Project: Creditworthiness

Step 1: Business and Data Understanding

Due to a financial scandal that hit a competitive bank last week, we suddenly have an influx of new people applying for loans. The manager sees this new influx as a great opportunity and wants a list of creditworthy customers to give a loan to, in the next two days.

Key Decisions:

1. What decisions needs to be made?

To which new loan applicant to give a loan to, based on the evaluation of creditworthiness of these new loan applicants.

2. What data is needed to inform those decisions?

We require the following data in order to inform the decision:

- All credit approvals from the past loan applicants the bank has ever completed;
- A list with the new set of people applying for loans.

For both datasets, we require the following information:

- Account balance;
- Age;
- Credit amount;
- Duration of credit;
- Length of employment;
- Payment status;
- Purpose;
- Total number of credits at this bank;
- Values of savings.

3. What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

In order to help make the decision, the model we need to use for this business problem is a binary model, as we are determining if a customer is creditworthy or non-creditworthy.

Step 2: Building the Training Set

1. In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

In the cleanup process the following changes occurred:

- Data fields that have been removed:
    - Concurrent-Credits – low variability, the data is entirely uniform and there are no other variations of the data;

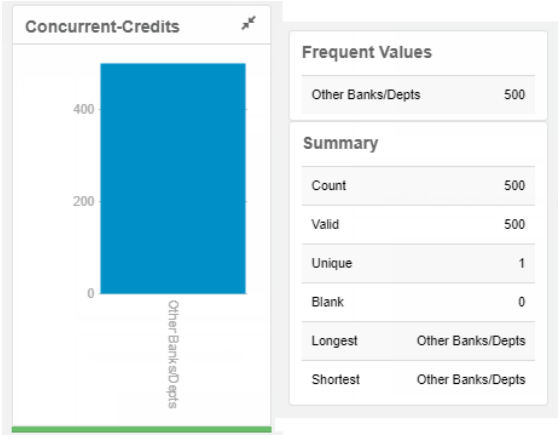


Fig. 2.1. Concurrent-Credits field data visualization.

    - Duration-in-Current-address – data field is missing a lot of values, 69% of the data are null values;

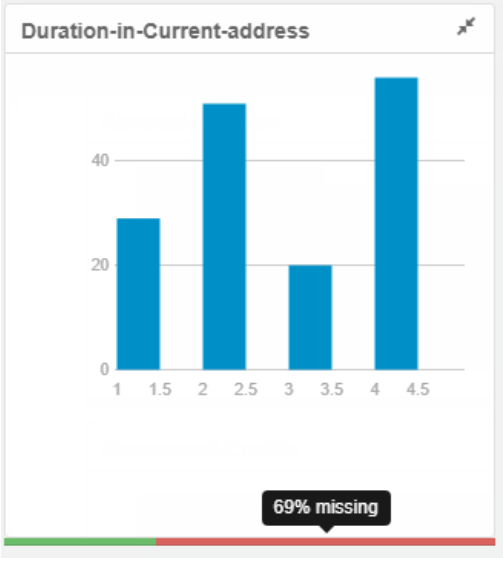


Fig. 2.2. Duration-in-Current-address field data visualization.

- Foreign-Worker – low variability, the majority of the data is skewed towards "1";

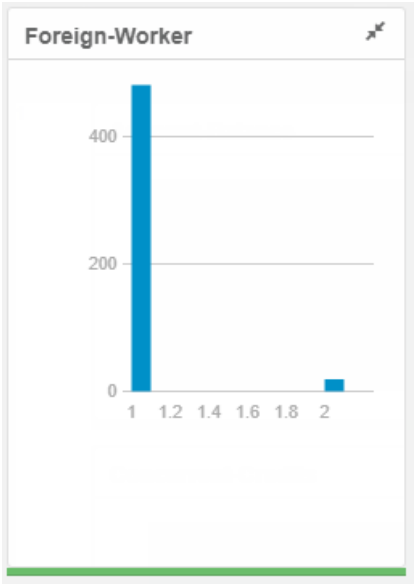


Fig. 2.3. Foreign-Worker field data visualization.

- Guarantors - low variability, the majority of the data is skewed towards "None";

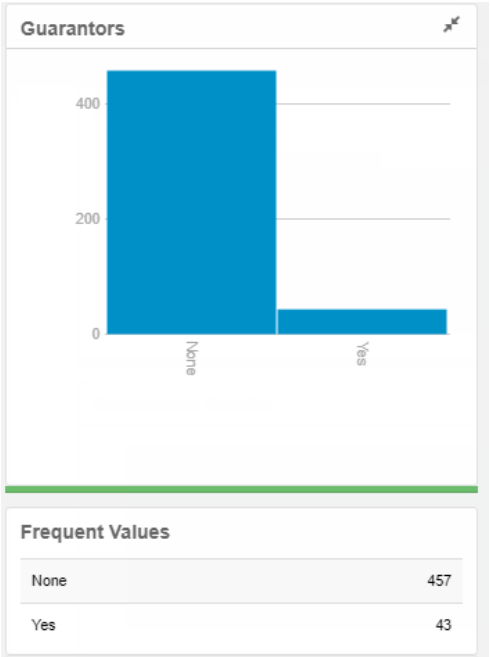


Fig. 2.4. Guarantors field data visualization.

