Christopher W. Hong

Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

- What decisions needs to be made?

  The company suddenly has an influx of nearly 500 loan applications to process this week. The manager needs to come up with an efficient and accurate classification model that can systematically evaluate all these loan applications within one week.

- What data is needed to inform those decisions?

  Data on all past applications.

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

  Since the credit application result consists of two categories, i.e., Creditworthy or Non-Creditworthy, a binary model is used to make these decisions.

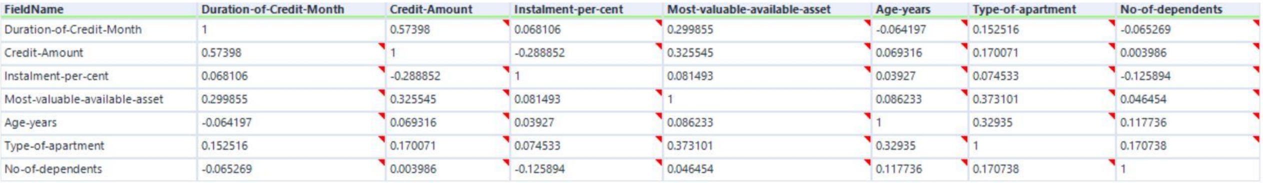
Christopher W. Hong

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.***

<u>Here are some guidelines to help guide your data cleanup:</u>

- 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".



FieldName	Duration-of-Credit-Month	Credit-Amount	Instalment-per-cent	Most-valuable-available-asset	Age-years	Type-of-apartment	No-of-dependents
Duration-of-Credit-Month	1	0.57398	0.068106	0.299855	-0.064197	0.152516	-0.065269
Credit-Amount	0.57398	1	-0.288852	0.325545	0.069316	0.170071	0.003986
Instalment-per-cent	0.068106	-0.288852	1	0.081493	0.03927	0.074533	-0.125894
Most-valuable-available-asset	0.299855	0.325545	0.081493	1	0.086233	0.373101	0.046454
Age-years	-0.064197	0.069316	0.03927	0.086233	1	0.32935	0.117736
Type-of-apartment	0.152516	0.170071	0.074533	0.373101	0.32935	1	0.170738
No-of-dependents	-0.065269	0.003986	-0.125894	0.046454	0.117736	0.170738	1

  Based on the correlation matrix shown above, no numerical data field has a high correlation with others.

- Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed

  By exploring the data using the Field Summary tool in Alteryx, 68.8% of field Duration In Current Address is missing. Besides, 2.4% of field Age Years is missing.

- 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.

  Using the same tool mentioned in the previous question, the result shows that fields including Occupation and Concurrent Credits look very uniform. Fields such as Guarantors, Foreign Worker, No-of-dependents show "low variability.

- Your clean data set should have 13 columns where the Average of **Age Years** should be 36 (rounded up)

