

## Project 4: Creditworthiness

By: Huong Ivy Nguyen

### Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

In this project, we need to determine which applicant is creditworthy to approve for a loan. Usually, there are 200 applications/week that need to be process; however, this number has increased for this week (500 applications). Hence, we need to build a classification model to speed up the approval process.

2. What data is needed to inform those decisions?

We will use information of a dataset from past loan applicants to build a classification model to predict the creditworthiness of the new applications. This dataset includes various features including occupation, concurrent credits, guarantors, etc....

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

In order to make the decision whether or not an applicant is creditworthy to get a loan, we need to use a binary classification model.

### Step 2: Building the Training Set

For this analysis, the target variable is the Credit-Application-Result field. In order to determine the distribution for each predictor variable, a field summary was established (Figure 1a and Figure 1a).

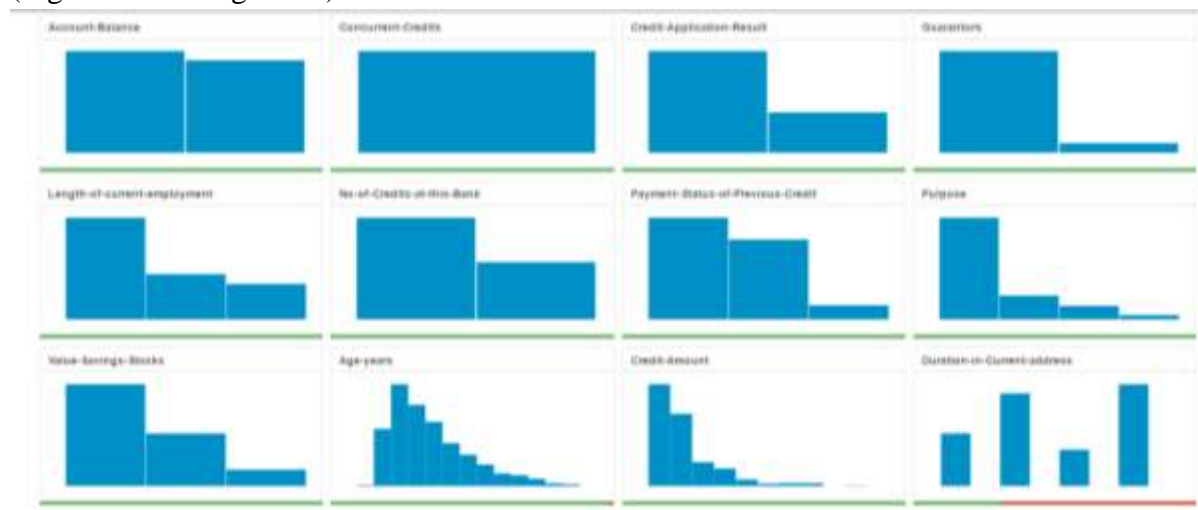


Figure 1a. Distribution of Account Balance, Concurrent-Credits, Credit-Application-Result, Guarantors, Length-of-current-employment, No-of-Credits-at-this-Bank,

Payment-Status-of-Previous-Credits, Purpose, Value-Savings-Stocks, Age-years, Credit-Amount, and Duration-at-Current-Address

