Creditworthiness

Business and Data Understanding

I work for a small bank and are responsible for determining if customers are creditworthy to give a loan to. We received 500 new loan applications this week and we need to select which customers are creditworthy by performing analysis on previous applications and applying the model to these 500 new applications.

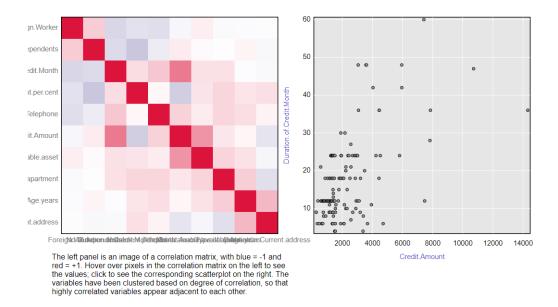
After we navigate the previous dataset, we picked below variables as predicted variables: Account Balance, Duration of Credit Month, Payment status of previous credit, Purpose, Credit-Amount, Value Savings Stocks, Length of current employment, Installment per cent, Most valuable available asset, Age years, Type of apartment and No of credits at this bank.

Since we need to categorize the outcome as creditworthy or noncreditworthy, I rule out the non-binary model. I picked up Binary model and tested the data on 4 different models: Logistic Regression, Decision Tree, Forest Model and Boosted Tree.

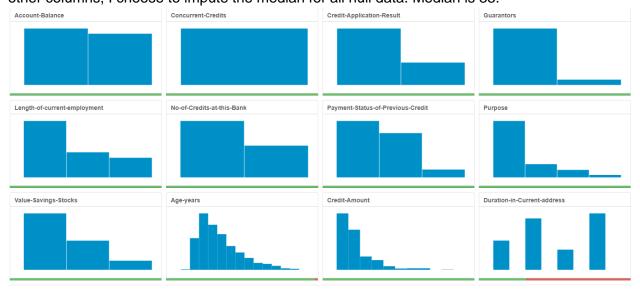
Building the Training Set

Before we build out model, we first need to clean the data set, remove unrelated values and select predicted variables.

• For numerical data fields, there are no field that highly-correlate with each other. All correlation is below 0.7. Please report from association analysis tool below.

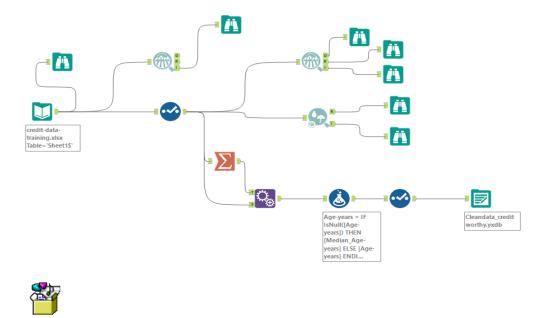


• There are some missing data in the dataset. Please see report from field summary. From below report we could see column Duration_to _current_address have large missing data. Thus, I choose to remove this field when we build our model. Another field Age_years also have 2% missing data. Since it's a small amount but it might impact other columns, I choose to impute the median for all null data. Median is 33.





- From report above, we also have some fields are low variability. We decided to remove these fields: Occupation, Concurrent-credits, Guarantors, Foreign-worker, No-ofdependents and Telephone.
- Alteryx screenshot and file below:



Train your Classification Models

wrangling.yxmd

First, I created my 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.

Logistic Regression:

From this model, Duration of credit month are most important.

The overall percent accuracy is 78%.

Bias calculation:

PPV = true positives/ (true positives+ false positives) = 95 / (95+23) = .8

NPV= true negatives \ (true negatives + false negatives) = $\frac{22}{(22+10)} = .68$

There's bias towards correctly predicting creditworthy.

