Project: Creditworthiness

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (250 word limit)

Key Decisions:

Answer these questions

- **What decisions needs to be made?**
  Find whether customers who applied for loan are creditworthy.
- **What data is needed to inform those decisions?**
  All past applications data, list of customers whose data need to be processed.
- **What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?**
  Binary model such as logistics regression, Decision Tree, Forest model and boosted tree will be used to find customer creditworthiness.

: Awesome: Good job identifying the key decision to be made.

: Suggestion: This list could have been more detailed, e.g. by explicitly listing some of the factors that might influence our decision, like the applicant's current length of employment, income, credit score, if the customer carries a credit balance from month to month, and their current savings.

: Awesome: Well done identifying the correct type of model to be used.

Step 2: Building the Training Set

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

A As per association analysis done on numerical variable shows that there are no variables which are highly correlated with each other.

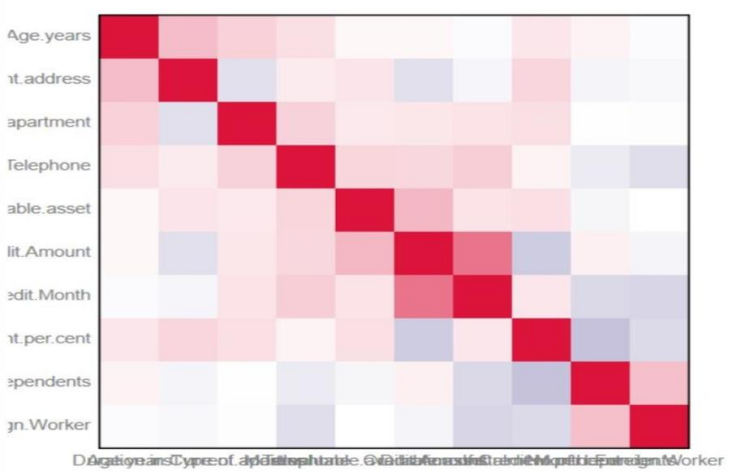


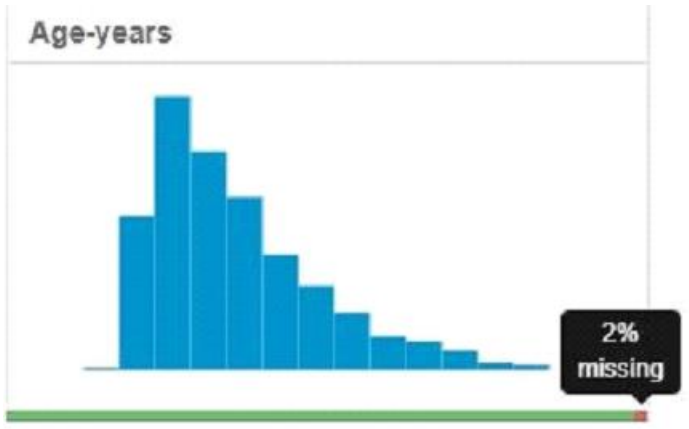
Fig. 2.1 Correlation Matrix with Scatterplot

Concurrent Credit as well as Occupation have one value only so we have to remove it. Telephone is irrelevant field for our problem with lowest correlation towards target variable hence we have to remove it. Foreign Worker and No of dependents show low variability. There is almost 80% data are skewed towards one data hence we have to remove them. Duration-in-current-address has 69% missing data hence we will remove it.

: Awesome: All the correct fields have been removed alongwith appropriate justification.

Age–years has 2% missing data hence we frame the missing data with median of age.

: Awesome: This decision is absolutely correct. We impute because not much data is missing for the age field. And we use the median to impute because of the presence of a slight left skew in it's distribution.



: Awesome: Nice work including the visualizations for each of the field's distributions.

Step 3: Train your Classification Models

*Answer these questions for **each model** you created:*

- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.
  Logistic Regression
  Account balance, Payment status, purpose, credit amount, length of current employment, instalment per cent, Most valuable available asset are the predictor variable as per Logistic Regression.

