**Project 3: Predicting Default Risk**

## **Step 1: Business and Data Understanding**

### **Key Decisions:**

1. What decisions needs to be made?

*Due to a financial scandal that hit a competitive bank last week, our bank suddenly has an influx of nearly 500 new customers applying for loans for our bank instead of the other bank in our city. As a loan officer at a young and small bank (been in operations for two years), I need to come up with an efficient solution to classify new customers on whether they can be approved for a loan or not. I'll use a series of classification models to figure out the best model and provide a list of creditworthy customers to bank manager.*

2. What data is needed to inform those decisions?

*We have two datasets, one for current customers data stored in ‘credit-data-training.xlsx’ file and another for new customers data stored in ‘customers-to-score.xlsx’ file.*

*Variables for the two datasets*

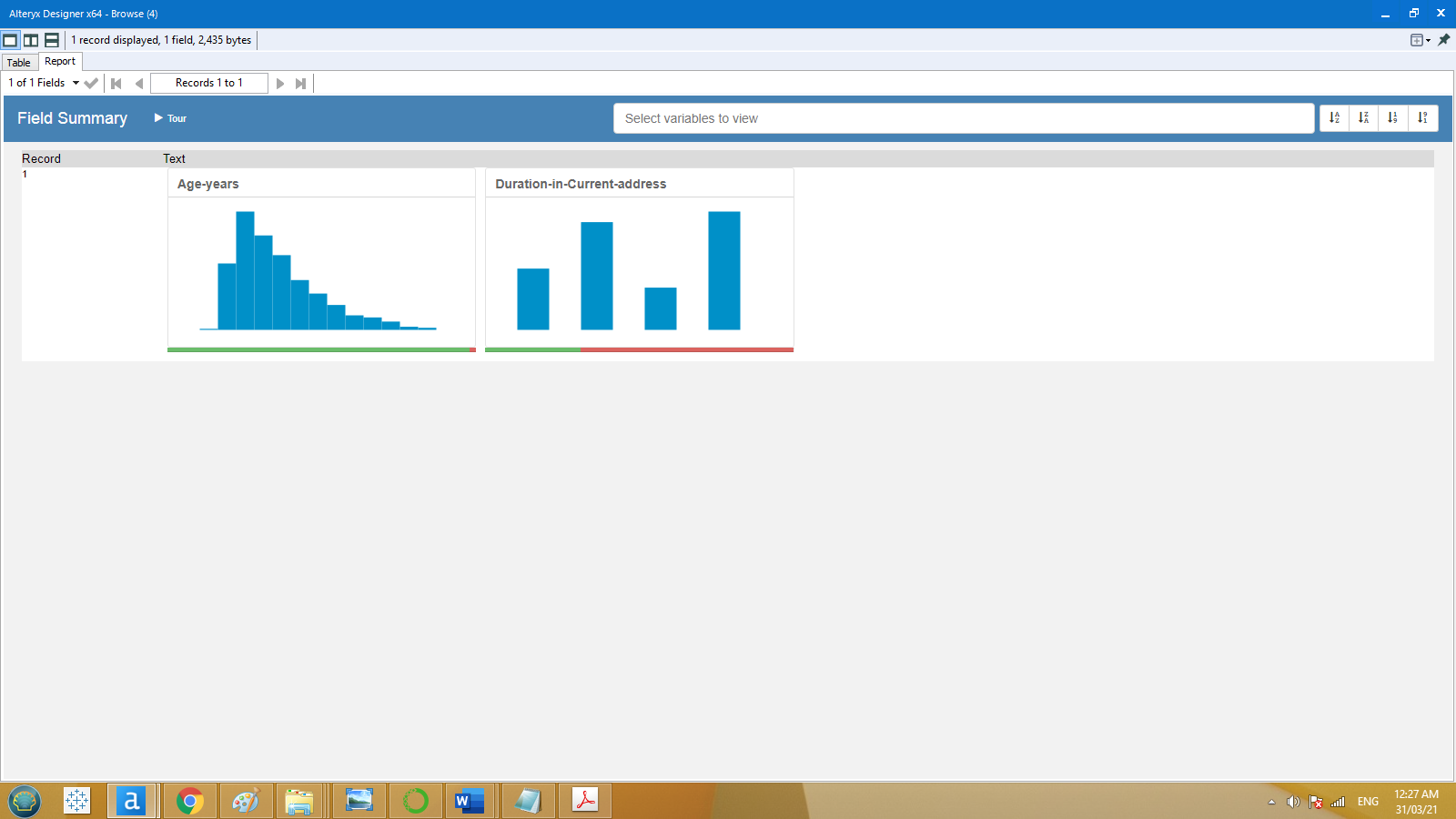
|  |  |  |
| --- | --- | --- |
|  | *‘credit-data-training.xlsx’* | *‘customers-to-score.xlsx’* |
| *Joint Fields* | *Account-Balance* | *Account-Balance* |
| *Duration-of-Credit-Month* | *Duration-of-Credit-Month* |
| *Payment-Status-of-Previous-Credit* | *Payment-Status-of-Previous-Credit* |
| *Purpose* | *Purpose* |
| *Credit-Amount* | *Credit-Amount* |
| *Value-Savings-Stocks* | *Value-Savings-Stocks* |
| *Length-of-current-employment* | *Length-of-current-employment* |
| *Instalment-per-cent* | *Instalment-per-cent* |
| *Guarantors* | *Guarantors* |
| *Duration-in-Current-address* | *Duration-in-Current-address* |
| *Most-valuable-available-asset* | *Most-valuable-available-asset* |
| *Age-years* | *Age-years* |
| *Concurrent-Credits* | *Concurrent-Credits* |
| *Type-of-apartment* | *Type-of-apartment* |
| *No-of-Credits-at-this-Bank* | *No-of-Credits-at-this-Bank* |
| *Occupation* | *Occupation* |
| *No-of-dependents* | *No-of-dependents* |
| *Telephone* | *Telephone* |
| *Foreign-Worker* | *Foreign-Worker* |
|  | *Credit-Application-Result* |  |

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

*Since we are interested in answering the question whether a customer is qualified to be approved for a loan or not, such problem needs a binary model building to answer it. The target field is ‘Credit-Application-Result’ contains two possible values (Creditworthy/Non-* *Creditworthy). I will compare 4 different binary classification models (Logistic, Decision Tree, Random Forest, and Boosted) to choose the one that best fit data.*

## **Step 2: Data Preparation**

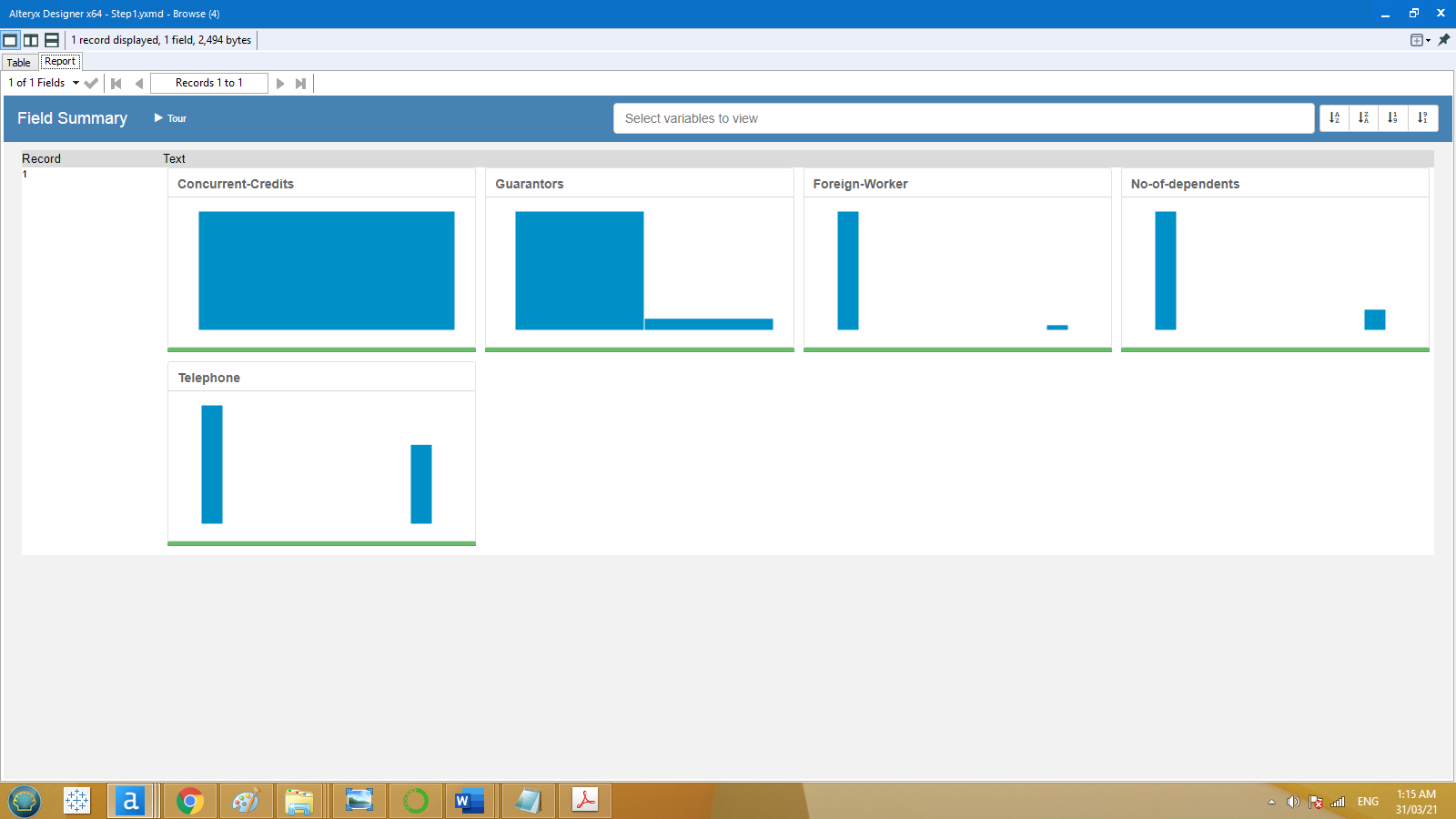
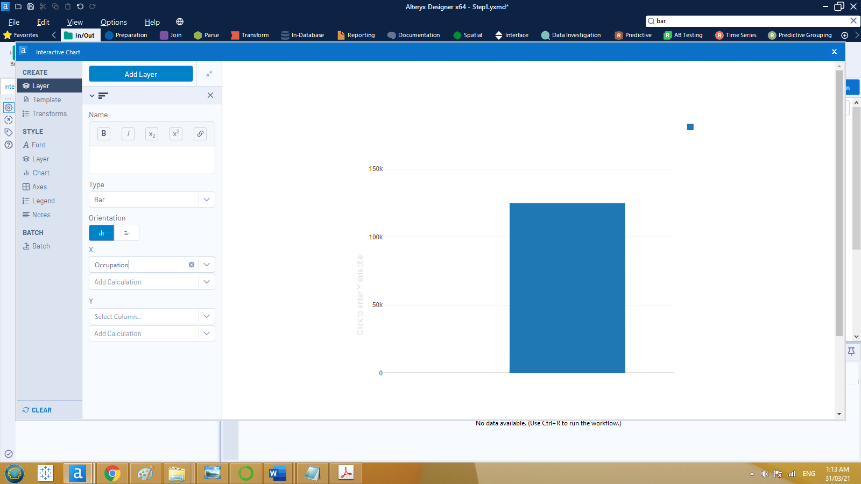
1. Fields with Missing Data



