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?

We need to decide which loans applications are to be approved using a model to be efficient given the influx in new loan applications.

- What data is needed to inform those decisions?

We need historical loan application data to build the model and new loan applications data to apply the model to.

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

We would need a Binary model since the output is either we approve the loan or not.

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:*

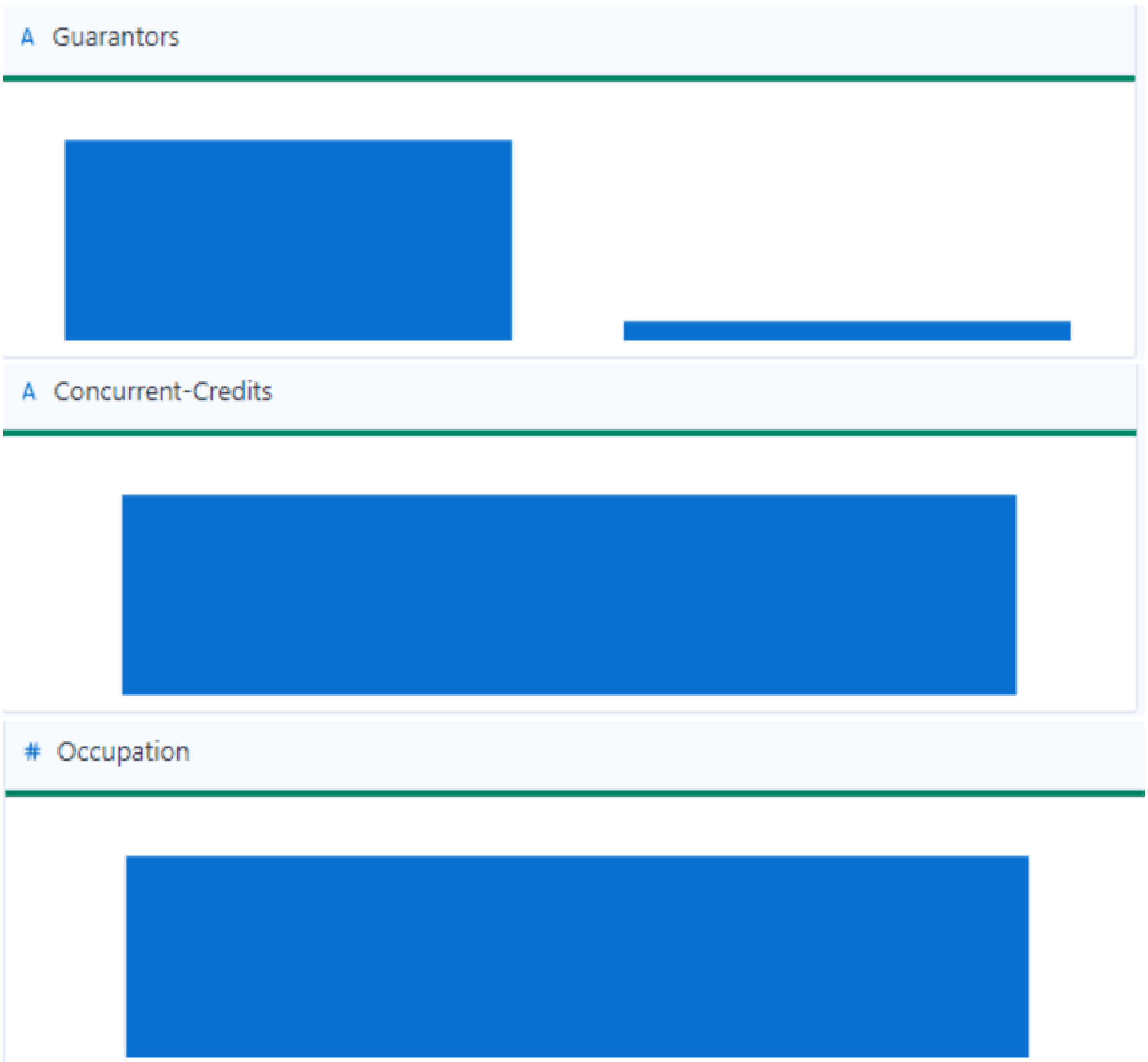
Variable	**Data Type**
Credit-Application-Result	String
Account-Balance	String
Duration-of-Credit-Month	Double
Payment-Status-of-Previous-Credit	String
Purpose	String
Credit-Amount	Double
Value-Savings-Stocks	String
Length-of-current-employment	String
Instalment-per-cent	Double
Guarantors	String
Duration-in-Current-address	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.