Christopher W. Hong

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:*

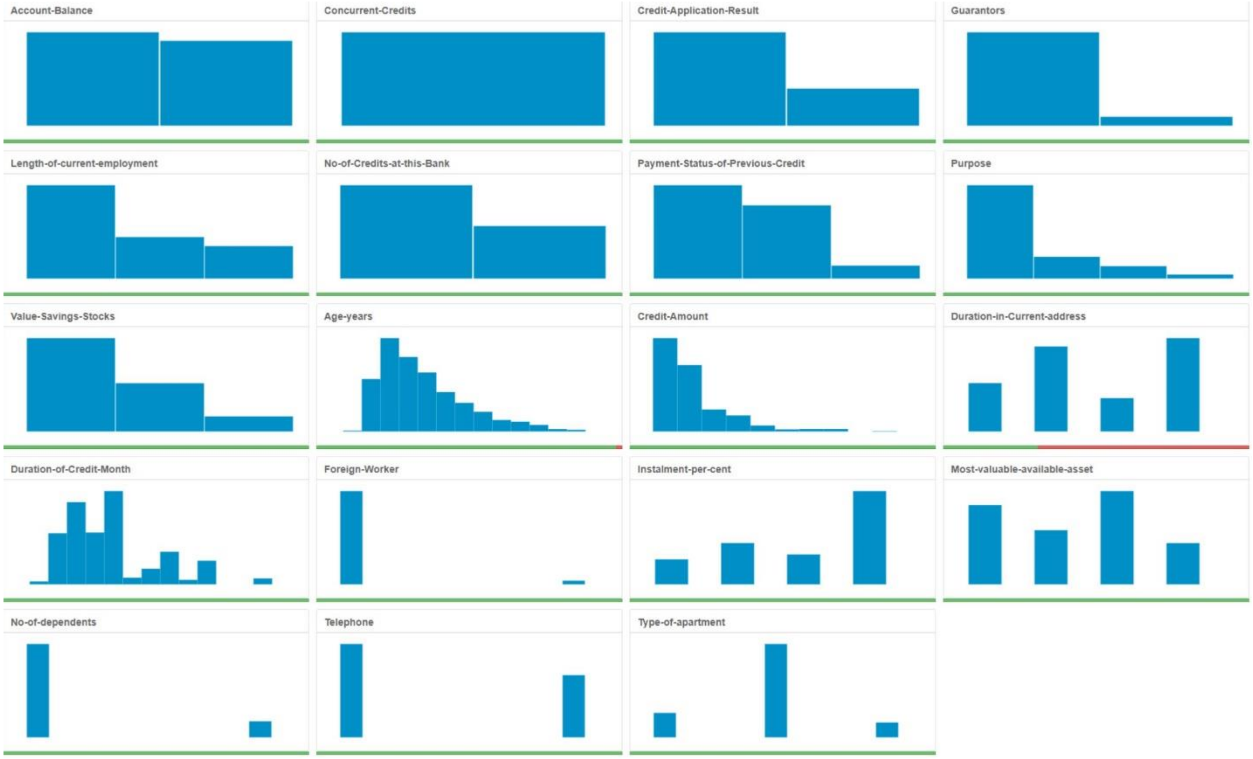
Variable	**Data Type**
Credit-Application-Result	String
Account-Balance	String
Duration-of-Credit-Month	Double
Payment-Status-of-Previous-Credit	String
Purpose	String
Credit-Amount	Double
Value-Savings-Stocks	String
Length-of-current-employment	String
Instalment-per-cent	Double
Guarantors	String
Duration-in-Current-address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

Christopher W. Hong

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.



As I mentioned before, there are too many missing values in the field Duration In Current Address to impute, so it is better to be removed. Fields with uniform or “low variability” as mentioned before are removed. Field Age Years is imputed with median strategy since there is only 2.4% missing values.

Christopher W. Hong

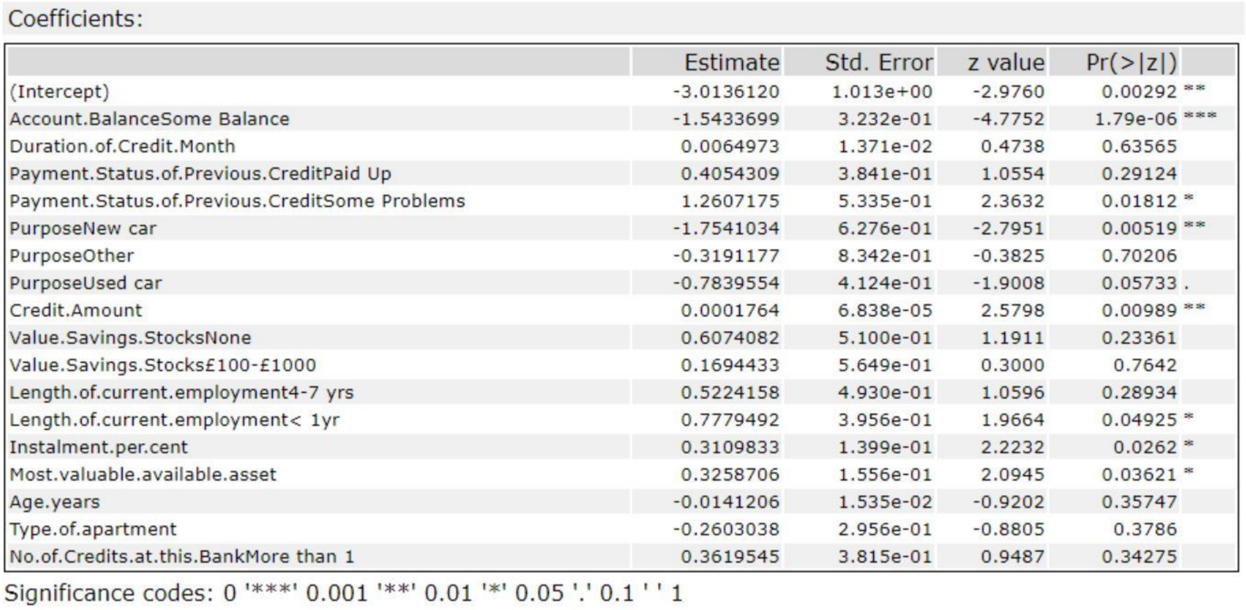
Step 3: Train your Classification Models

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.