- No-of-dependents - low variability, the majority of the data is skewed towards "1";

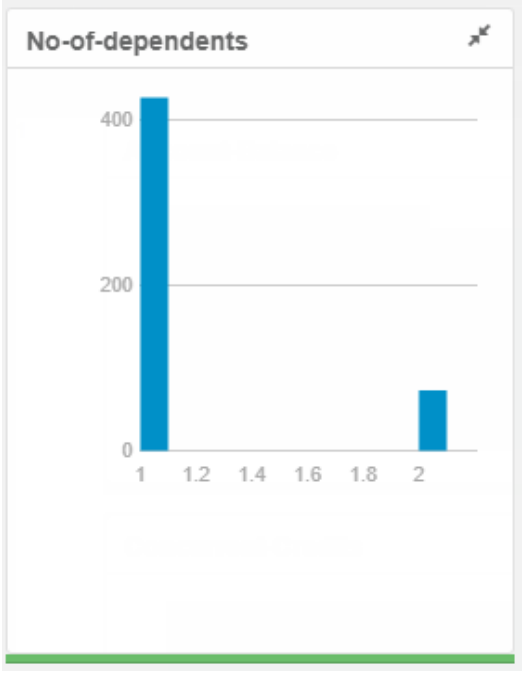


Fig. 2.5. No-of-dependents field data visualization.

- Occupation - low variability, the data is entirely uniform and there are no other variations of the data;

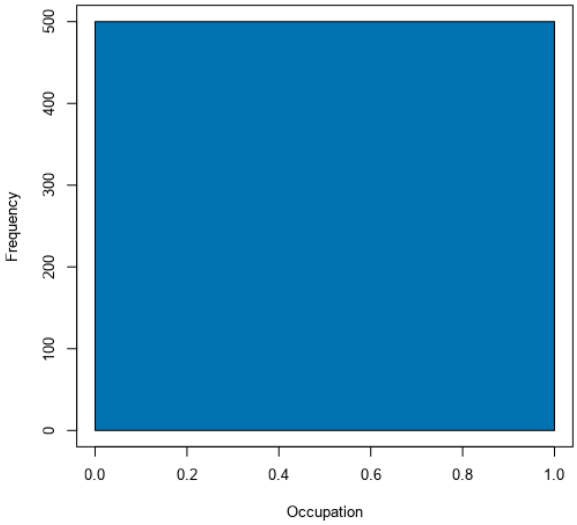


Fig. 2.6. Occupation histogram visualization.

- Telephone - there is no logical reason for including the variable.

- Data fields that have been imputed:

  - Age-years - for the sake of consistency in the data cleanup process and because the data is slightly skewed to the left, instead of removing the null data points, I imputed the data using the median of the entire data field;

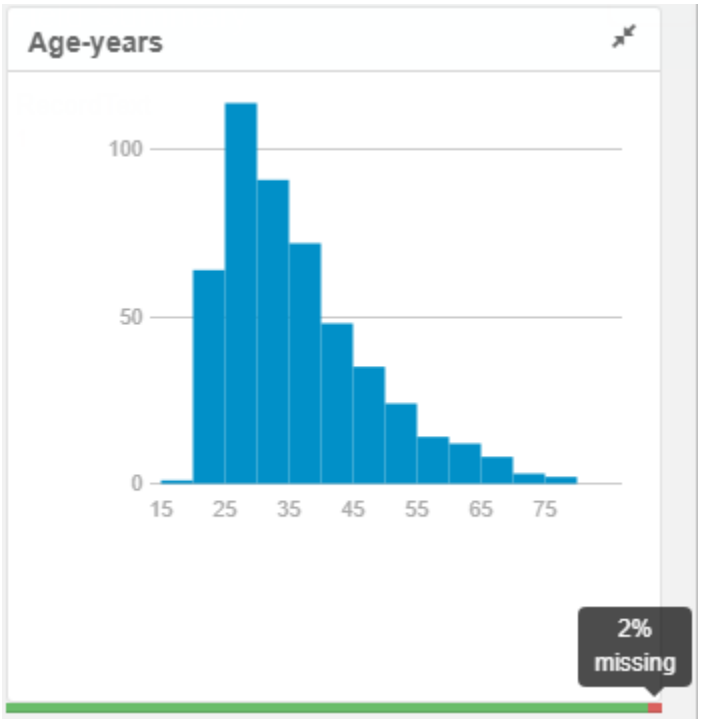


Fig. 2.7. Age-years field data visualization.

Step 3: Train your Classification Models

1. Logistic model:

    a. The predictor variables that are significant for this model are listed below, as for the p-values, are available in Pic.3.1.
        i. Account.Balance;
        ii. Credit.Amount;
        iii. Instalment.per.cent;

Record	Report
1	**Report for Logistic Regression Model Stepwise_Logistic_Model**
2	*Basic Summary*
3	Call: glm(formula = Credit.Application.Result ~ Account.Balance + Credit.Amount + Instalment.per.cent, family = binomial(logit), data = the.data)
4	Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.590	-0.835	-0.494	0.898	2.484

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.9042500	4.914e-01	-3.875	0.00011	***
Account.BalanceSome Balance	-1.6841104	2.895e-01	-5.816	6.01e-09	***
Credit.Amount	0.0001785	4.795e-05	3.722	2e-04	***
Instalment.per.cent	0.3288306	1.283e-01	2.564	0.01036	*

