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?

Ans. For each loan application that we receive we have to evaluate the creditworthiness of the customer. We have to determine whether we should give loan to the customer or not.

- What data is needed to inform those decisions?

Ans. We have data on all the past applicants. We can use this data to build a classification model to classify each loan applicant as creditworthy or noncredit worthy.

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

Ans. Since we have to classify whether a customer is creditworthy or not, we need to use the Binary Classification model. Our target variable 'Credit Application Result' has only two outcomes i.e Creditworthy and Noncredit Worthy .Therefore, Binary Classification Model.

Step 2: Building the Training Set

- For numerical data fields, are there any fields that highly-correlate with each other?

Ans. In Alteryx, an Association Analysis is performed on numerical variables. And there are no two numerical variables that are highly- correlated with each other i.e correlation of higher than 0.7.

Pearson Correlation Analysis

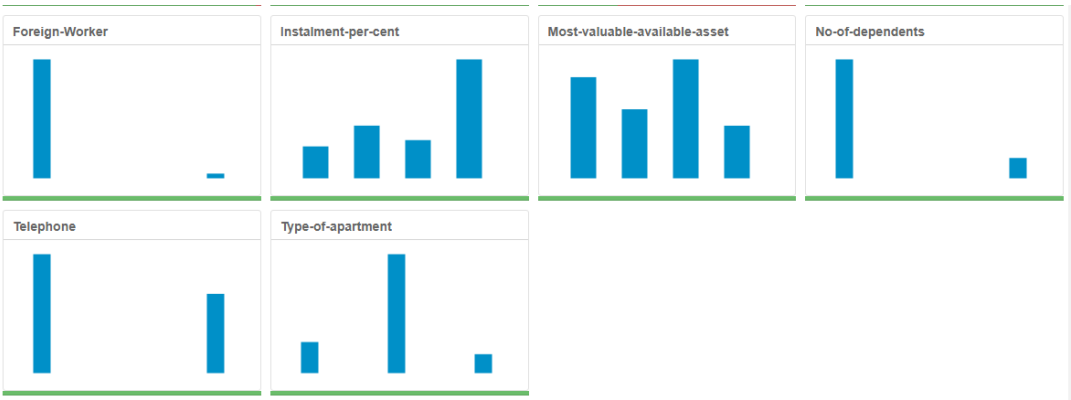
Full Correlation Matrix

	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Most.valuable.available.asset	Age.years	Type.of.apartment
Duration.of.Credit.Month	1.000000	0.570441	0.079515	0.304734	-0.066319	0.153141
Credit.Amount	0.570441	1.000000	-0.285631	0.327762	0.068643	0.168683
Instalment.per.cent	0.079515	-0.285631	1.000000	0.078110	0.040540	0.082936
Most.valuable.available.asset	0.304734	0.327762	0.078110	1.000000	0.085437	0.379650
Age.years	-0.066319	0.068643	0.040540	0.085437	1.000000	0.333075
Type.of.apartment	0.153141	0.168683	0.082936	0.379650	0.333075	1.000000

- Are there any missing data for each of the data fields?

Ans. Of all the fields in the dataset , the field '*Duration in Current Address'* has 68.8% missing data .Therefore, this field must be removed.
The field 'Age Years' has 2.4% missing data. Therefore, we impute the missing field values with the median age. Median Age is used instead of mean age for imputation as data skewed to the right as shown below.

- Are there only a few values in a subset of your data field?
  1. **Ans.** The fields ' Guarantors ','Foreign Worker' and 'No of Dependents ' are removed from our dataset as these fields show low variability i.e these fields are heavily skewed towards one type of data. These fields should be removed to prevent skewness in our analysis results.
  2. The fields 'Concurrent Credits' and 'Occupation' shows low variability as data in these fields is entirely uniform and there is no other variations of data. These fields are removed from the dataset.
  3. *'Telephone'* field is also removed due to its irrelevancy to the customer creditworthy.



- Our cleaned data set have 13 columns. We'll use this cleaned data set to build our classification model.

Step 3: Train your Classification Models

We created our Estimation and Validation samples where 70% of our dataset is used for Estimation and 30% of our entire dataset is reserved for Validation

1). Logistic Regression(Stepwise)

Using 'Credit Application Result' as the target variable , *Account Balance , Purpose , Credit amount* are most significant variables with p-value lesser than 0.05.