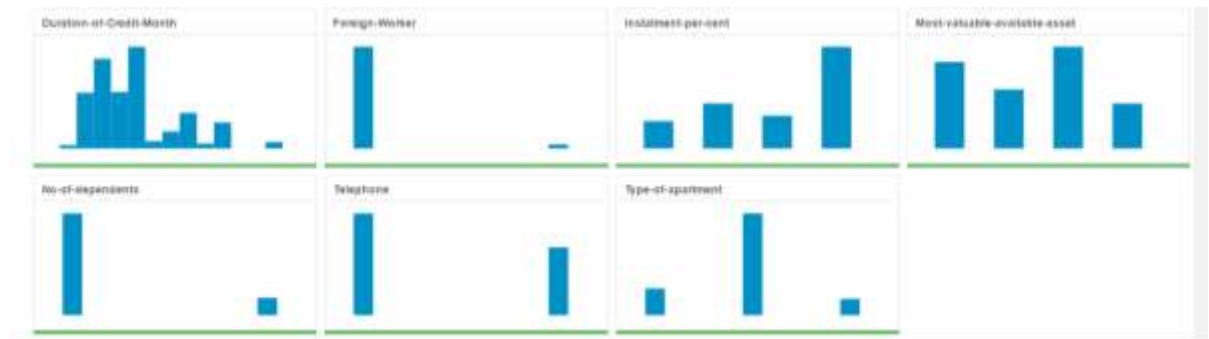


Figure 1b. Distribution of Duration-of-Credit-Month, Foreign-Worker, Instalment-per-cent, Most-valuable-available-asset, No-of-dependents, Telephone, and Type-of-apartments.

The field summary can help us determine if it is necessary to remove a particular variable due to its low variability (too uniform), its percentage of missing data, or its bias towards a single variable. Based on the field summary visualization, the following fields were removed:

- Concurrent-credits (low variability)
- Occupation (low-variability)
- Guarantors (bias towards the “none” category)
- Duration-in-current-address (69% data is missing)
- Foreign-worker (bias towards category “2”)
- Telephone (eliminate due to being irrelevant to the target variable)
- No-of-dependents (bias towards category “1-1.1”)

In addition, the Age-years field also has some missing data (2%) which would need to be imputed with the median value of this field. The median value is more appropriate to use in the imputation process than the mean value since the dataset of Age-years is positively skewed to the left.

Finally, an association analysis was also established to determine if there is any correlation between any two predictor variables. (Figure 2). If the association score is greater than 0.7 then that means correlation exists.

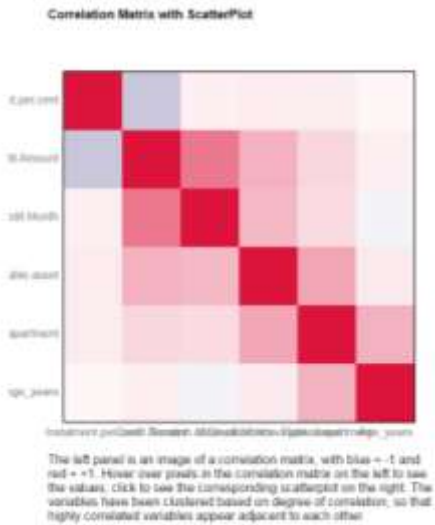


Figure 2. Correlation matrix of predictor variables

According to Figure 2, there is no variable that has a correlation score of 0.7 or greater. Thus, no inter-correlation is present. The cleaned-up dataset includes 13 columns with the average of the Age\_years column is 35.574.

### Step 3: Train your Classification Models

#### Logistic Regression-Stepwise:

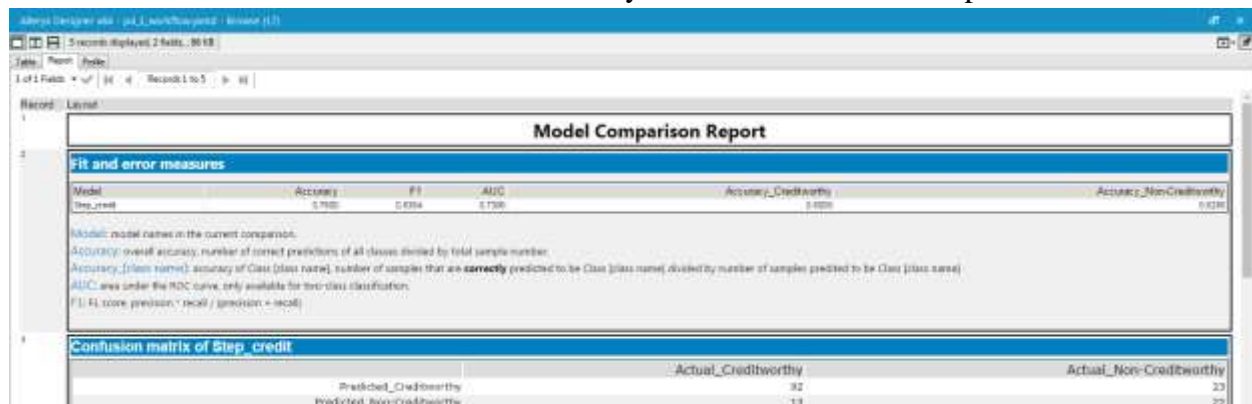
1. Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