Report for Logistic Regression Model LR_creditworthy

Basic Summary

igim(formula = Credit.Application.Result ~ Account.Balance + Duration.of.Credit.Month + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Value.Savings.Stocks + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset + Age.years + Type.of.apartment + No.of.Credits.at.this.Bank, family = binomial(logit), data = the.data)
Devlance Residuals:

(Intercept) -3.0136120 1.018+00 -2.9760 0.00202 Account BalanceSome Balance -1.5433699 3.232e-01 -4.7752 1.79e-06 Duration-of. Credit. Month 0.0064973 1.371e-02 0.4738 0.63556 Payment. Status. of. Previous. Creditsome Problems 0.4054009 3.841e-01 1.0554 0.29124 Payment. Status. of. Previous. Creditsome Problems 1.2607175 5.335e-01 2.3632 0.10112 PurposeOther -0.319117 8.342e-01 -3.825 0.70200 PurposeUsed car -0.7839554 4.124e-01 -1,9008 0.05733 Credit. Amount 0.0001764 6.88e-05 2.5798 0.00980 Value. Savings. Stocks Into-£1000 0.6074082 5.100e-01 1.1911 0.238574 Value, Savings. Stocks Elon-£1000 0.1694433 5.649e-01 0.3000 0.7642 Length. of. current. employment-7 yrs 0.522418 4,90e-01 1.0596 0.28834 Length. of. current. employment-7 yrs 0.310893 3.396e-01 1.9664 0.4925 <	b c riance residuals					
Coefficients: Estimate Std. Error z value Pr(≥ z)	Min	1Q	Median		3Q	Max
Std. Error Z value P(Z Z)	-2.088	-0.719	-0.430		0.686	2.542
(Intercept) -3,0136120 1,013e-00 -2,9760 0,00292 Account, BalanceSome Balance -1,5438699 3,232e-01 -4,7752 1,79e-06 Duration, of Credit, Honth 0,0064973 1,371e-02 0,4738 0,5315-02 Payment, Status, of Previous, Credit/Some Problems 0,4054309 3,81e-01 1,0554 0,9124-02 Purpose, Offeredit, Offeredit/Some Problems 1,2607175 5,335e-01 2,352 0,10112-00 Purpose, Offeredit/Some Problems -0,191177 8,342e-01 -2,7951 0,00519-00 Purpose Offeredit/Some Problems -0,3191177 8,342e-01 -2,7951 0,00519-00 Purpose Offeredit/Some Problems -0,3191177 8,342e-01 -3,825 0,00519-00 Purpose Offeredit/Some Problems -0,3191177 8,342e-01 -1,908 0,5273 Credit, Amount -0,0001764 6,83e-05 2,5798 0,00589-0 Value, Savings, Stocks Hone -0,0001764 6,83e-05 2,5798 0,00589-0 Value, Savings, Stocks Flore -0,0001764 6,83e-05 1,0064 0,0089-0	Coefficients:					
Account,BalanceSome Balance -1.5433699 3.22e-01 -4.7752 1.79e-05 Duration of,Credit. Month 0.0064973 1.71e-02 0.4738 0.3555 Payment. Status. of, Previous. CreditSome Problems 1.2607175 5.335-01 2.3632 0.1812 Payment. Status. of, Previous. CreditSome Problems 1.2607175 5.335-01 2.3632 0.1812 PurposeRever -0.739174 6.276e-01 -2.7991 0.0012 PurposeOther -0.839554 4.124e-01 -0.9082 0.0720 PurposeUsed Gar -0.8095954 4.124e-01 -0.908 0.0737 Credit. Amount 0.0001764 6.38e-05 2.798 0.0988 Value. Savings. Stocks Isone 0.0001764 6.38e-05 2.5798 0.0988 Value. Savings. Stocks Isone 0.0001764 6.38e-05 2.5798 0.0988 Value. Savings. Stocks Isone 0.0001764 6.38e-05 2.5798 0.0988 Value. Savings. Stocks Isone 0.0070482 5.10e-01 1.911 0.2302 Value. Savings. Stocks Isone			Estimate	Std. Error	z value	Pr(> z)
