Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

   a) We have a business problem which involves predictive analysis to provide a list of creditworthy customers due to a sudden influx of 500 loan applications up from a normal range of 200 per week.
   b) We have a lot of data in this problem. However, we need to properly format the data to form the training data set.
   c) Therefore, the data needs to be treated with appropriate data cleaning in proper order such as removing or imputing fields to obtain the final training dataset.

2. What data is needed to inform those decisions?

   - Data on all past applications
   - The list of customers that need to be processed in the next few days

3. What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
   - Since the answer to the business problem is in the form of binary choice, hence binary model is needed.

Step 2: Building the Training Set

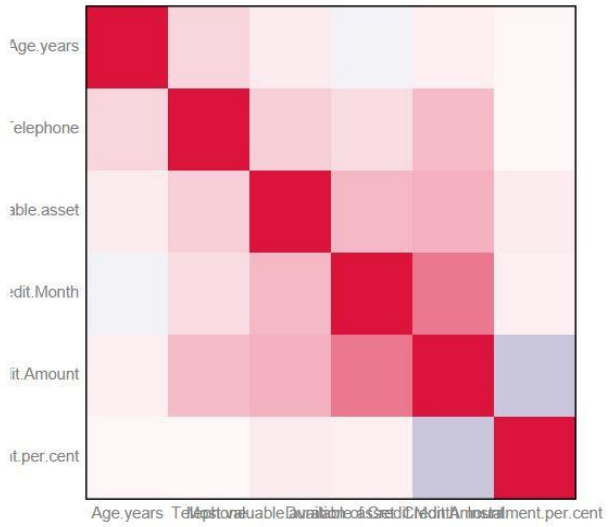
The data set consist of 20 fields including both numerical and non-numerical fields. On running the field summary tool, it was found that the 'Duration in current address field has a high percentage of null values, hence it was removed.

Fields Guarantors, Concurrent Credit, Foreign workers, Occupation & No of dependents have very low variability and hence rejected.

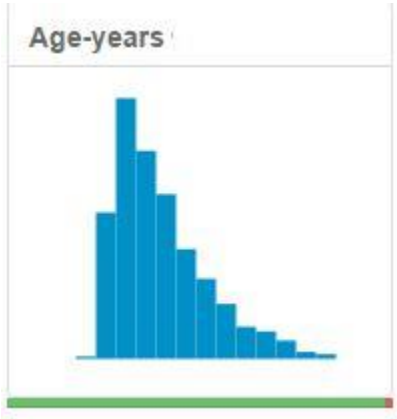
Also Telephone needs to be removed as we certainly do not need that kind of information on an applicant in order to decide whether to grant him loan or not.

Hence, the modelling is run with the rest 13 variables which remain available for use.

A correlation matrix was set up to check highly correlated fields after deselecting the above mentioned fields with high correlation limit set to >= 0.7. None were found to be of that figure.



Age Years field has few missing values and those values were replaced by the average value determined for the rest of the values in the null field.



