Project Creditworthiness

Dung Nguyen

Step 1: Business and Data Understanding

What decisions needs to be made?

Determining whether customers are creditworthy to give a loan to.

What data is needed to inform those decisions?

- Data set of past applications for training models:
  - Outcome: Credit-Application-Result.
  - Potention predictors: Duration_of_Credit_Month, Credit_Amount, Age_years, Instalment_per_cent, Most_valuable_available_asset, Type_of_apartment, and so on.
- New data set (without outcome variable) to apply the chosen model and make decisions.

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

Binary model.

Step 2: Building the Training Set

After importing, the data set were formatted as suggested.

Identifying and imputing missing data

The number of missing values in each field:

```
##                              field no_of_missing
## 1          Credit_Application_Result             0
## 2                    Account_Balance             0
## 3           Duration_of_Credit_Month             0
## 4  Payment_Status_of_Previous_Credit             0
## 5                            Purpose             0
## 6                      Credit_Amount             0
## 7                Value_Savings_Stocks             0
## 8        Length_of_current_employment             0
## 9                Instalment_per_cent             0
## 10                        Guarantors             0
## 11        Duration_in_Current_address           344
## 12      Most_valuable_available_asset             0
## 13                         Age_years            12
```

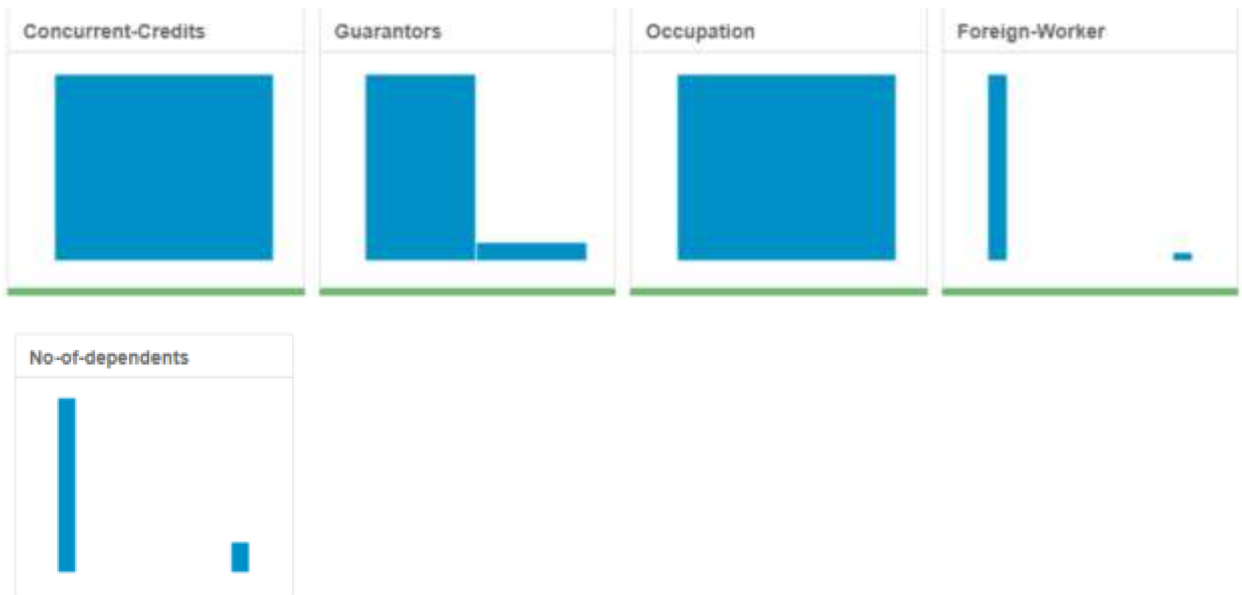
```
## 14             Concurrent_Credits            0
## 15              Type_of_apartment            0
## 16        No_of_Credits_at_this_Bank        0
## 17                     Occupation            0
## 18               No_of_dependents            0
## 19                      Telephone            0
## 20                 Foreign_Worker            0
```

`Duration_in_Current_address` has 344 missing values so I dropped it. `Age_years` has 12 missing values so I imputed it by median.

```
## [1] 35.574
```

Identifying low-variability fields and removing them

Histogram of low-variability fields:



After removing those fields plus `Telephone`, no high correlation was detected among numeric fields:

	FieldName	Duration-of-Credit-Month	Credit-Amount	Instalment-per-cent	Most-valuable-available-asset	Age-years	Type-of-apartment
1	Duration-of-Credit-Month	1	0.57398	0.068106	0.299855	-0.06519	0.152516
2	Credit-Amount	0.57398	1	-0.288852	0.325545	0.068262	0.170071
3	OK: 100.00% cent	0.068106	-0.288852	1	0.081493	0.04008	0.074533
4	Most-valuable-available-asset	0.299855	0.325545	0.081493	1	0.083963	0.373101
5	Age-years	-0.06519	0.068262	0.04008	0.083963	1	0.327718
6	Type-of-apartment	0.152516	0.170071	0.074533	0.373101	0.327718	1

The final data set has 13 columns.

	Name	Type
1	Credit-Application-Result	V_String
2	Account-Balance	V_String
3	Duration-of-Credit-Month	Double
4	Payment-Status-of-Previous-Credit	V_String
5	Purpose	V_String
6	Credit-Amount	Double
7	Value-Savings-Stocks	V_String
8	Length-of-current-employment	V_String
9	Instalment-per-cent	Double
10	Most-valuable-available-asset	Double
11	Age-years	Double
12	Type-of-apartment	Double
13	No-of-Credits-at-this-Bank	V_String

Step 3: Train your Classification Models

Create Estimation and Validation samples

Confusion matrices

Confusion matrix of Boosted_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	27
Predicted_Non-Creditworthy	4	18

Confusion matrix of Decision_Tree

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	26
Predicted_Non-Creditworthy	12	19

Confusion matrix of Forest_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	29
Predicted_Non-Creditworthy	3	16

Confusion matrix of Logistic_Regression

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

Overall accuracy

Among the 4 models, Boosted Model has the highest overall accuracy agains the Validation set (0.7933). Forest Model is the least biased with PPV = 0.78 and NPV = 0.84.

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_Regression	0.7800	0.8520	0.7310	0.9048	0.4889
Decision_Tree	0.7467	0.8304	0.7035	0.8857	0.4222
Forest_Model	0.7867	0.8644	0.7389	0.9714	0.3556
Boosted_Model	0.7933	0.8670	0.7539	0.9619	0.4000