We remove the following fields: **Concurrent-Credits[1], Guarantors[2], Occupation[3], Foreign Worker[4]**, **number of dependents[5]** since there is little to no variation for the answers within these fields, i.e. “low variability.”



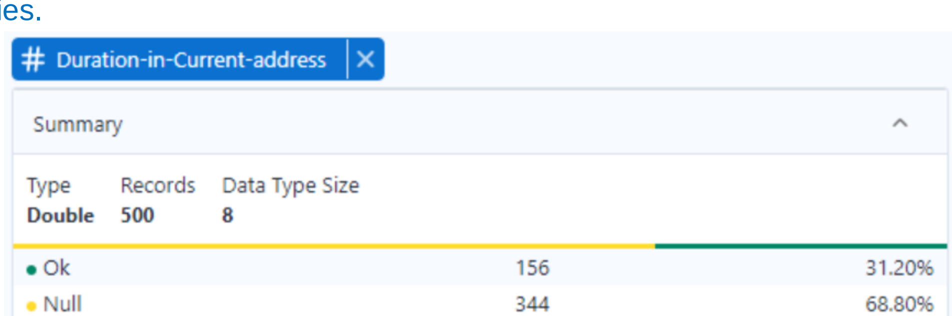
We will also remove other fields that likely do not have a relationship the loan application such as the existence of a **telephone[6].** Association analysis confirms there is little correlation to the result and not even significant.

Pearson Correlation Analysis

Focused Analysis on Field Credit.Application.Result.num

	Association Measure
Most.valuable.available.asset	-0.232248
Duration.of.Credit.Month	-0.215149
Instalment.per.cent	-0.130496
Age.years	0.123088
Credit.Amount	-0.092205
Duration.in.Current.address	0.067284
Type.of.apartment	-0.039360
No.of.dependents	0.038037
Telephone	0.030838

Finally, we remove **Duration in Current Address[7]** as there are too many null values at 68.8% of the entries.



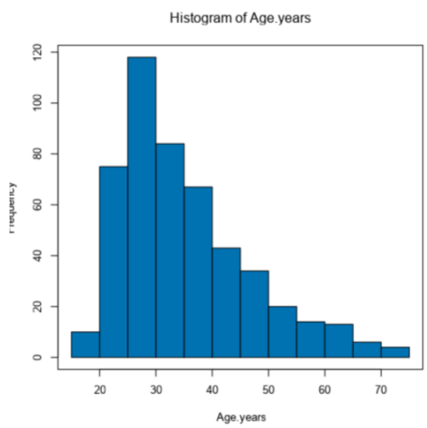
There is no multicollinearity, i.e. none of the variables are highly correlated to each other so we do not have to worry about possible errors later in fit and prediction.

Full Correlation Matrix

	Credit.Application.Result.num	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Duration.in.Current.address	Most.valuable.available.asse
Credit.Application.Result.num	1.000000	-0.215149	-0.092205	-0.130496	0.067284	-0.23224
Duration.of.Credit.Month	-0.215149	1.000000	0.565054	0.145637	-0.032494	0.12881
Credit.Amount	-0.092205	0.565054	1.000000	-0.253286	-0.136621	0.45714
Instalment.per.cent	-0.130496	0.145637	-0.253286	1.000000	0.131231	0.11511
Duration.in.Current.address	0.067284	-0.032494	-0.136621	0.131231	1.000000	-0.04738
Most.valuable.available.asset	-0.232248	0.128814	0.457147	0.115114	-0.047386	1.00000
Age.years	0.123088	-0.018171	0.040486	0.111456	0.301966	0.12357
Type.of.apartment	-0.039360	0.126967	0.100413	0.178926	-0.163386	0.18274
No.of.dependents	0.038037	-0.185180	0.082721	-0.293380	-0.036814	0.01943
Telephone	0.030838	0.238437	0.192532	0.038515	0.055112	0.08339