*The above visualization identifies missing data with two fields:*

* *‘Duration-in-Current-address’ field has about 69% of its data are missing, so with a high missing data we should remove this field forever.*
* *‘Age-years’ field has about 2% of its data are missing, by taking into consideration the logical impact of age as a variable in our decision, we should impute the missing ages by replacing them with age median.*

2. Fields with Low Variability

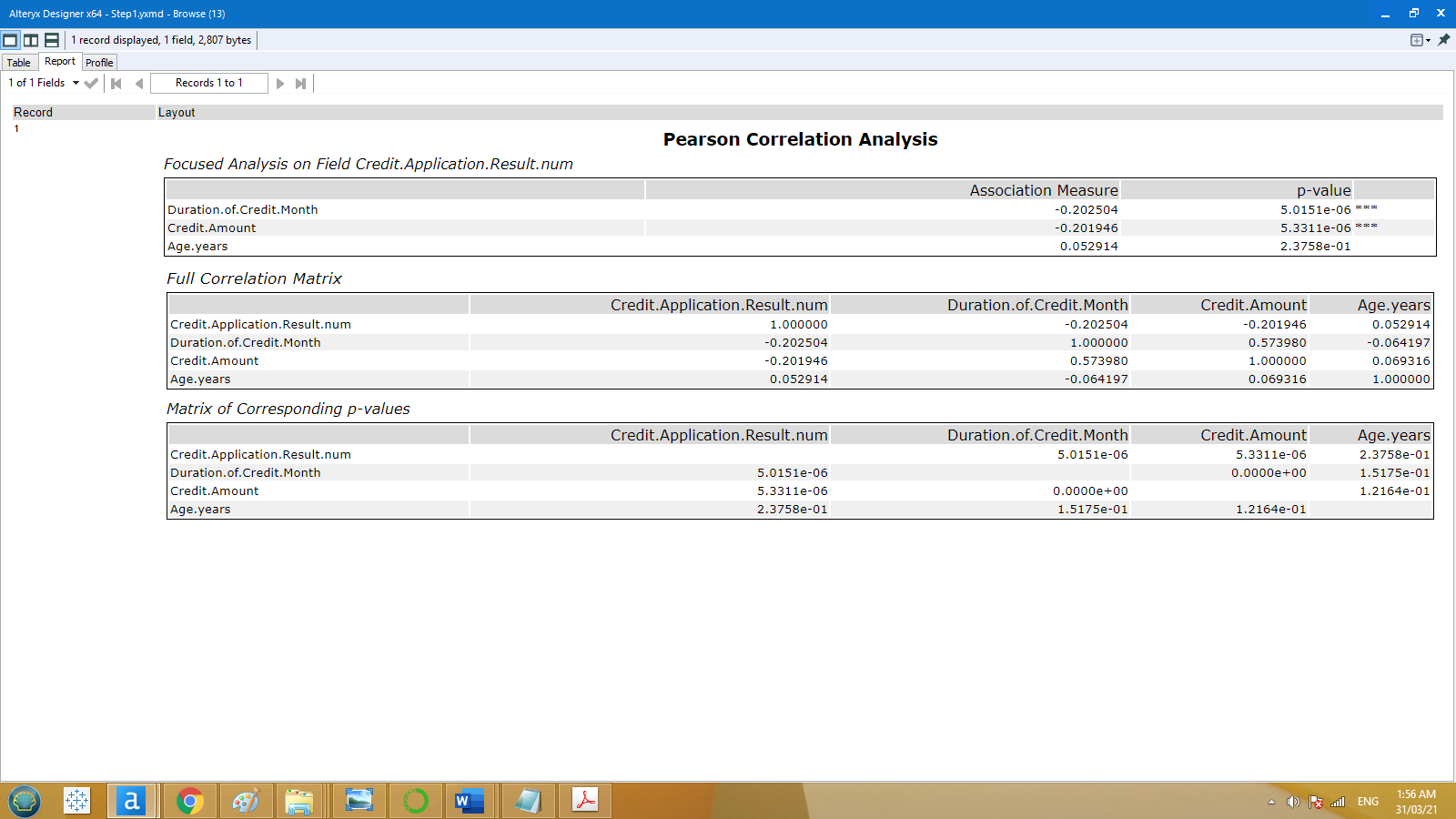


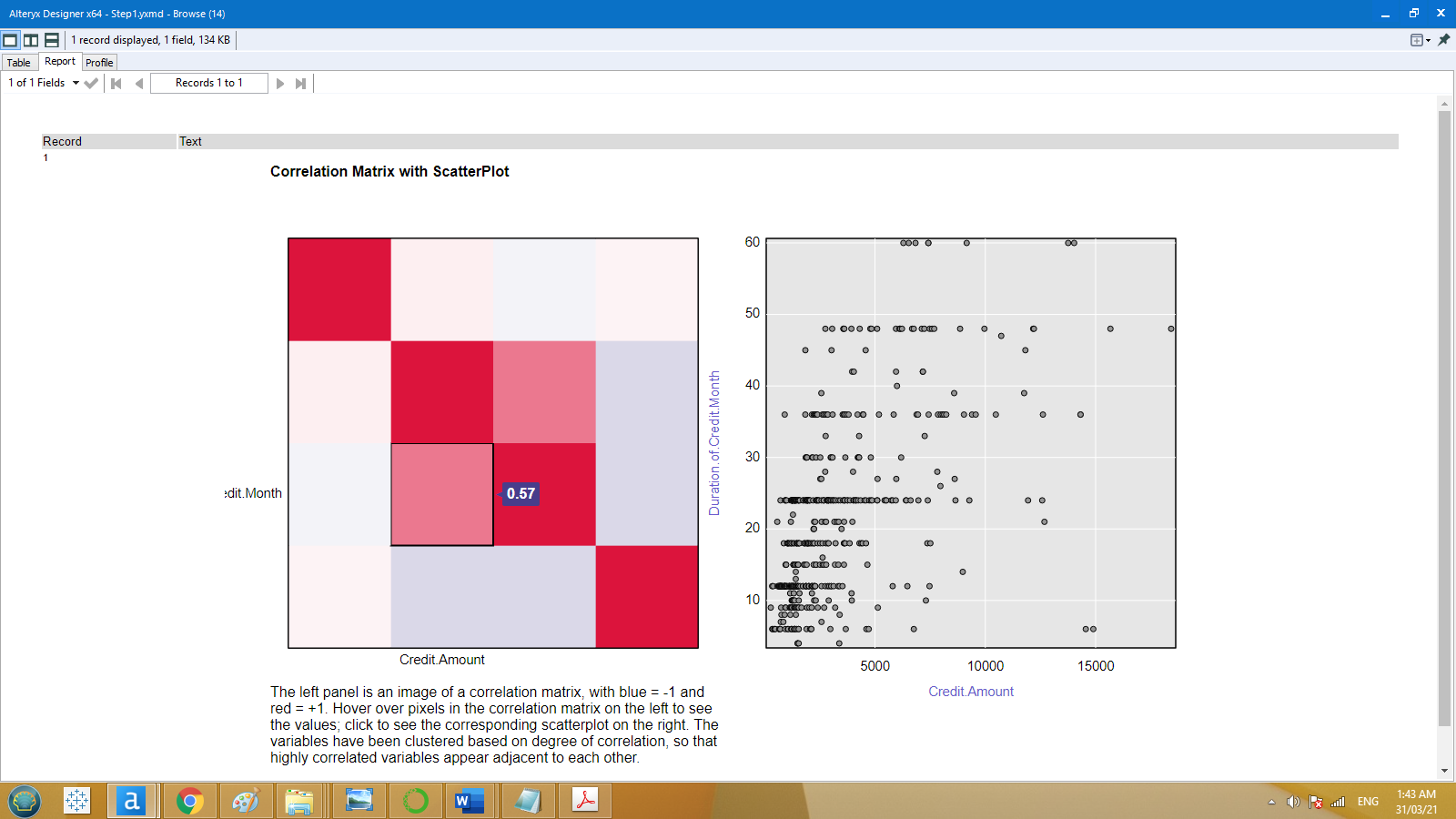
Occupation

*The above visualization identifies low variability with 6 fields in which we should remove all of them.*

3. Multicollinearity Identification

*We need to check whether any group of the possible predictors are highly correlated or not. The correlation plot matrix between all possible predictor variable is given below*