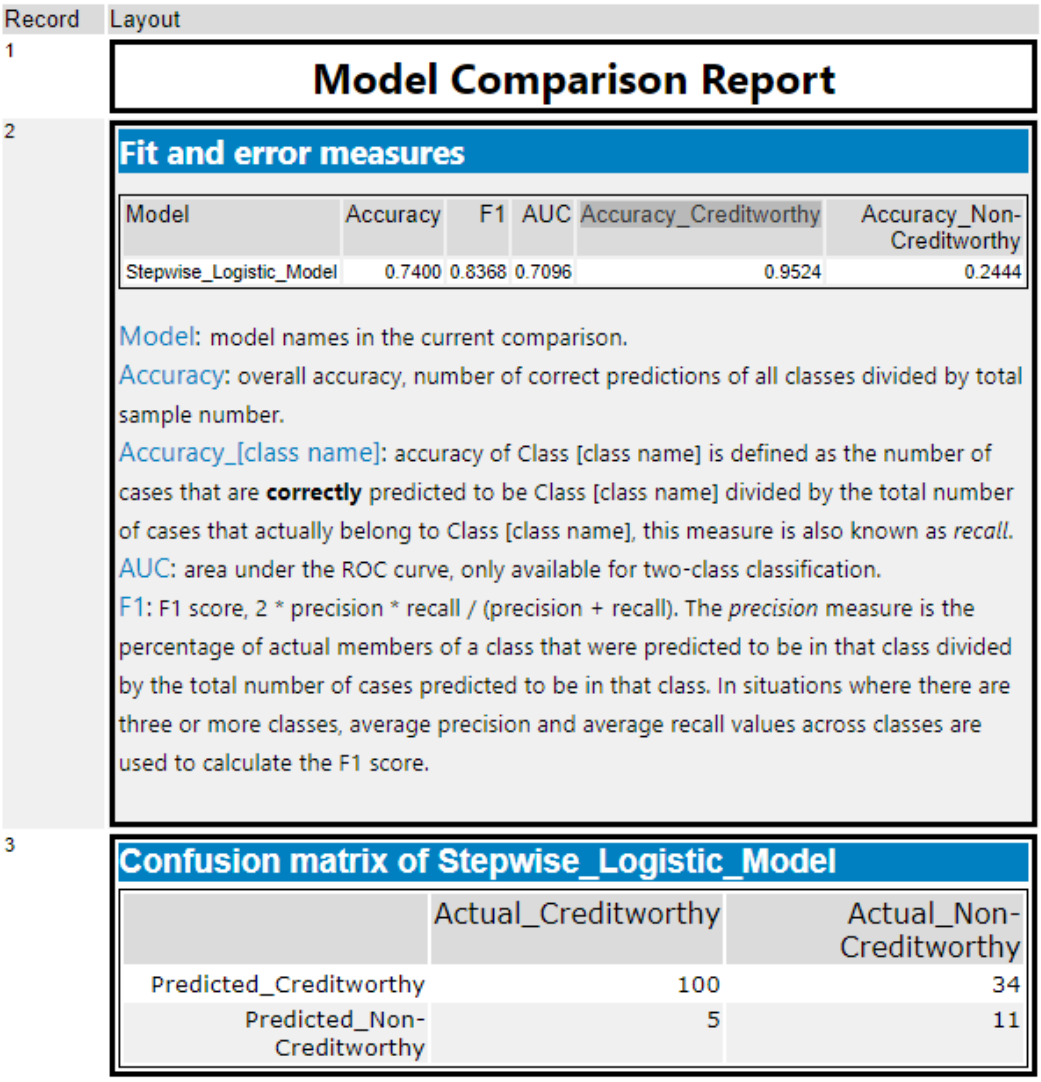
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial taken to be 1)

8	Null deviance: 413.16 on 349 degrees of freedom Residual deviance: 355.04 on 346 degrees of freedom McFadden R-Squared: 0.1407, Akaike Information Criterion 363
9	Number of Fisher Scoring iterations: 4
10	*Type II Analysis of Deviance Tests*

Fig. 3.1. Report of the Stepwise Logistic Model.

    b. Against the validation set the overall percent of accuracy is 74%, although, the Accuracy_Creditworthy is up to 95%, meaning that it does a very good job predicting if a person is Creditworthy. As for the bias, the positive predictive value (PPV) is 75% and the negative predictive value (NPV) is 69%; because the values are close to each other, this model has no bias.



Record | Layout

1

Model Comparison Report

2

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise_Logistic_Model	0.7400	0.8368	0.7096	0.9524	0.2444

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

3

Confusion matrix of Stepwise_Logistic_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	34
Predicted_Non-Creditworthy	5	11

Fig. 3.2. Model comparison report of the Stepwise Logistic Model.

2. Tree model:

    a. The variable importance charts for all of the predictor variables is available in Pic. 3.3., in Pic. 3.4. I have included only the most important predictor variables importance from which the model had a higher overall accuracy, which are:
        i. Account.Balance;
        ii. Duration.of.Credit.Month
        iii. Credit.Amount

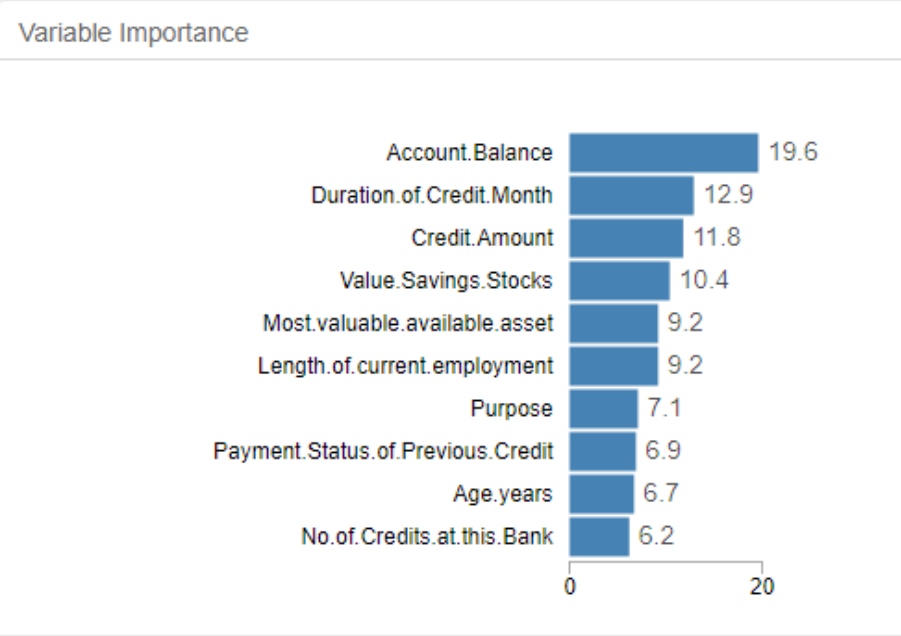


Fig. 3.3. All variables importance of the Tree Model.