Report for Logistic Regression Model Step_credit				
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(link), data = the.data)				
Deviance Residuals:				
	Min	1Q	Median	3Q
	-2.289	-0.713	-0.448	0.722
Coefficients:				
(Intercept)	-2.9921814	Std. Error	z value	Pr(> z )
Account.Balance	-1.8953228	3.667e-01	-5.168	1.85e-07 ***
Payment.Status.of.Previous.CreditPaid Up	0.2366657	2.677e-01	0.793	0.42776
Payment.Status.of.Previous.CreditSame.Problem	1.2154514	5.151e-01	2.359	0.0182 *
Purposeused car	-1.0991164	5.142e-01	-2.136	0.0336 **
PurposeOther	-0.3257637	5.179e-01	-0.629	0.5282
Purposeused car	-0.7641020	4.004e-01	-1.908	0.0561 *
Credit.Amount	0.0001798	1.733e-01	0.001	0.9998
Length.of.current.employment-7.yrs	0.3127022	4.587e-01	0.681	0.4940
Length.of.current.employment-1yr	0.8128789	3.894e-01	2.085	0.0398 **
Instalment.per.cent	0.3016791	1.260e-01	2.390	0.0180 **
Most.valuable.available.asset	0.2850267	1.425e-01	1.998	0.0469 *
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 336 degrees of freedom				
McFadden R-Squared: 0.2048, AIC: 352.5				
Number of Fisher Scoring iterations: 5				
Type II Analysis of Deviance Tests				

Figure 3. Report summary for the logistic regression-stepwise model

Based on the p-value (Figure 3) obtained for all the predictor variables used in the Logistic Regression-Stepwise Model, the following variables are significant:

- Account-Balance
  - Payment Status
  - Purpose
  - Credit Amount
  - Length of current employment
  - Instalment-per-cent
2. 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?



Model Comparison Report				
Fit and error measures				
Model	Accuracy	F1	AUC	
Stepwise	0.76	0.83	0.73	
Accuracy_Creditworthy: 0.80				
Accuracy_NonCreditworthy: 0.63				
<small>           Model: model names in the current comparison.            Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.            Accuracy (from name): accuracy of Class (class name), number of samples that are <b>correctly</b> predicted to be Class (class name) divided by number of samples predicted to be Class (class name).            AUC: area under the ROC curve, only available for two-class classification.            F1: F1 score: precision * recall / (precision + recall)         </small>				
Confusion matrix of Stepwise				
	Actual_Creditworthy		Actual_Non-Creditworthy	
Predicted_Creditworthy	32	23		
Predicted_Non-Creditworthy	11	22		

Figure 4. Model comparison report for the logistic regression-stepwise model

The overall percent accuracy is 76% with the confusion matrix listed above in Figure 4. The model is a bit bias towards the 'creditworthy' category since the accuracy of the creditworthy group is 80%, which is approximately 17% higher than the accuracy of the non-creditworthy 63%.

### Decision Tree Model:

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.

