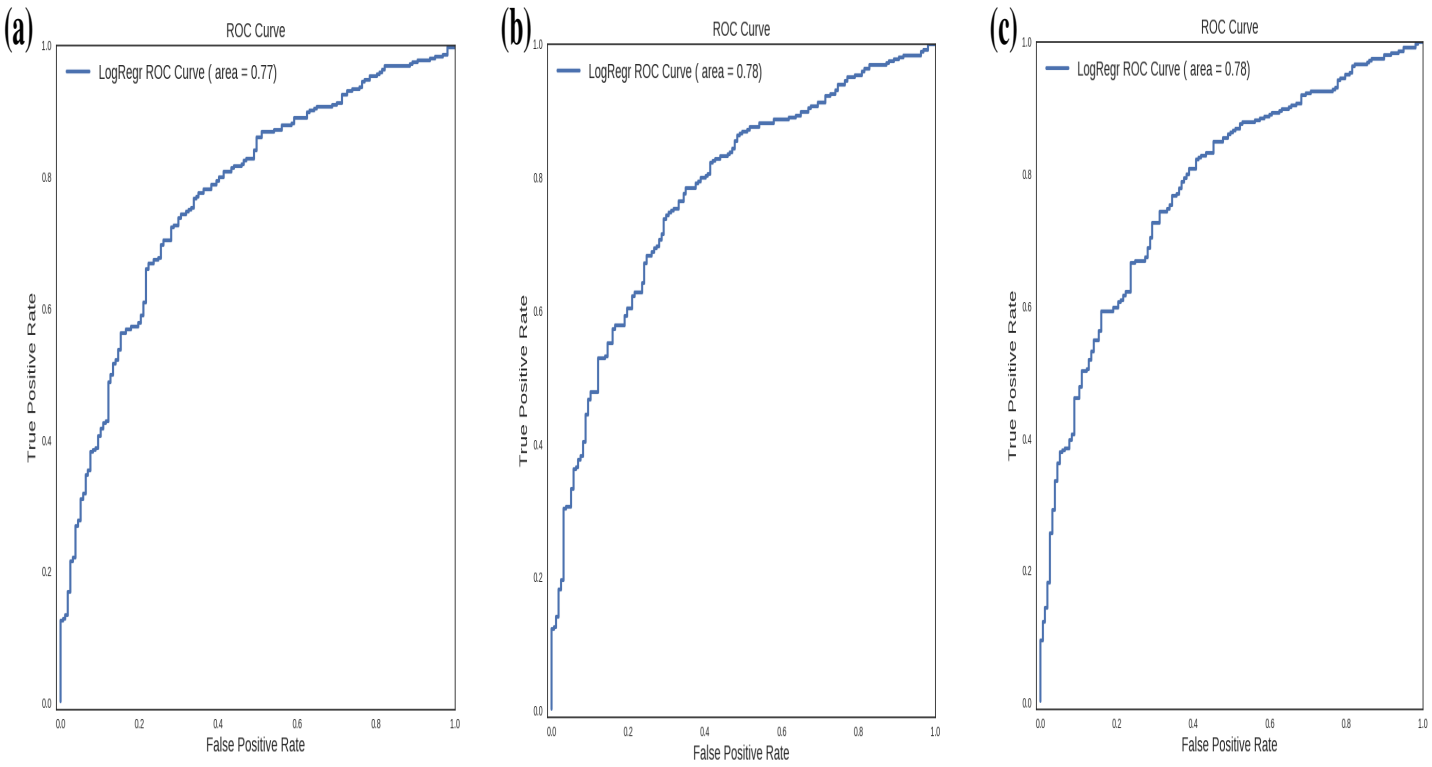


FIG. 19: ROC Curve using Logistic Regression : (a) C = 0.1, (b) C = 1.0, (c) C = 100.0

## VI. DECISION TREE CLASSIFIER

A decision tree classifier starts with a root node which splits into branches based on the attributes. The branches end in leaves which ultimatley classifies the customers as creditable or non-creditable. A grid search is carried out with 5-fold cross validation and two parameters : 'max depth' : the maximum number of nodes, 'min samples split' : minimum number of samples required to split an internal node.

