Project: Creditworthiness

# Step 1: Business and Data Understanding

## Key Decisions:

Answer these questions

* What decisions needs to be made?

First, we need to decide which of the variables in the dataset will stay and which will be removed based on the p-values and the correlation between the variables, then we need to decide how to impute the missing values in the Age-Years attribute, with the data cleaned, we need to build the models and verify which performed better in the validation set based on what the manager cares more, in this case accuracy, then with the model that performed better, score the data in “*customers\_to\_score.xlsx”.*

* What data is needed to inform those decisions?

The data needed to inform those decisions are 2 different sets of data, 14 columns in *“credit-data-training.xlsx*”, the columns are: *Credit-Application-Result, Account-Balance, Duration-of-Credit-Month, Payment-Status-of-Previous-Credit, Purpose, Credit-Amount, Value-Saving-Stocks, Length-of-current-employment, Instalment-per-cent, Most-valuable-available-asset, Type-of-apartment, No-of-Credits-at-this-Bank, Occupation, Age-Year,* and the entire datain “*customers\_to\_score.xlsx*”.

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

We need to predict an outcome, we have data enough to train a statistical model, it is a classification task since the model has to decide whether the customers is creditworthy or not.

# 

# 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.***

*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.



It was decided to remove the columns *Concurrent-Credits, Guarantors, No-of-dependents, Foreign-Worker* for their low variability, *Duration-in-Current-address* for having over 69% of missing values and *Telephone* for not giving any information about the customer’s creditworthiness. It was decided to impute the median in the missing values of the *Age-Years* attribute for being a whole number.