Duration of Credit. Month 0.0064973 1.71e-02 0.4738 0.5356 Payment. Status. of. Previous. Credit/some Problems 1.007175 5.336-01 2.3632 0.1812 Purpose New Car -1.7941034 6.276e-01 -2.7951 0.00519 Purpose Observe -0.3191177 8.342e-01 -3.825 0.70206 Purpose Observe -0.3191177 8.342e-01 -1.9082 0.05733 Purpose Observed -0.7899554 4.124e-01 -1.9008 0.05733 Credit. Amount 0.0001764 6.838e-05 2.5798 0.00899 Value. Savings. Stocks None 0.0001764 6.838e-05 2.5798 0.00899 Value. Savings. Stocks 100-1000 0.1694433 5.49e-01 0.3000 0.7642 Length of Current. employment < 1yr	(Intercept)		-3.0136120	1.013e+00	-2.9760	0.00292 **
Payment Status of Previous CreditPaid Up 0.4054009 3.841e-01 1.0554 0.29124 Payment Status of Previous CreditSome Problems 1.2607175 5.335e-01 2.3632 0.01812 Purposelder Gr -1.7541034 6.276e-01 2.75951 0.00512 Purposelder Gr -0.8191177 8.342e-01 -0.3825 0.70206 Purposelder Gr -0.7839554 4.124e-01 -1.9008 0.0572 Credit.Amount Value. Savings. Stockstone 0.6074082 5.100e-01 1.1911 0.2334 Value. Savings. Stockstone 0.1694433 5.649e-01 0.3000 0.7642 Length of.current.employment-1 yr 0.522418 4.930e-01 1.0596 0.2828 Length of.current.employment-2 yr 0.9777992 3.95e-01 1.9664 0.4925 Instalment.per.cent 0.3109833 1.399e-01 1.9664 0.04225 Most. valuable.available.	Account.BalanceSome Balance		-1.5433699	3.232e-01	-4.7752	1.79e-06 ***
Payment, Status. of Previous. CreditSome Problems 1.2607175 5.335e-01 2.3632 0.1812 Purposellew care -1.7951034 6.26e-01 2.7951 0.00319 Purposelber care -0.8789954 4.124e-01 -1.9008 0.0573 Credit, Amount 0.0001764 6.838e-05 2.5798 0.0088 Value, Savings, Stockstone 0.001764 6.838e-05 2.5798 0.0088 Value, Savings, Stockst 100-c1000 0.001764 6.838e-05 2.5798 0.0088 Length, of Current, employment + 7 yrs 0.1644433 5.69e-01 1.0566 0.28894 Length, of Current, employment + 2 yr 0.07524 3.05e-01 1.9664 0.40423 Installment, ercent 0.109833 1.199e-01 2.2232 0.0262 Most, valuabl	Duration.of.Credit.Month		0.0064973	1.371e-02	0.4738	0.63565
Purposelew car -1,7541034 6,26e-01 -2,7951 0,0519 PurposeUsder -0,3191177 8,342e-01 -0,3825 0,70206 PurposeUsda car -0,7893954 4,124e-01 1,908 0,05734 Alube-Savings-Stockstore 0,0001764 6,838e-05 2,5798 0,0088 Value-Savings-Stockstore 0,05074082 5,100e-01 1,1911 0,2334 Value-Savings-Stockstore 0,1694433 5,649e-01 0,3000 0,7642 Length-of-Current-employment-7 yrs 0,5224158 4,930e-01 1,0596 0,2883 Length-of-Current-employment-2 tyr 0,7777492 3,95e-01 1,9664 0,4925 Most.valuable-available-	Payment.Status.of.Previous.CreditPaid Up		0.4054309	3.841e-01	1.0554	0.29124
PurposeOther -0,3191177 8,342-01 -0,3825 0,70206 PurposeOther -0,7899954 4,124-01 1,908 0,5733 Cradit.Amount 0,0001764 6,88e-05 2,5798 0,00989 Value.Savings.Stockstone 0,6074082 5,100e-01 1,191 0,2324 Value.Savings.Stockstine 0,1644433 5,64e-01 0,300 0,7642 Length.of.current.employment-Y yrs 0,5224158 4,90e-01 1,0564 0,28824 Instalment.per.cent 0,2109833 1,396-01 2,2232 0,0562 Most. valuable -asset 0,2238706 1,556-01 2,2342 0,0562 Age, years -0,0141206 1,356-02 -9,9202 0,3747 Type. of apartment -0,2603038 2,956-01 -0,8805 0,3784	Payment.Status.of.Previous.CreditSome Problems		1.2607175	5.335e-01	2.3632	0.01812 **
PurposeUsed car -0,7839554 4,124e-01 -1,9008 0,5733 Credit_Amount 0,0001764 6,838e-05 2,5798 0,00989 Value_Savings_Stockst100e 0,000174822 5,100e-01 1,1911 0,2330 Value_Savings_Stockst100e 0,1694433 5,649e-01 0,3000 0,7642 Length.of.current_employment-Y yrs 0,5224158 4,930e-01 1,0596 0,2893 Length.of.current_employment-C yr 0,7779492 3,956e-01 1,9664 0,04925 Instalment_per.ott 0,1010933 1,399e-01 2,2322 0,0262 Most.valuable_available.asset 0,2328706 1,556e-01 2,0945 0,03621 Age,years -0,10141206 1,535e-02 -0,9202 0,35747 Yippe_of_apartment -0,2603038 2,956e-01 -0,8805 0,3786	PurposeNew car		-1.7541034	6.276e-01	-2.7951	0.00519 ***