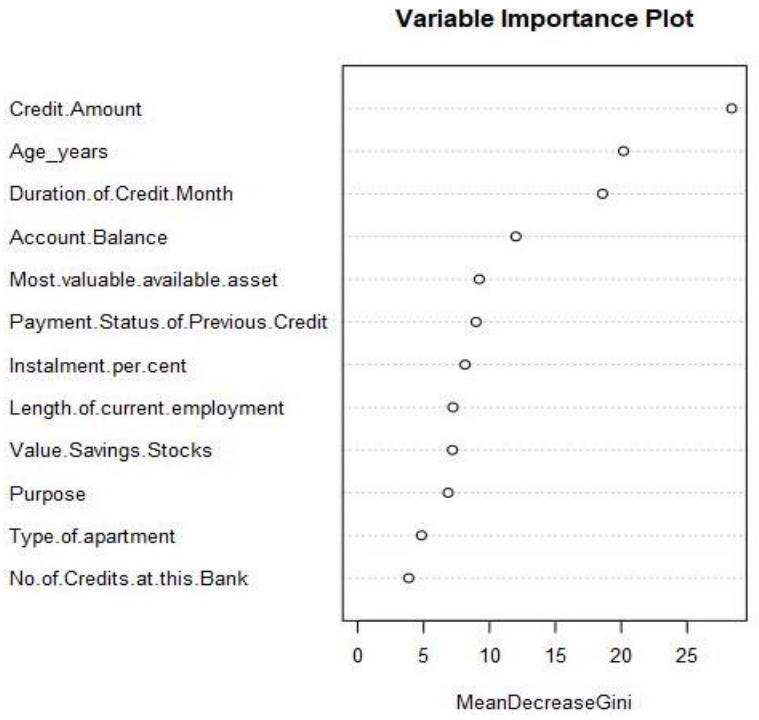
: Suggestion: It's also suggested to mention if you used the stepwise tool of the logistic regression model. Recall from the lessons that stepwise automates the process of coming up with the best predictor variables for us.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.0136120	1.013e+00	-2.9760	0.00292 **
Account.BalanceSome Balance	-1.5433699	3.232e-01	-4.7752	1.79e-06 ***
Duration.of.Credit.Month	0.0064973	1.371e-02	0.4738	0.63565
Payment.Status.of.Previous.CreditPaid Up	0.4054309	3.841e-01	1.0554	0.29124
Payment.Status.of.Previous.CreditSome Problems	1.2607175	5.335e-01	2.3632	0.01812 *
PurposeNew car	-1.7541034	6.276e-01	-2.7951	0.00519 **
PurposeOther	-0.3191177	8.342e-01	-0.3825	0.70206
PurposeUsed car	-0.7839554	4.124e-01	-1.9008	0.05733 .
Credit.Amount	0.0001764	6.838e-05	2.5798	0.00989 **
Value.Savings.StocksNone	0.6074082	5.100e-01	1.1911	0.23361
Value.Savings.Stocks£100-£1000	0.1694433	5.649e-01	0.3000	0.7642
Length.of.current.employment4-7 yrs	0.5224158	4.930e-01	1.0596	0.28934
Length.of.current.employment< 1yr	0.7779492	3.956e-01	1.9664	0.04925 *
Instalment.per.cent	0.3109833	1.399e-01	2.2232	0.0262 *
Most.valuable.available.asset	0.3258706	1.556e-01	2.0945	0.03621 *
Type.of.apartment	-0.2603038	2.956e-01	-0.8805	0.3786
No.of.Credits.at.this.BankMore than 1	0.3619545	3.815e-01	0.9487	0.34275
Age_years	-0.0141206	1.535e-02	-0.9202	0.35747

Decision Tree
Account Balance, Value savings stocks, duration of credit month, credit amount, most valuable available asset, payment status of previous credit, no of credit at this bank, length of current employment and age_years are the predictor variable as per Decesion tree model.