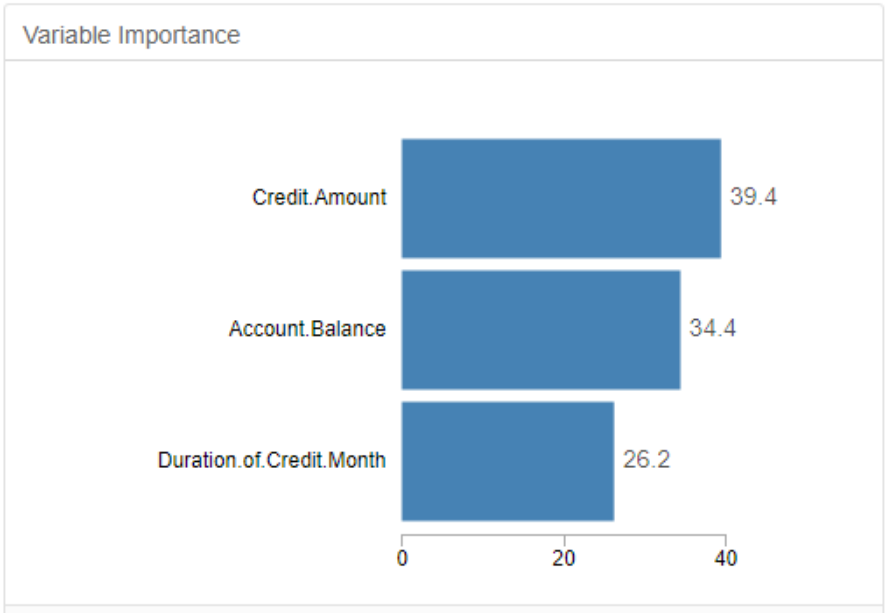


Fig. 3.4. Top variables by importance of the Tree Model.

b. Against the validation set the overall percent of accuracy is 72%, although, the Accuracy_Creditworthy is 85.7%, meaning that it does a good job predicting if a person is Creditworthy. As for the bias, the positive predictive value (PPV) is 77% and the negative predictive value (NPV) is 55%; because there is a big difference, the model has a high bias to creditworthy.

Record	Layout
1	**Model Comparison Report**
2	**Fit and error measures**

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree	0.7200	0.8108	0.6801	0.8571	0.4000

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

3 **Confusion matrix of Decision_Tree**

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	90	27
Predicted_Non-Creditworthy	15	18

Fig. 3.5. Model comparison report of the Tree Model.

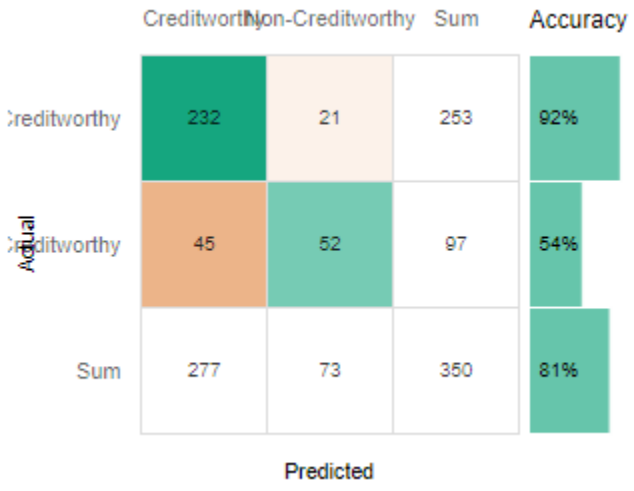


Fig. 3.6. Confusion Matrix of the Tree Model.

3. Forest model:

    a. The variable importance plots for all of the predictor variables is available in Pic. 3.7., in Pic. 3.8. I have included only the most important predictor variables importance from which the model had a higher overall accuracy and a OOB estimate of the error rate of 24%, which are:
        i. Credit.Amount;
        ii. Age.years;
        iii. Duration.if.Credit.Month;
        iv. Account.Balance;
        v. Most.valuable.available.asset;
        vi. Payment.Status.of.Previous.Credit;
        vii. Instalment.per.cent;
        viii. Value.Savings.Stocks;
        ix. Purpose;
        x. Length.of.current.employment.

Note: using only the first 3 variables the Accuracy_Creditworthy is the same as using all the above variables, although, the overall accuracy is lower, 79.3% instead of 82%.