*Credit.Amount*

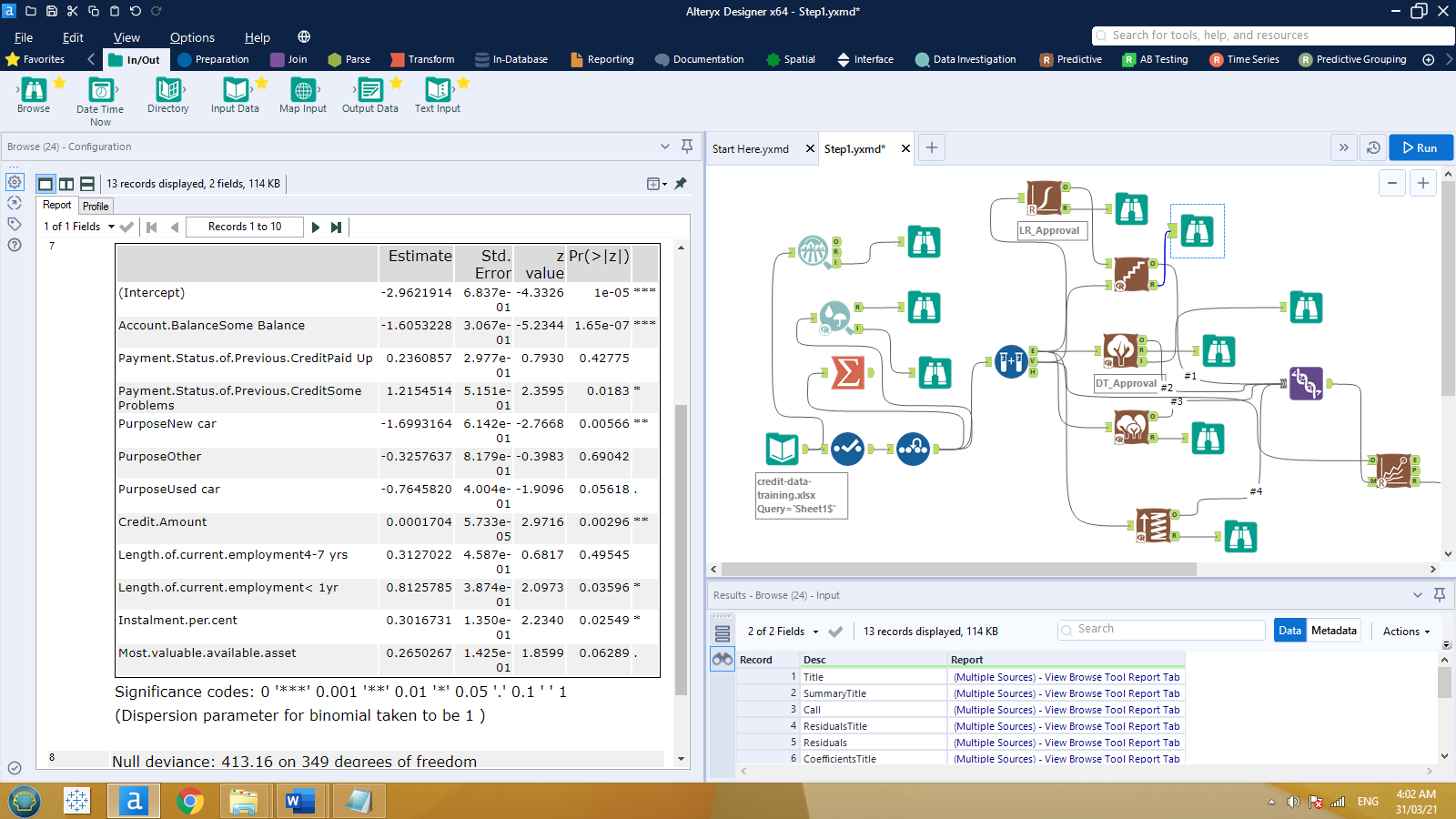
*It is clear that there isn’t a high correlation between any two possible predictor variables.*

## **Step 3: Training Classification Models**

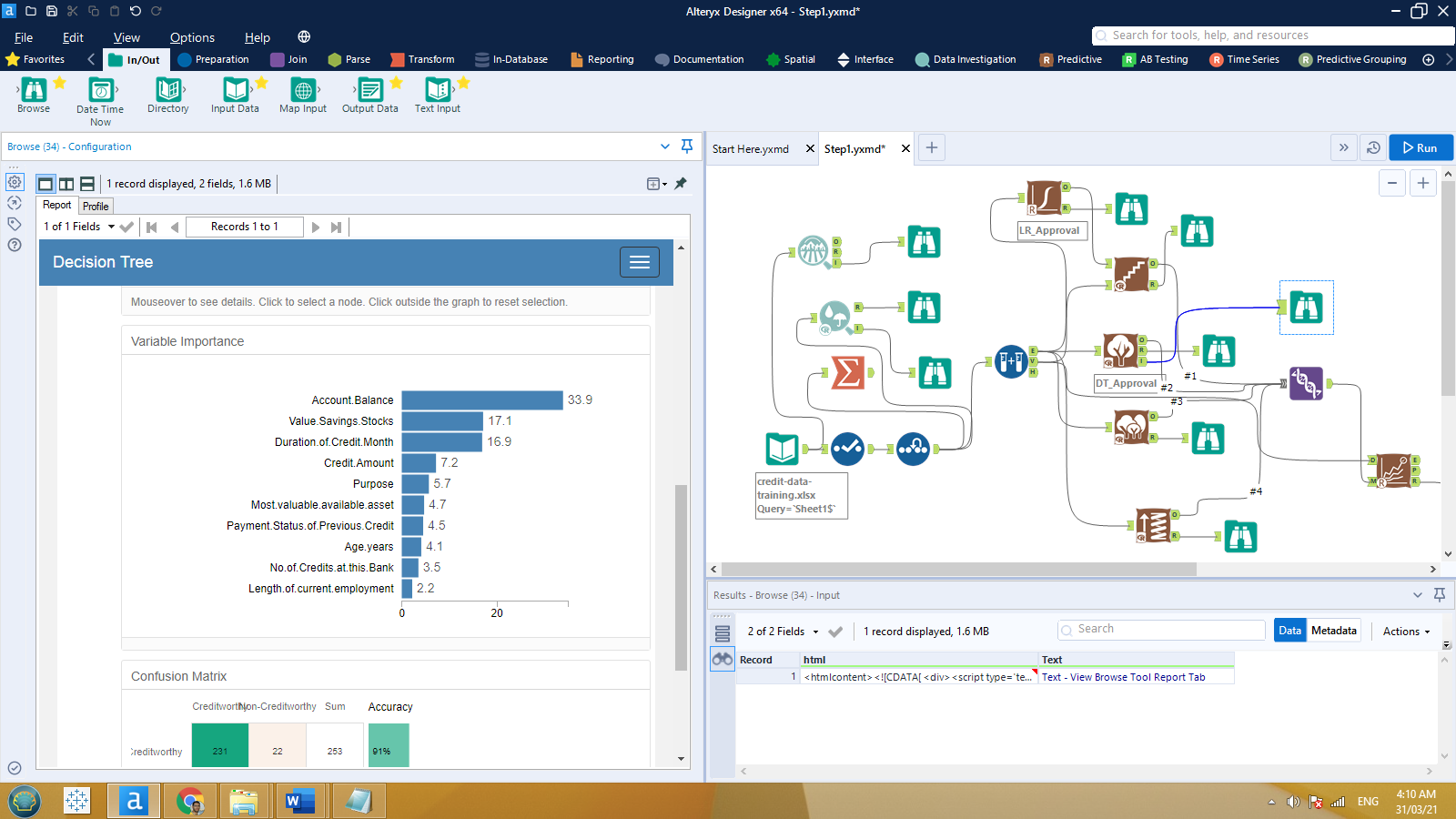
*First, I have randomly split dataset into two subsets (70% Estimation and 30% Validation). Then I have trained the 4 models (Logistic, Decision Tree, Random Forest, and Boosted) on Estimation group. Finally, I have used the validation group to test each model accuracy.*

1. Which predictor variables are significant or the most important?

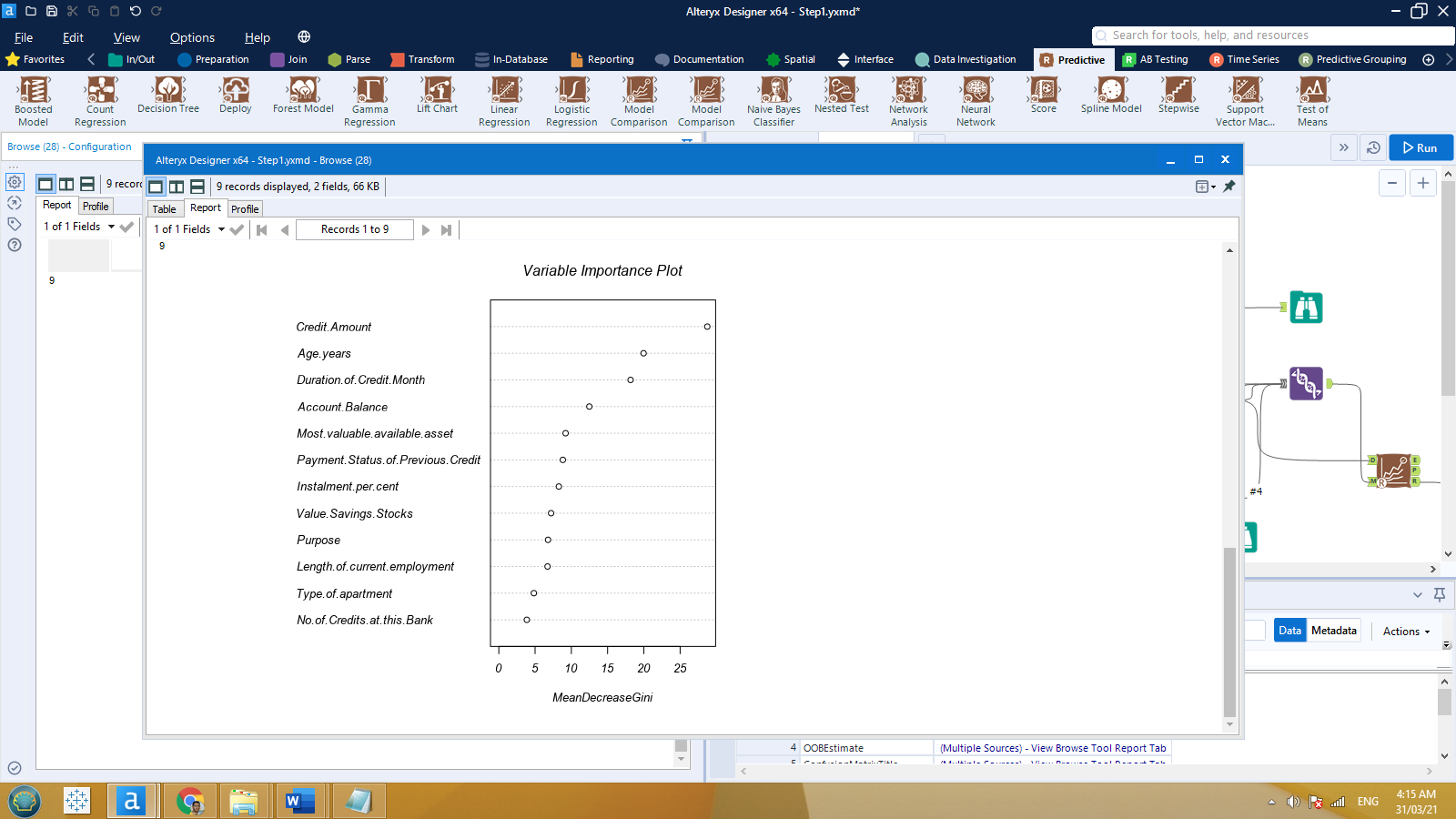
* *with* ***Logistic Model****: ‘Account Balance’, ‘Credit Amount’, and ‘Purpose’ are the top significant variables descendingly.*