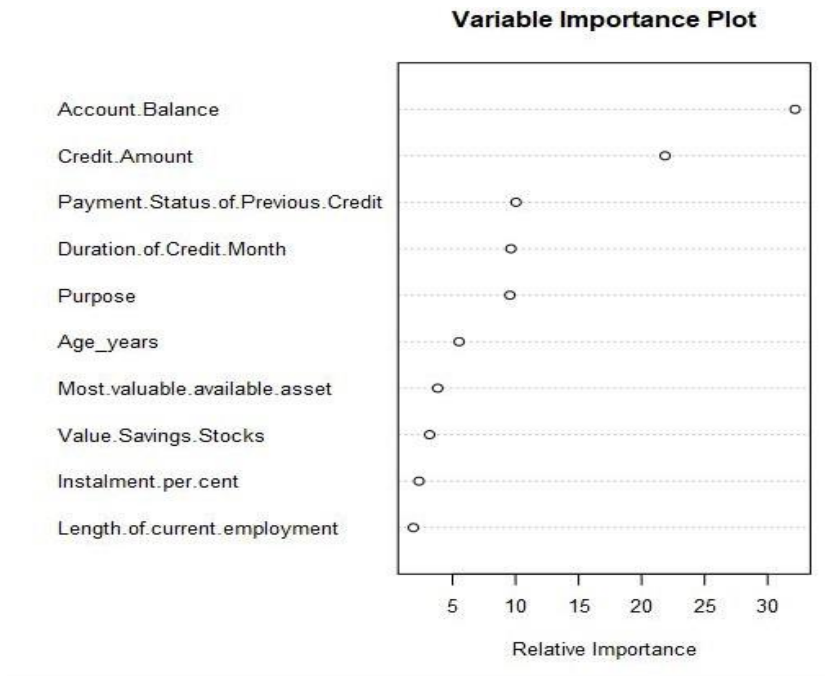
Forest Model
Credit Amount, age_years, duration of credit month, account balance most valuable available asset, payment status of previous credit, instalment per cent, length of current

employment, value saving stocks, purpose, type of apartment, no of credit at this bank are the predictor variable for forest model.



Boosted Model

Account balance, credit amount, payment status of previous credit, duration of credit month, purpose, Age_years, most valuable available asset, value saving stocks, instalment per cent, length of current employment are the predictor variable in boosted model.



: Awesome: Good job correctly setting up the four models to come up with the correct list of the most important variables for each of them.

- Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model's predictions?

Accuracy

Accuracy of logistic regression model against validation set is 0.7800
Accuracy of Decision tree model against validation set is 0.7467
Accuracy of Forest model against validation set is 0.8000
Accuracy of Boosted Model against validation set is 0.7867

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LogisticRegression	0.7800	0.8520	0.7314	0.9048	0.4889
DecisionTree	0.7467	0.8273	0.7054	0.8667	0.4667
ForestModel	0.8000	0.8707	0.7361	0.9619	0.4222
BoostedModel	0.7867	0.8632	0.7524	0.9619	0.3778

Fig.3.1 Model Comparison

Confusion Matrices

1. Logistic Regression
   Here we can see that it is slight bias towards predicting non-creditworthy status for clients who are actually creditworthy

Confusion matrix of LogisticRegression

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

Fig.3.2 Confusion matrix of Logistic Regression Model

2. Decision Tree
   We can see that it is also a bias towards predicting non-creditworthy for clients who are creditworthy.

Confusion matrix of DecisionTree

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

Fig.3.3 Confusion matrix of Decision Tree Model

3. Forest model
   Here we can see that this model better predicted the Non-creditworthy applicants (83%) and in case of Creditworthy applicants it is 80% which is good which make this model unbiased .

Confusion matrix of ForestModel

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

Fig.3.4 Confusion matrix of Forest Model

4. Boosted Model
   In this model we can see it predicted creditworthy 78% applicants correctly and non-creditworthy 81% applicants. Hence we can say it make this model unbiased.