Overall accuracy of this model is 76%. The accuracy for predicting customer as creditworthy i.e. 87.62% is higher than the accuracy for predicting customer as noncredit worthy i.e.48.89%. From the confusion matrix it can be seen that this model is biased towards predicting customers as noncredit worthy.

Records 1 to 10

Report for Logistic Regression Model Stepwise_Project

Basic Summary

Call:

glm(formula = Credit.Application.Result ~ Account.Balance + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset, family = binomial("logit"), data = the.data)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.289	-0.713	-0.448	0.722	2.454

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	.

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial taken to be 1)

Null deviance: 413.16 on 349 degrees of freedom

Residual deviance: 328.55 on 338 degrees of freedom

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise_Project	0.7600	0.8364	0.7306	0.8762	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_[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.

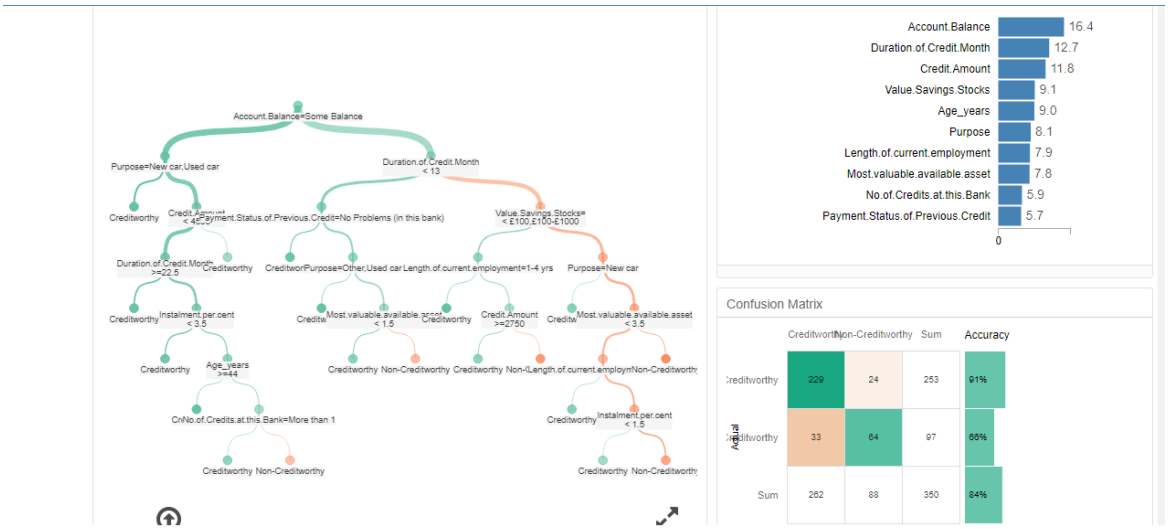
Confusion matrix of Stepwise_Project

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

2). Decision Tree

Using ‘Credit-Application Result’ as the target variable, *Account Balance , Duration of Credit Month and Credit Amount are the top 3 most significant predictor variables.* The decision tree has an overall accuracy of estimation sample of 84%. But when we validated our model using Model Comparison tool, the overall accuracy of the model is 66.66%. This means that decision tree has overfitted our model.

From the confusion matrix it can be seen that this model is biased towards predicting customers as noncredit worthy.



Layout

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Project	0.6667	0.7685	0.6272	0.7905	0.3778

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] 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.