TABLE III: Decision Tree Classifier

| | max depth = 5; min samples split = 20 |
|---|---|
| **Accuracy - Training Set** | 0.802 |
| **Accuracy - Test Set** | 0.664 |
| **Precision** | 0.720 |
| **Recall** | 0.834 |
| **F1-score** | 0.773 |

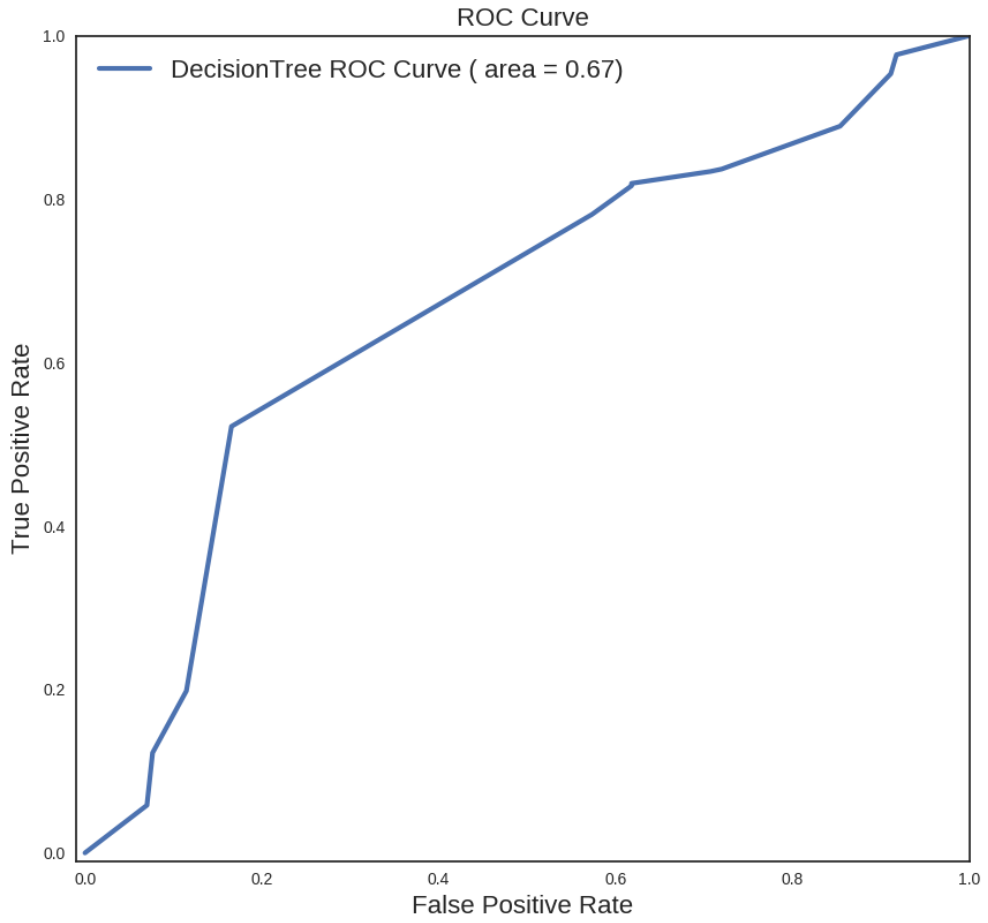


FIG. 20: ROC curve using Decision Tree Classifier : max depth = 5; min samples split = 20

## VII. RANDOM FOREST CLASSIFIER

A Random Forest uses an ensemble of decision trees to train different subsets of the training data with each tree classifying the customers as creditable or non-creditable. The predictions of the individual trees are collected and creates an aggregate model. Random Forest are less prone to overfitting. Grid search cross validation (5-fold) is carried out with maximum depth, minimum samples split and number of estimators (number of trees in the random forest) as the parameters.

