Project: Creditworthiness

Complete each section. When you are ready, save your file as a PDF document and submit it here: https://classroom.udacity.com/nanodegrees/nd008/parts/11a7bf4c-2b69-47f3-9aec-108ce847f855/project

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?

  The bank has a list of 500 new loan applications to process in a week and approves the applications if the customers are creditworthy. An accurate prediction model is required to classify the new customers' applications on whether the customers can be approved for a loan or not, i.e. to detect if the customer is Non-Creditworthy.

- What data is needed to inform those decisions?

  We need a dataset of all past applications to create and train a model (For training and validation purposes) and a dataset that contains a list of customers (For testing/scoring purpose) who have applied to get a loan.

  Specifically, the following variables/fields are required for the final modelling:-
  1. Credit-Application-Result (Labelling required for the training/validation dataset)
  2. Account-Balance
  3. Duration-of-Credit-Month
  4. Payment-Status-of-Previous-Credit
  5. Purpose
  6. Credit-Amount
  7. Value-Savings-Stocks
  8. Length-of-current-employment
  9. Instalment-per-cent
  10. Most-valuable-available-asset
  11. Age-years Concurrent-Credits
  12. Type-of-apartment
  13. No-of-Credits-at-this-Bank

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

  A binary classification model is required to determine if a customer is Creditworthy or Non-creditworthy.

Step 2: Building the Training Set

Build your training set given the data provided to you. The data has been cleaned up for you already so you shouldn't ***need to convert any data fields to the appropriate data types.***

Here are some guidelines to help guide your data cleanup:

- For numerical data fields, are there any fields that highly-correlate with each other? The correlation should be at least .70 to be considered "high".

- Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed

- Are there only a few values in a subset of your data field? Does the data field look very uniform (there is only one value for the entire field?). This is called "low variability" and you should remove fields that have low variability. Refer to the "Tips" section to find examples of data fields with low-variability.

- Your clean data set should have 13 columns where the Average of **Age Years** should be 36 (rounded up)

Note: *For the sake of consistency in the data cleanup process, impute data using the median of the entire data field instead of removing a few data points. (100 word limit)*

Note: *For students using software other than Alteryx, please format each variable as:*

Variable	**Data Type**
Credit-Application-Result	String
Account-Balance	String
Duration-of-Credit-Month	Double
Payment-Status-of-Previous-Credit	String
Purpose	String

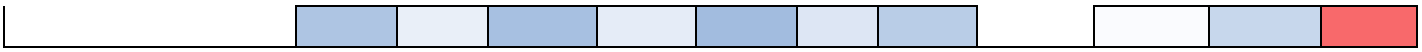
Credit-Amount	Double
Value-Savings-Stocks	String
Length-of-current-employment	String
Instalment-per-cent	Double
Guarantors	String
Duration-in-Current-address	Double
Most-valuable-available-asset	Double
Age-years	Double
Concurrent-Credits	String
Type-of-apartment	Double
No-of-Credits-at-this-Bank	String
Occupation	Double
No-of-dependents	Double
Telephone	Double
Foreign-Worker	Double

To achieve consistent results reviewers expect.

Answer this question:

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

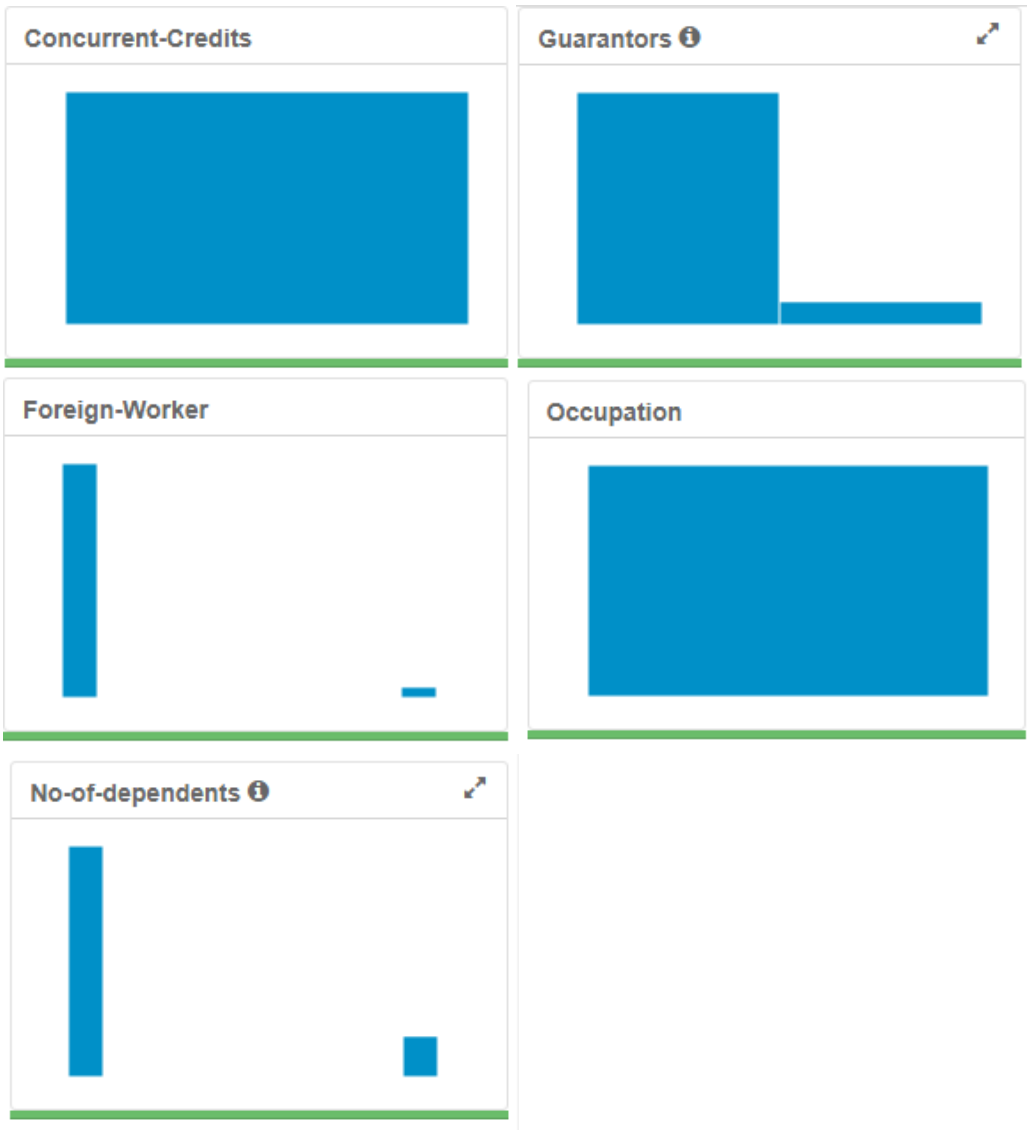
• FieldName	Duration-of-Credit-Month	Credit-Amount	Instalment-per-cent	Duration-in-Current-address	Most-valuable-available-asset	Age-years	Type-of-apartment	Occupation	No-of-dependents	Telephone	Foreign-Worker
Duration-of-Credit-Month	1.00	0.57	0.07	-0.12	0.30	-0.06	0.15		-0.07	0.14	-0.12
Credit-Amount	0.57	1.00	-0.29	-0.15	0.33	0.07	0.17		0.00	0.29	0.03
Instalment-per-cent	0.07	-0.29	1.00	0.08	0.08	0.04	0.07		-0.13	0.03	-0.13
Duration-in-Current-address	-0.12	-0.15	0.08	1.00	-0.32	0.14	-0.12		-0.03	-0.02	0.02
Most-valuable-available-asset	0.30	0.33	0.08	-0.32	1.00	0.09	0.37		0.05	0.20	-0.15
Age-years	-0.06	0.07	0.04	0.14	0.09	1.00	0.33		0.12	0.18	0.00
Type-of-apartment	0.15	0.17	0.07	-0.12	0.37	0.33	1.00		0.17	0.10	-0.09
Occupation								1.00			
No-of-dependents	-0.07	0.00	-0.13	-0.03	0.05	0.12	0.17		1.00	-0.05	0.07
Telephone	0.14	0.29	0.03	-0.02	0.20	0.18	0.10		-0.05	1.00	-0.06
Foreign-Worker	-0.12	0.03	-0.13	0.02	-0.15	0.00	-0.09		0.07	-0.06	1.00



There is no highly correlated (i.e. pearson correlation > 0.7) numerical variables.