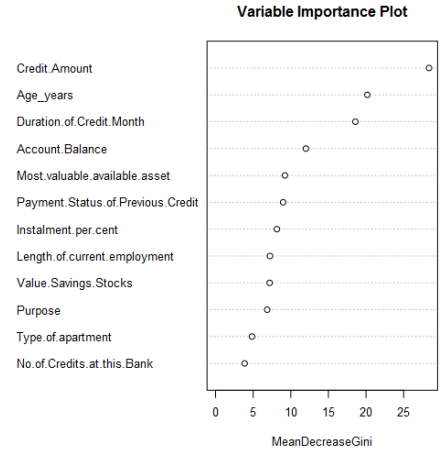
Confusion matrix of DT_Project

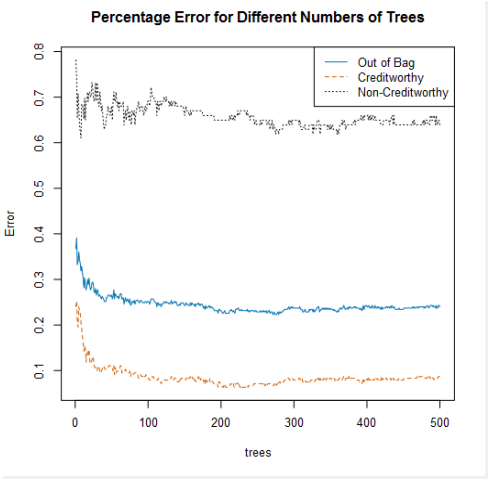
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	28
Predicted_Non-Creditworthy	22	17

3). Forest Model

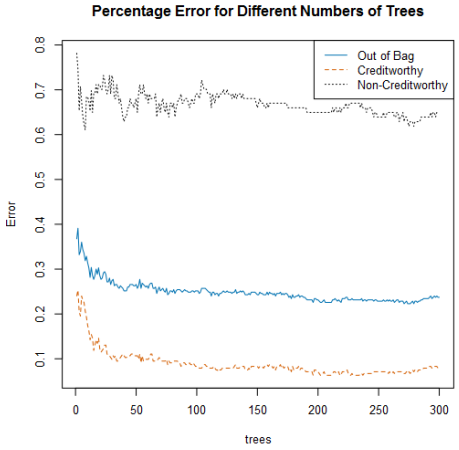
Using ‘Credit-Application Result’ as the target variable, *Credit Amount, Age Years and Duration of Credit Month* are the 3 most important variables. From the plot *Percentage Error For Different Number Of Trees* it can be observed that we achieved the flatline almost after 300 trees. Therefore, we can reduce the total number of trees used in the model from 500 to 300 to avoid over computing and to minimize the error.

The overall accuracy (after reducing the number of trees to 300) is 82%. From the confusion matrix it can be seen that this model is biased towards predicting customers as noncredit worthy.





Variable Importance Plot



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
FM_Project	0.8200	0.8831	0.7387	0.9714	0.4667

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] 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 FM_Project

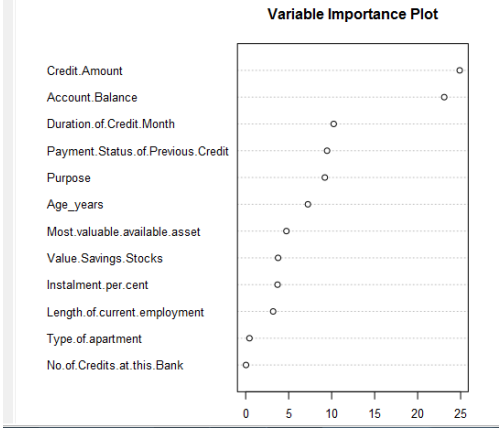
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	24
Predicted_Non-Creditworthy	3	21

4). Boosted Model

Using ‘Credit-Application Result’ as the target variable, *Credit Amount , Account Balance and Duration of Credit Month are top 3 most significant variables.*

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 3940

Plots:



Overall accuracy of the model is 79.43%.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
BM_Project	0.7943	0.8705	0.8351	0.9565	0.3711

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] 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 BM_Project

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	242	61
Predicted_Non-Creditworthy	11	36

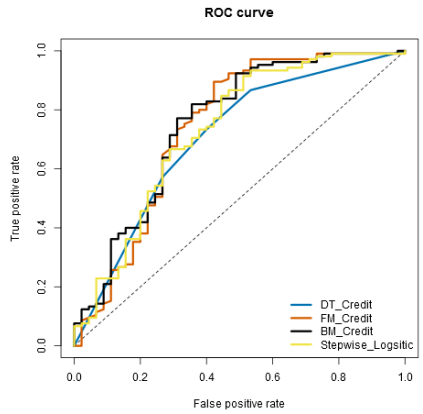
From the confusion matrix it can be seen that this model is biased towards predicting customers as noncredit worthy.

Step 4: Writeup

- Which model did you choose to use?

  ANS. Forest model is chosen as it offers the highest accuracy at 82% against validation set. Its accuracy for creditworthy is highest of all i.e. 97.14% and accuracy for non-creditworthy is 46.7%

Forest model reaches the true positive rate at the fastest rate.



.

- How many individuals are creditworthy?

ANS. 410