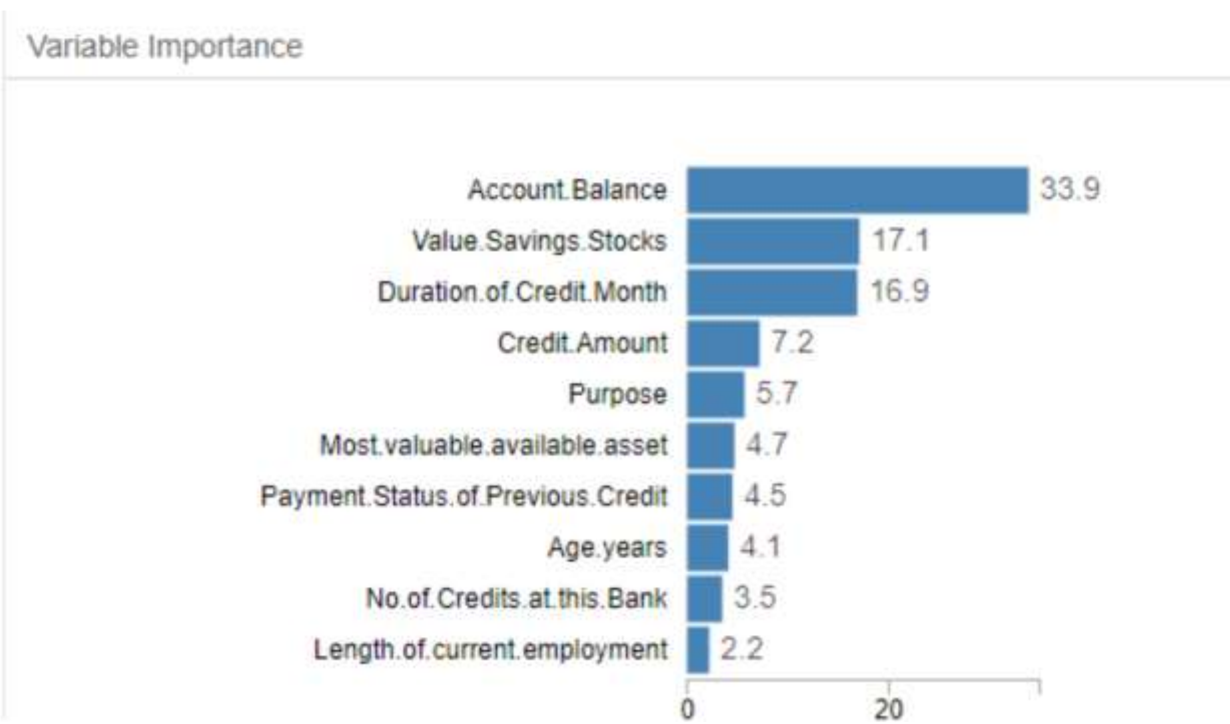
Plot variable importance

Most important predictors: - For Logistic regression: Account_Balance, Purpose, Payment_Status_of_Previous_Credit (Paid up), Length_of_current_employment (<1yr), Instalment_per_cent, and Most_valuable_available_asset. - For Decision tree: Acocunt_Balance, Value_Savings_Stocks, Duration_of_Credit_Month, Credit_Amount, and Purpose. - For Forest Model: Credit_Amount, Age_years, Duration_of_Credit_Month, Account_Balance, and Most_valuable_available_asset. - For Boosted tree: Account_Balance, Credit_Amount, Duration_of_Credit_Month, Paymen_Status_of_Previous_Credit, and Purpose. The difference is not remarkable among the fields.

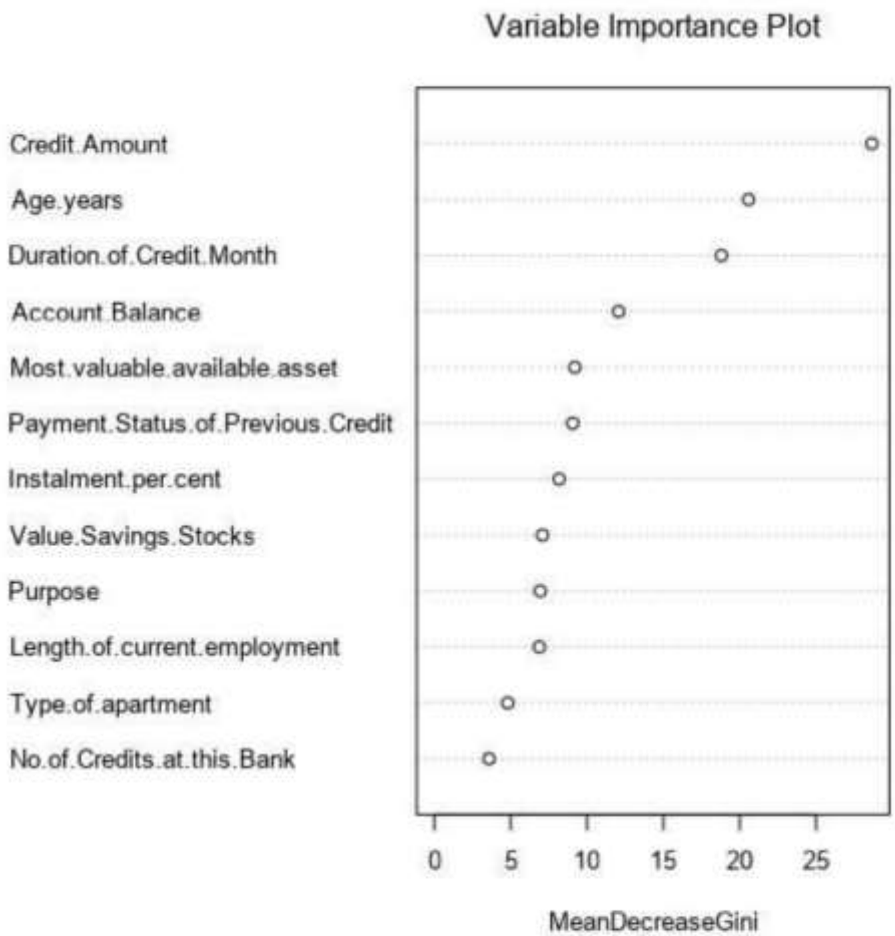
- Logistic Regression:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.990817	1.013e+00	-2.9527	0.00315 **
Account.BalanceSome Balance	-1.543669	3.233e-01	-4.7745	1.80e-06 ***
Duration.of.Credit.Month	0.006391	1.371e-02	0.4660	0.6412
Payment.Status.of.Previous.CreditPaid Up	0.402974	3.843e-01	1.0487	0.2943
Payment.Status.of.Previous.CreditSome Problems	1.259683	5.334e-01	2.3616	0.0182 *
PurposeNew car	-1.755074	6.278e-01	-2.7954	0.00518 **
PurposeOther	-0.290165	8.359e-01	-0.3471	0.72848
PurposeUsed car	-0.785627	4.124e-01	-1.9049	0.05679 .
Credit.Amount	0.000177	6.841e-05	2.5879	0.00966 **
Value.Savings.StocksNone	0.609298	5.099e-01	1.1949	0.23213
Value.Savings.Stocks£100-£1000	0.172241	5.649e-01	0.3049	0.76046
Length.of.current.employment4-7 yrs	0.530959	4.932e-01	1.0767	0.28163
Length.of.current.employment< 1yr	0.777372	3.957e-01	1.9646	0.04946 *
Instalment.per.cent	0.310524	1.399e-01	2.2197	0.02644 *
Most.valuable.available.asset	0.325606	1.557e-01	2.0918	0.03645 *
Age.years	-0.015092	1.539e-02	-0.9809	0.32666
Type.of.apartment	-0.254565	2.958e-01	-0.8605	0.38949
No.of.Credits.at.this.BankMore than 1	0.362688	3.816e-01	0.9505	0.34184