	Age.years	Type.of.apartment	No.of.dependents	Telephone		
Credit.Application.Result.num	0.123088	-0.039360	0.038037	0.030838		
Duration.of.Credit.Month	-0.018171	0.126967	-0.185180	0.238437		
Credit.Amount	0.040486	0.100413	0.082721	0.192532		
Instalment.per.cent	0.111456	0.178926	-0.293380	0.038515		
Duration.in.Current.address	0.301966	-0.163386	-0.036814	0.055112		
Most.valuable.available.asset	0.123579	0.182744	0.019435	0.083395		
Age.years	1.000000	0.208552	0.046996	0.141103		
Type.of.apartment	0.208552	1.000000	-0.010189	0.179688		
No.of.dependents	0.046996	-0.010189	1.000000	-0.097632		
Telephone	0.141103	0.179688	-0.097632	1.000000		

There are 12 rows with missing Age values. The histogram indicates a positive skew with median at 33 and mean at 36. Using the median vs. the average is more representative of the entire data set given the skewness.



As such we impute the median age on the null rows and check the correlations again. The numbers below indicate the relationships were preserved.

Original:

Pearson Correlation Analysis

Focused Analysis on Field Credit.Application.Result.num

	Association Measure	p-value
Duration.of.Credit.Month	-0.204317	5.3642e-06 ***
Credit.Amount	-0.200990	7.6638e-06 ***
Most.valuable.available.asset	-0.137917	2.2621e-03 **
Instalment.per.cent	-0.065345	1.4948e-01
Age.years	0.056737	2.1088e-01
Type.of.apartment	-0.021860	6.3000e-01

When the 12 rows with null age values are imputed with the median age:

Focused Analysis on Field Credit.Application.Result.num

	Association Measure	p-value
Duration.of.Credit.Month	-0.202504	5.0151e-06 ***
Credit.Amount	-0.201946	5.3311e-06 ***
Most.valuable.available.asset	-0.141332	1.5334e-03 **
Instalment.per.cent	-0.062107	1.6556e-01
Age.years	0.052914	2.3758e-01
Type.of.apartment	-0.026516	5.5417e-01

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)

Model: Logistic Regression

Significant predictors are: Account Balance, Payment Status of Previous Credit, Purpose, Credit Amount, Length of Current Employment, InstallmentPercent, Most Valuable Available Asset

	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	

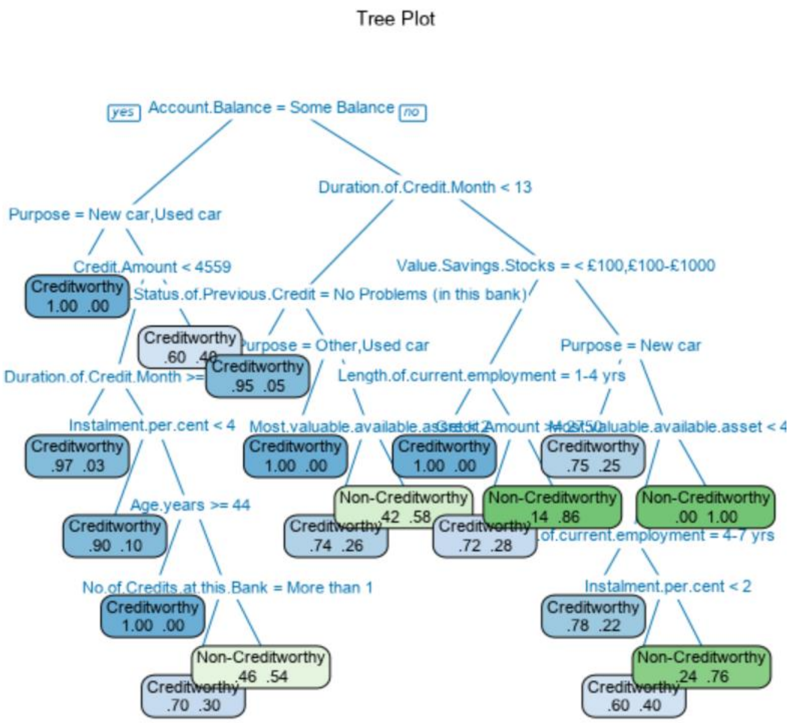
From the initial results, we then add stepwise to logistic regression to just get the significant predictors,

	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	.