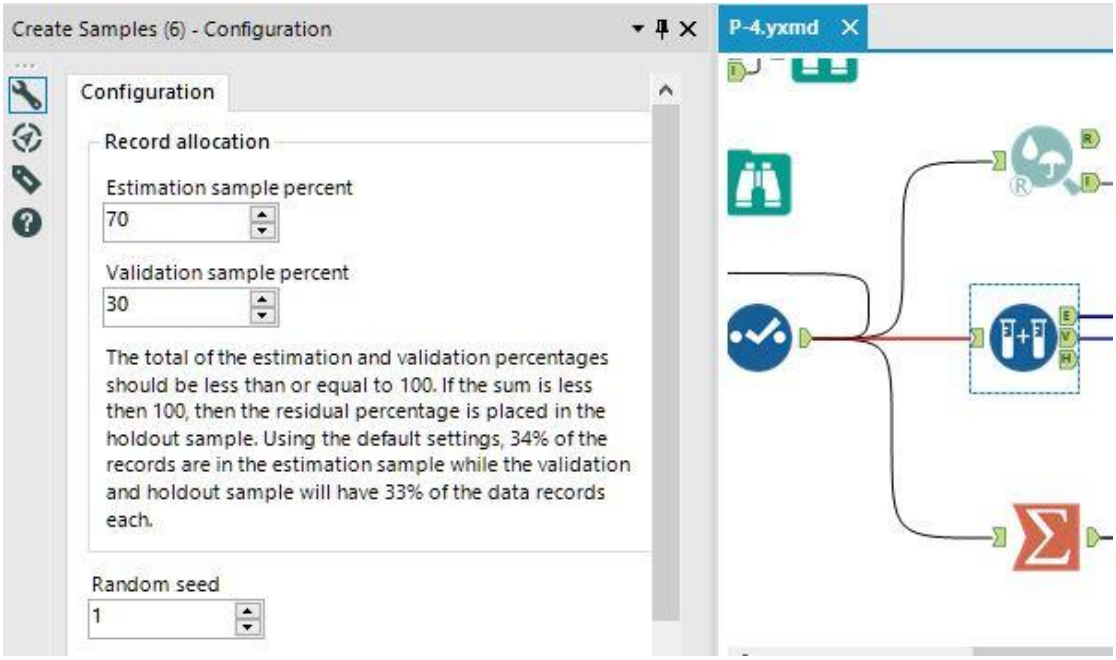
Record #	Avg_Age-years
1	35.637295

Record #	Avg_Age-years
1	36

So we end up with 13 columns and the Age Years field has an average of 36(after rounding off).

Step 3: Train your Classification Models

Estimation and Validation samples where 70% of dataset is Estimation and 30% is for Validation is created. Now, Logistic Regression, Decision Tree, Forest Model and Boosted Model were created.



*Answer these questions for **each model** you created:*

1. Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

   Predictor variable significance is mapped by running the various models with all the 13 variables and the following results were obtained.

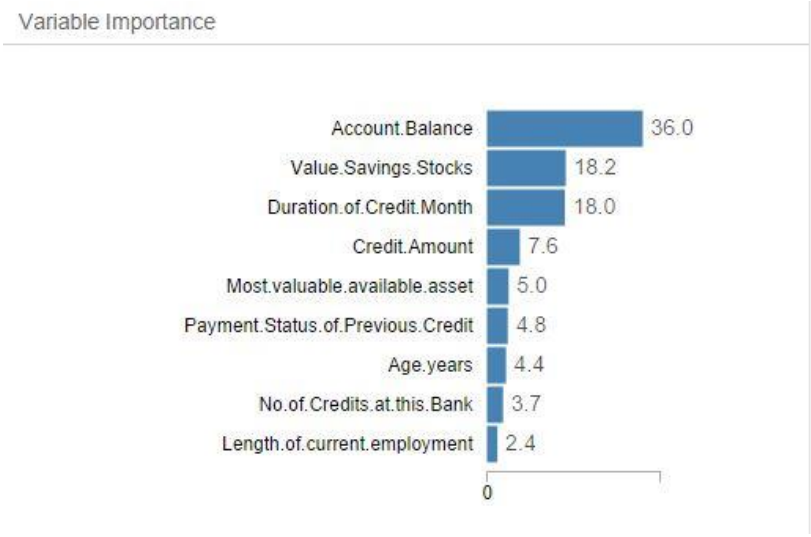
   Logistic regression:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.990817	1.013e+00	-2.9527	0.00315	**
Account.BalanceSome Balance	-1.543669	3.233e-01	-4.7745	1.80e-06	***
Duration.of.Credit.Month	0.006391	1.371e-02	0.4660	0.6412	
Payment.Status.of.Previous.CreditPaid Up	0.402974	3.843e-01	1.0487	0.2943	
Payment.Status.of.Previous.CreditSome Problems	1.259683	5.334e-01	2.3616	0.0182	*
PurposeNew car	-1.755074	6.278e-01	-2.7954	0.00518	**
PurposeOther	-0.290165	8.359e-01	-0.3471	0.72848	
PurposeUsed car	-0.785627	4.124e-01	-1.9049	0.05679	.
Credit.Amount	0.000177	6.841e-05	2.5879	0.00966	**
Value.Savings.StocksNone	0.609298	5.099e-01	1.1949	0.23213	
Value.Savings.Stocks£100-£1000	0.172241	5.649e-01	0.3049	0.76046	
Length.of.current.employment4-7 yrs	0.530959	4.932e-01	1.0767	0.28163	
Length.of.current.employment< 1yr	0.777372	3.957e-01	1.9646	0.04946	*
Instalment.per.cent	0.310524	1.399e-01	2.2197	0.02644	*
Most.valuable.available.asset	0.325606	1.557e-01	2.0918	0.03645	*
Age.years	-0.015092	1.539e-02	-0.9809	0.32666	
Type.of.apartment	-0.254565	2.958e-01	-0.8605	0.38949	
No.of.Credits.at.this.BankMore than 1	0.362688	3.816e-01	0.9505	0.34184	