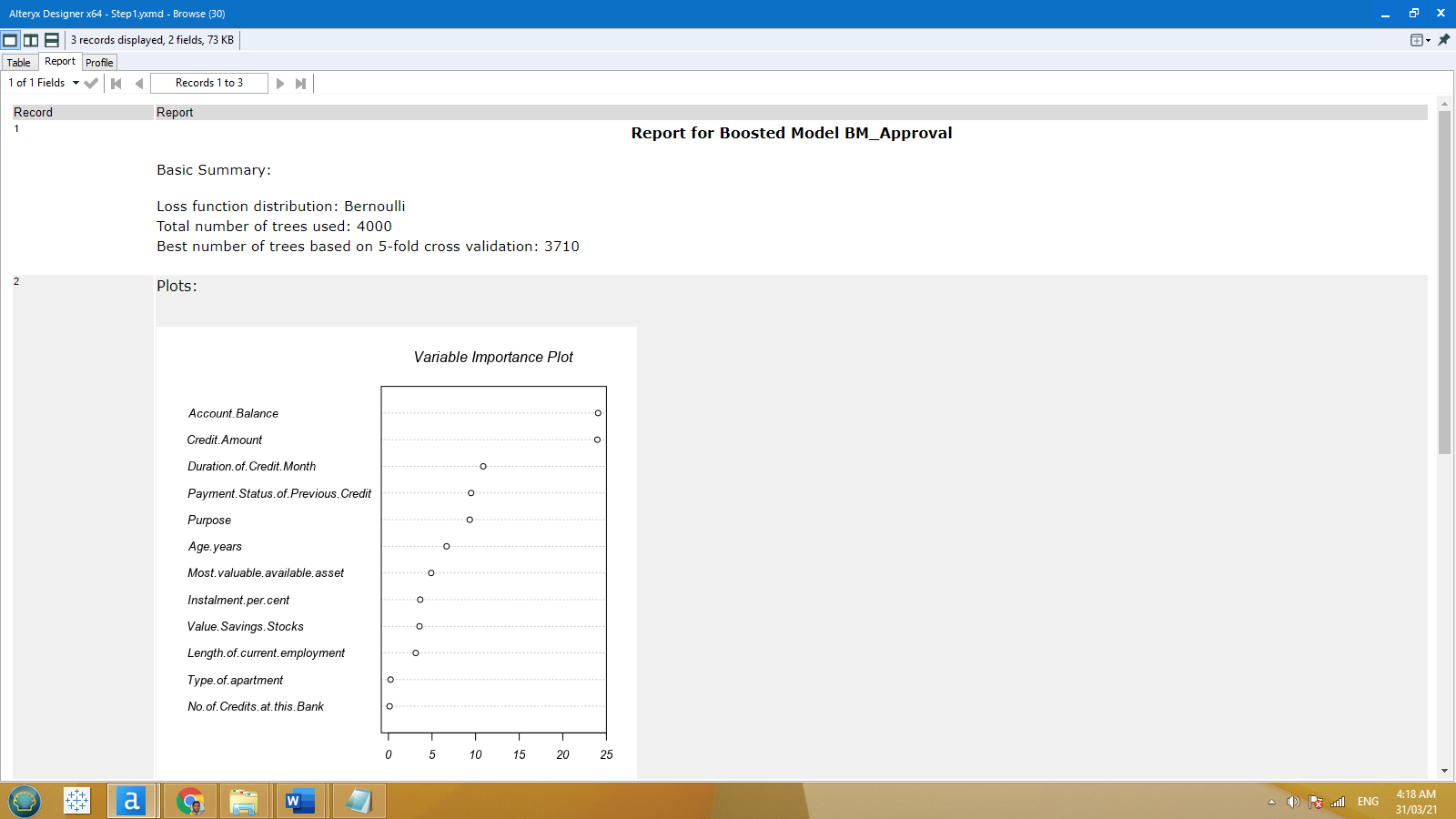
* *with* ***Decision Tree****: ‘Account Balance’, ‘Value Savings Stocks’, and ‘Duration of Credit Month’ are the top important variables descendingly.*



* *with* ***Forest Model****: ‘Credit Amount’, ‘Age Years’, and ‘Duration of Credit Month’ are the top important variables descendingly.*