Model: Decision Tree

Significant predictors are: Account Balance, Duration of CreditMonth, Installment Percent, Length of Current Employment, Most Valuable Available Asset , No. of Credits at Bank, Payment Status of Previous Credit, and Value Savings Stocks.

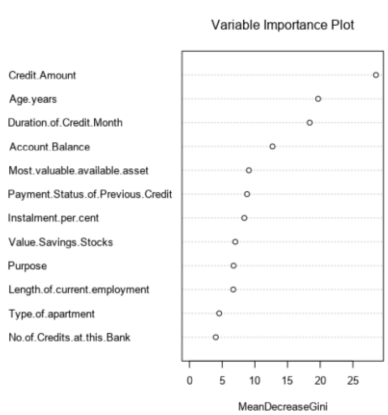
```
Model Summary
Variables actually used in tree construction:
[1] Account.Balance Age.years
[3] Credit.Amount Duration.of.Credit.Month
[5] Instalment.per.cent Length.of.current.employment
[7] Most.valuable.available.asset No.of.Credits.at.this.Bank
[9] Payment.Status.of.Previous.Credit Purpose
[11] Value.Savings.Stocks
Root node error: 97/350 = 0.27714
n= 350
```



We reduce the tree size to 200 for Random Forest since this is where the error rate became flat.

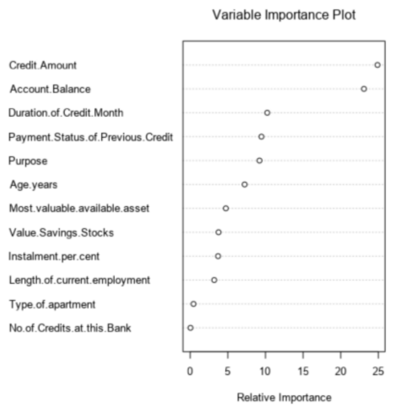
Model: Random Forest

Most important variables are: Credit Amount, Age, Duration of CreditMonth, Account Balance, Most Valuable Available Asset, Payment Status of Previous Credit



Model: Boosted Model

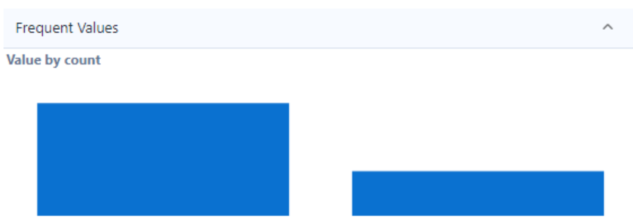
Most important variables are: Credit Amount, Account Balance, Duration of CreditMonth, Payment Status of Previous Credit, Purpose



Validation the models against the validation set, we note that the accuracy for the non-creditworthy is much lower overall likely because majority if our data is composed of "creditworthy" applicants.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT	0.6733	0.7721	0.6296	0.7905	0.4000
RF	0.8067	0.8755	0.7459	0.9714	0.4222
BM	0.7933	0.8670	0.7509	0.9619	0.4000
LogReg	0.7600	0.8364	0.7306	0.8762	0.4889



Top Values

Creditworthy	358
Non-Creditworthy	142

The overall percent accuracy is highest for the Random Forest model at 0.81 with a bias for creditworthy status. The confusion matrix for each model is shown below:

Recall: $Precision = \frac{TP}{TP+FP}$, $Recall = \frac{TP}{TP+FN}$

Model: Logistic Regression

For logistic regression, the confusion matrix indicates precision (positive predictive value) of 0.80, a recall (sensitivity) of 0.88, and an overall accuracy of 0.76.