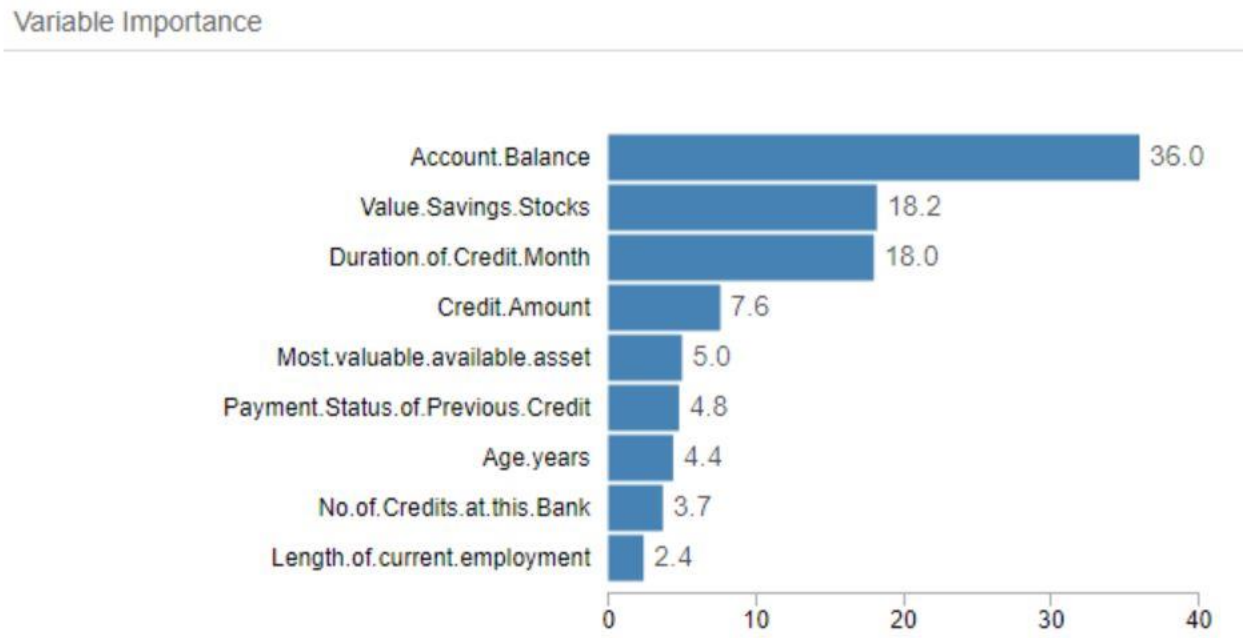


Coefficients:

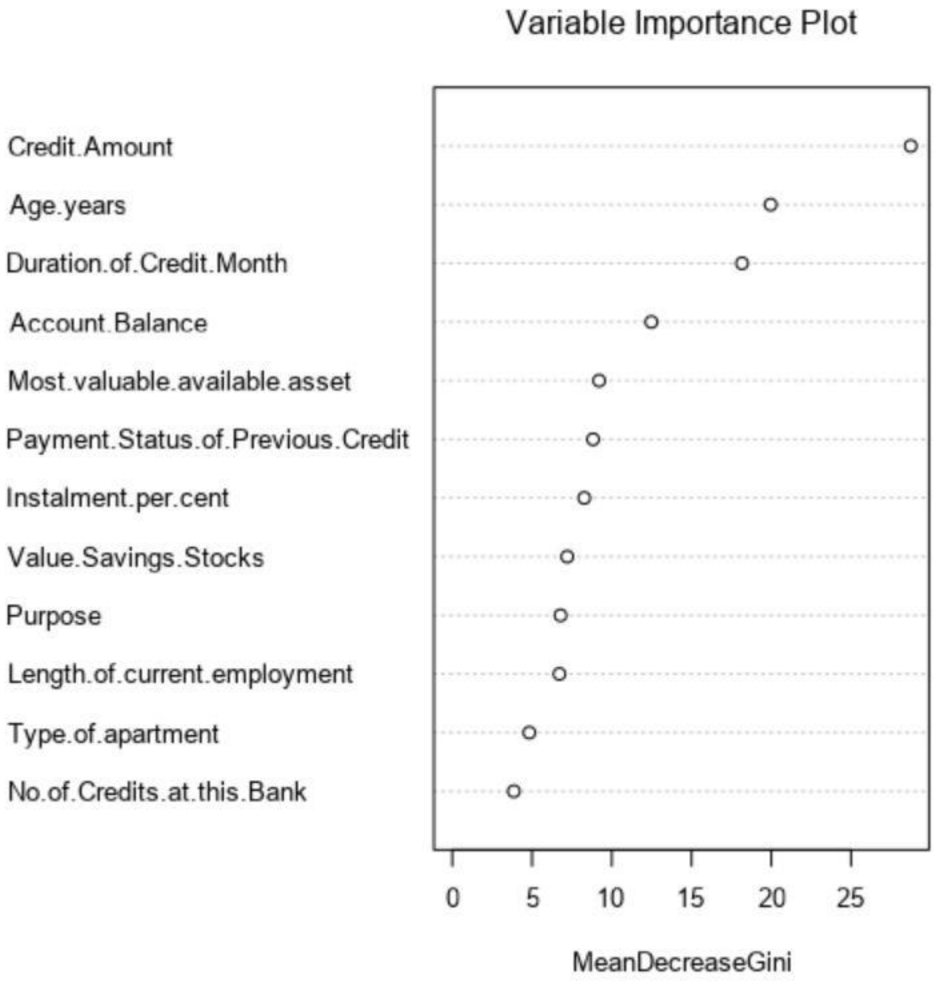
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.0136120	1.013e+00	-2.9760	0.00292	**
Account.BalanceSome Balance	-1.5433699	3.232e-01	-4.7752	1.79e-06	***
Duration.of.Credit.Month	0.0064973	1.371e-02	0.4738	0.63565	
Payment.Status.of.Previous.CreditPaid Up	0.4054309	3.841e-01	1.0554	0.29124	
Payment.Status.of.Previous.CreditSome 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	

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

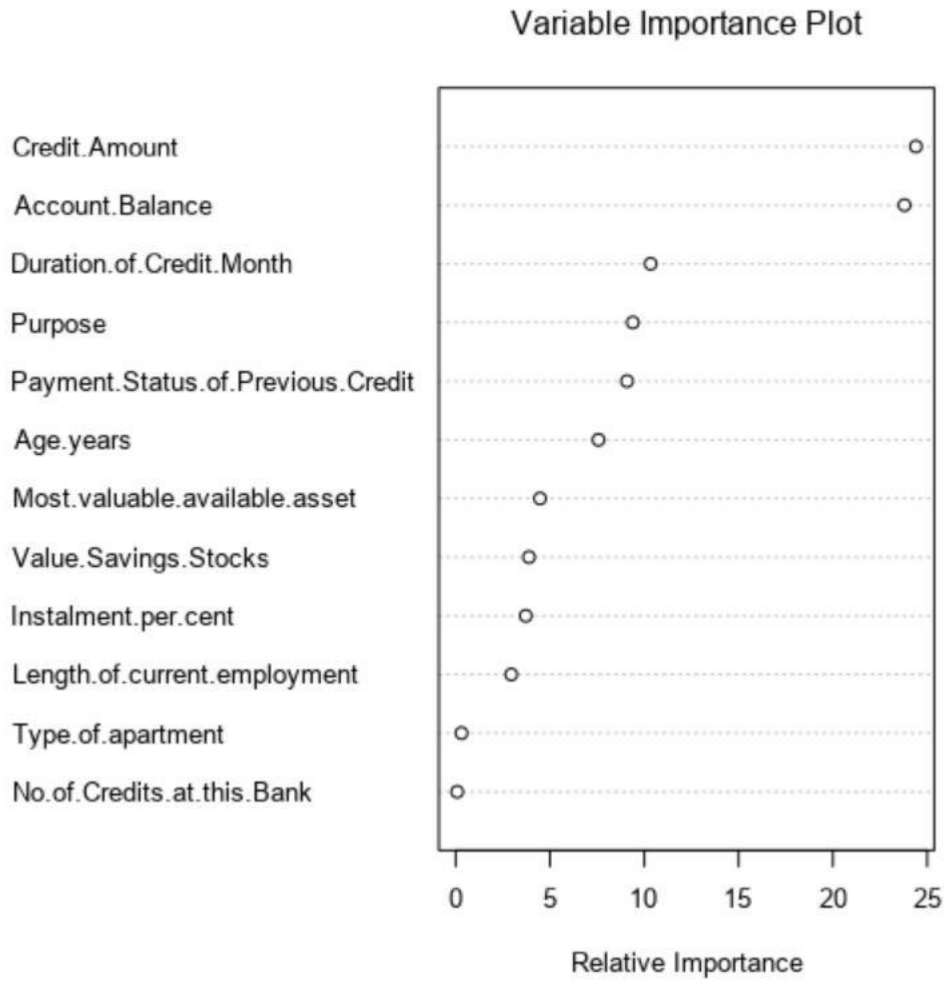
Based on the result of the Logistic Regression model shown above, predictors including subcategory Some Balance of Account Balance, subcategory Some Problems of Payment Status of Previous Credit, subcategory New Car of Purpose, Credit Amount, subcategory < 1yr of Length of Current Employment, Instalment per cent and Most Valuable Available Asset are statistically significant at the level of 0.05.