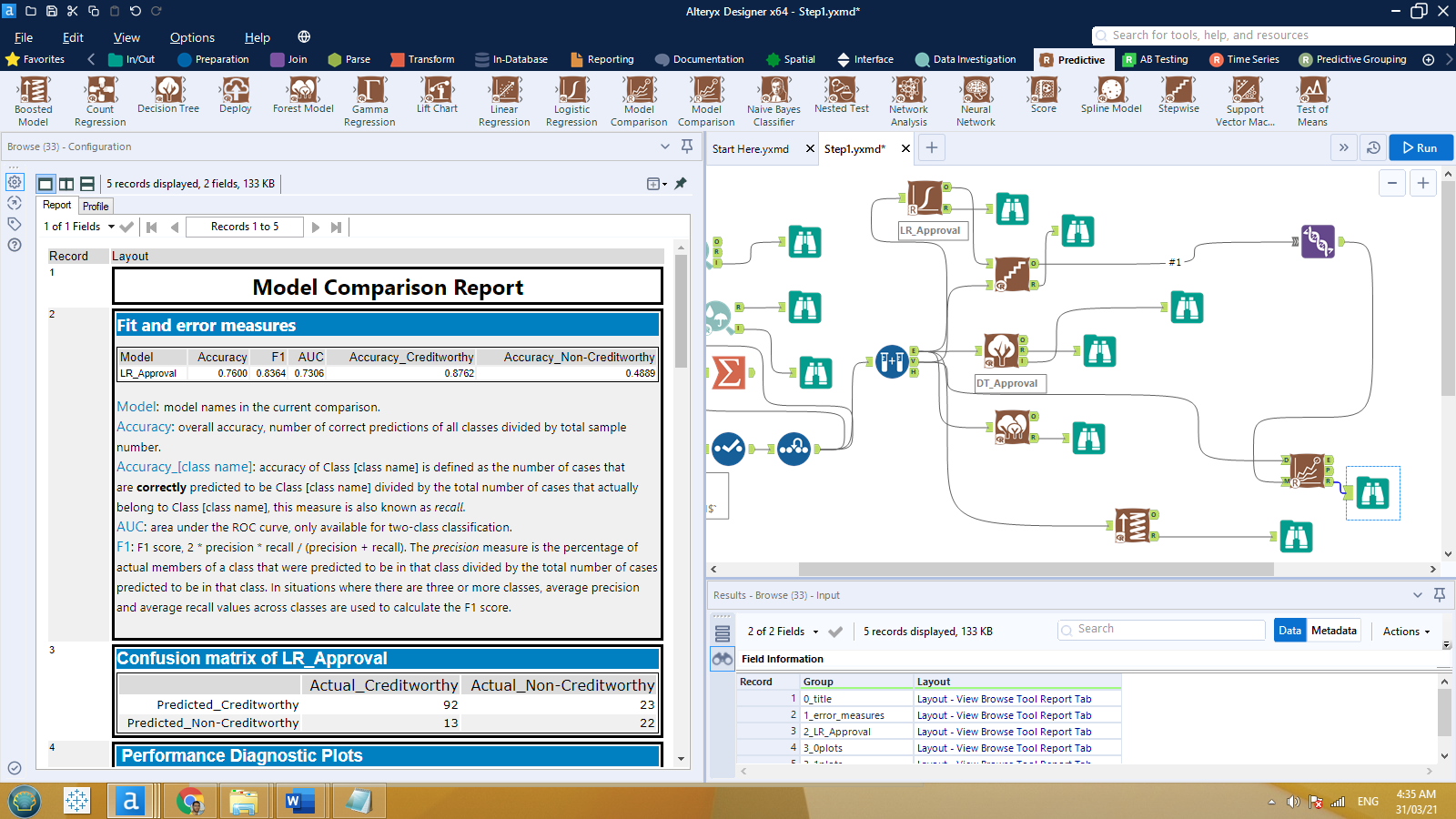


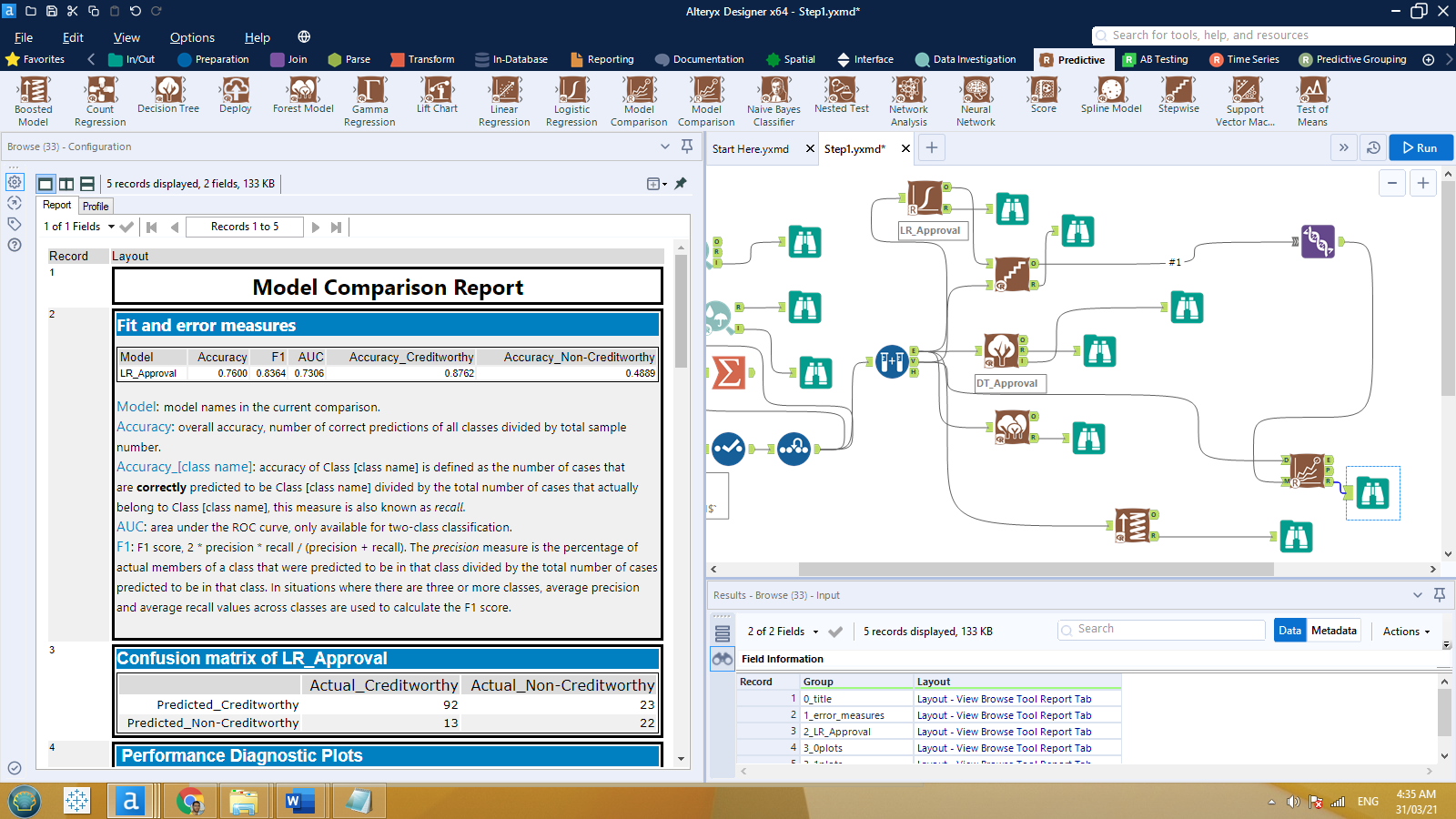
* *with* ***Boosted Model****: ‘Account Balance’, ‘Credit Amount’, and ‘Duration of Credit Month’ are the top important variables descendingly.*



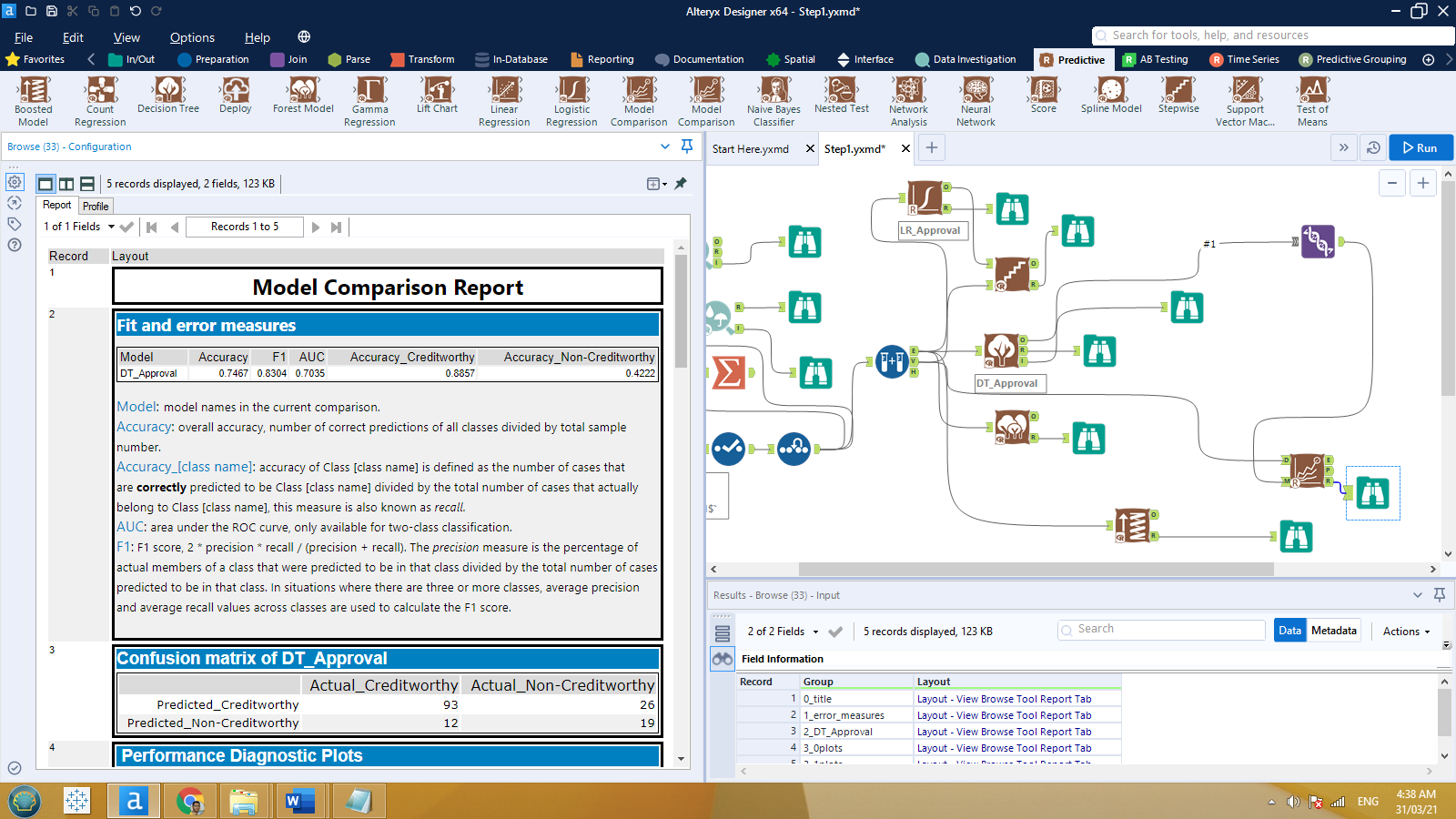
2. What was the overall percent accuracy? Are there any bias seen in the model’s predictions?

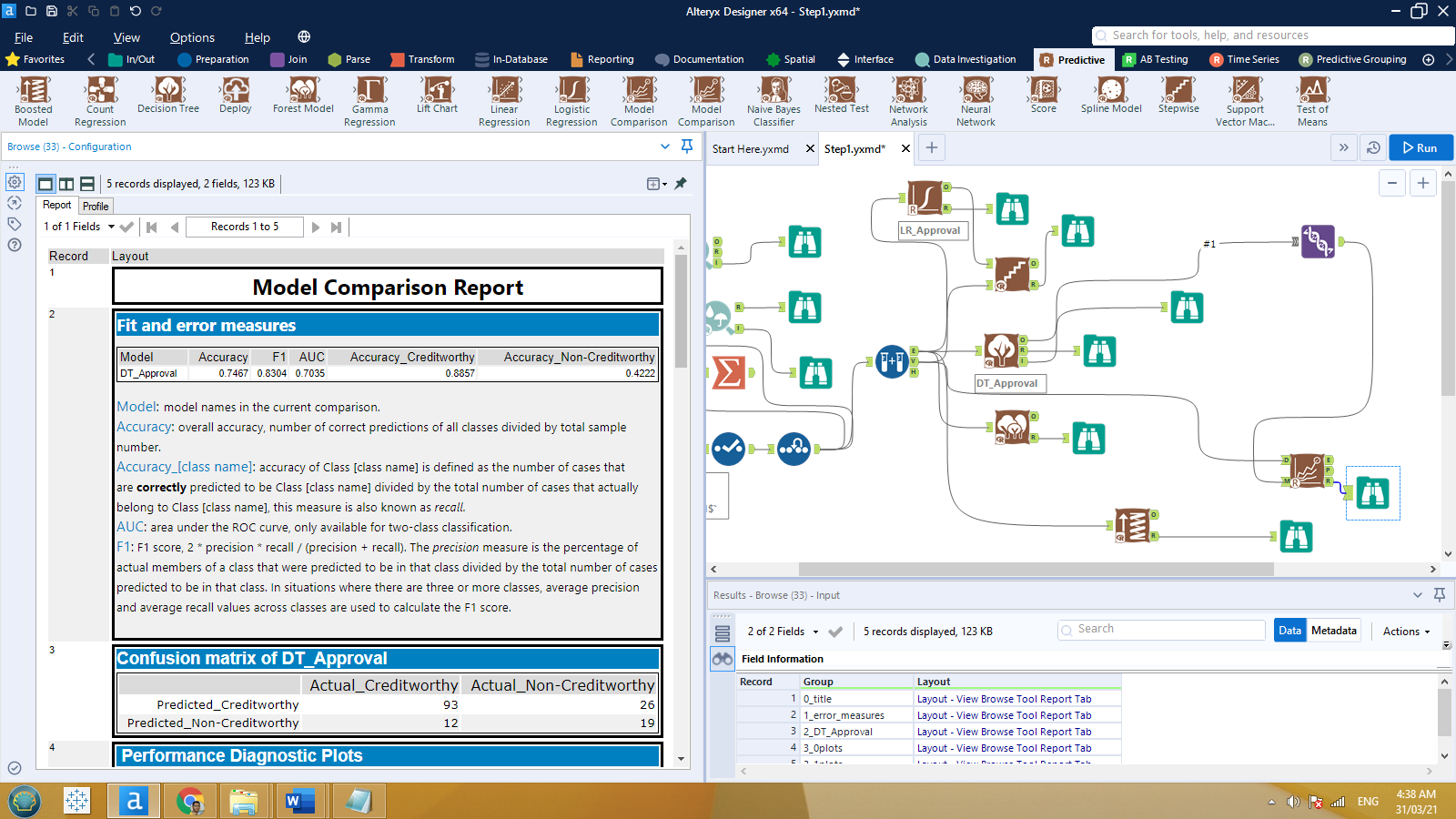
* *with* ***Logistic Model****: the overall accuracy is 76% while the accuracy of predicting Creditwothy is 88% and the accuracy of predicting Non-Creditwothy is 49%. In such case, we can say that this model is biased to Creditworthy than Non-Creditwothy.*