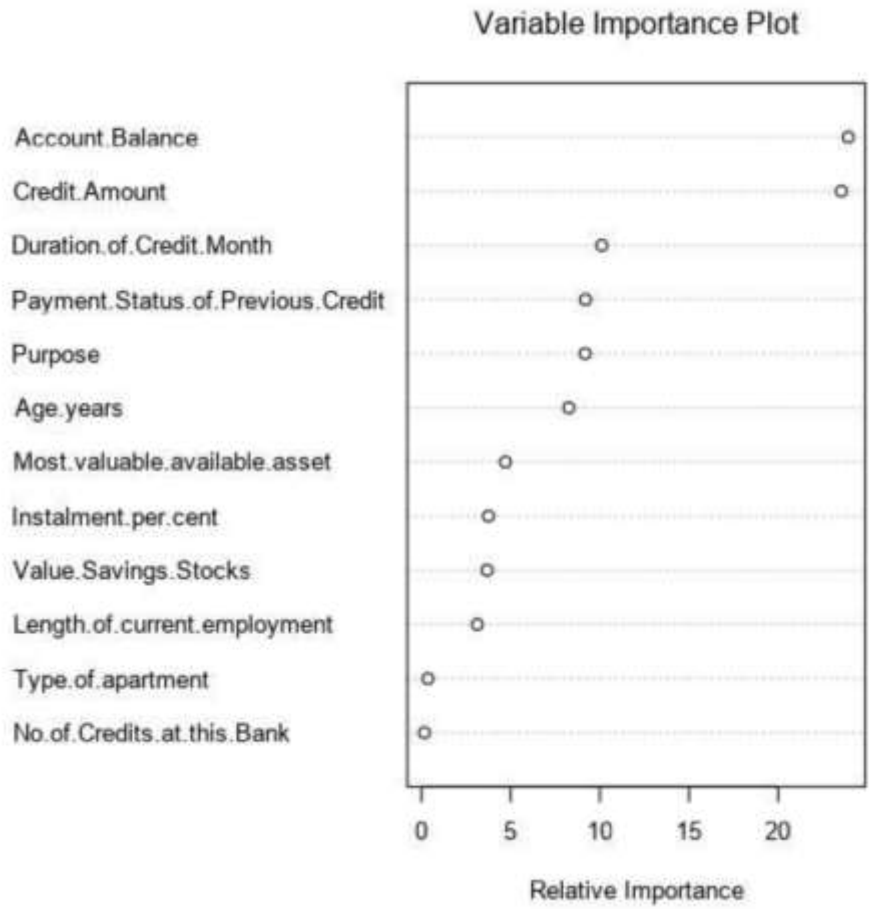
- Decision Tree:



- Forest Model:



- Boosted Model:

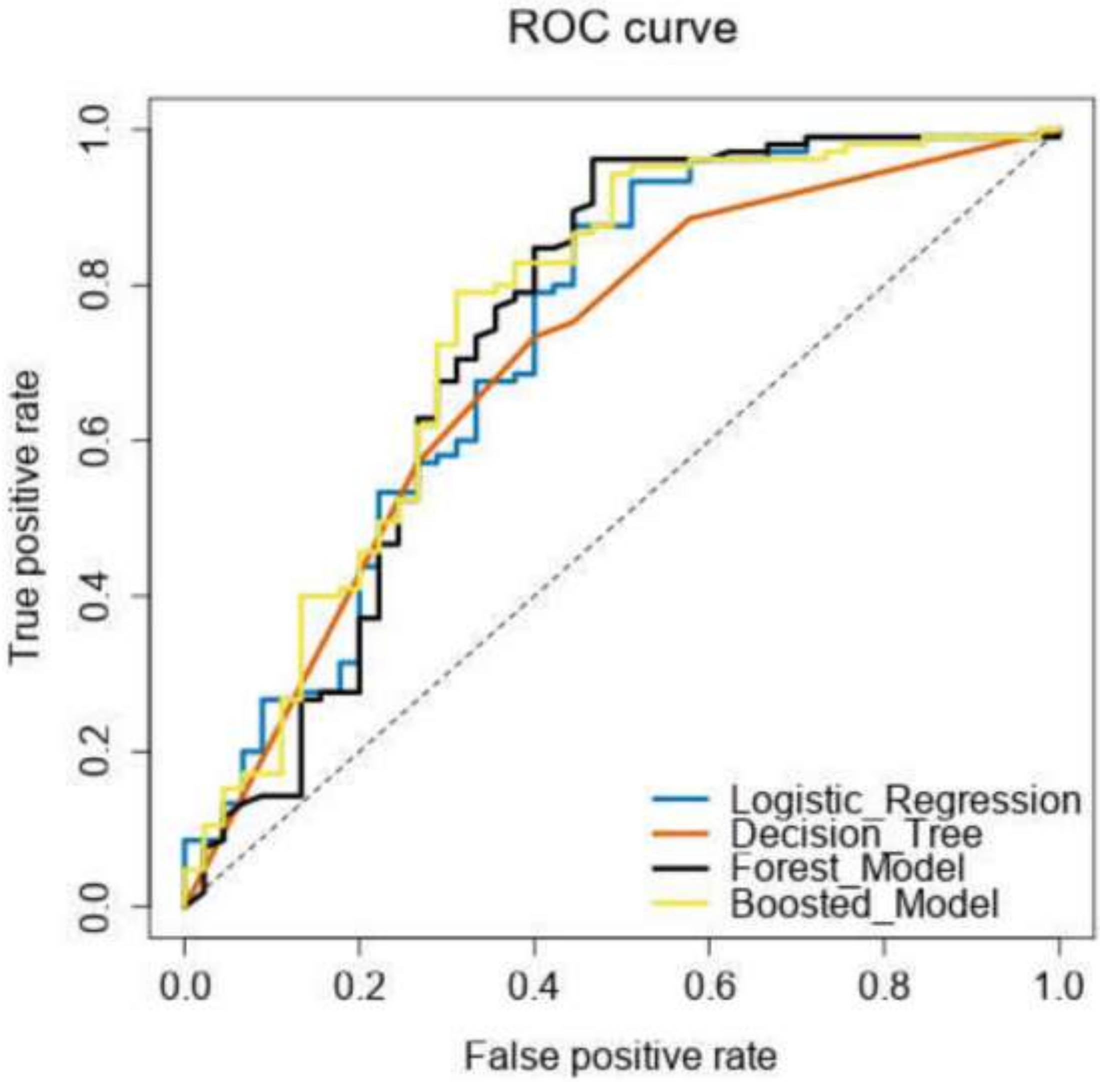


Step 4: Writeup

Accuracy table

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_Regression	0.7800	0.8520	0.7310	0.9048	0.4889
Decision_Tree	0.7467	0.8304	0.7035	0.8857	0.4222
Forest_Model	0.7867	0.8644	0.7389	0.9714	0.3556
Boosted_Model	0.7933	0.8670	0.7539	0.9619	0.4000

Plot ROCs



Forest Model is the least biased and has the nearly highest overall accuracy against the Validation set. Although its accuracy within Non-creditworthy segment is not high, its respective index in Creditworthy segment is the highest. Forest Model also has the most optimal ROC which reaching the top quicker than Boosted Model. Therefore, I choose to use Forest Model. Applying the chosen model, the number of individuals are creditworthy is 412.