   Here significant predictor variables are: Balance, Purpose, Payment status, Credit amount, length of current employment, credit amount, installment per cent, most valuable available asset.

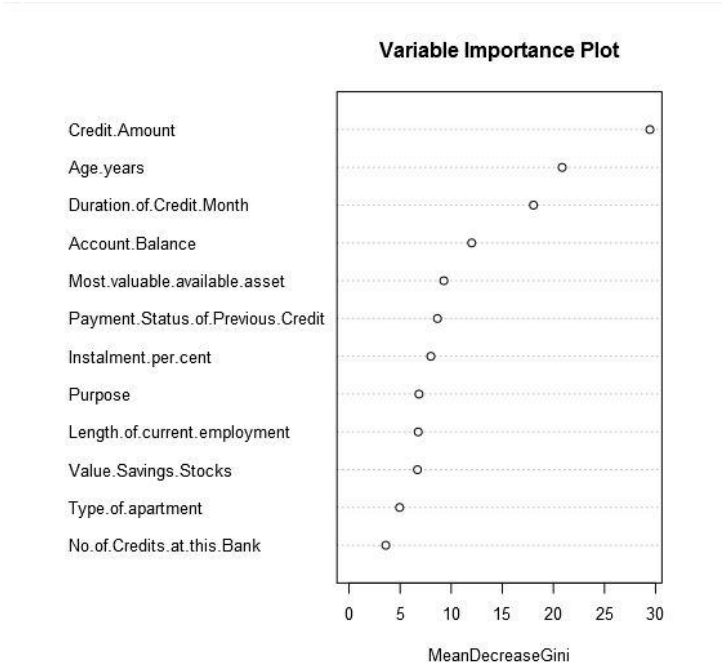
   Decision tree

Here significant predictor variables are: Balance, Value of savings stocks, Duration of credit month, Credit Amount, most valuable available asset, payment status of previous credit