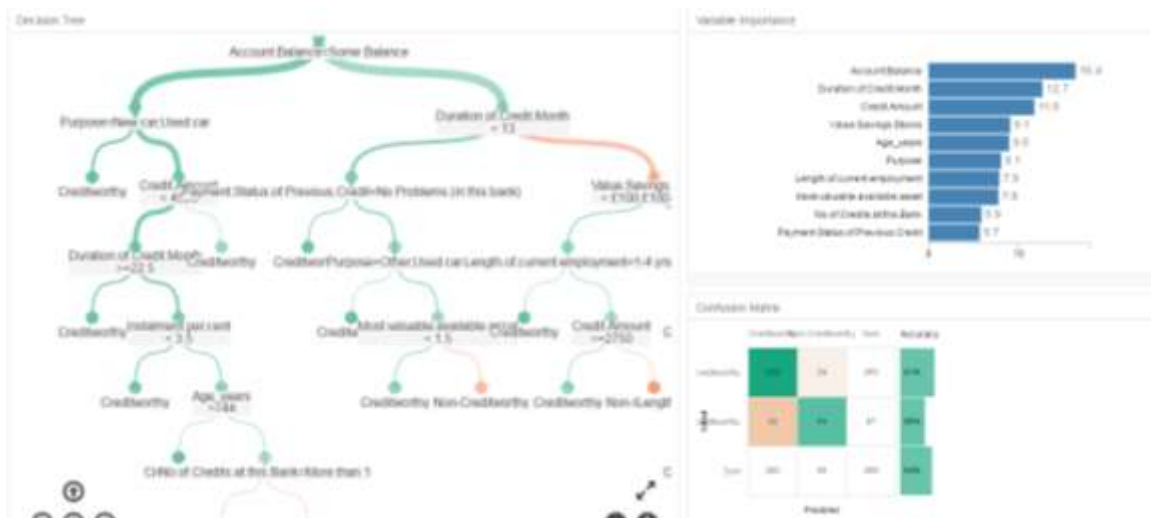


Figure 5. Report summary for the decision-tree model

The variance importance chart is shown above for the decision-tree model. The most important predictor variable is 'Account Balance' followed by 'Duration of credit month'.

2. 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?

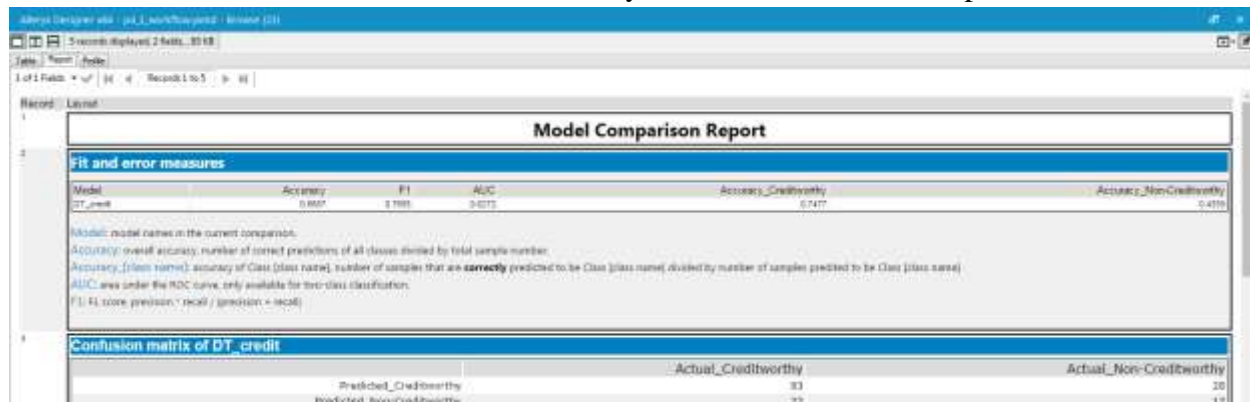


Figure 6. Model comparison report for the decision-tree model

The overall percent accuracy is 66.7% with the confusion matrix is shown above. The model seems to bias towards the 'creditworthy' category since the accuracy of creditworthy is 75%, which is much higher than the accuracy of the non-creditworthy group, 44%.