TABLE IV: Random Forest Classifier

| | max depth = 5; min samples split = 20; number of estimators = 20 |
|---|---|
| **Accuracy - Training Set** | 0.814 |
| **Accuracy - Test Set** | 0.708 |
| **Precision** | 0.726 |
| **Recall** | 0.921 |
| **F1-score** | 0.812 |

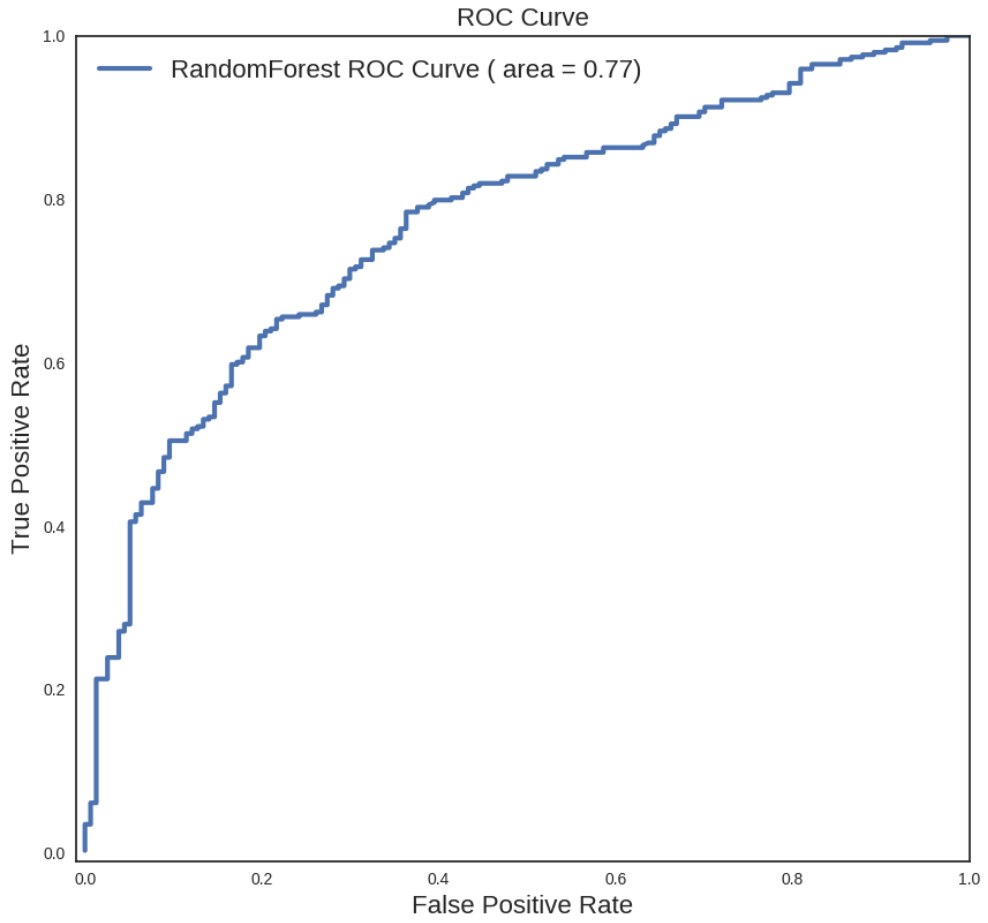


FIG. 21: ROC curve using Random Forest Classifier: max depth = 5; min samples split = 10; number of estimators = 20

## VIII. GRADIENT BOOSTING CLASSIFIER

Gradient Boosting is similar to Random Forests which trains the dataset iteratively by building a series of decision trees that grow smaller in size. Each tree attempts to correct errors from the previous stage. The learning rate controls emphasis on fixing errors from previous iteration. High (low) learning rate means more complex (simple) trees.