Creditmount 0.0001764 6.88e-05 2.5798 0.00898 ValueSavings.Stockstone 0.00704082 5.100e-01 1.1911 0.2336 ValueSavings.Stockstione 0.1504433 5.649e-01 0.3000 0.7642 Lengthfocurrent.employment4-7 yrs 0.5224158 4.930e-01 1.0596 0.28934 Instalment.per.cent 0.2109833 1.399e-01 2.2232 0.0562 Most. valuableasset 0.2238706 1.556e-01 2.0453 0.0362 Age.years -0.0141206 1.356e-02 -0.9202 0.3747 Type. of.apartment -0.2603038 2.956e-01 -0.8055 0.3786	PurposeOther		-0.3191177	8.342e-01	-0.3825	0.70206
Value.Savings.Stocksflooe 0.6074082 5.100e-01 1.1911 0.23361 Value.Savings.Stocksflooe 0.1694433 5.649e-01 0.3000 0.7642 Length.of.current.employment-/ yrs 0.5224158 4.90e-01 1.9664 0.28293 Length.of.current.employment- 0.7779492 3.956e-01 1.9664 0.04925 Instalment.per.ont 0.3100893 1.399e-01 2.222 0.0062 Most.valuable.asset 0.328706 1.556e-01 2.0945 0.03621 Age.years -0.041206 1.535e-02 -0.9202 0.37747 Type.of.apartment -0.2603038 2.956e-01 -0.805 0.3786	PurposeUsed car		-0.7839554	4.124e-01	-1.9008	0.05733 .
Value Savings_Stocks(100-E1000 0.1694433 5.64e-01 0.3000 0.7642 Length of.current.employment4-7 yrs 0.5224158 4.90e-01 1.0566 0.2893 Length.of.current.employment4-1 yr 0.777992 3.956e-01 1.9664 0.04925 Instalment.per.cent 0.109833 1.99e-01 2.2232 0.0262 Most. valuable.asset 0.222870c 1.556e-01 2.0455 0.0362 Age, years -0.161206 1.535e-02 -0.9202 0.37747 Type. of.apartment -0.2603038 2.956e-01 -0.8805 0.3786	Credit.Amount		0.0001764	6.838e-05	2.5798	0.00989 **
Length.of.current.employment4-7 yrs 0.5224158 4,930e-01 1.0596 0.28934 Length.of.current.employment 0.7779492 3,950e-01 1.9664 0.4925 Instalment.pecnet 0.310993 1.99e-01 2.232 0.0262 Most. valuable.available.asset 0.3258706 1.556e-01 2.945 0.03621 Age-years -0.0141206 1.535e-02 -0.9202 0.37547 Type.of.apartment -0.603038 2.956e-01 -0.8805 0.3786	Value.Savings.StocksNone		0.6074082	5.100e-01	1.1911	0.23361
Length of current employment < 1yr	Value.Savings.Stocks£100-£1000		0.1694433	5.649e-01	0.3000	0.7642
Instalment.per.cent 0.3109833 1.399e-01 2.2232 0.0262 Most.valuable.asset 0.3298706 1.556e-01 2.0945 0.03621 Age.years -0.0141206 1.535e-02 -0.922 0.3574 Type.of.apartment -0.263038 2.956e-01 -0.805 0.3786	Length.of.current.employment4-7 yrs		0.5224158	4.930e-01	1.0596	0.28934
Most valuable asset 0.3258706 1.556-01 2.0945 0.03621 Age years -0.0141206 1.535-02 -0.9202 0.35747 Type, or Lapartment -0.2603038 2.956-01 -0.8805 0.3748	Length.of.current.employment< 1yr		0.7779492	3.956e-01	1.9664	0.04925 **
Age, years -0.0141206 1.535e-02 -0.9202 0.35747 Type, of, apartment -0.2603038 2.956e-01 -0.8805 0.3786	Instalment.per.cent		0.3109833	1.399e-01	2.2232	0.0262 **
Type.of.apartment -0.2603038 2.956e-01 -0.8805 0.3786	Most.valuable.available.asset		0.3258706	1.556e-01	2.0945	0.03621 *
	Age.years		-0.0141206	1.535e-02	-0.9202	0.35747
No.of. Credits, at. this. Bank More than 1 0.3619545 3.815e-01 0.9487 0.34275	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

