Wilson

Project 4 Classification Model – Creditworthiness

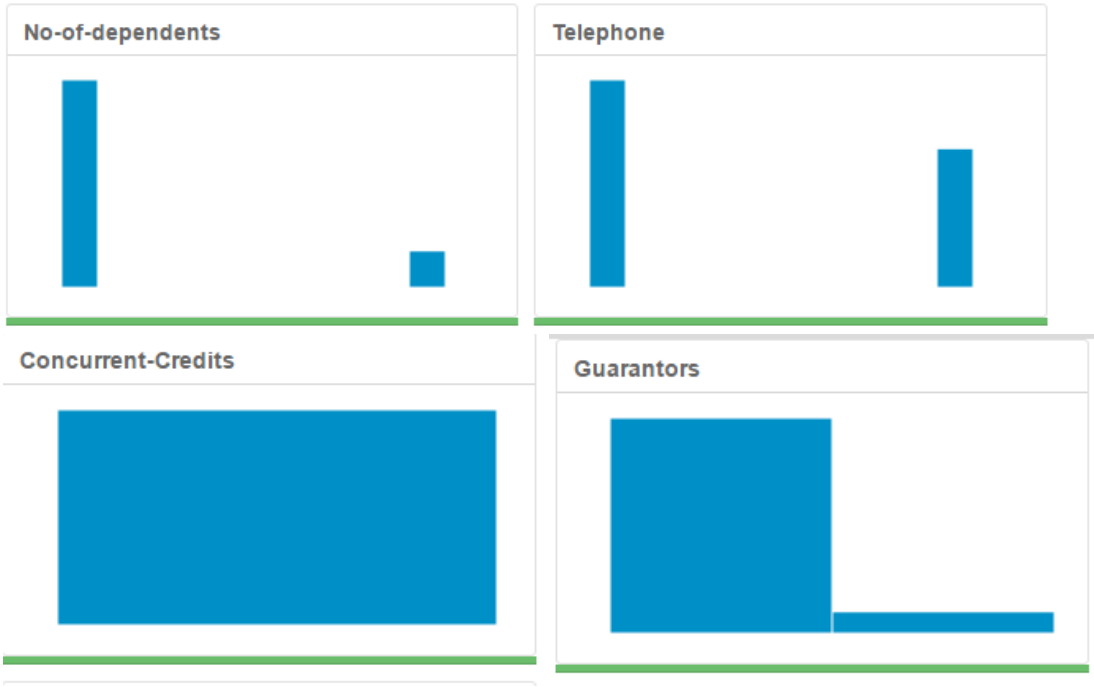
1. What decisions need to be made?
    a. Decision is to determine whether these 500 loan applications are creditworthy or not through data analysis and classification model building. Usually this is done by hand but due to volume of new applicants, it's best to systematically evaluate them.

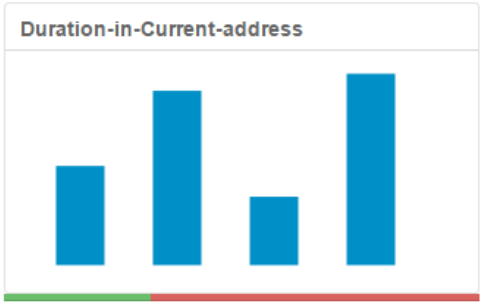
2. What data is needed to inform those decisions?
    a. Data on all past applications
    b. The list of customers that need to be processed in the next few days
    c. Field Summary Data, Modeling data such as Logistic Regression, Stepwise, Model Comparison, Decision Tree, Forested Model, Boosted Model and Score Tool.

3. What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
    a. Since we're trying to determine whether applicants are deemed creditworthy or not, a Binary Classification Model is required.