TABLE V: Gradient Boosting Classifier

| | max depth = 5; min samples split = 15; number of estimators = 40; learning rate = 0.1 |
|---|---|
| **Accuracy - Training Set** | 0.942 |
| **Accuracy - Test Set** | 0.678 |
| **Precision** | 0.753 |
| **Recall** | 0.790 |
| **F1-score** | 0.771 |

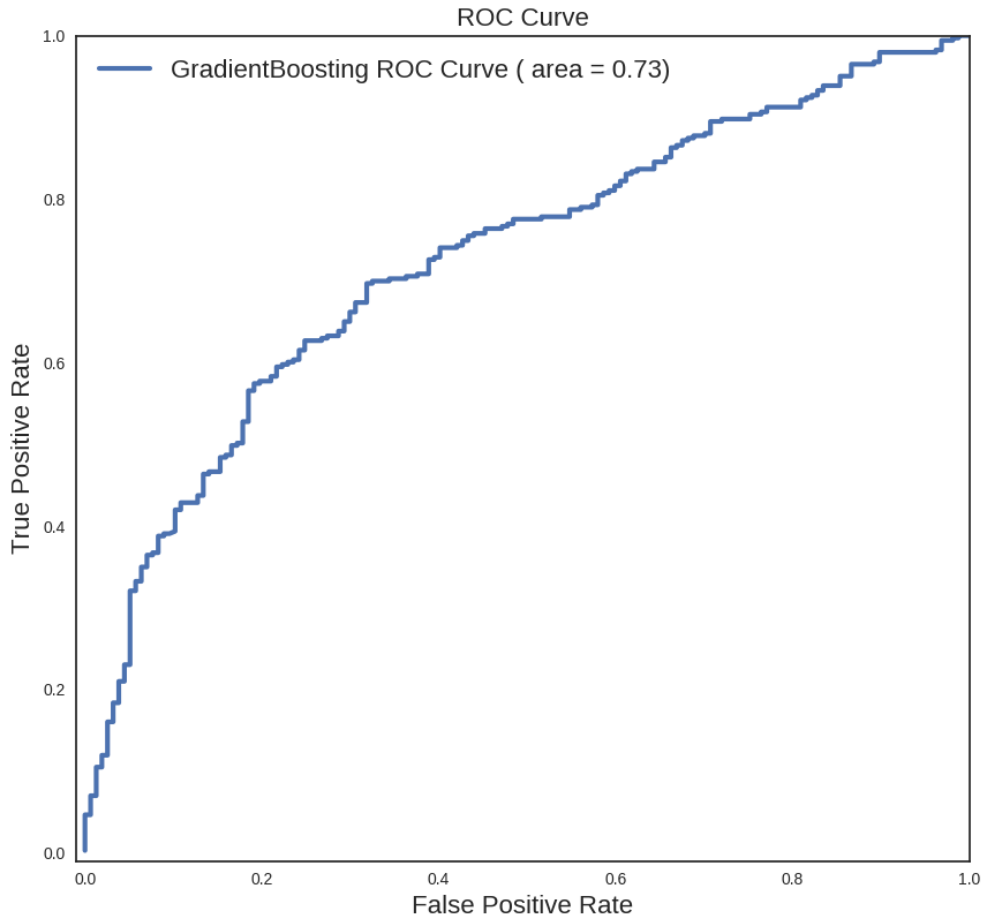


FIG. 22: ROC curve using Gradient Boosting Classifier: max depth = 5; min samples split = 15; number of estimators = 40; learning rate = 0.1

## IX. CONCLUSION

Among all the models, Logistic Regression with C = 1.0 (Precision = 0.797) and C = 100.0 (Precision = 0.809) yield the most optimal results. Area under the ROC curve (AUC) yields 0.78 for both the parameters. Random Forest with the following parameters : max depth = 5; min samples split = 10; number of estimators = 20 result in an AUC of 0.77.