For the Decision Tree model, based o the variable importance shown above, predictors including Account Balance, Value Savings Stocks and Duration of Credit Month are the three most important ones.

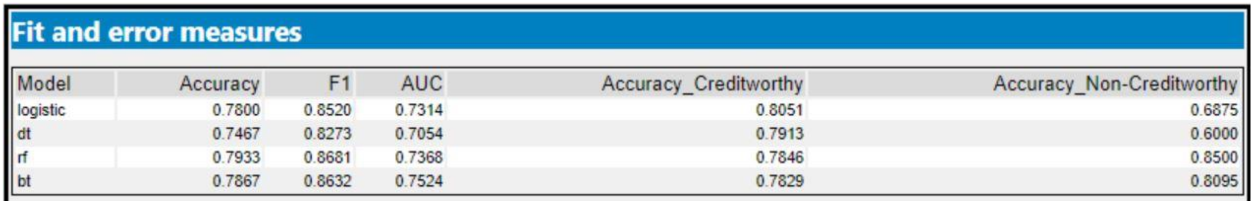


For the Random Forest model, the variable importance plot shows that predictors such as Credit Amount, Age Years and duration of Credit Month are the three most important ones.



For the Boosted model, the variable importance plot shows that predictors such as Credit Amount, Account Balance and Duration of Credit Month are the three most important ones.

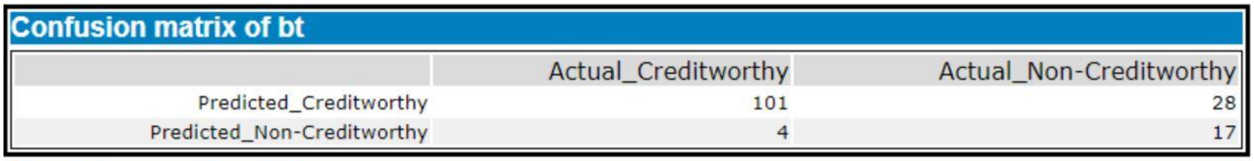
- Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model’s predictions?



Fit and error measures

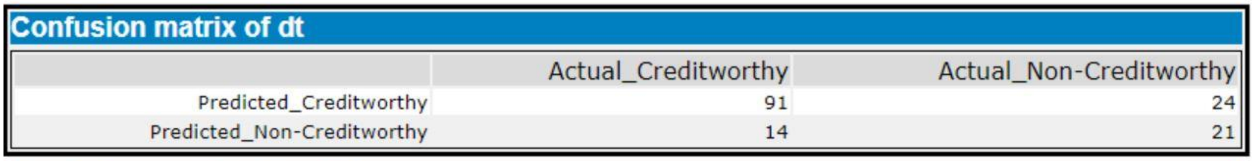
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
logistic	0.7800	0.8520	0.7314	0.8051	0.6875
dt	0.7467	0.8273	0.7054	0.7913	0.6000
rf	0.7933	0.8681	0.7368	0.7846	0.8500
bt	0.7867	0.8632	0.7524	0.7829	0.8095

The overall percent accuracy is around 78% for all the models but Decision Tree model (around 75%).