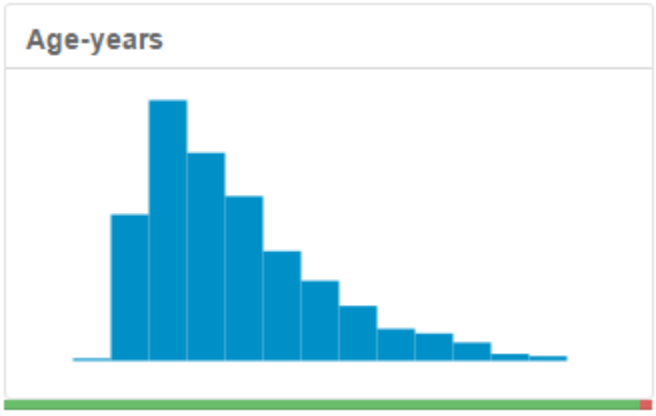
4. In your cleanup process, which field(s) did you impute or remove?
    a. I use Select Tool to remove Guarantors, Duration in Current Address, Concurrent-Credits, Occupation, No-of-dependents, Telephone, and Foreign-Worker. I use Imputation Tool to impute 'Age-Years' field replacing null values with median (33).

5. Please justify why you imputed or removed these fields. Visualizations are encouraged.





a. I removed Guarantors, Concurrent-Credits, Occupation, No-of-dependents, Telephone and Foreign-Worker because of low variability. Concurrent-Credits has no variability while the others have only 2 variability heavily skewing towards one result. Leaving them in would make model building less accurate and increase bias. Duration in Current Address had 69% missing values leaving too little existing data to recommend imputing therefore removed.

b. 

For age-years field there's 2% missing values therefore I used imputation. The graph is right skewed therefore using median (33) to impute would keep the graph more centralized within the distribution compared to average (35.6).

I built 4 models (Logistic, Decision Tree, Forest Model, Boosted Model) to test which one is best used to solve our binary classification problem. For Logistic Regression I use Stepwise to help filter out the best predictor variables for our model.

	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)

The most important predictor variables are Account Balance, Payment Status of Previous Credit, Purpose, Credit Amount and Length of current employment.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
SW_Creditworthy	0.7600	0.8364	0.7306	0.8000	0.6286

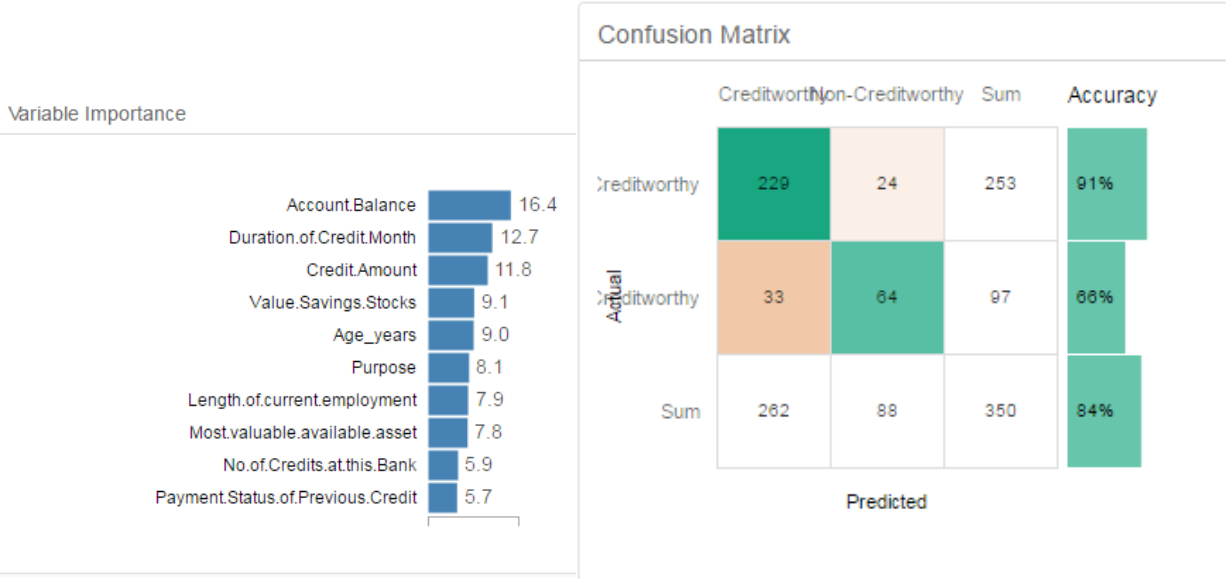
To validate, I use Model Comparison to compare Stepwise with validation set. Overall Accuracy is 76%, Creditworthy accuracy is 80% and Non-Creditworthy Accuracy is 62.8%.