Name	Plot	% Missing	Unique Values	Min	Mean	Median	Max	Std Dev	Remarks
Age-years		2.4%	54	19.000	35.637	33.000	75.000	11.502	
Duration-in-Current-address		68.8%	5	1.000	2.660	2.000	4.000	1.150	This field has over 10% missing values. Consider imputing these values. This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".

There are 2.4% and 68.8% missing data in the Age-years and Duration-in-Current-address Fields, respectively. The missing Age-years were imputed with Median of Non-zero values. The Duration-in-Current-address carries too many missing values - to be removed from modelling.



Besides, the Concurrent-Credits (Only contains 'other bank/dept' values), Guarantors (457 "Yes" vs 43 "None"), Foreign-Worker (481 "1" vs 19 "2"), Occupation (Only contains '1' values), and No-of-dependents (427 "1" vs 73 "2") Fields were removed due to low variability. Lastly, the Telephone Field was removed due to irrelevancy.

Step 3: Train your Classification Models

First, create your Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. Set the Random Seed to 1.

Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model

*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.

- 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?

You should have four sets of questions answered. (500 word limit)

Logistic Regression

Type II Analysis of Deviance Tests

Response: Credit.Application.Result

	LR Chi-Sq	DF	Pr(>Chi-Sq)
Account.Balance	25.507	1	4.40e-07 ***
Duration.of.Credit.Month	0.223	1	0.63679
Payment.Status.of.Previous.Credit	5.75	2	0.05641 .
Purpose	12.479	3	0.00591 **
Credit.Amount	7.434	1	0.0064 **
Value.Savings.Stocks	2.737	2	0.25451
Length.of.current.employment	4.155	2	0.12524
Instalment.per.cent	5.148	1	0.02327 *
Most.valuable.available.asset	4.486	1	0.03418 *
Age.years	0.865	1	0.35245
Type.of.apartment	0.774	1	0.37904
No.of.Credits.at.this.Bank	0.904	1	0.34164

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

Logistic regression coefficient:

Coefficients:

	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 *
Age.years	-0.0141206	1.535e-02	-0.9202	0.35747
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

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

Type II Analysis of Deviance and Logistic Regression suggested most significant variables as 1) Account Balance, 2) Purpose, 3) Credit Amount, 4) Instalment per.cent, 5) Most.valuable.available.asset.

Fit and error measures

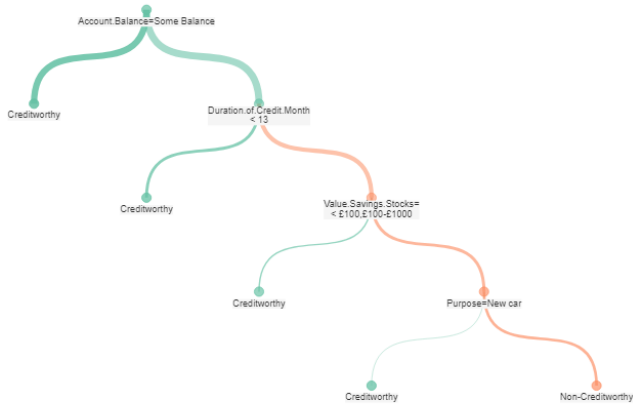
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LogisticRegression	0.7800	0.8520	0.7314	0.9048	0.4889