# Step 3: Train your Classification Models

*Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model*

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

1. Logistic Regression

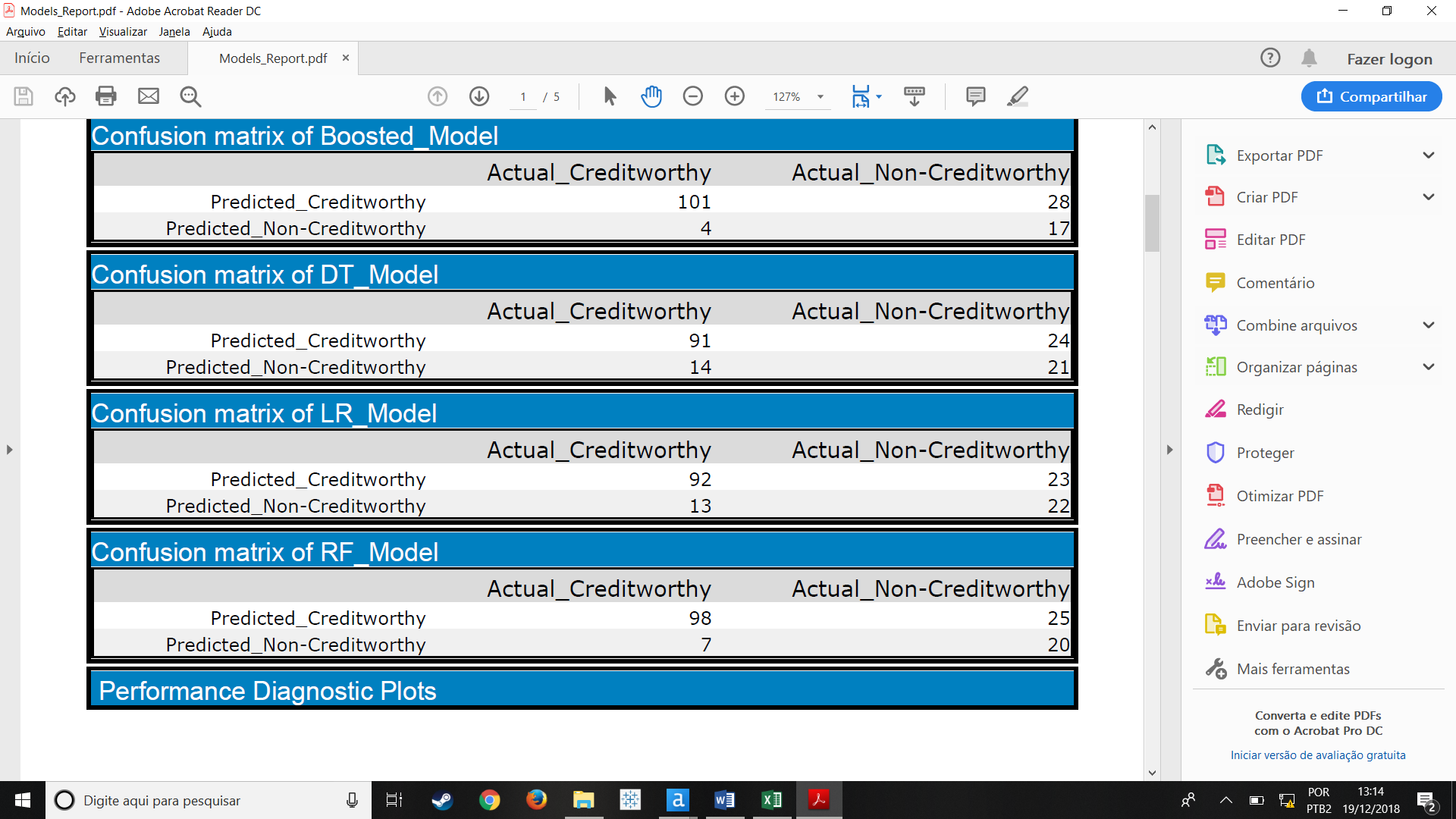
* Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Coefficients | | | | | |
|  | Estimate | Std. Error | z value | Pr(>|z|) |  |
| (Intercept) | -29.621.914 | 6,84E+02 | -43.326 | 1,00E-05 | \*\*\* |
| Account.BalanceSome Balance | -16.053.228 | 3,07E+02 | -52.344 | 1.65e-07 | \*\*\* |
| Payment.Status.of.Previous.CreditPaid Up | 0.2360857 | 2,98E+02 | 0.7930 | 0.42775 |  |
| Payment.Status.of.Previous.CreditSome Problems | 12.154.514 | 5,15E+02 | 23.595 | 0.0183 | \* |
| PurposeNew car | -16.993.164 | 6,14E+02 | -27.668 | 0.00566 | \*\* |
| PurposeOther | -0.3257637 | 8,18E+02 | -0.3983 | 0.69042 |  |
| PurposeUsed car | -0.7645820 | 4,00E+02 | -19.096 | 0.05618 | . |
| Credit.Amount | 0.0001704 | 5,73E-02 | 29.716 | 0.00296 | \*\* |
| Length.of.current.employment4-7 yrs | 0.3127022 | 4,59E+02 | 0.6817 | 0.49545 |  |
| Length.of.current.employment< 1yr | 0.8125785 | 3,87E+02 | 20.973 | 0.03596 | \* |
| Instalment.per.cent | 0.3016731 | 1,35E+02 | 22.340 | 0.02549 | \* |
| Most.valuable.available.asset | 0.2650267 | 1,43E+02 | 18.599 | 0.06289 | . |
| Significance codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1 | | | | | |

The most important variables according to the Logistic Regression method are *Account Balance, Payment Status of Previous Credit, Purpose, Credit Amount, Length of current employment, Installment per cent and Most valuable available asset.*