Confusion matrix of SW_Creditworthy

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

This model has some bias as it doesn't seem to predict non-creditworthy as well compared to creditworthy.

Decision Tree



The most important variable for this model is Account Balance, Duration of Credit Month and Credit Amount. Creditworthy Accuracy is 91%, Non-creditworthy accuracy is 66% and Overall is 84%.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Creditworthy	0.6667	0.7685	0.6272	0.7477	0.4359

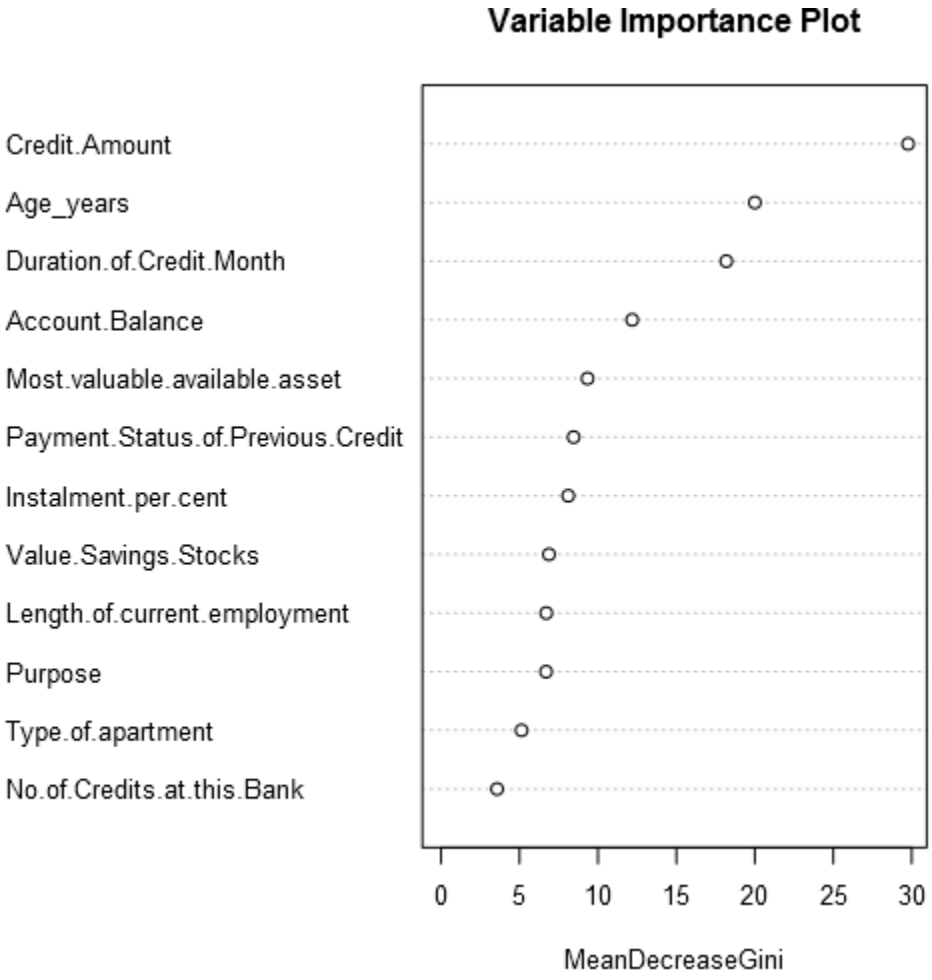
Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of DT_Creditworthy

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

Validating the model against validation samples aren't the same as Estimation sample. 74% Creditworthy Accuracy, 43% Non-creditworthy Accuracy and Overall is 66%. This model is also bias towards creditworthy and doesn't predict non-creditworthy accuracy well.