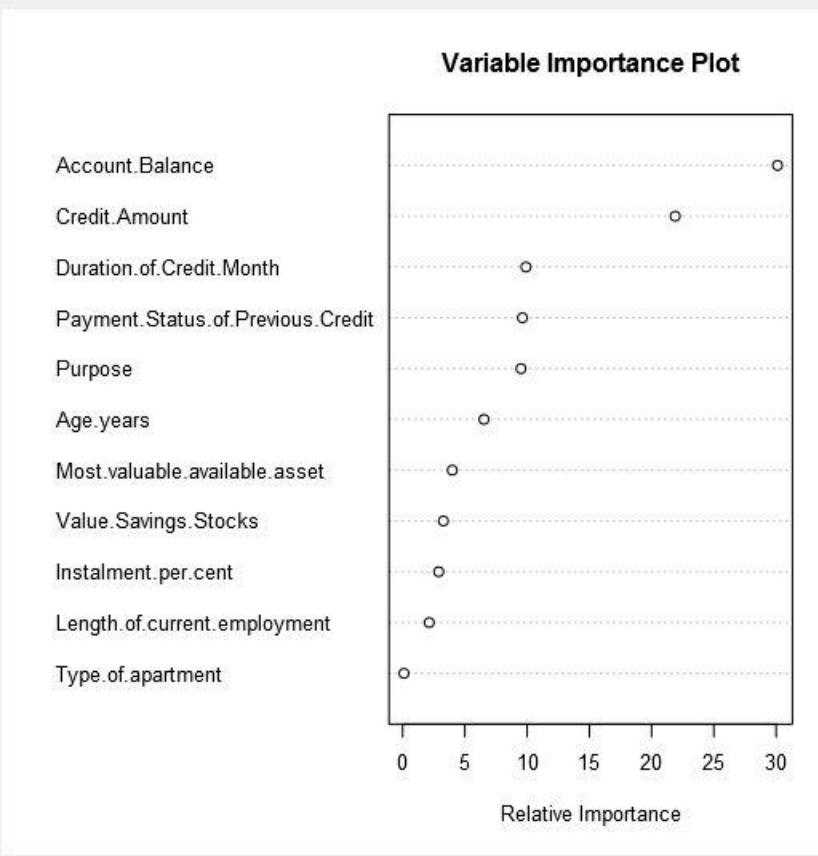
Forest Model

Here significant predictor variables are: Credit Amount, Age Years, Duration of Credit Month, Account Balance



Boosted Model

Here significant predictor variables are: Account Balance, credit amount, Duration of Credit Month , payment status, purpose.



The model chosen for final analysis is the Forest model which was based on the following parameters:

1. The best accuracy overall is for Forest Model with the 13 variables

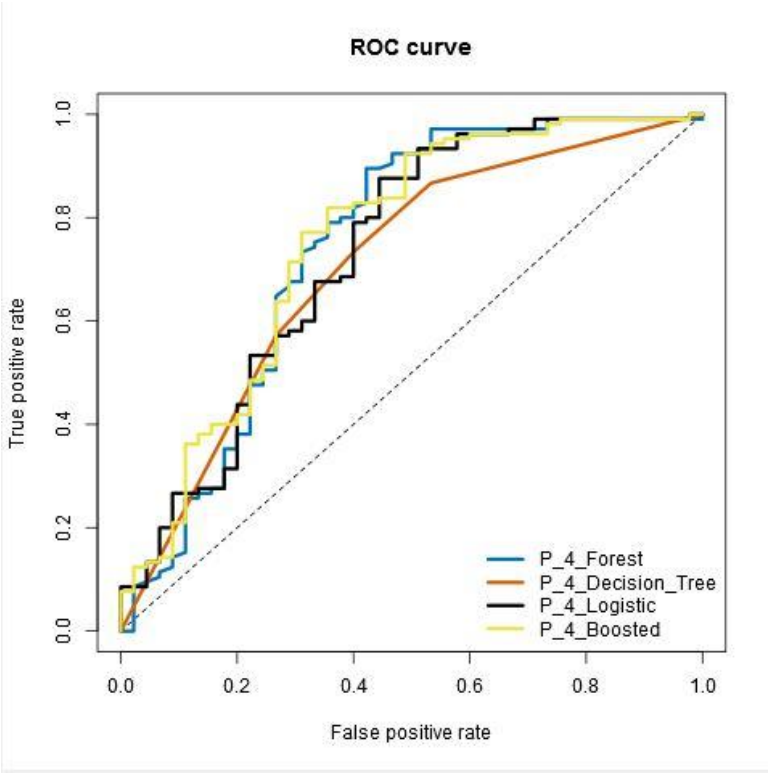
Fit and error measures

Model	Accuracy	F1	AUC
P_4_Forest	0.8000	0.8718	0.7426
P_4_Decision_Tree	0.7467	0.8273	0.7054
P_4_Logistic	0.7800	0.8520	0.7310
P_4_Boosted	0.7867	0.8621	0.7526

2. For predicting non-creditworthy, best model was Forest model, and logistic tree model is the best for predicting Creditworthy correctly.

Accuracy_Creditworthy	Accuracy_Non-Creditworthy
0.7907	0.8571
0.7913	0.6000
0.8051	0.6875
0.7874	0.7826

3. The ROC curve shows that the Forest model has the best overall true positive rates.



4. With the confusion matrix, we see that Forest model predicts better both creditworthy and non_creditworthy.

Confusion matrix of P_4_Boosted

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	27
Predicted_Non-Creditworthy	5	18

Confusion matrix of P_4_Decision_Tree

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

Confusion matrix of P_4_Forest

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	27
Predicted_Non-Creditworthy	3	18

Confusion matrix of P_4_Logistic

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

Hence, the Forest model is used to score the new customers as it is the better fit overall. As per criteria, if Score_Creditworthy is greater than Score_NonCreditworthy, the person is labeled as “Creditworthy” and the number determined is 413.

Record #	Count
1	413