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?

ANS: To determine the creditworthiness of new clients applying for loans at our bank systematically.

- What data is needed to inform those decisions?

ANS: Historical data on all past loan applications & List of 500 customers who applied loans recently.

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

ANS: Binary model

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)*

Answer this question:

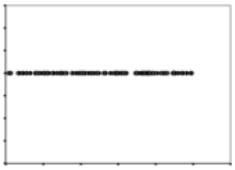
- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

Fields to be removed:

- A lot of missing data: "*Duration-in-Current-address*" field to be removed due to 60% of data is missing

- Two fields with Uniform values: "*Concurrent-Credits*", "Occupation"

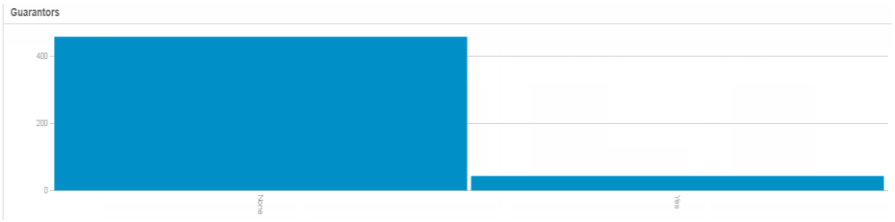
Occupation	0.0%	1	1.000



- Field "*Telephone*" is not relevant to the modeling of creditworthiness;
- Low variabilities fields: "*No-of-dependents*", "*Foreign-Worker*", "*Guarantors*"
  - One dependents vs. two dependents = 427:73 (data heavily skewed to the right, positive skewed, low variability)

  - "Foreign-Worker" field: 1:2 = 481:19 (heavily skewed to the right, positive skewed, low variability)

- “Guarantors”: None : Yes = 457:43



Fields to be imputed: Age-years (2% of missing data) wiith median value 33.

To the numerical predicted variable fields selected, a Pearson correlation analysis is performed.

As we can see below, none of the fields have correlation coefficient r >0.7. Thus, all the numerical predicted variables are independent (not highly correlated) of each other.

Pearson Correlation Analysis

Full Correlation Matrix

	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Most.valuable.available.asset	Age.years	Type.of.apartment
Duration.of.Credit.Month	1.000000	0.570441	0.079515	0.304734	-0.066319	0.153141
Credit.Amount	0.570441	1.000000	-0.285631	0.327762	0.068643	0.168683
Instalment.per.cent	0.079515	-0.285631	1.000000	0.078110	0.040540	0.082936
Most.valuable.available.asset	0.304734	0.327762	0.078110	1.000000	0.085437	0.379650
Age.years	-0.066319	0.068643	0.040540	0.085437	1.000000	0.333075
Type.of.apartment	0.153141	0.168683	0.082936	0.379650	0.333075	1.000000

Matrix of Corresponding p-values

	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Most.valuable.available.asset	Age.years	Type.of.apartment
Duration.of.Credit.Month		0.0000e+00	7.9292e-02	6.0352e-12	1.4350e-01	6.8791e-04
Credit.Amount	0.0000e+00		1.2929e-10	1.1013e-13	1.2996e-01	1.8138e-04
Instalment.per.cent	7.9292e-02	1.2929e-10		8.4757e-02	3.7152e-01	6.7164e-02
Most.valuable.available.asset	6.0352e-12	1.1013e-13	8.4757e-02		5.9299e-02	0.0000e+00
Age.years	1.4350e-01	1.2996e-01	3.7152e-01	5.9299e-02		4.1744e-14
Type.of.apartment	6.8791e-04	1.8138e-04	6.7164e-02	0.0000e+00	4.1744e-14	

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 y*our entire dataset should be reserved for Validation. Set the Random Seed to 1.

Create all of the following mo*dels: 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 Stepwise Model

- Significant predictor variables (p<0.05):
  - Account Balance(Some Balance): p = 1.65 x10^{-7}
  - Purpose (New Car): p = 0.00566
  - Payment Status of Previous Credit (Some Problems): p = 0.0183
  - Credit Amount: p = 0.00296
  - Length of current employment (<1yr): p = 0.03596
  - Instalment per cent: p = 0.0254

Report for Logistic Regression Model Stepwise_Creditworthynss

Basic Summary

Call:
glm(formula = Credit.Application.Result ~ Account.Balance + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset, family = binomial("logit"), data = the.data)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.289	-0.713	-0.448	0.722	2.454