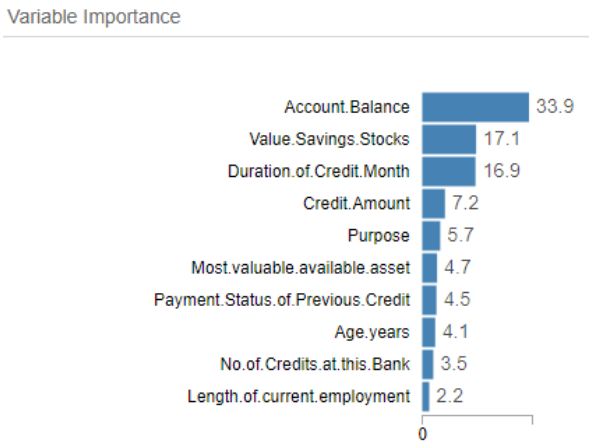
Confusion matrix of LogisticRegression

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

The overall accuracy of Logistic Regression is 0.7800. The Accuracy to predicting Creditworthy (0.9048) is much higher than predicting Non-Creditworthy (0.4889).

Decision Tree





The top 4 most important features suggested by the Decision Tree are Account Balance, Value.Savings.Stocks, Duration.of.Credit.Month, and Credit.Amount.

Fit and error measures

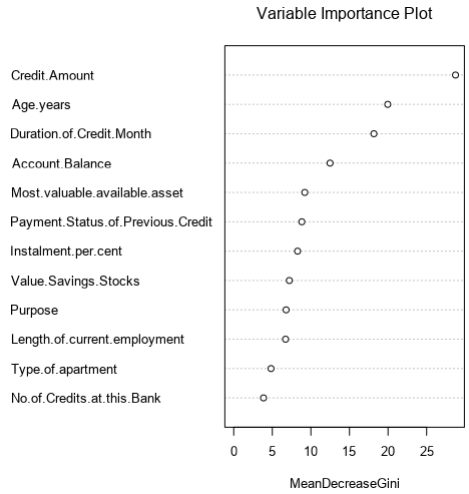
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT	0.7467	0.8304	0.7035	0.8857	0.4222

Confusion matrix of DT

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

The overall accuracy of Decision Tree is 0.7467. The Accuracy to predicting Creditworthy (0.8857) is much higher than predicting Non-Creditworthy (0.4222).

Random Forest



The top 4 most important features suggested by the Random Forest are Credit_Amount, Age.years, Duration.of.Credit.Month, and Account.Balance.

Fit and error measures

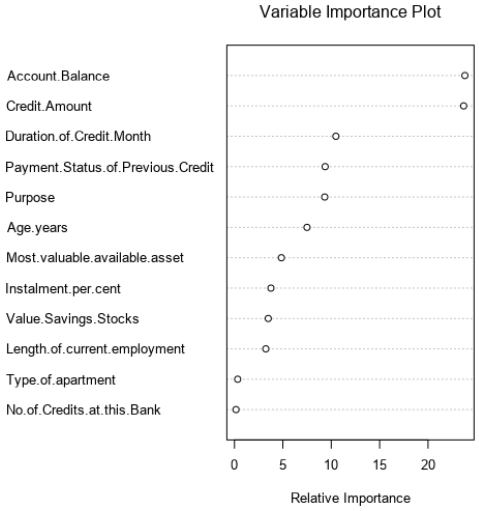
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
RF	0.7933	0.8681	0.7368	0.9714	0.3778

Confusion matrix of RF

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

The overall accuracy of Random Forest is 0.7933. The Accuracy to predicting Creditworthy (0.9714) is much higher than predicting Non-Creditworthy (0.3778).

Boosted Model



The top 4 most important features suggested by the Random Forest are Account.Balance, Credit_Amount, Duration.of.Credit.Month, and Payment.Status.of.Previous.Credit.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted	0.7867	0.8632	0.7507	0.9619	0.3778

Confusion matrix of Boosted

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

The overall accuracy of Boosted Model is 0.7867. The Accuracy to predicting Creditworthy (0.9619) is much higher than predicting Non-Creditworthy (0.3778).

Step 4: Writeup

Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as "Creditworthy"

Write a brief report on how you came up with your classification model and write down how many of the new customers would qualify for a loan. (250 word limit)

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

Note: Remember that your boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments.