Forest Model



The most important variables for this model is Credit Amount, Age-years and Duration of Credit Month.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
FM_Creditworthy	0.8000	0.8707	0.7419	0.7953	0.8261

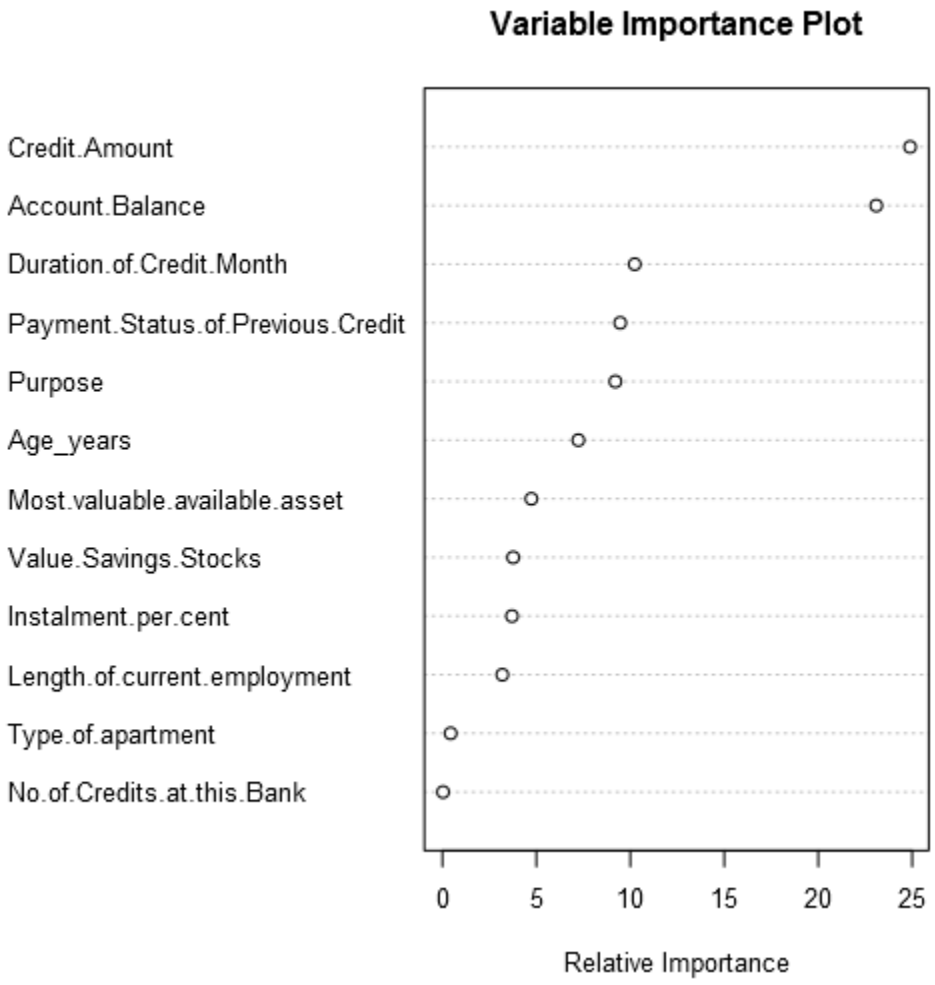
Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of FM_Creditworthy

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

Creditworthy Accuracy is 79%, Non-creditworthy accuracy is 82% and Overall is 80%. This model seems to be the least bias and has the highest overall accuracy.