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.9621914	6.837e-01	-4.3326	1e-05	***
Account.BalanceSome Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07	***
Payment.Status.of.Previous.CreditPaid Up	0.2360857	2.977e-01	0.7930	0.42775	
Payment.Status.of.Previous.CreditSome Problems	1.2154514	5.151e-01	2.3595	0.0183	*
PurposeNew car	-1.6993164	6.142e-01	-2.7668	0.00566	**
PurposeOther	-0.3257637	8.179e-01	-0.3983	0.69042	
PurposeUsed car	-0.7645820	4.004e-01	-1.9096	0.05618	.
Credit.Amount	0.0001704	5.733e-05	2.9716	0.00296	**
Length.of.current.employment4-7 yrs	0.3127022	4.587e-01	0.6817	0.49545	
Length.of.current.employment< 1yr	0.8125785	3.874e-01	2.0973	0.03596	*
Instalment.per.cent	0.3016731	1.350e-01	2.2340	0.02549	*
Most.valuable.available.asset	0.2650267	1.425e-01	1.8599	0.06289	.

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial taken to be 1)

Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 328.55 on 338 degrees of freedom
McFadden R-Squared: 0.2048, Akaike Information Criterion 352.5

- Overall model accuracy for validation: 0.7600
- The logistic model is biased towards creditworthy. It is not quite accurate when it comes to determine non-creditworthy.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise_Creditworthynss	0.7600	0.8364	0.7306	0.8762	0.4889

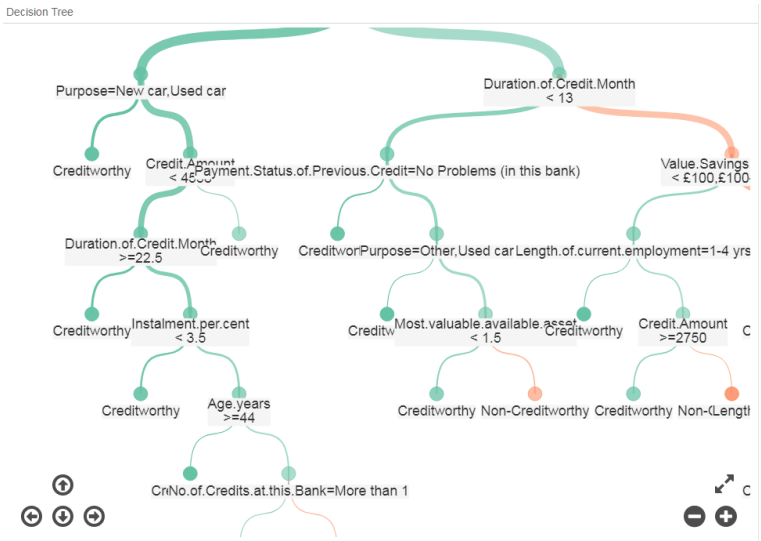
Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided 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.

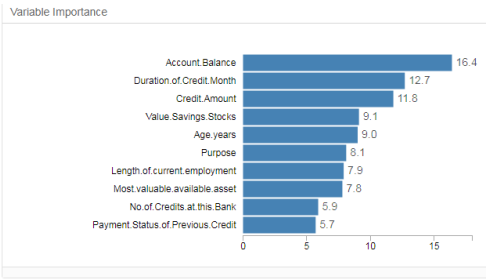
Confusion matrix of Stepwise_Creditworthynss

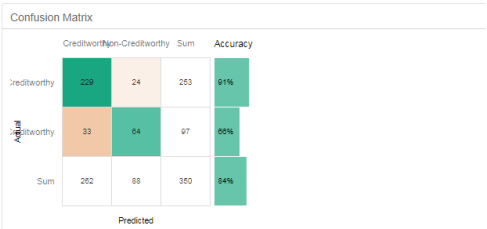
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

Decision Tree

- Significant predictor variables: Account Balance, Duration of Credit Month, Credit Amount, Value Savings Stocks







- The decision tree model has an overall accuracy 0.6667.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Creditworthiness	0.6667	0.7685	0.6272	0.7905	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.

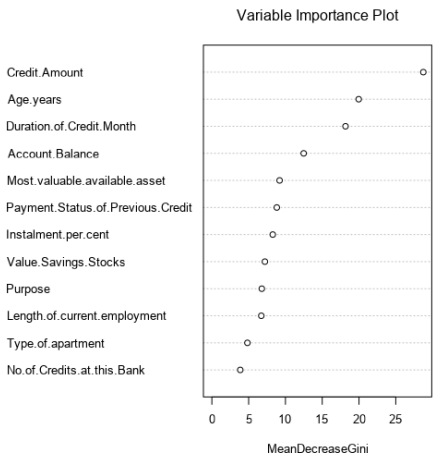
Confusion matrix of DT_Creditworthiness

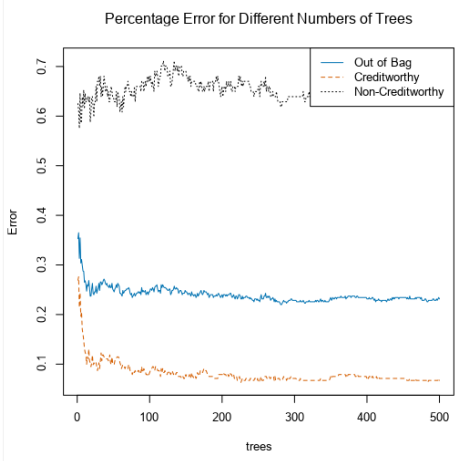
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	28
Predicted_Non-Creditworthy	22	17