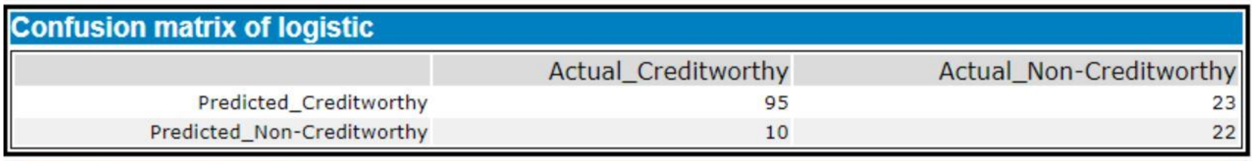
Confusion matrix of bt

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



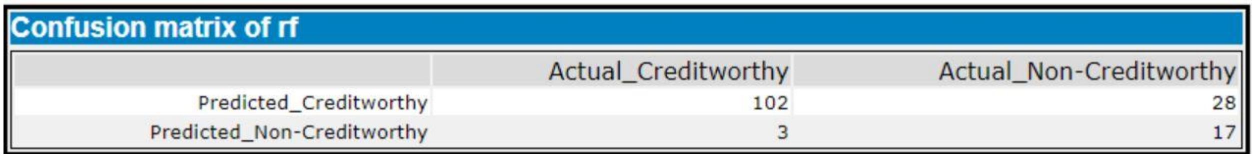
Confusion matrix of dt

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



Confusion matrix of logistic

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



Confusion matrix of rf

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

Based on the confusion matrices of all the models shown above, it is more difficult to predict Non-Creditworthy than Creditworthy since the false positive rate is higher than the false negative rate for each model.

Christopher W. Hong

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.

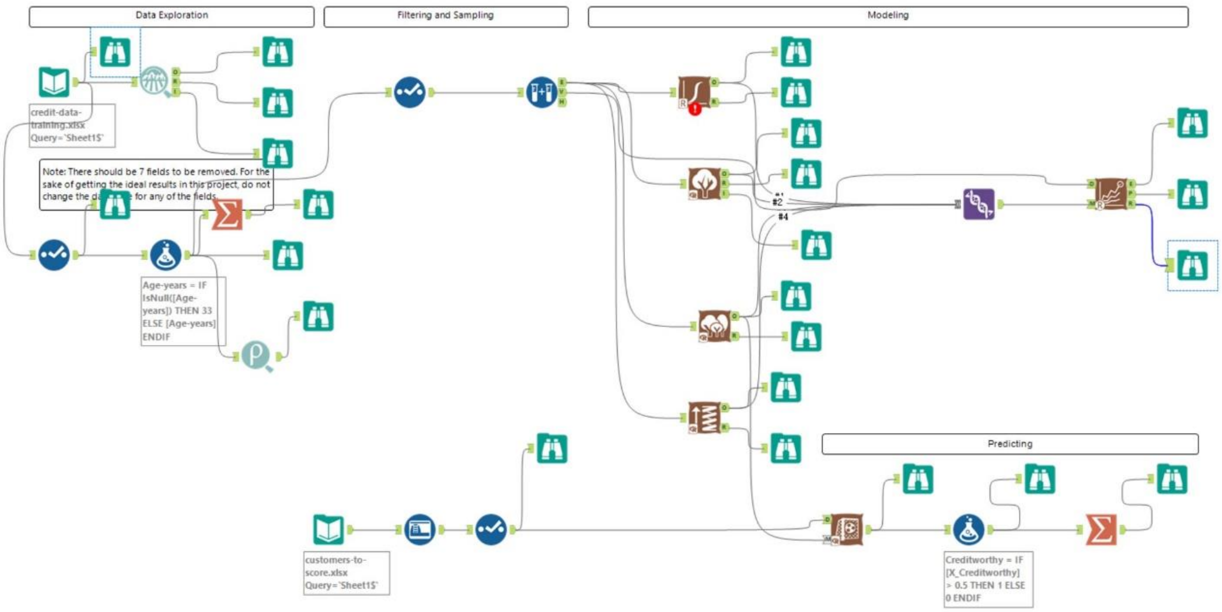
Since the manager cares about prediction accuracy for both Creditworthy and Non-Creditworthy segments and the model was trained in a dataset which there are more Creditworthy instances than Non-Creditworthy ones, the F1 measure should be the appropriate measure in this context. Among all these models, the Random Forest model is the best since it has the highest F1 measure against the Validation dataset. In addition, the Random Forest model has the highest accuracy rate and AUC score, too.

- How many individuals are creditworthy?

There are 408 individuals are classified as creditworthy based on the Random Forest model.

Christopher W. Hong

Alteryx workflow:



Note: There should be 7 fields to be removed. For the sake of getting the ideal results in this project, do not change the da[illegible]e for any of the fields.

```
Age-years = IF
IsNull([Age-
years]) THEN 33
ELSE [Age-years]
ENDIF
```

```
Creditworthy = IF
[X_Creditworthy]
> 0.5 THEN 1 ELSE
0 ENDIF
```