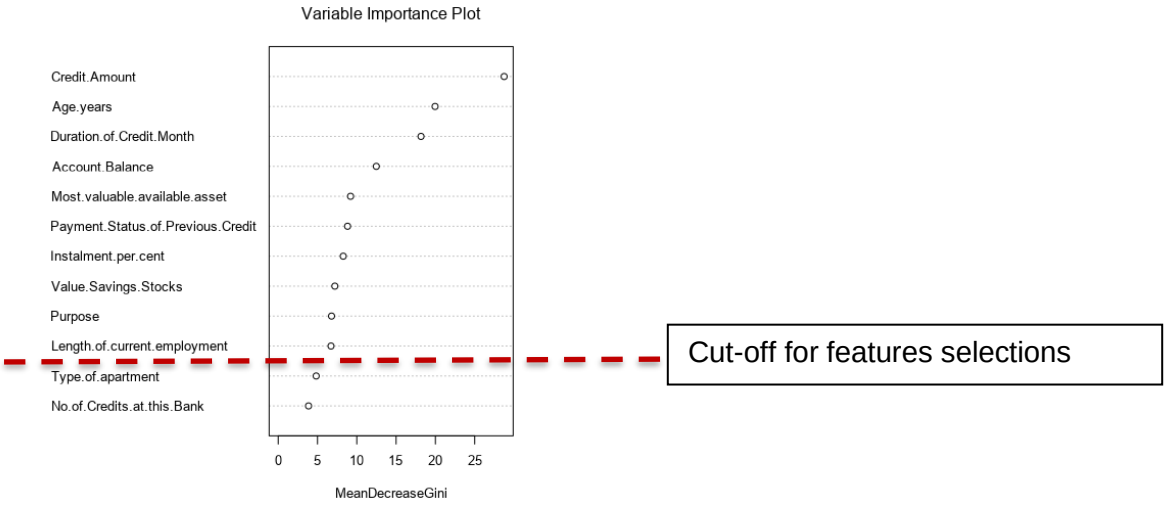
Overall Accuracy against your Validation set & Accuracies within "Creditworthy" and "Non-Creditworthy" segments

Fit and error measures

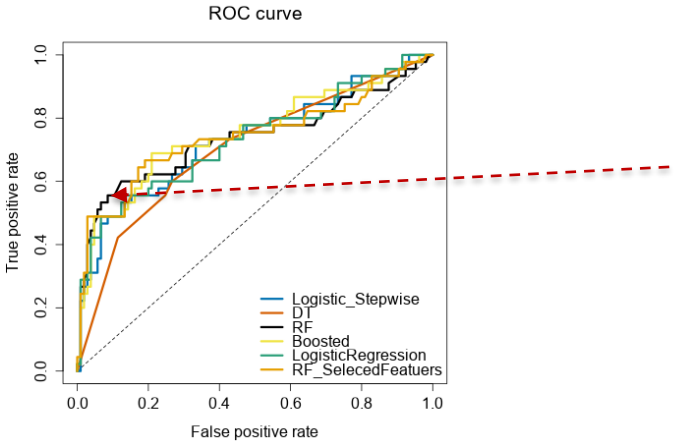
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_Stepwise	0.7600	0.8364	0.7306	0.8762	0.4889
DT	0.7467	0.8304	0.7035	0.8857	0.4222
RF	0.7933	0.8681	0.7368	0.9714	0.3778
Boosted	0.7867	0.8632	0.7507	0.9619	0.3778
LogisticRegression	0.7600	0.8520	0.7314	0.9048	0.4889
RF_SelecedFeatuers	0.8200	0.8831	0.7420	0.9714	0.4667

As stated, the boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments. The Random Forest (RF) has the highest validation accuracy among the Logistic Regression (LR), Decision Tree (DT), RF, and Boosted Model (BM) are the RF and BM. It also has highest accuracy in detecting Creditworthy, i.e. 0.9714. However, admittedly, it has lowest accuracy in detecting Non-Creditworthy.

In order to improve the model performance, feature selections were done using Stepwise method on LR and Elbow point on RF (cutoff indicated below). The RF with the 10 selected features outperformed all models, with overall accuracy on validation dataset of 0.82, with accuracy to predict creditworthy and non-creditworthy on validation dataset reaching 0.9714 and 0.4667 respectively.



