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

Analyze the problem using the Problem Solving Framework and provide a list of creditworthy customers to manager.

Provide an explanation of the key decisions that need to be made. (250 word limit)

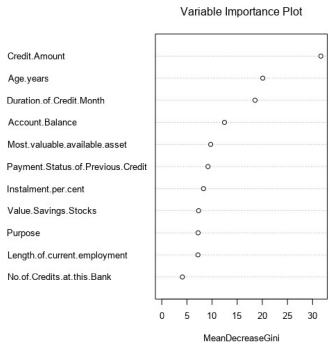
Key Decisions:

Answer these questions

- What decisions needs to be made?
  Provide list of creditworthy customer out of 500 application

- What data is needed to inform those decisions?

  Looking at the significance of data, we need credit_amount, age.year, Duration of credit month, Account Balance, Purpose, Length.of.current.employment< 1yr



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 .

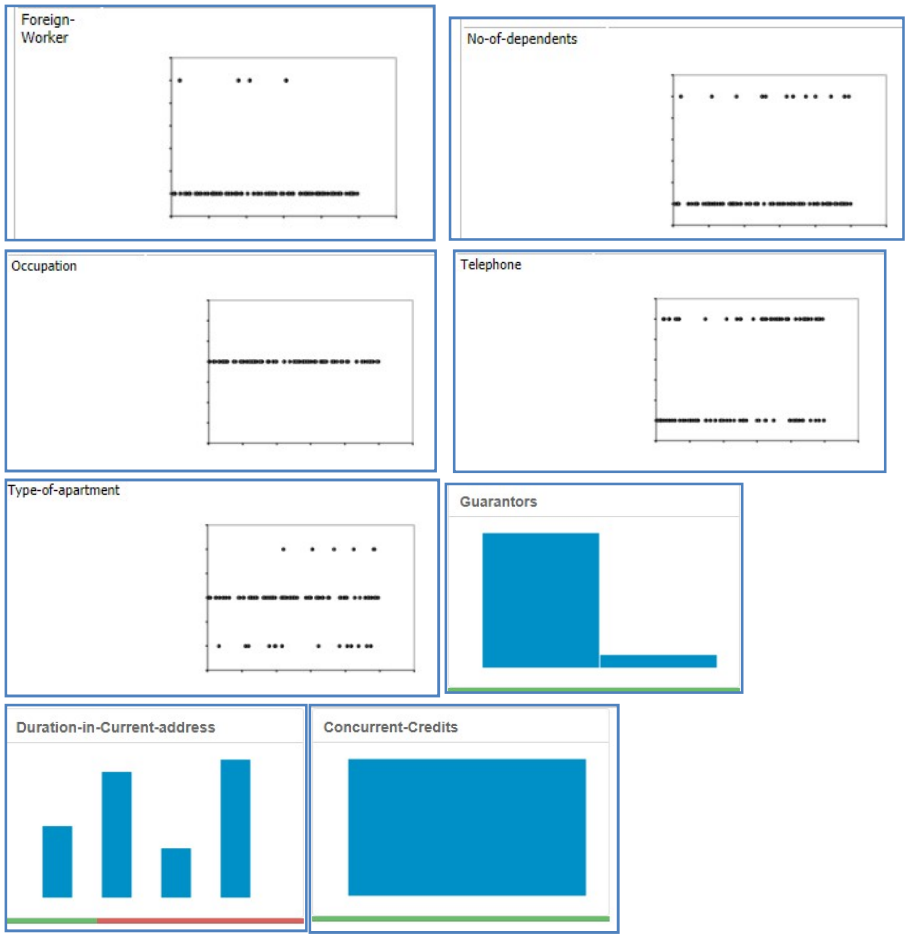
- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

  As result is in binary i,e yes or no, so we need to use Binary classification model to make decision of creditworthy customers

Step 2: Building the Training Set

Answer this question:

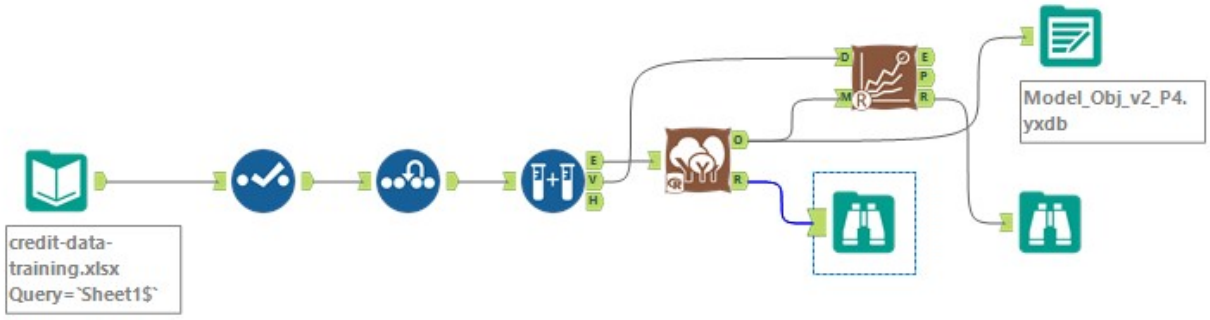
- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.
  Following fields has been removed
  Duration In Current Address field has been removed due to lot's of missing data in this field,
  Concurrent-Credits and **Occupation** field has been removed as it is completely uniform.
  Other fields **foreign workers, no-of-dependents, guarantors** and **Type-of-apartment** are removed due to low variability



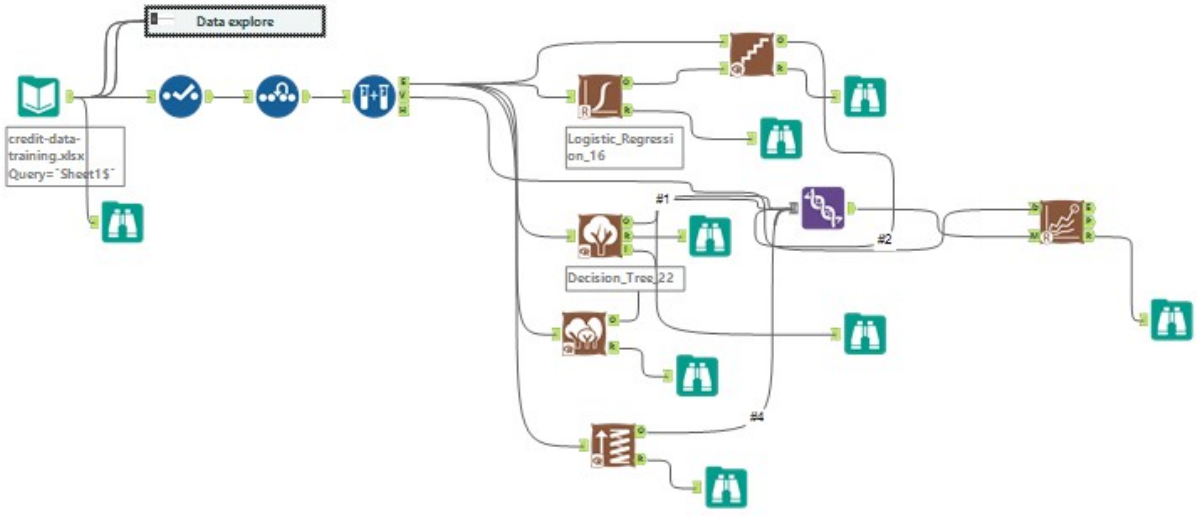
As there are only few records were missing in **age.year** field, so we decided to impute using median value as using median value doesn't change variance of data.

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.
  After looking into all model, it's been found that **Credit amount, Account Balance, Duration of credit month** are predictor variables are significant to make decision.

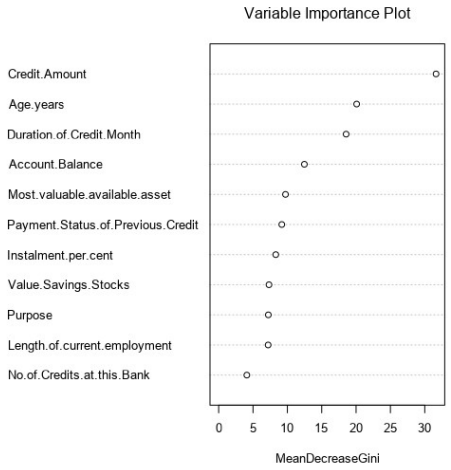
  Please find below model and their respective important variable
  For logistic regression: Account Balance, Purpose, Credit amount, Length of current employment

Coefficients:

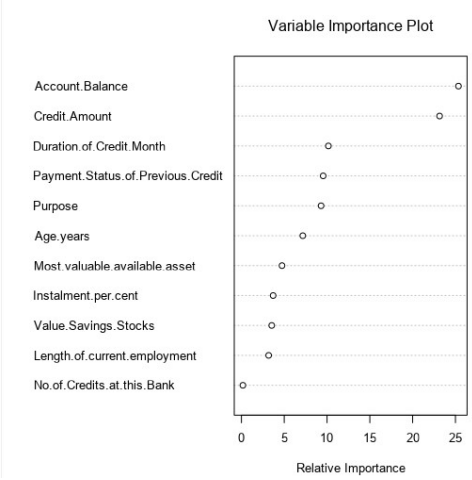
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.2290394	9.845e-01	-3.2800	0.00104	**
Account.BalanceSome Balance	-1.5843791	3.200e-01	-4.9511	7.38e-07	***
Duration.of.Credit.Month	0.0058321	1.365e-02	0.4272	0.6692	
Payment.Status.of.Previous.CreditPaid Up	0.4306851	3.847e-01	1.1195	0.26294	
Payment.Status.of.Previous.CreditSome Problems	1.2872278	5.339e-01	2.4109	0.01591	*
PurposeNew car	-1.7472435	6.271e-01	-2.7862	0.00533	**
PurposeOther	-0.2780516	8.305e-01	-0.3348	0.73778	
PurposeUsed car	-0.7651003	4.108e-01	-1.8624	0.06255	.
Credit.Amount	0.0001734	6.833e-05	2.5375	0.01116	*
Value.Savings.StocksNone	0.5996934	5.065e-01	1.1840	0.2364	
Value.Savings.Stocks£100-£1000	0.1818563	5.621e-01	0.3236	0.74628	
Length.of.current.employment4-7 yrs	0.5259720	4.934e-01	1.0660	0.28642	
Length.of.current.employment< 1yr	0.7776684	3.951e-01	1.9681	0.04906	*
Instalment.per.cent	0.2969774	1.384e-01	2.1457	0.0319	*
Most.valuable.available.asset	0.2877408	1.488e-01	1.9337	0.05315	.
Age.years	-0.0180861	1.475e-02	-1.2259	0.22022	
No.of.Credits.at.this.BankMore than 1	0.3918288	3.812e-01	1.0280	0.30397	

Decision tree: Account Balance, Duration of credit month, credit amount

Forest Model: Credit amount, age.years, Duration of credit month, Account Balance



Boost Model: Credit amount, Account Balance, Duration of credit month,



- 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?

  Please find below accuracy of model

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree_P4	0.6667	0.7685	0.6272	0.7905	0.3778
Logistic_stepwise	0.7600	0.8364	0.7306	0.8762	0.4889
Forest_model_P4	0.8000	0.8707	0.7421	0.9619	0.4222
Boost_P4	0.7867	0.8632	0.7513	0.9619	0.3778

  Confusion Matrix:

Confusion matrix of Boost_P4

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

Confusion matrix of Decision_Tree_P4

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

Confusion matrix of Forest_model_P4

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

Confusion matrix of Logistic_stepwise

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

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

Step 4: Writeup

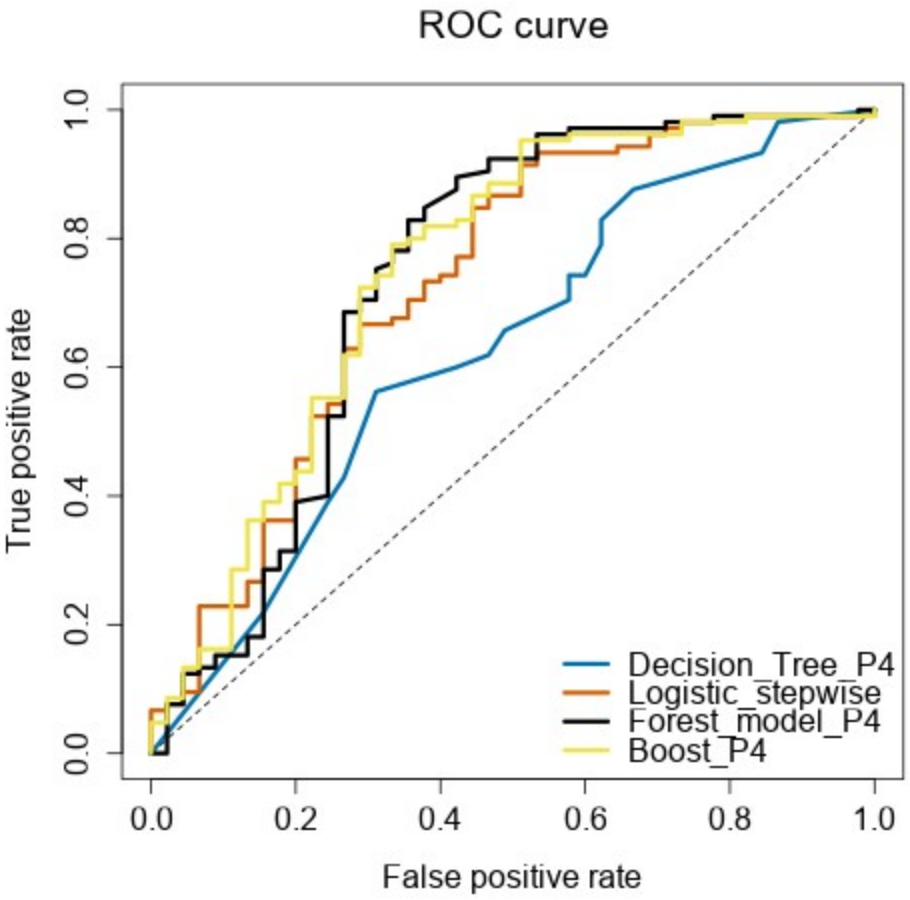
Answer these questions:

- 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

Forest model is selected among four models (logistic, decision tree, forest, boost) as its accuracy is 80% in which 96.19% for Accuracy_Creditworthy and 42.22% for Accuracy_Non_Creditworthy.

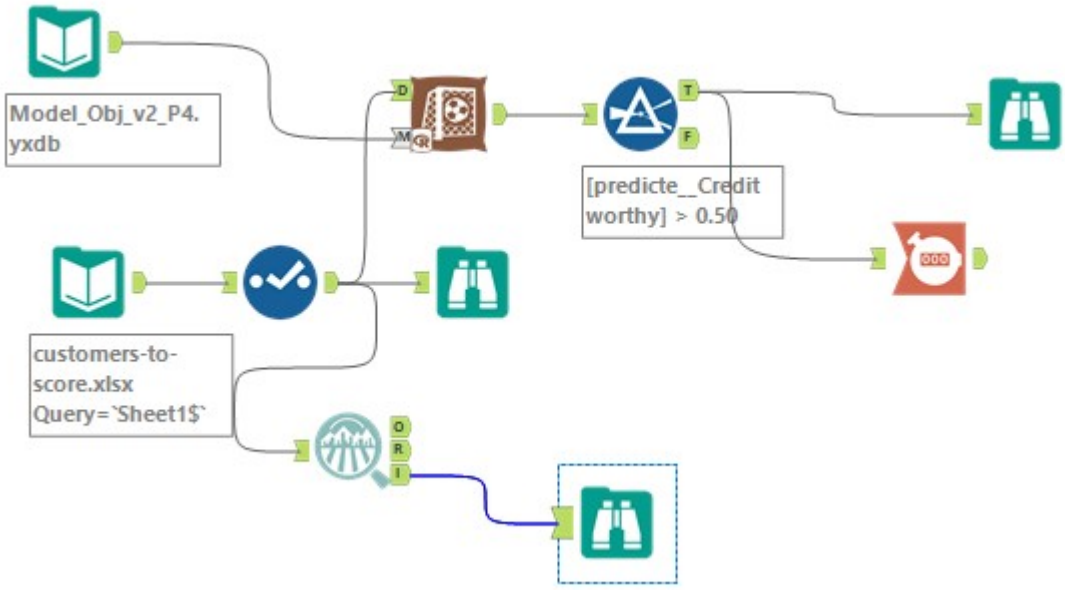
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree_P4	0.6667	0.7685	0.6272	0.7905	0.3778
Logistic_stepwise	0.7600	0.8364	0.7306	0.8762	0.4889
Forest_model_P4	0.8000	0.8707	0.7421	0.9619	0.4222
Boost_P4	0.7867	0.8632	0.7513	0.9619	0.3778

Please find below ROC graph comparing all models where forest model producing best result



:

- How many individuals are creditworthy?



After selecting *Score_Creditworthy>0.50, we got* **408**