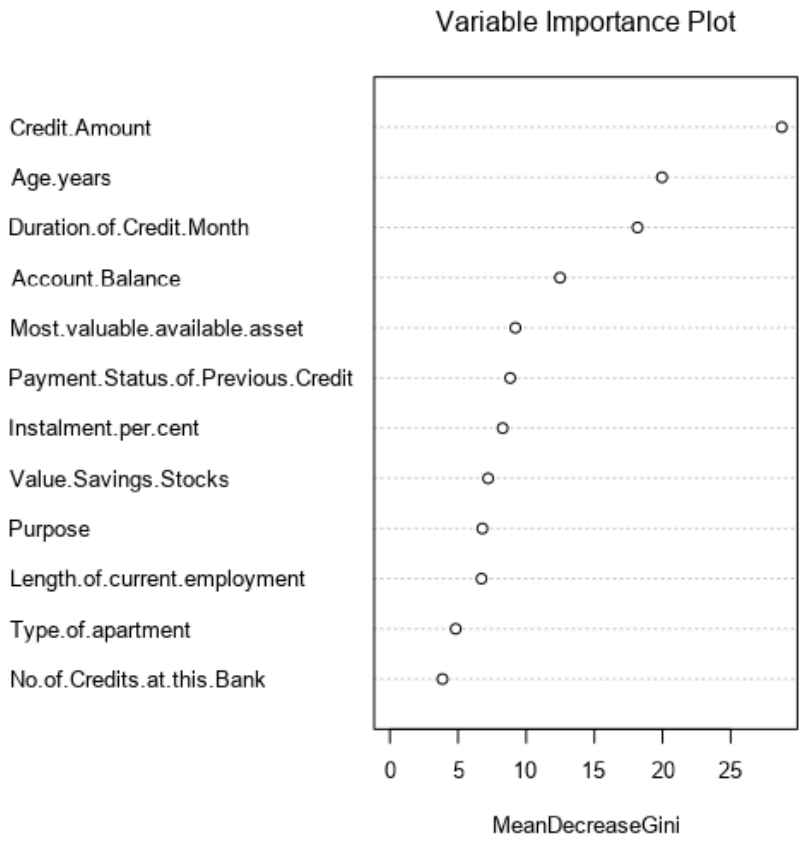


Fig. 3.7. All variables importance plot of the Forest Model.

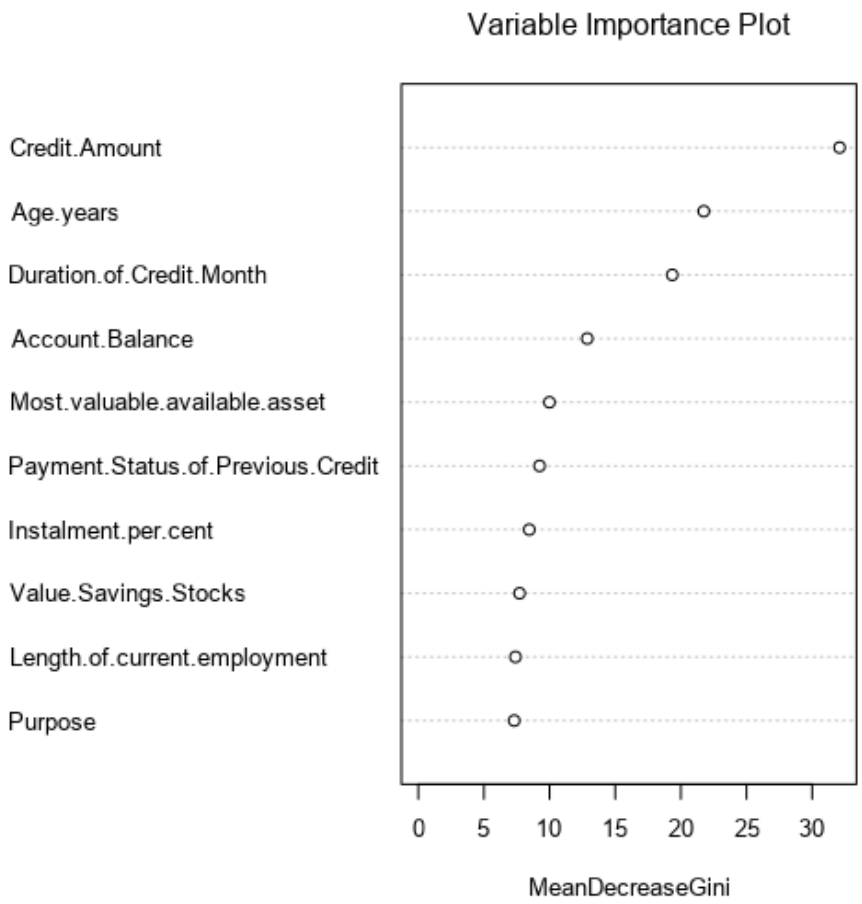


Fig. 1.8. Top variables importance plot of the Forest Model.

b. Against the validation set the overall percent of accuracy is 82%, were the Accuracy_Creditworthy is 97%, meaning that it does a very good job predicting if a person is Creditworthy. As for the bias, the positive predictive value (PPV) is 81% and the negative predictive value (NPV) is 88%; because the values are close to each other, this model has no bias.