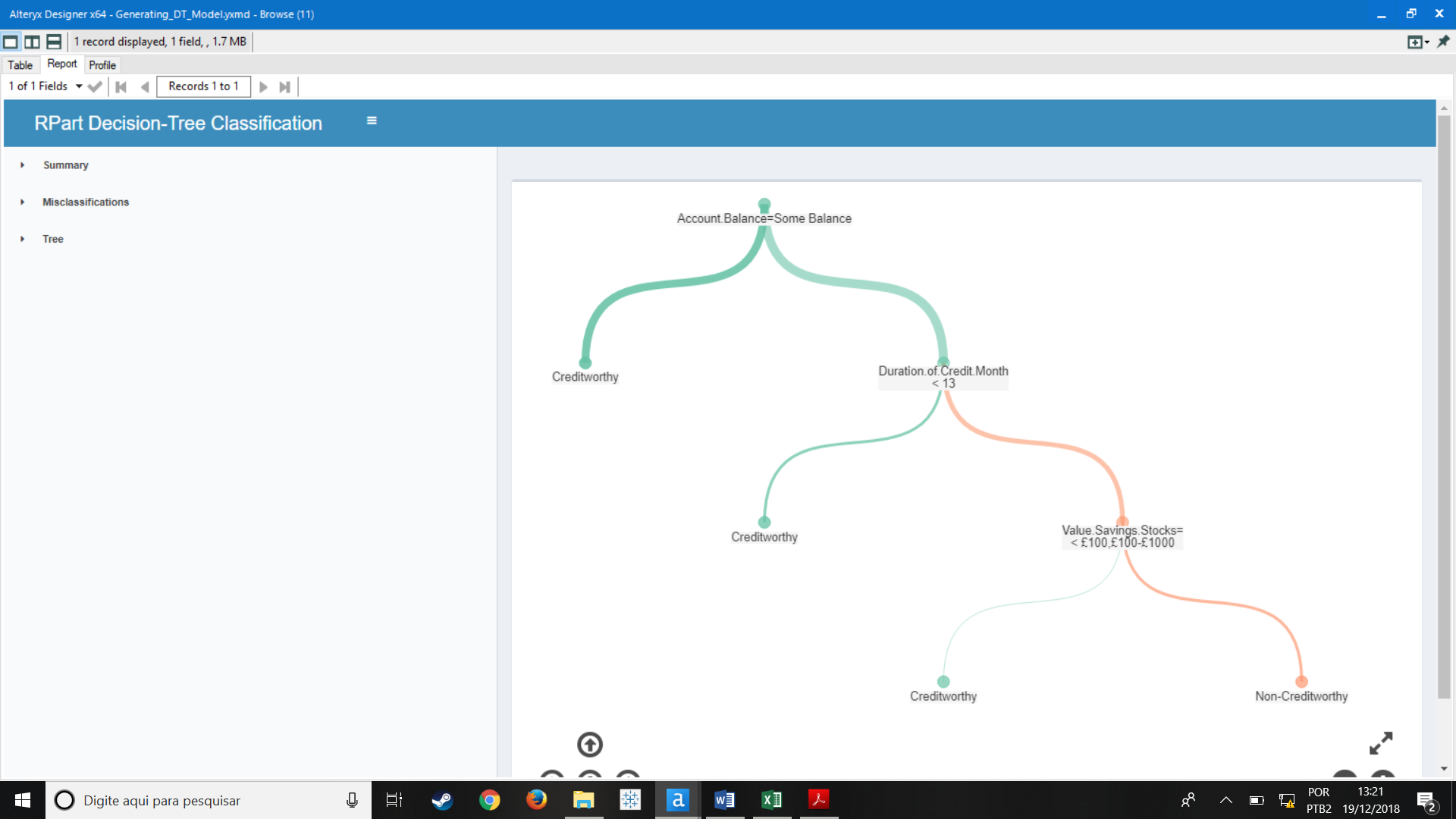
* Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model’s predictions?

The Logistic Regression Model achieved an overall accuracy of 76%. The model is bit biased to predict that the customer is creditworthy, comparing to the other models, it shows the best Accuracy for the Non-Creditworthy customers.