Confusion matrix of LogReg

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

$$Precision = \frac{92}{92 + 23} = 0.80$$

$$Recall = \frac{92}{92 + 13} = 0.88$$

Model: Decision Tree

For Decision Tree, the confusion matrix indicates precision (positive predictive value) of 0.75, a recall (sensitivity) of 0.79, and an overall accuracy of 0.67. These are lower than for Logistic Regression and is thus not our chosen final model.

Confusion matrix of DT

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	27
Predicted_Non-Creditworthy	22	18

$$Precision = \frac{83}{83 + 27} = 0.75$$

$$Recall = \frac{83}{83 + 22} = 0.79$$

Model: RandomForest

For Random Forest Model, the confusion matrix indicates precision (positive predictive value) of 0.80, a recall (sensitivity) of 0.97, and an overall accuracy of 0.81. These are higher than for Logistic Regression and a candidate as final model.

Confusion matrix of RF

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	26
Predicted_Non-Creditworthy	3	19

$$Precision = \frac{102}{102 + 26} = 0.80$$

$$Recall = \frac{102}{102 + 3} = 0.97$$

Model: Boosted Model

For Boosted Model, the confusion matrix indicates precision (positive predictive value) of 0.79, a recall (sensitivity) of 0.96 (higher than RandomForest), and an overall accuracy of 0.79. Overall accuracy is lower than the RandomForest model due to the higher False Positives, i.e. predicted as creditworthy when it is non-creditworthy.

Confusion matrix of BM

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

$$Precision = \frac{101}{101 + 27} = 0.79$$

$$Recall = \frac{101}{101 + 4} = 0.96$$

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.

- How many individuals are creditworthy?

As final model, we choose the **Random Forest model which had the highest overall accuracy**. This was closely followed by the Boosted Model, but the Boosted Model had higher False Positives, i.e. predicted as creditworthy when it is non-creditworthy, which is dangerous for loan application modelling purposes.

The Random Forest model has higher accuracy in creditworthy segment vs. Logistic Regression but has lower accuracy in the non-creditworthy segment.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT	0.6733	0.7721	0.6296	0.7905	0.4000
RF	0.8067	0.8755	0.7459	0.9714	0.4222
BM	0.7933	0.8670	0.7509	0.9619	0.4000
LogReg	0.7600	0.8364	0.7306	0.8762	0.4889

A deeper dive into these 2 models confusion matrices shows the following tendencies:

Predicting as creditworthy when it is non-creditworthy is especially important for a bank because it would not want any defaults. Using the confusion matrix, the false positive rate for logistic regression is at 51% vs. 58% for RandomForest. This is a significant difference with preference for Logistic Regression.

Confusion matrix of LogReg

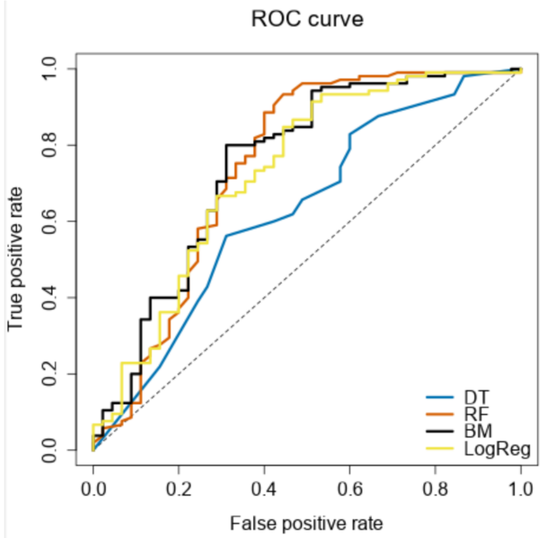
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

Confusion matrix of RF

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	26
Predicted_Non-Creditworthy	3	19

Desiring profitability, the bank would also not want to reject too many applications when these should have been approved, i.e. predicted as non-creditworthy when they are creditworthy. The false negative rate/miss rates for logistic regression is 12% vs 3% for RandomForest. Again, this is a significant difference with preference for RandomForest.

The ROC curve also indicates a preference for the RandomForest model as it is farthest from the random guess dotted-line and closest to the perfect classification of (x,y) as (0,1).



Scoring the new data set, 409 individuals are creditworthy.

	Decision	Count
1	Creditworthy	409
2	Not	91

Before you Submit

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