### Forest Model:

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.

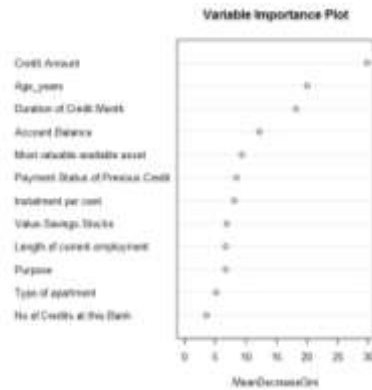


Figure 7. Variance importance plot for the forest model

Based on Figure 6, the most important variable is Credit-Amount, followed by Age\_years.

2. 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?

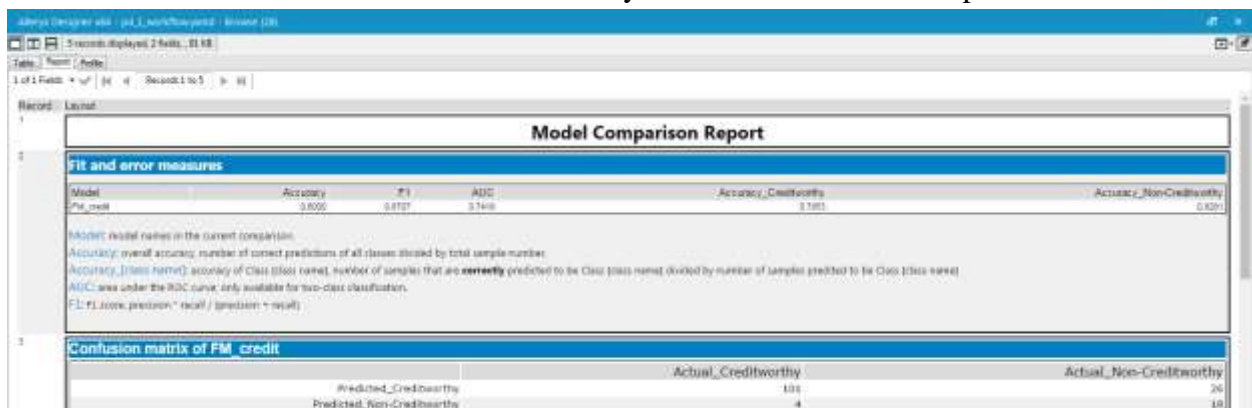


Figure 8. Model Comparison Report for the forest model

The overall accuracy is 80% with the confusion matrix is shown above. The model is not biased since the accuracies of creditworthy and non-creditworthy are comparable, 80% and 83% respectively.

### **Boosted Model:**

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.

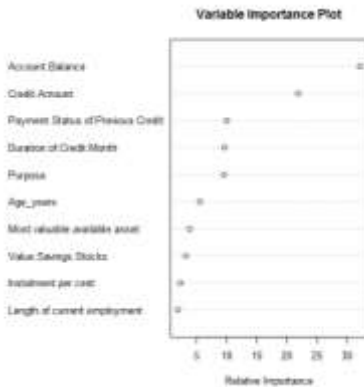


Figure 9. Variance importance plot for the boosted model

The most important variable is the Account-Balance, followed by Credit-Amount.

2. 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?