1. Decision Tree

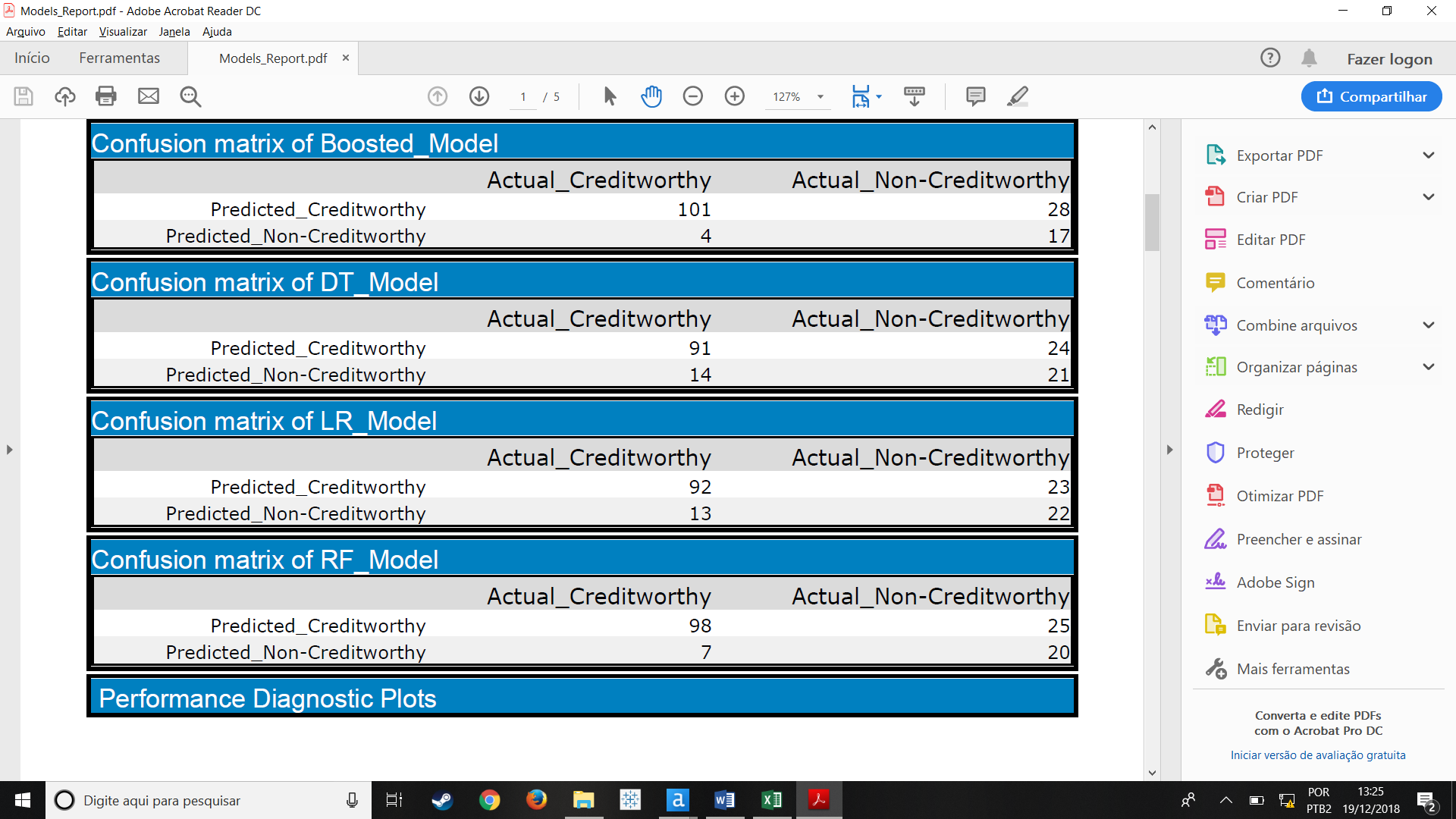
* Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.