- Bias: The model is biased towards Creditworthy applicant. More creditworthy applicants will be refused for loan to application due to misclassification.

Random Forest

- Significant predictor variables: Credit Amount, Age Years, Duration of Credit Month, Account Balance





- Biased: Random Forest has an overall accuracy as 0.79833. The model is very accurate in predicting creditworthy.

Fit and error measures

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

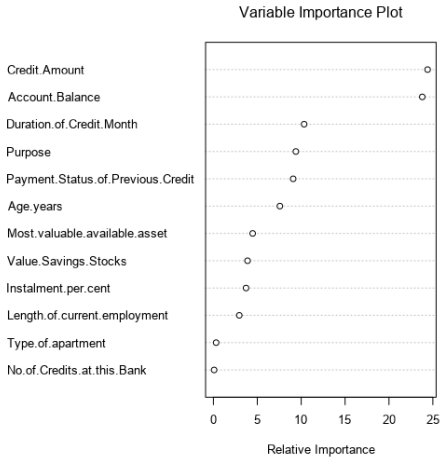
Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided 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.

Confusion matrix of FM_Creditworthiness

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

Boosted Model

- Significant predictor variables: Credit Amount, Account balance, Duration of Credit Month, Purpose.



- The overall accuracy is 0.7867. The model is biased towards creditworthy.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted_Creditworthiness	0.7867	0.8632	0.7524	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.

Confusion matrix of Boosted_Creditworthiness

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

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

ANS: I chose to use Radom Forest Model based on the Model Comparison Report. Among the four models, the random forest model

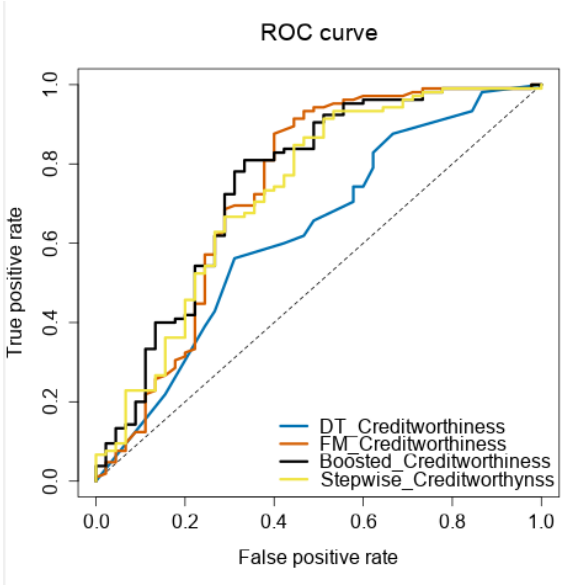
- Has the highest overall accuracy (0.7933)
- Has the highest accuracy with “Creditworthy” (0.9714)
- Non-creditworthy accuracy: Forest Model is the second highest (0.3778)

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Creditworthiness	0.6667	0.7685	0.6272	0.7905	0.3778
FM_Creditworthiness	0.7933	0.8681	0.7368	0.9714	0.3778
Boosted_Creditworthiness	0.7867	0.8632	0.7524	0.9619	0.3778
Stepwise_Creditworthynss	0.7600	0.8364	0.7306	0.8762	0.4889

- **ROC curve**: It looks like the Area Under the Curve (AUC) of Forest Model and Boosted Model are closest to each other; however, the AUC for Decision Tree Model is the lowest, which indicates it is the worst model to measure separability between creditworthy and non-creditworthy. Looking deep, it is Forest Model has better sensitivity than Boosted Model.