(Dispersion parameter for binomial taken to be ${\bf 1}$)

Null deviance: 413.16 on 349 degrees of freedom Residual deviance: 322.31 on 332 degrees of freedom McFadden R-Squared: 0.2199, Akaike Information Criterion 358.3

Number of Fisher Scoring Iterations: 5

Model Comparison Report							
Fit and error measures							
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy		
LR_creditworthy	0.7800	0.8520	0.7314	0.9048	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_closs 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 recoil. 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 LR_c	editworthy						
				Actual_Creditworthy	Actual_Non-Creditworthy		
	Predicted_Cre	editworthy		95	23		
	Predicted Non-Cre	editworthy		10	22		

Decision Tree:

From this model, we can see account balance and age years are most important.

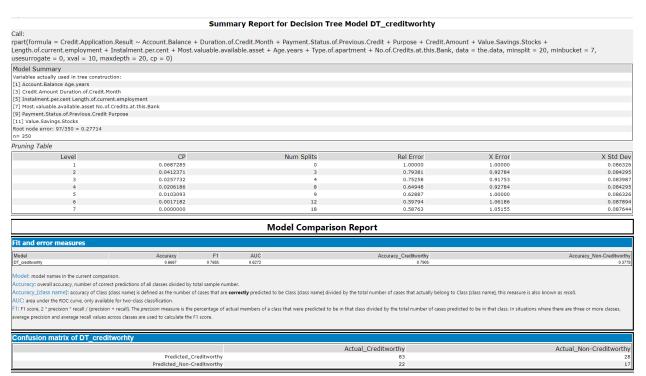
The overall percent accuracy is 67%.

Bias calculation:

PPV = true positives/ (true positives+ false positives) = 83 / (83+28) = .75

NPV= true negatives \ (true negatives + false negatives) = 17/(17+23) = .43

In this model, there's also bias towards correctly predicting creditworthy.



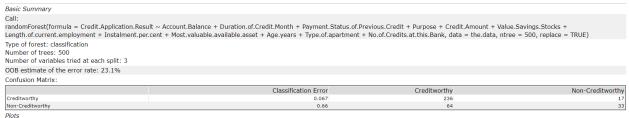
Forest Model:

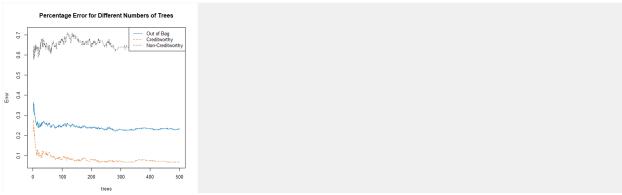
From this model, we can see credit amount is most important.