The most important variables according to the model are *Account Balance,* *Duration of Credit Month* and *Value Savings Stocks.*

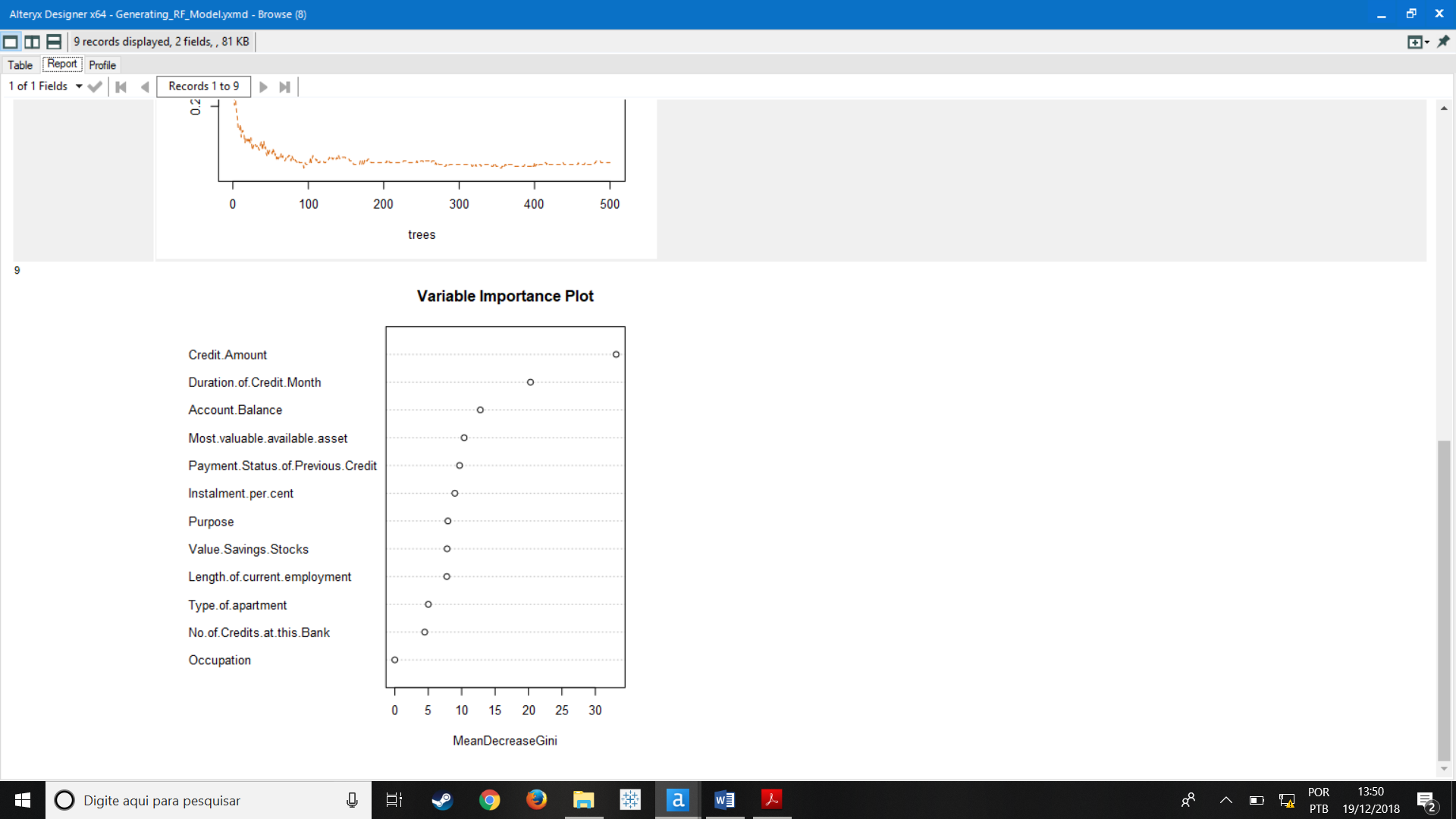
* Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model’s predictions?

The Decision Tree model achieved an overall accuracy of 74.67% in the validation set, comparing to the other model, this is the one with the less accurate. This model again shows to be biased to predict that the customer is creditworthy.