Boosted Model



Top 3 most important variable is Credit Amount, Account Balance and Duration of Credit Month.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
BM_creditworthy	0.2067	0.0630	0.2491	0.1818	0.2109

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], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of BM_creditworthy

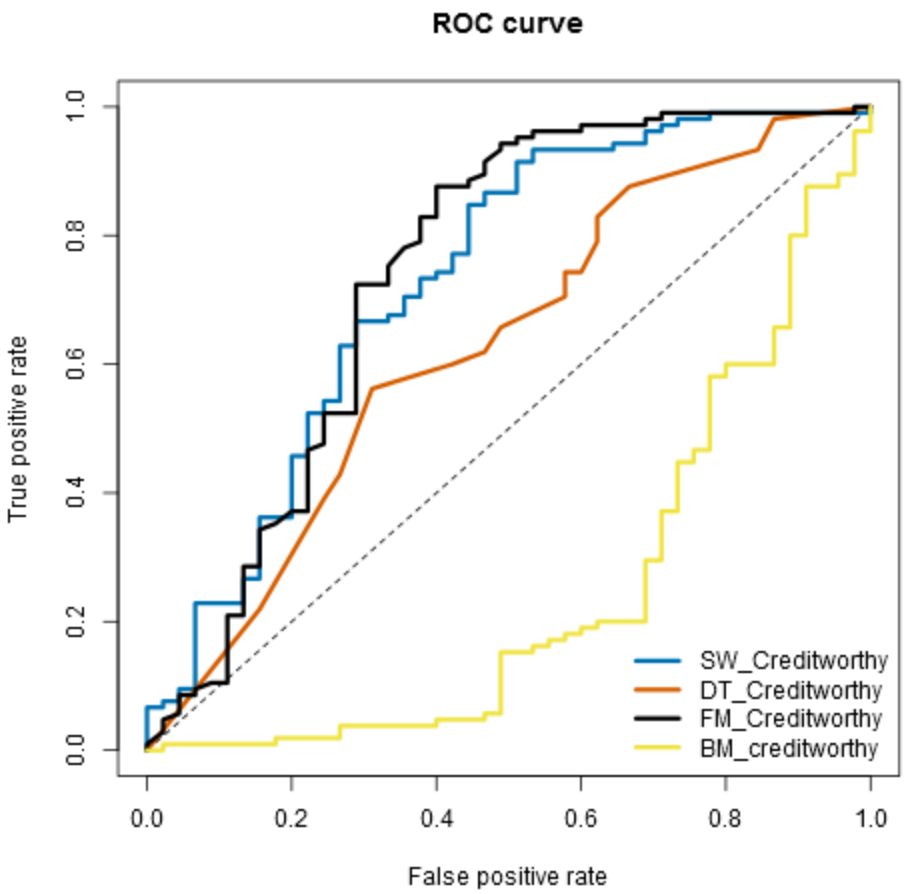
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	4	18
Predicted_Non-Creditworthy	101	27

Boosted model seems to perform the least accurate of all models. 18% creditworthy accuracy, 21% non-creditworthy accuracy and overall accuracy of 20%. This model has a bias for predicting inaccurately.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
SW_Creditworthy	0.7600	0.8364	0.7306	0.8000	0.6286
DT_Creditworthy	0.6667	0.7685	0.6272	0.7477	0.4359
FM_Creditworthy	0.8000	0.8707	0.7419	0.7953	0.8261
BM_creditworthy	0.2067	0.0630	0.2491	0.1818	0.2109

Looking at all the data I chose the Forest Model to calculate our new applications. It has the highest overall accuracy, high creditworthy and non-creditworthy accuracy while least bias towards a particular outcome.



ROC graph shows Forest Model being the most accurate as it's further away from black diagonal line. It solidifies our accuracy as the model isn't randomly guessing the predicted and actual outcomes.

Using the Forest Model to score 500 new applications, 415 applicants are deemed creditworthy.