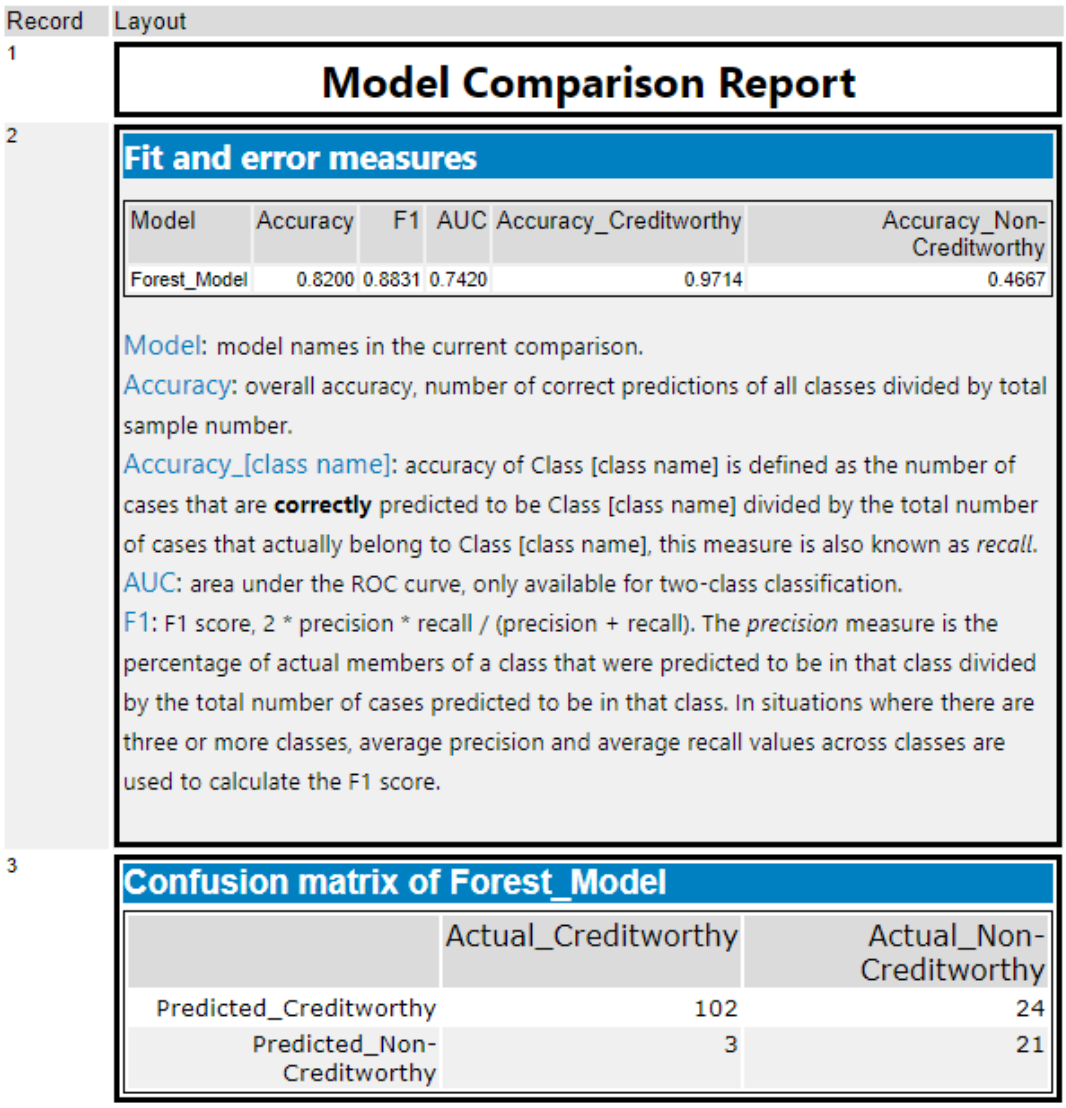


Fig. 3.9. Model comparison report of the Forest Model.

4. Boosted model:

    a. The variable importance plots for all of the predictor variables is available in Pic. 3.10., in Pic. 3.11. I have included only the most important predictor variables importance from which the model had a higher overall accuracy, which are:
        i. Account.Balance;
        ii. Credit.Amount;
        iii. Duration.of.Credit.Month;
        iv. Payment.Status.of.Previous.Credit;
        v. Purpose;
        vi. Age.years;
        vii. Most.valuable.available.asset.

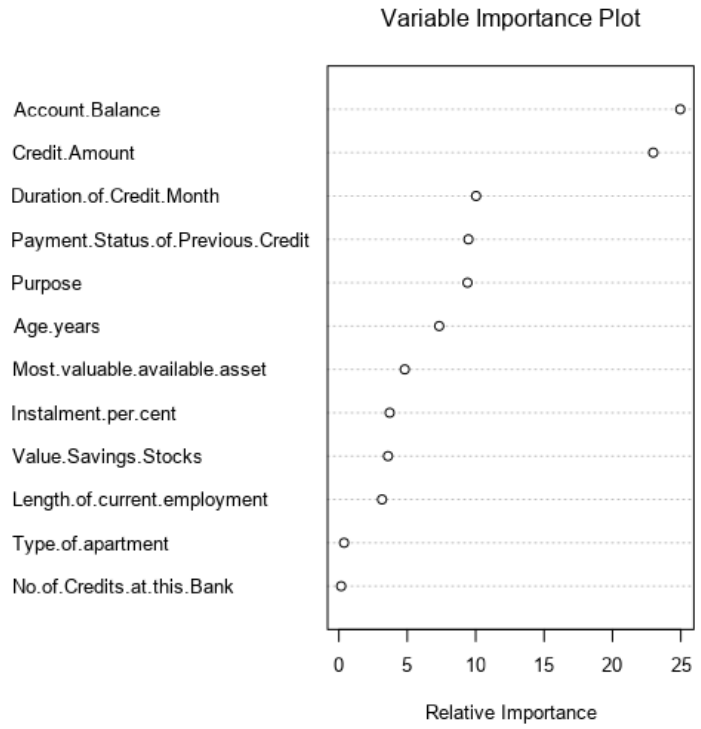


Fig. 3.10. All variables importance plot of the Boosted Model.