ROC graph



When comparing ROC curves (by setting the Non-Creditworthy as the positive) between the models, the highest AUC among the LR, DT, RF, BM, stepwise-LR, and features-selected RF were achieved by the RF, while the lowest were achieved by the DT. Also, the RF is also the quickest model (As pointed in red) to have highest True Positive with a given same False Positive rate.

Bias in the Confusion Matrices

Confusion matrix of Boosted

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

Confusion matrix of DT

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

Confusion matrix of LogisticRegression

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

Confusion matrix of Logistic_Stepwise

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

Confusion matrix of RF

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

Confusion matrix of RF_SelecedFeatuers

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	24
Predicted_Non-Creditworthy	3	21

Among the LR, DT, RF, and BM, the RF is the best in identifying Actual Creditworthy but not the Non-Creditworthy. The features-selected RF successfully increased the accuracy in the detecting the Non-Creditworthy. As observed above, the features-selected RF seems to be effectively reducing the overfitting of the non-features selected RF, i.e. produced highest overall accuracy.

For the purpose of this tutorial as well as to meet the boss expectation that he only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments, we chose the non-features-selected RF, which has the highest Overall Accuracy and F1 (A metric that weight both recall and precision equally) scores among the LR, DT, RF, and BM.

- How many individuals are creditworthy?

Record	Sum_Score_Creditworthy	Sum_Score_Non-Creditworthy
1	408	90

By using the trained non-features-selected RF model, 408 and 90 individuals are predicted to be Creditworthy and Non-Creditworthy, respectively.

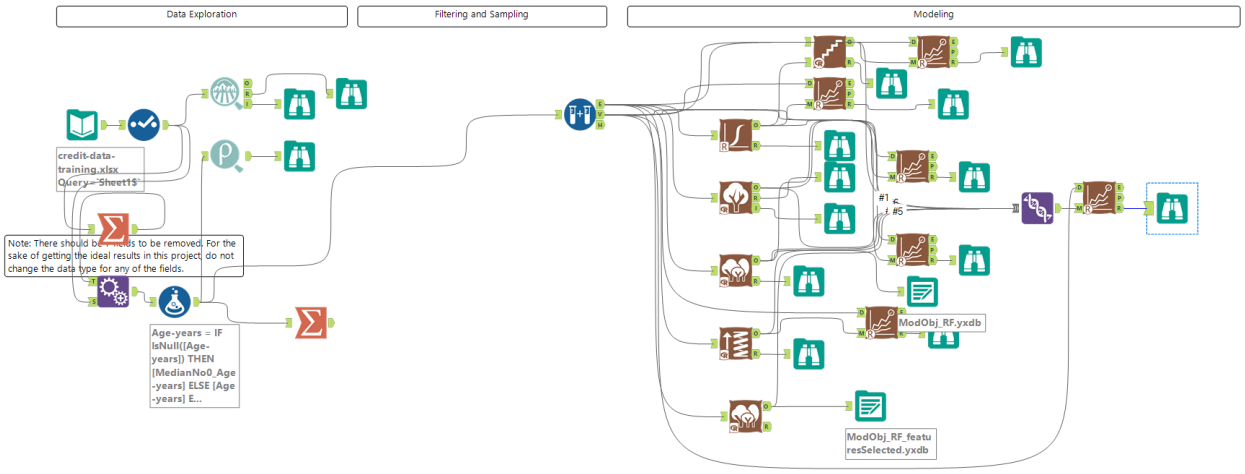
Lastly, author's opinion -- although RF was selected as the final model for prediction, to approve for a loan, I will be particularly care about/putting higher weight on the accuracy in identifying non-creditworthy, i.e. to be more sensitive in detecting the Non-creditworthy, in order to minimize the loss due to credit default. The highest of the Accuracy in Non-Creditworthy metric is currently achieved by the LR among the LR, DT, RF, and BM, i.e. 0.4889. A good metric to evaluate this would be F1.5 or F2 etc, where recall is given 1.5x or 2x higher weight than the precision score.

Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.

Appendixes:

Models training and selection workflow



Scoring workflow with trained model object