The overall percent accuracy is 79%.

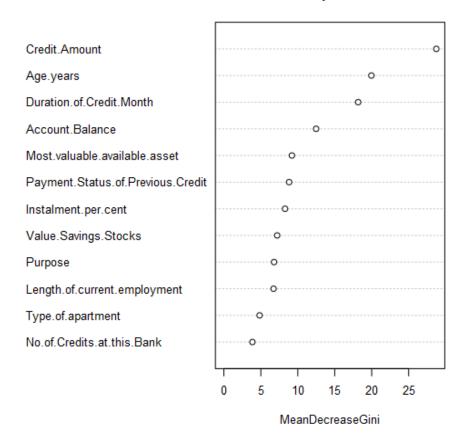
Bias calculation:

PPV = true positives/ (true positives+ false positives) = 102 / (102+28) = .78 NPV= true negatives \ (true negatives + false negatives) = 17/ (17+3) = .85 There's almost no in this model.





Variable Importance Plot



Model Comparison Report Fit and error measures							
RF_creditworthy	0.7933	0.8681	0.7368	0.9714	0.377		
UC: area under the ROC curve, only ava 1: F1 score, 2 " precision " recall / (preci- verage precision and average recall valu	illable for two-class classification. sion + recall). The <i>precision</i> measure is the perc es across classes are used to calculate the F1 sc	entage of actual i		lass (class name) divided by the total number of cases that actually belong to Class (class at were predicted to be in that class divided by the total number of cases predicted to be			
Confusion matrix of RF_cre	editworthy						
				Actual_Creditworthy	Actual_Non-Creditworth		
	Predicted_Credi			102	2		
	Predicted Non-Credi	itworthy		2			

Boosted model:

From this model, we can see credit amount is most important.

The overall percent accuracy is 79%.

Bias calculation:

PPV = true positives/ (true positives+ false positives) = 101 / (101+28) = .78

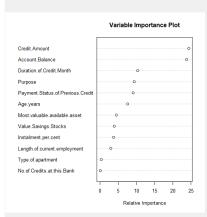
NPV= true negatives \ (true negatives + false negatives) = 17/ (17+4) = .81

There's almost no bias in this model.

Basic Summary:

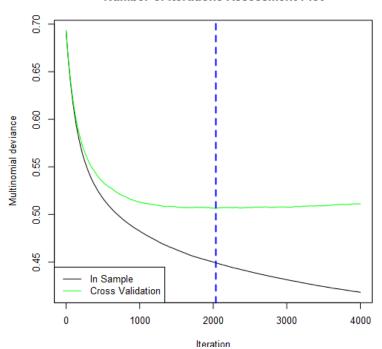
Loss function distribution: Bernoulli

Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 2036



The Variable Importance Plot provides information about the relative importance of each predictor field. The measures are normalized to sum to 100, and the value for each field gives the relative percentage importance of that field to the overall model.

Number of Iterations Assessment Plot



Model Comparison Report

Tit dild ciror incubarcs					
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted_creditworthy	0.7867	0.8632	0.7524	0.9619	0.3778

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.

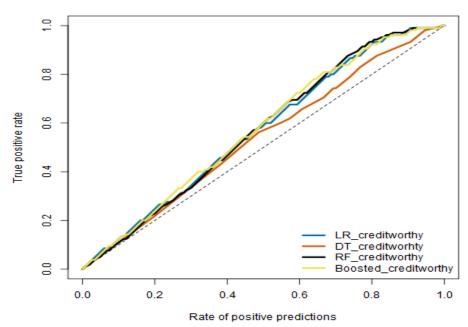
Confusion matrix of Boosted_creditworthy		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

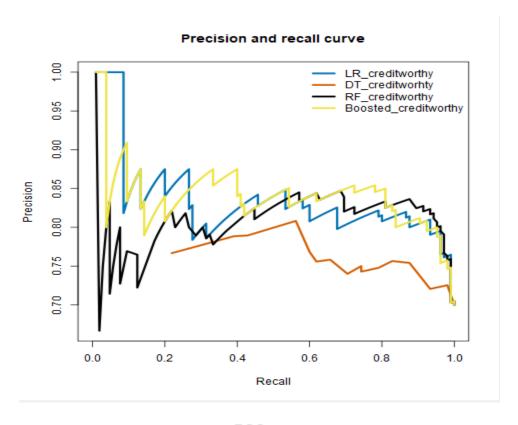
We can see there are some bias in the model's predictions since the accuracy is different between Estimate report and validation report for each model.