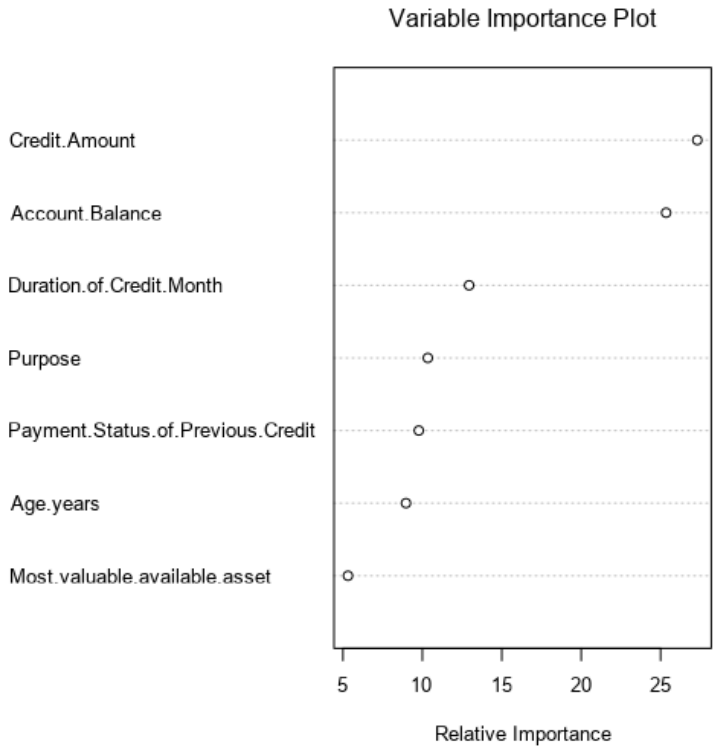
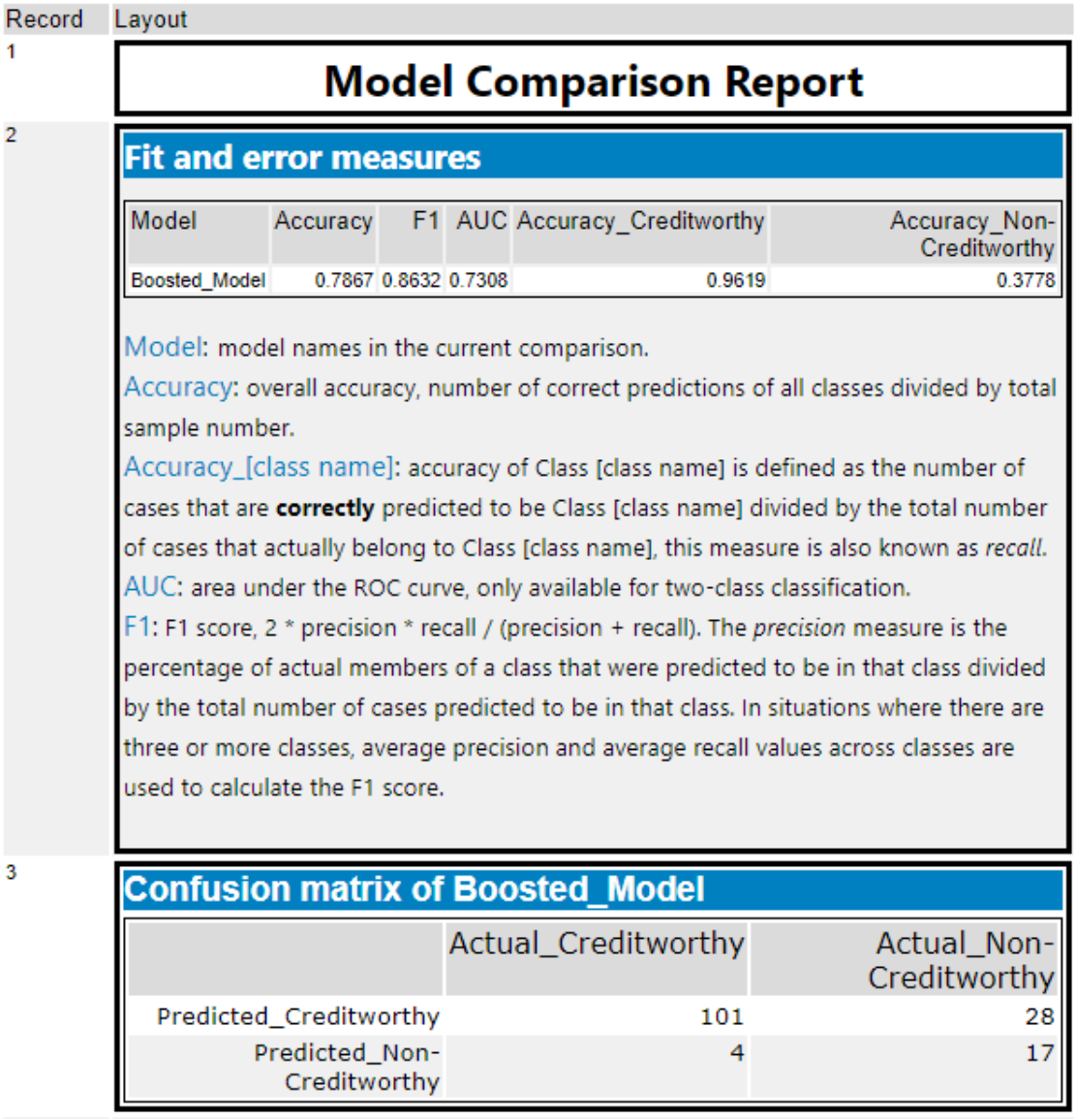


Fig. 3.11. Top variables importance plot of the Boosted Model.

b. Against the validation set the overall percent of accuracy is 78.67%, were the Accuracy_Creditworthy is 96%, meaning that it does a very good job predicting if a person is Creditworthy. As for the bias, the positive predictive value (PPV) is 78% and the negative predictive value (NPV) is 81%; because the values are close to each other, this model has no bias.



Record | Layout

1

Model Comparison Report

2

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted_Model	0.7867	0.8632	0.7308	0.9619	0.3778

Model: model names in the current comparison.

Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.

Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.

AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

3

Confusion matrix of Boosted_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

Fig. 3.12. Model comparison report of the Boosted Model.

Step 4: Writeup

I came up with the Forest model being the best suited for this business problem by taking the following steps:

- I first identified the field that needed to be removed from our dataset because of their low variability or were missing a lot of data fields, and imputed a field for the sake of consistency in the data cleanup process;
- Then, I have created a sample set were 70% of values represented the estimation set and 30% the validation set and I choose the best suited variables by using the variable importance plot, confusion matrix and the overall accuracy of each model;
- Afterwards, I have compared all of the models performance against each other and decided that the Forest model is the best-suited based on the overall accuracy against the validation set, accuracy within both segments, Creditworthy and Non-Creditworthy, ROC graph and the bias in the confusion matrices;
- In the end, I have scored the new customers and predicted that 400 out of the 500 applicants are Creditworthy.