1. Random Forest

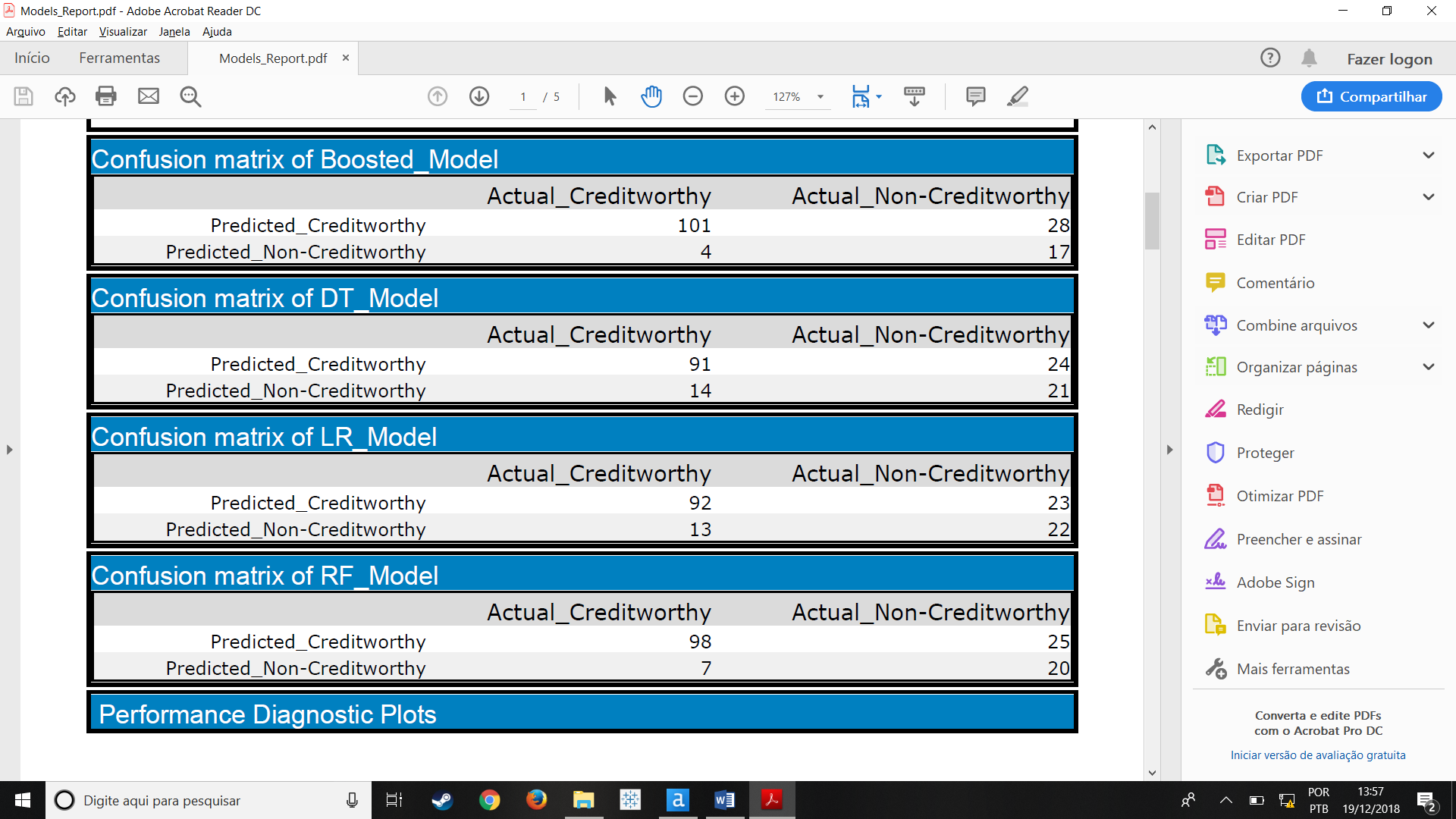
* Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.