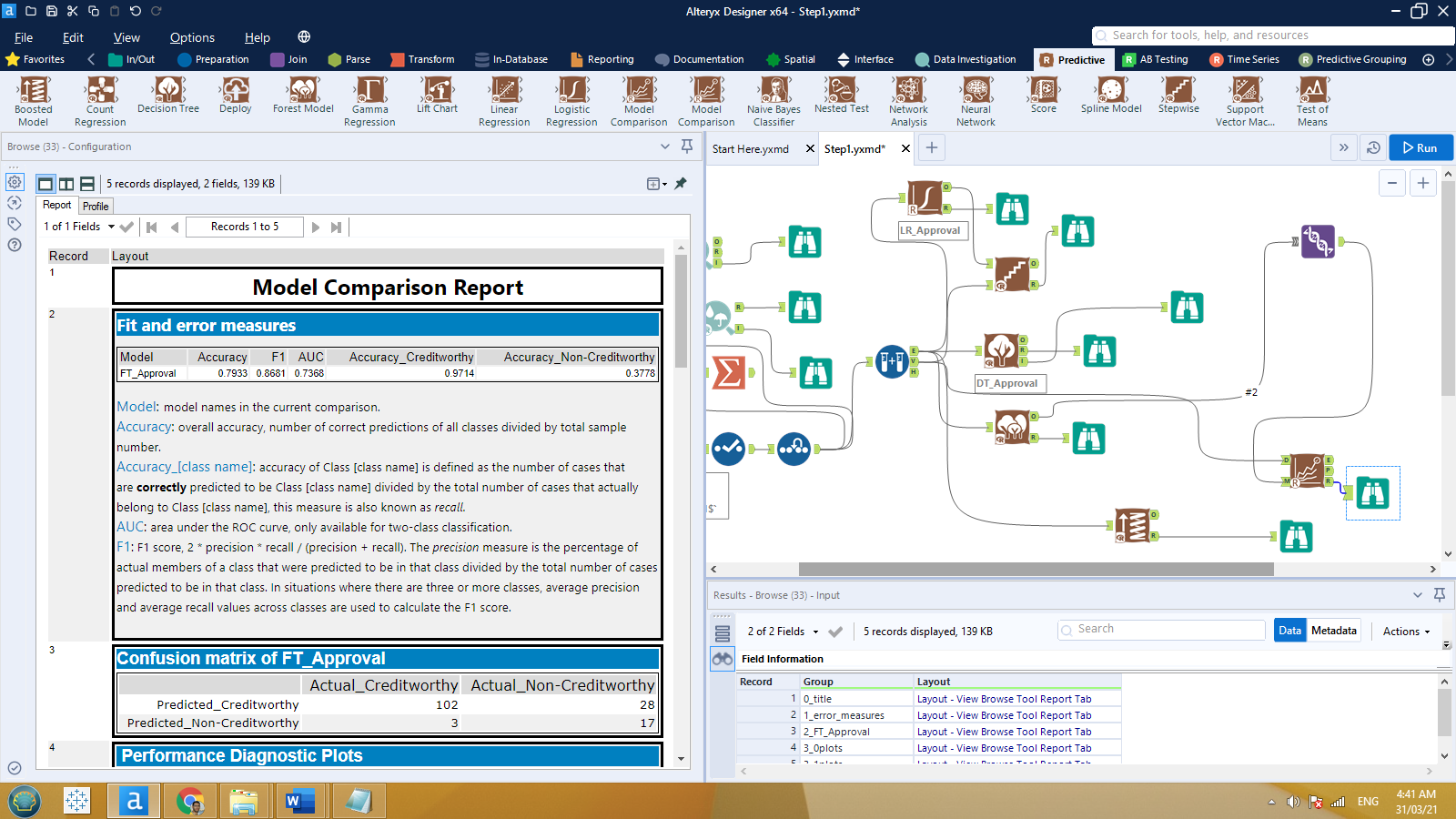


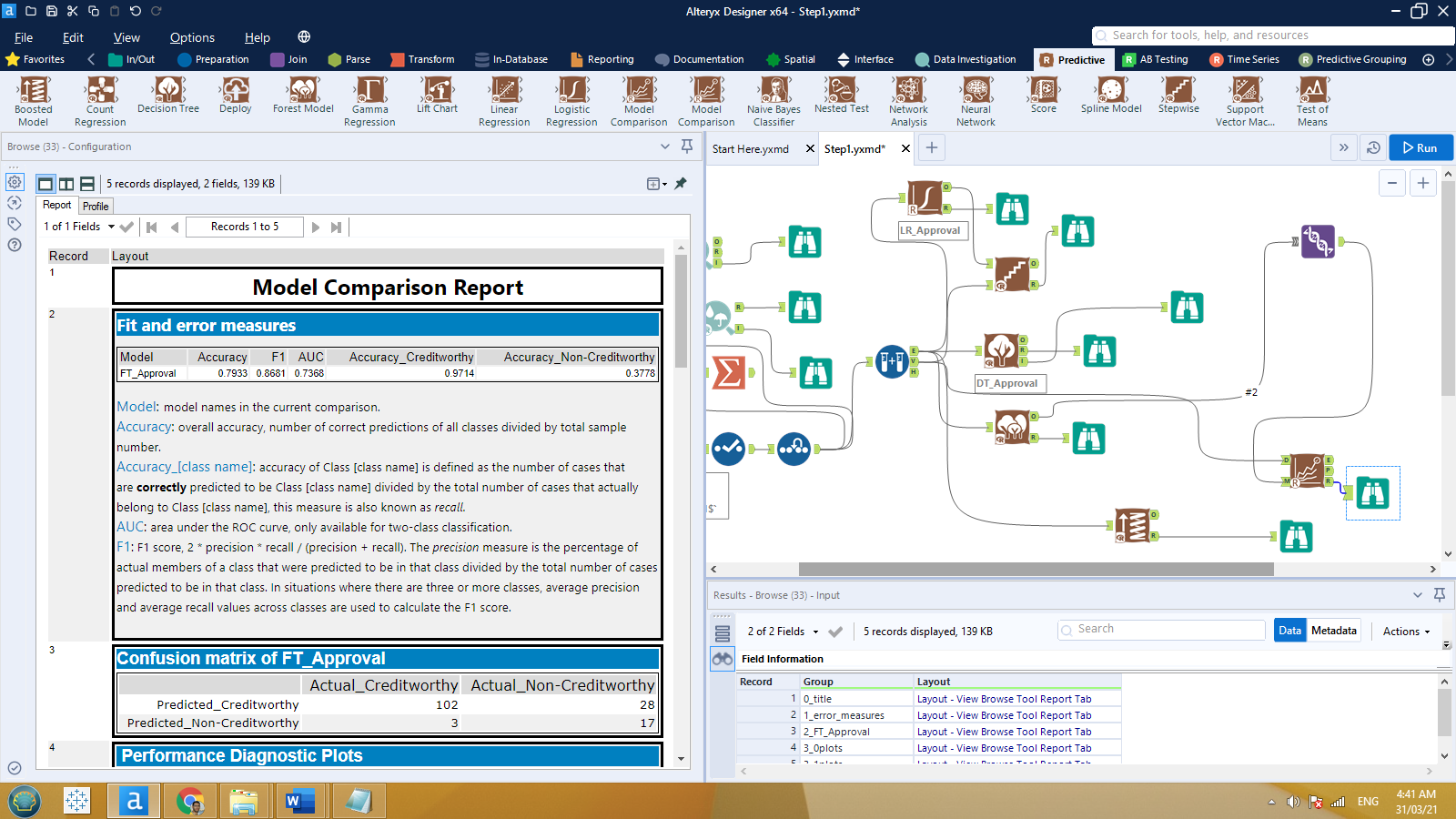
* *with* ***Decision Tree****: the overall accuracy is 75% while the accuracy of predicting Creditwothy is 89% and the accuracy of predicting Non-Creditwothy is 42%. In such case, we can say that this model is biased to Creditworthy than Non-Creditwothy.*