---

1. The model I opted to use for this business problem is the Forest Model, based on the followings:
    a. Against the validation set, the overall accuracy of the Forest Model is higher than the other models, which is 82%;
    b. Within the two segments the Forest Model has the highest accuracies, 97.14% at predicting if a person is creditworthy, and 46.67% if non-creditworthy, which ensures that loans are provided to people that are creditworthy and not provided to those that are non-creditworthy;

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Tree_Model	0.7133	0.8000	0.6965	0.8190	0.4667
Forest_Model	0.8200	0.8831	0.7420	0.9714	0.4667
Boosted_Model	0.7867	0.8632	0.7308	0.9619	0.3778
Stepwise_Logistic_Model	0.7400	0.8368	0.7096	0.9524	0.2444

Fig. 4.1. Overall and segment accuracy of all models.

c. By further verifying the ROC graph, we can see that the Forest Model is the highest of the models, meaning that it is a better classifier;

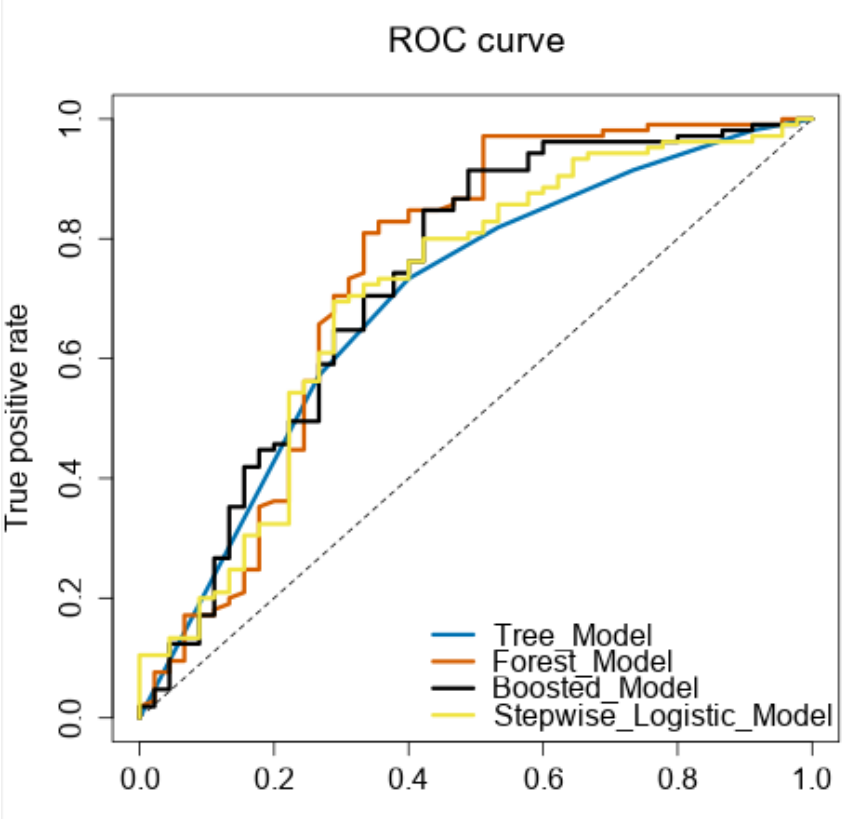


Fig. 4.2. ROC curve of all models.

d. The bias analysis comparing PPV and NPV of the four models indicates that only the Tree model has a high bias to creditworthy, where the other three models have no bias.

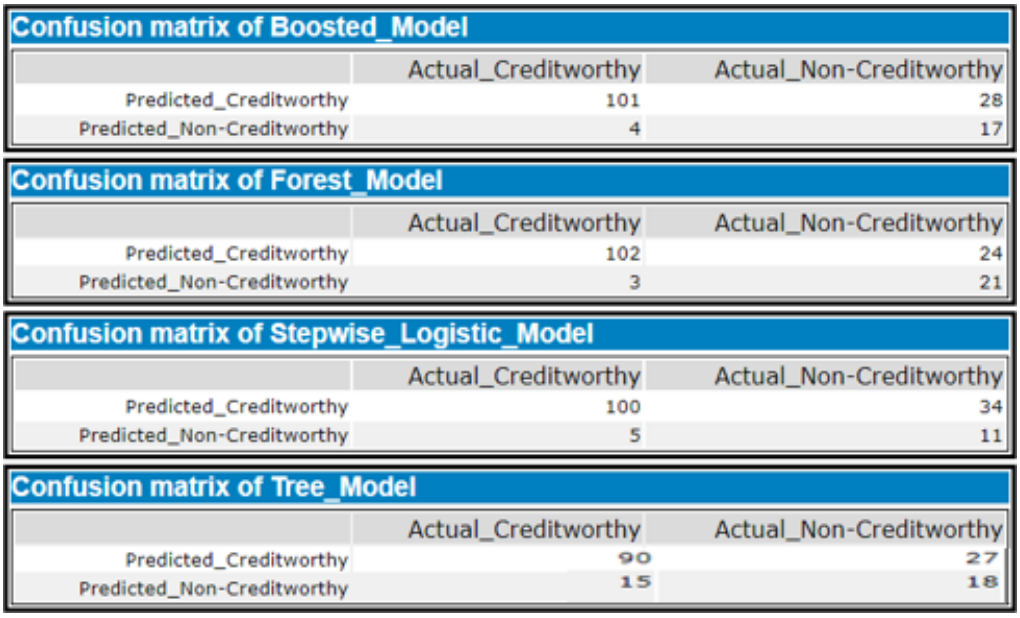


Fig. 3.3. Confusion matrices of all models.

Tab. 3.1. Bias analysis of the four models.

Model	PPV	NPV
Boosted Model	78%	81%
Forest Model	81%	88%
Stepwise_Logistic Model	75%	69%
Tree Model	77%	55%

2. Using the model that provides the highest prediction accuracy for Creditworthy segment, the Forest Model, there are a total of 400 individuals that are creditworthy.