Confusion matrix of BoostedModel

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

Fig.3.5 Confusion matrix for Boosted Model

: Awesome: Nice work correctly identifying the presence of bias, if any in the models. As you noted, we wouldn't want to use the logistic regression or the decision tree models, or else we would deny loans to many individuals who are creditworthy.

: Awesome: All the four models have been correctly validated to come up with the correct confusion matrices.

Step 4: Writeup

Answer these questions:

- Which model did you choose to use? Please justify your decision using **all** of the following techniques. Please only use these techniques to justify your decision:
    - Overall Accuracy against your Validation set
    - Accuracies within “Creditworthy” and “Non-Creditworthy” segments
    - ROC graph
    - Bias in the Confusion Matrices

  As per Model comparison shown in Fig. 3.1 we can see that forest model has highest accuracy of 0.8000 followed by boosted model with 0.7867. Logistic Regression Model has 0.7800 accuracy and Decision tree model with 0.7467. In case of Actual Creditworthy category almost all model has performed same around 0.78 to 0.80 to be maximum. As per confusion matrix Figure 3.2, 3.3, 3.4, 3.5, for Non-creditworthy category predication Logistic Regression Model and Decision Tree Model are stand reliable. Forest Model with 0.8261 & Boosted Model with 0.8095 has highest accuracy for Non-Creditworthy Category. We can say Forest Model and Boosted Models are not biased.
  Receiver operating characteristic(ROC) curve help us to visualize the performance of binary classifier. Area under the curve depict better performing classifier. From the ROC Graph we can observe that Forest and Boosted Model perform well. AUC values in fig. 3.1 for forest model’s is 0.7361 & Boosted model has 0.7524. Hence we can say that Forest model performs best.

: Awesome: The final model chosen is absolutely correct and has been appropriately justified.

: Awesome: Nice work with this detailed explanation! From the ROC curve, we can indeed see that the forest model is the highest line along the graph for most of the chart, and it rises the fastest of all models – meaning that we are getting a higher rate of true positive rates vs. false positives. This is important because we do not want to extend loans to people who are not creditworthy.

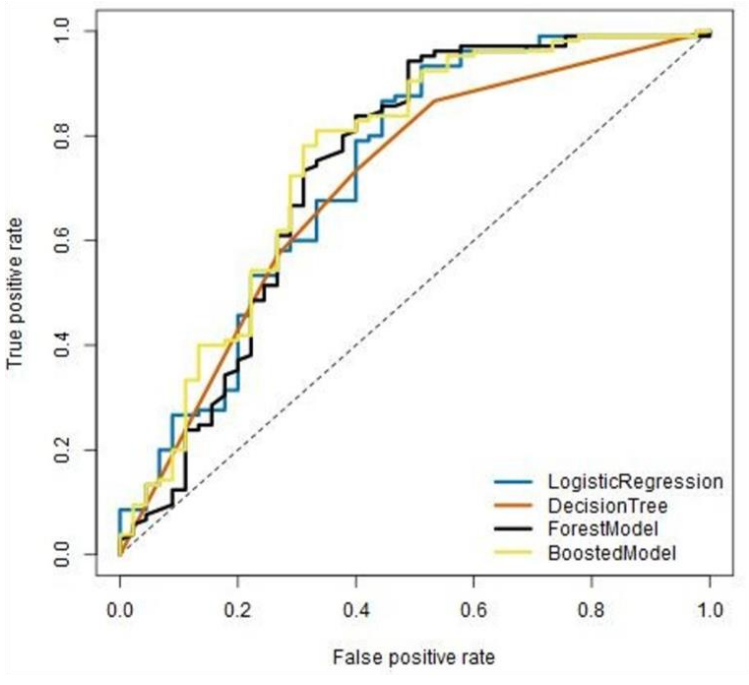


Fig.4.1 ROC Curve

- How many individuals are creditworthy?
  By using Score tool, we can predict that 406 applicants out of 500 are creditworthy.

Sum_X_Creditworthy
406

: Awesome: Great job scoring the model to come up with the correct number of creditworthy individuals.