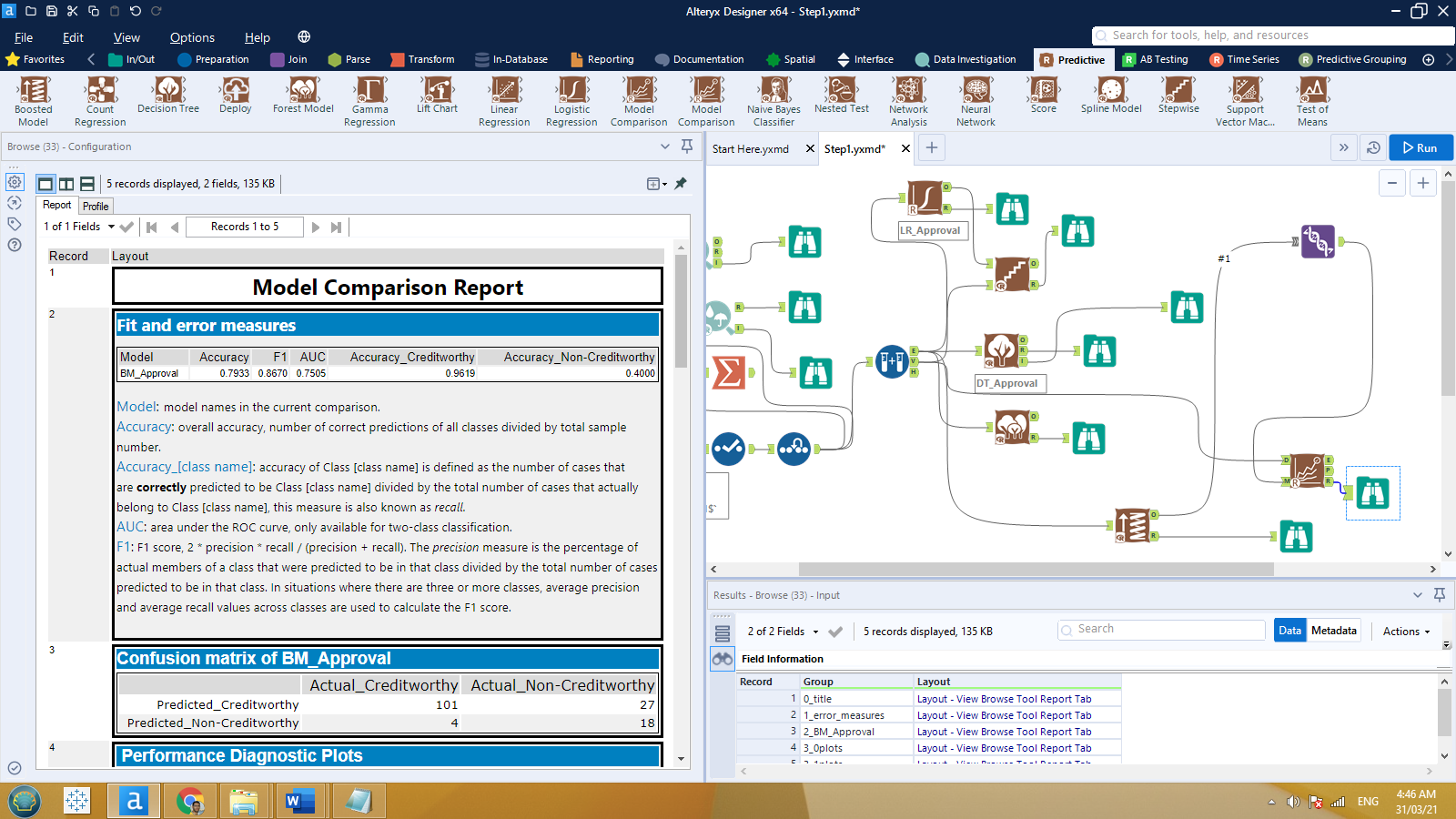


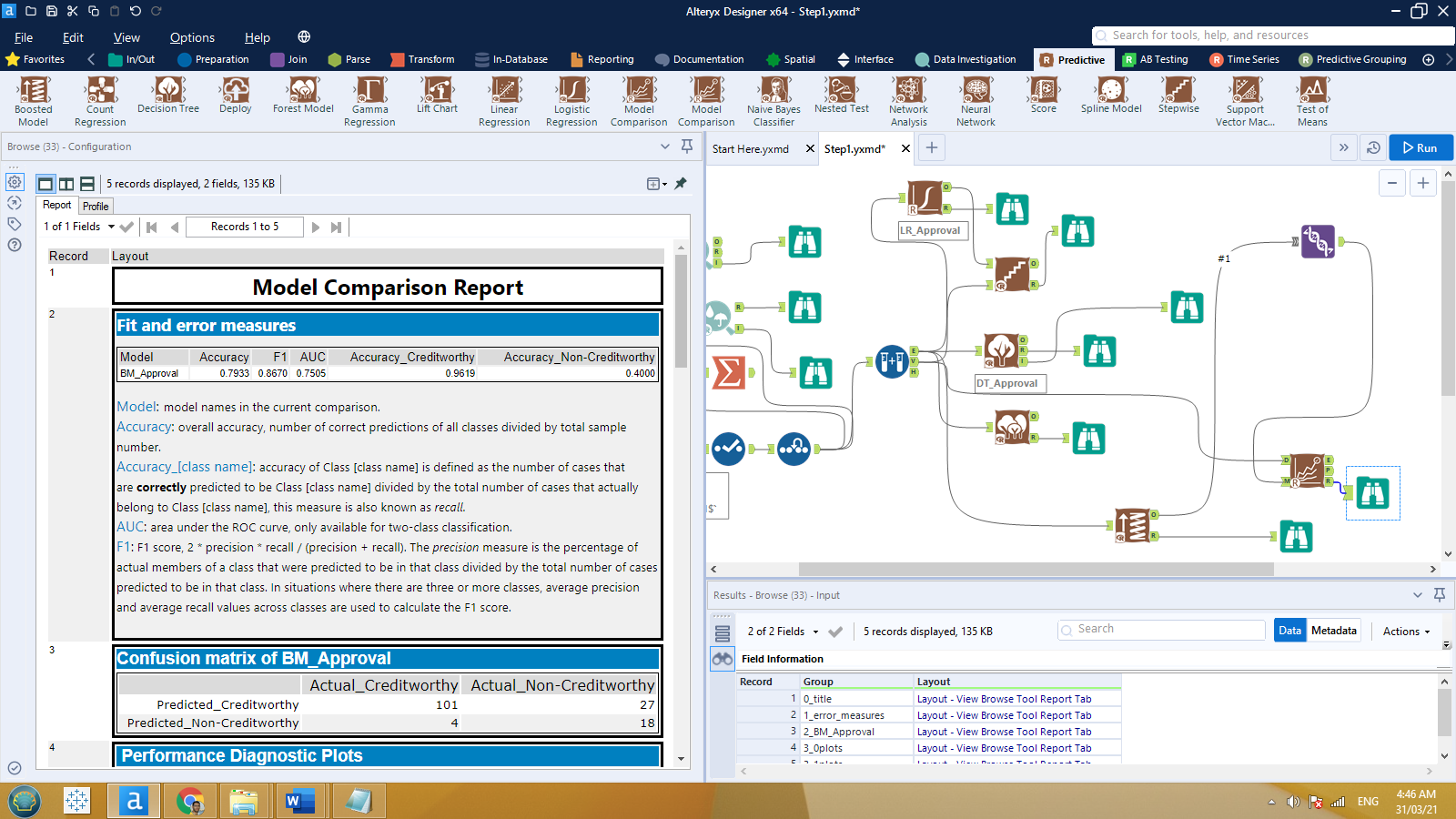
* *with* ***Forest Model****: the overall accuracy is 79% while the accuracy of predicting Creditwothy is 97% and the accuracy of predicting Non-Creditwothy is 38%. In such case, we can say that this model is mostly biased to Creditworthy than Non-Creditwothy.*





* *with* ***Boosted Model****: the overall accuracy is 79% while the accuracy of predicting Creditwothy is 96% and the accuracy of predicting Non-Creditwothy is 40%. In such case, we can say that this model is mostly biased to Creditworthy than Non-Creditwothy.*