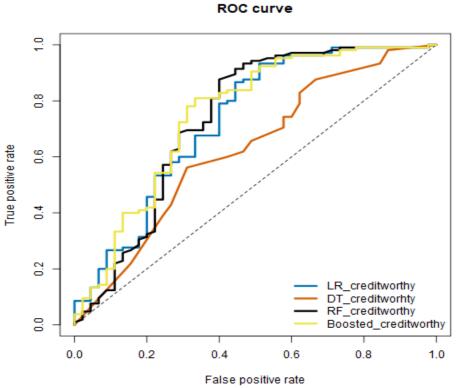
Writeup

Model Comparison Report Fit and error measures							
LR_creditvorthy	0.7800	0.8520	0.7314	0.9048	0.4885		
DT_creditworhty	0.6667	0.7685	0.6272	0.7905	0.377		
RF_creditworthy Boosted creditworthy	0.7933 0.7867	0.8681 0.8632	0.7368 0.7524	0.9714 0.9619	0.377 0.377		
Model: model names in the current comparison.							
Accuracy: overall accuracy, number of correct predictions of al	alaccae divided by total camp	la numbar					
Accuracy_[class name]: accuracy of Class [class name] is define			andisted to be Clear falous and divided by the t				
		at are correctly	predicted to be class [class name] divided by the t	otal number of cases that actually belong to class (class ha	amej, uns measure is also known as recall.		
AUC: area under the ROC curve, only available for two-class cla							
F1: F1 score, 2 * precision * recall / (precision + recall). The prec		e of actual mem	bers of a class that were predicted to be in that cla	ass 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 us	sed to calculate the F1 score.						
Confusion matrix of Boosted_creditworthy	/						
			Actua	al_Creditworthy	Actual_Non-Creditworthy		
	Predicted_Creditwor	thy		101	28		
	Predicted_Non-Creditwor	thy		4	11		
Confusion matrix of DT creditworhty							
			Actua	al_Creditworthy	Actual_Non-Creditworthy		
	Predicted_Creditwor	Phys	Actu	83	Actual_Non creatmorth		
	Predicted Non-Creditwor			22	1		
	Fredicted_Non CreditWor	any			17		
Confusion matrix of LR_creditworthy							
			Actua	al_Creditworthy	Actual_Non-Creditworthy		
	Predicted_Creditwor	thy		95	23		
	Predicted_Non-Creditwor	thy		10	22		
Confusion matrix of RF_creditworthy							
			Actua	al_Creditworthy	Actual Non-Creditworthy		
	Predicted_Creditwor	de	Accus				
				102	28		
	Predicted_Non-Creditwor	tny		3	1		

Gain chart







After testing on 4 different models, I chose Random Forest model for below 4 reason:

- 1. It provided highest overall validation accuracy 79% which is highest among 4 different models.
- This model has almost no bias. From below calculation, we see that 0.78 and 0.85 are quite close. Thus, forest model also works well on confusion metrix.
 Bias calculation:
 - PPV = true positives/ (true positives+ false positives) = 102 / (102+28) = .78NPV= true negatives \ (true negatives + false negatives) = 17/ (17+3) = .85
- 3. Forest model also has highest accuracies with creditworthy and non-creditworthy prediction, as we can see this model correctly predicted 102 for actual creditworthy and 17 for non-actual creditworthy.
- 4. From the ROC graph, Forest model has the highest curve in these models and boosted model rises the fastest. After compare 2 models in chart, we still choose Forest model as it reached top left corner which means that for a given amount of false positive predictions, this model will give the best number of true positive predictions.

For final calculation, after we applied the Forest model to 500 new applications, we had 407 individuals are creditworthy.