- **Bias in the confusion matrix:**

Confusion matrix of Boosted_Creditworthiness

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

Confusion matrix of DT_Creditworthiness

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	28
Predicted_Non-Creditworthy	22	17

Confusion matrix of FM_Creditworthiness

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

Confusion matrix of Stepwise_Creditworthynss

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

Model	TP	FP	FN	TN	PPV	NPV
Boosted	101	28	4	17	0.78	0.81
Decision Tree	83	28	22	17	0.75	0.44
Random Forest	102	28	3	17	0.78	0.85
Stepwise	92	23	13	22	0.80	0.63

PPV = TP / (TP + FP)
NPV = TN / (TN + FN)

Based on the calculation above, Decision Tree model is biased towards PPV (0.75 > 0.44). Creditworthy clients will be more likely to categorize as non-creditworthy. Stepwise model is also biased towards Positive Predict Value (0.8 > 0.63), which leads to more creditworthy persons grouped as non-creditworthy. Decision Tree and Logistic Stepwise Regression models will deny a loan to many creditworthy individuals since those models classify many creditworthy applicants as non-creditworthy.

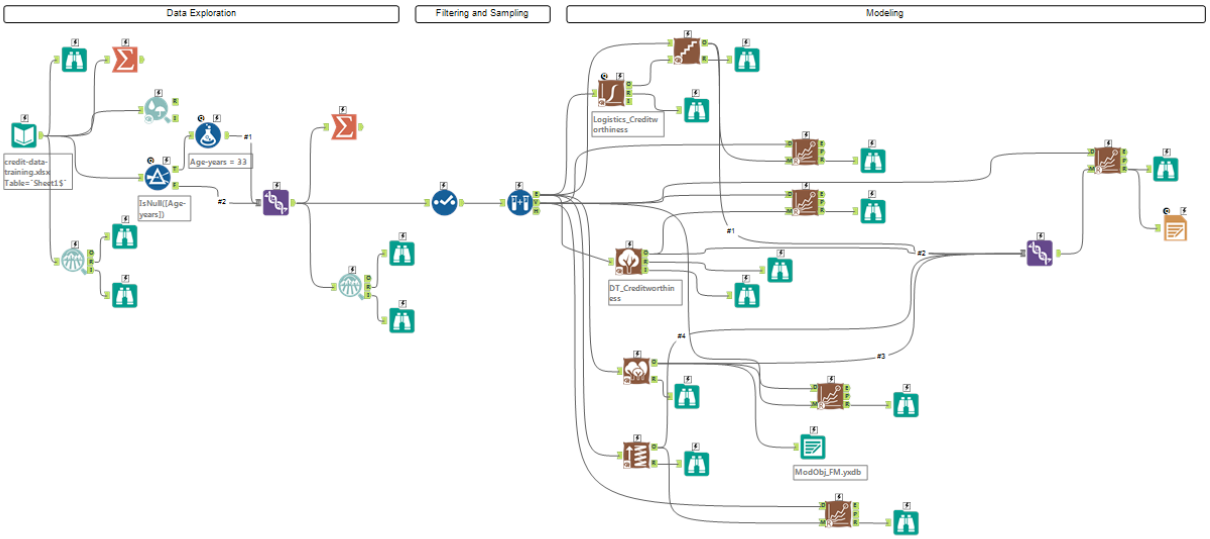
On the other hand, both Random Forest and Boosted models have higher NPV than PPV (0.85 > 0.78), which means both model predict non-creditworthy applicants with higher accuracy. The higher the PPV and NPV values, the more accurate the model prediction is. Thus, Boosted and Random Forest are better. However, Boosted model is no better than Random Forest.

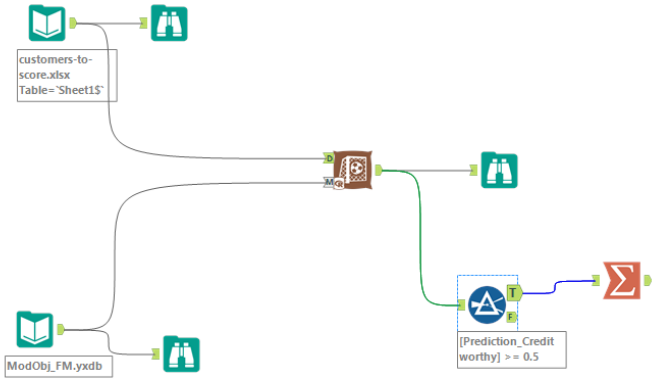
Looking on the F1 Score (the highest F1 score, the more accurate a model is. Thus, Random Forest model has the highest F1 score (0.8681) among all.

To sum up, from the aspects of overall accuracy, PPV, NPV and F1 Score, Random Forest is the best model to predict applicant creditworthiness.

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

- How many individuals are creditworthy?
  ANS: 410 individuals are creditworthy using the forest model to score new customers.


Data Exploration
Filtering and Sampling
Modeling
credit-data-training.xlsx Table='Sheet1$'
IsNull([Age-years])
Age-years = 33
Logistics_Creditworthiness
DT_Creditworthiness
ModObj_FM.yxdb


customers-to-score.xlsx Table='Sheet1$'
ModObj_FM.yxdb
[Prediction_Creditworthy] >= 0.5