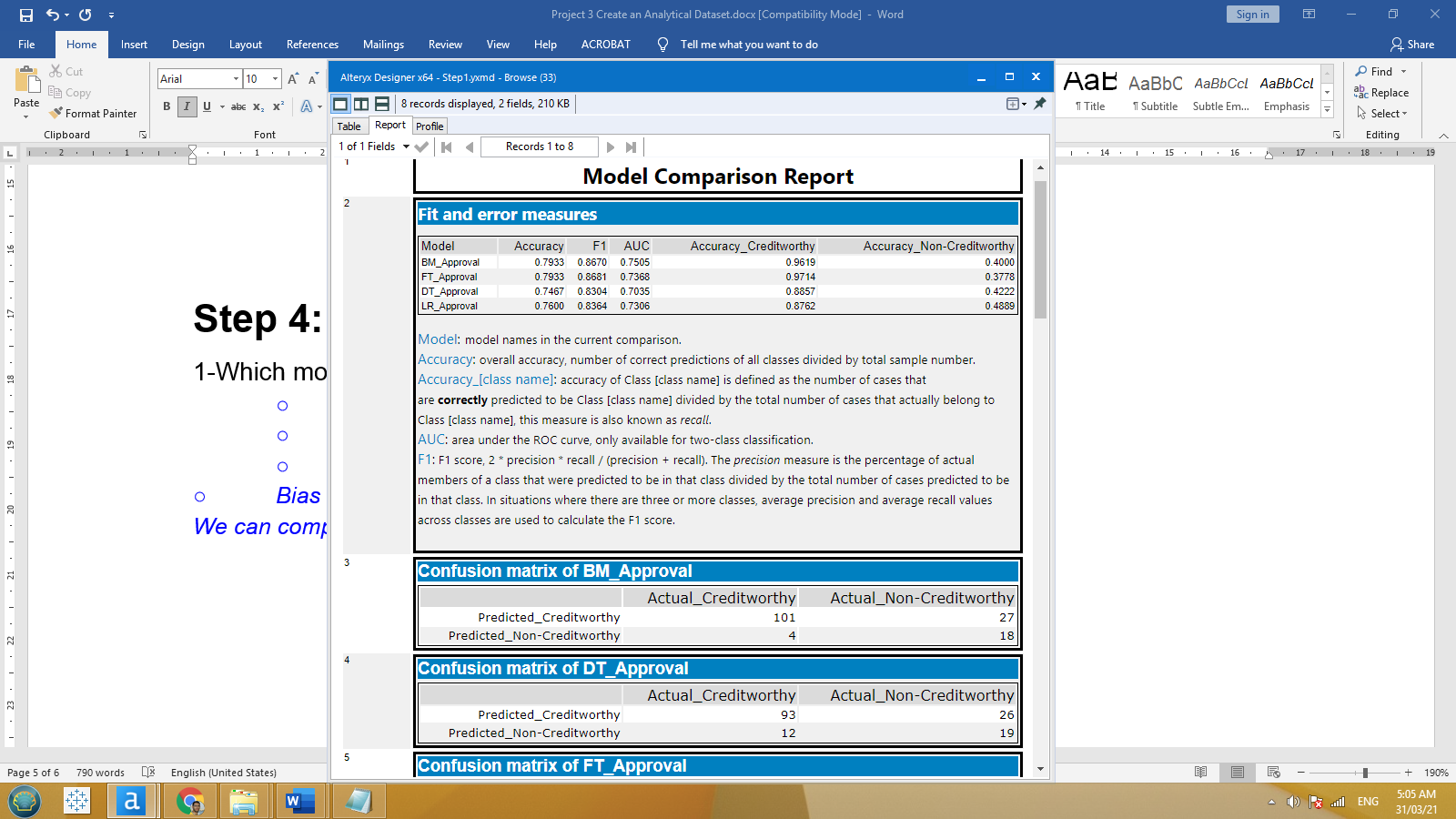


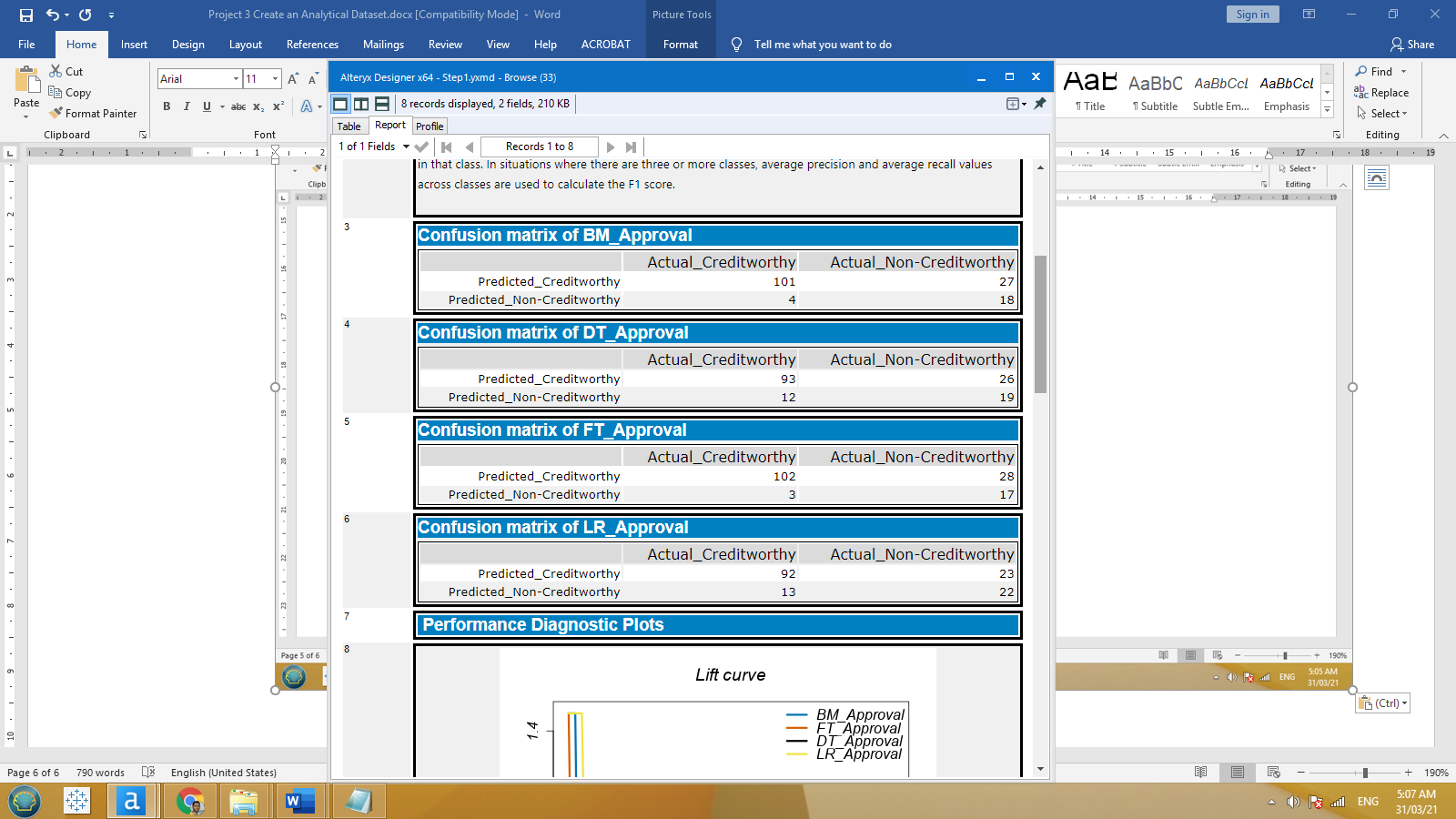


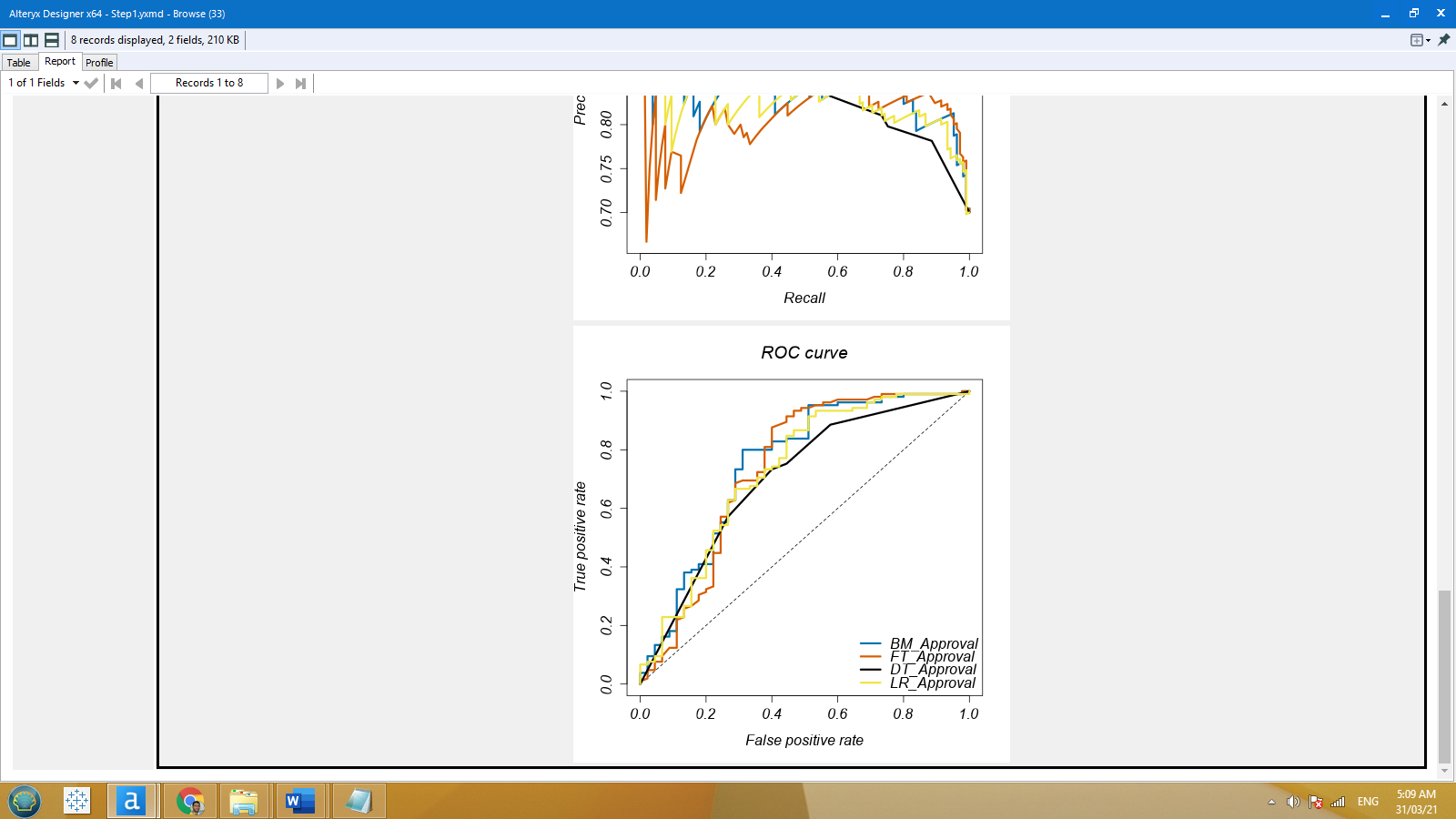
## **Step 4: Writeup**

1-Which model did I choose to use?

*We can compare all 4 models side by side by looking to the following*







*Taking into consideration the overall accuracy, both Forest and Boosted models have the highest overall accuracy of 79.33%, as well we can see that Forest model has the highest Accuracy of predicting Creditworthy at 97.14%, while Boosted model is more accurate in predicting Non-Creditworthy than Forest do.*

*Also using ROC graph, we can say that Forest model has the highest value with top true positive side of the graph.*

*Since we are interested in predicting Creditworthy we should choose Forest as the best fit model.*

2-How many individuals are creditworthy?

*Once we have come up to the best fit model, we could apply that model with our new dataset and the results as follows:*

|  |  |
| --- | --- |
| Sum\_Score\_Creditworthy | Sum\_Score\_Non-Creditworthy |
| 408 | 92 |
| 81.6% | 18.4% |

*408 customers will be approved and 92 customers will be disapproved*