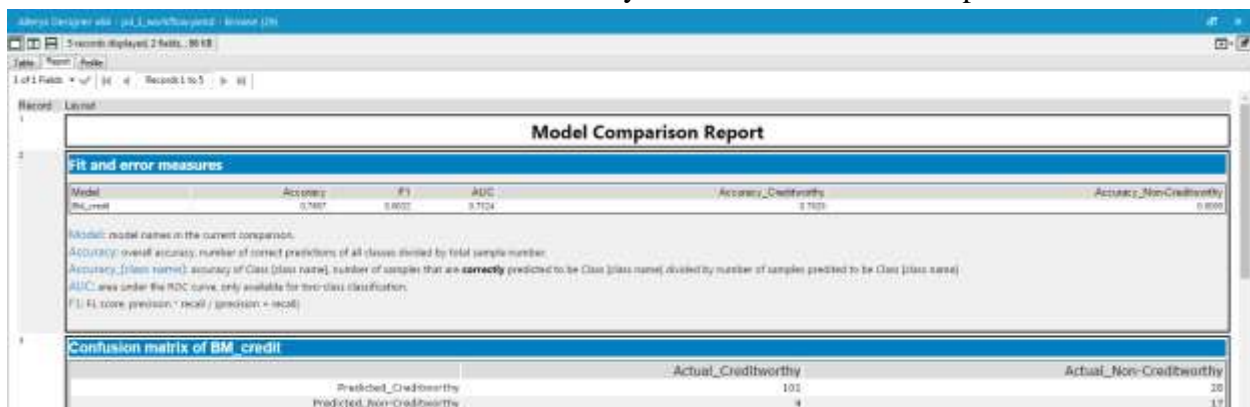


Figure 10. Model Comparison Report for the boosted model

The overall accuracy of the boosted model is 78.7% with the confusion matrix is shown above. The model is not biased since the accuracies of creditworthy and non-creditworthy are comparable, 78% and 81% respectively.

#### Step 4: Write-up

*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"*

1. Which model did you choose to use? Please justify your decision using only the following techniques:
  - a. Overall Accuracy against your Validation set

- b. Accuracies within “Creditworthy” and “Non-Creditworthy” segments
- c. ROC graph
- d. Bias in the Confusion Matrices

I chose forest model since it offers the highest accuracy (80%) in comparison to the other models. Moreover, the forest model does not seem to be bias towards a single category. The accuracies obtained for both creditworthy and non-creditworthy groups are among the highest of all accuracies obtained by the different models (Figure 11). This fact is really important since that means the model will not overlook the data.

Based on the ROC graph of all models (Figure 12), the forest model seems to reach the true positive rate at the fastest rate. This confirms that this model is more appropriate in use to predict the category for the loan applicant than the others.

Model Comparison Report					
Fit and error measures					
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_NonCreditworthy
logit_model	0.700	0.694	0.709	0.693	0.699
DT_model	0.697	0.780	0.670	0.707	0.693
FM_model	0.693	0.677	0.749	0.700	0.691
RF_model	0.797	0.800	0.794	0.793	0.800

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), number of samples that are **correctly** predicted to be Class (class name) divided by number of samples predicted to be Class (class name).  
AUC: area under the ROC curve, only available for two-class classification.  
F1: F1 score, precision \* recall / (precision + recall).

Confusion matrix of SM_credit		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	28
Predicted_Non-Creditworthy	4	17

Confusion matrix of DT_credit		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	69	28
Predicted_Non-Creditworthy	22	17

Confusion matrix of FM_credit		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	28
Predicted_Non-Creditworthy	4	19

Confusion matrix of Stgp_credit		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	90	21
Predicted_Non-Creditworthy	63	20

Figure 11. Model comparison report for all models

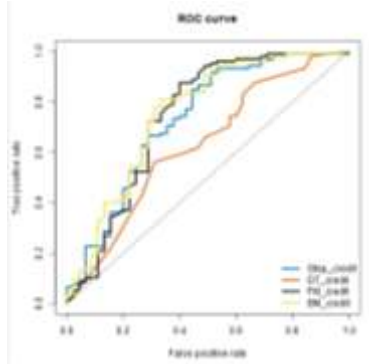


Figure 12. ROC graphs for all models

## 2. How many individuals are creditworthy?

There are 415 individuals who are creditworthy to get a loan. This result is based on the score of creditworthy obtained from using the forest model for the new customers.