Differently of the results of the single Decision Tree model, the Random Forest approach reached that the three most important variables are *Credit Amount, Duration of Credit Month and Account Balance.*

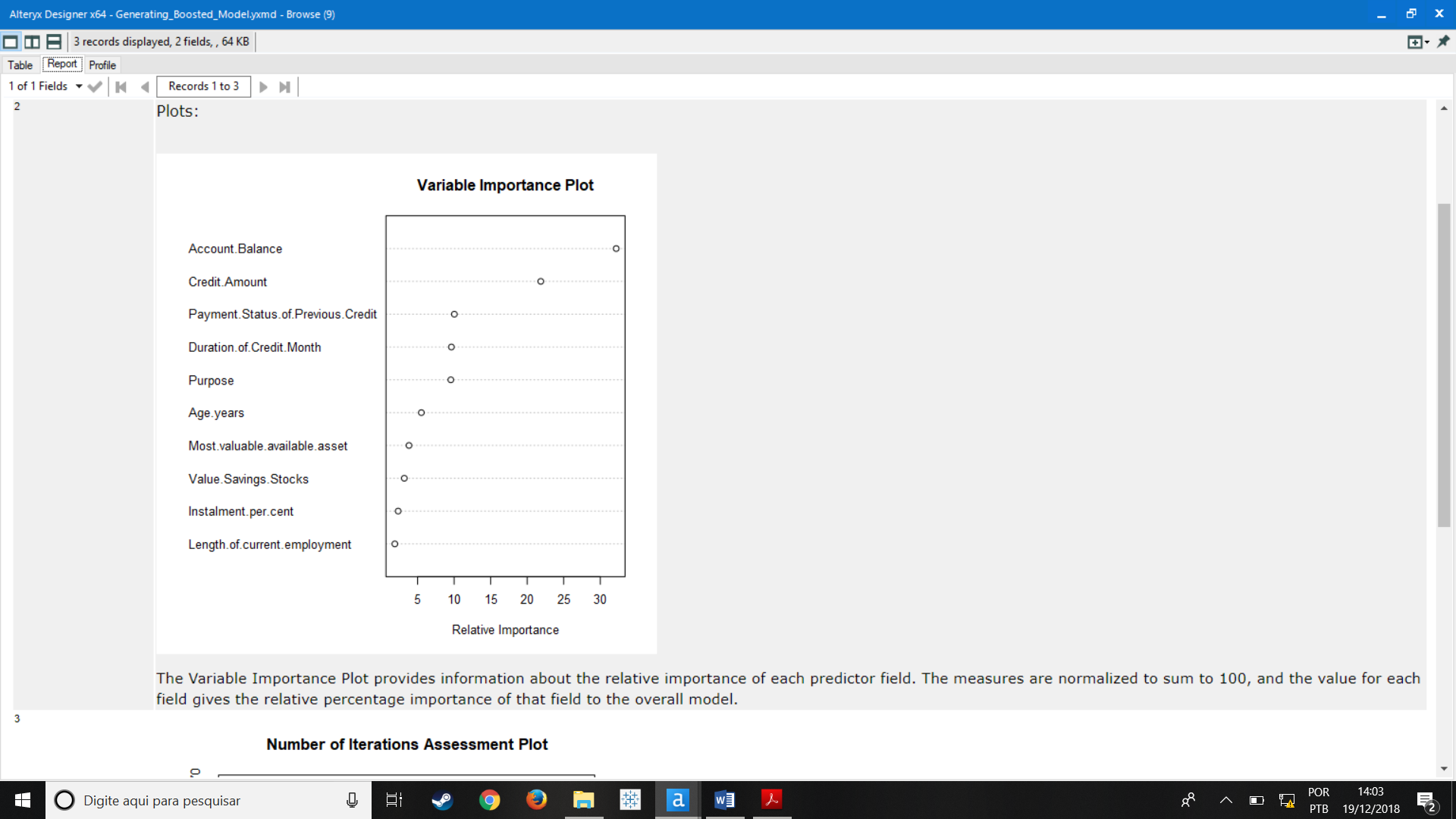
* Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model’s predictions?

The model achieved an overall accuracy of 78.67% and it is the one chosen be used in the test set.



1. Boosted Model

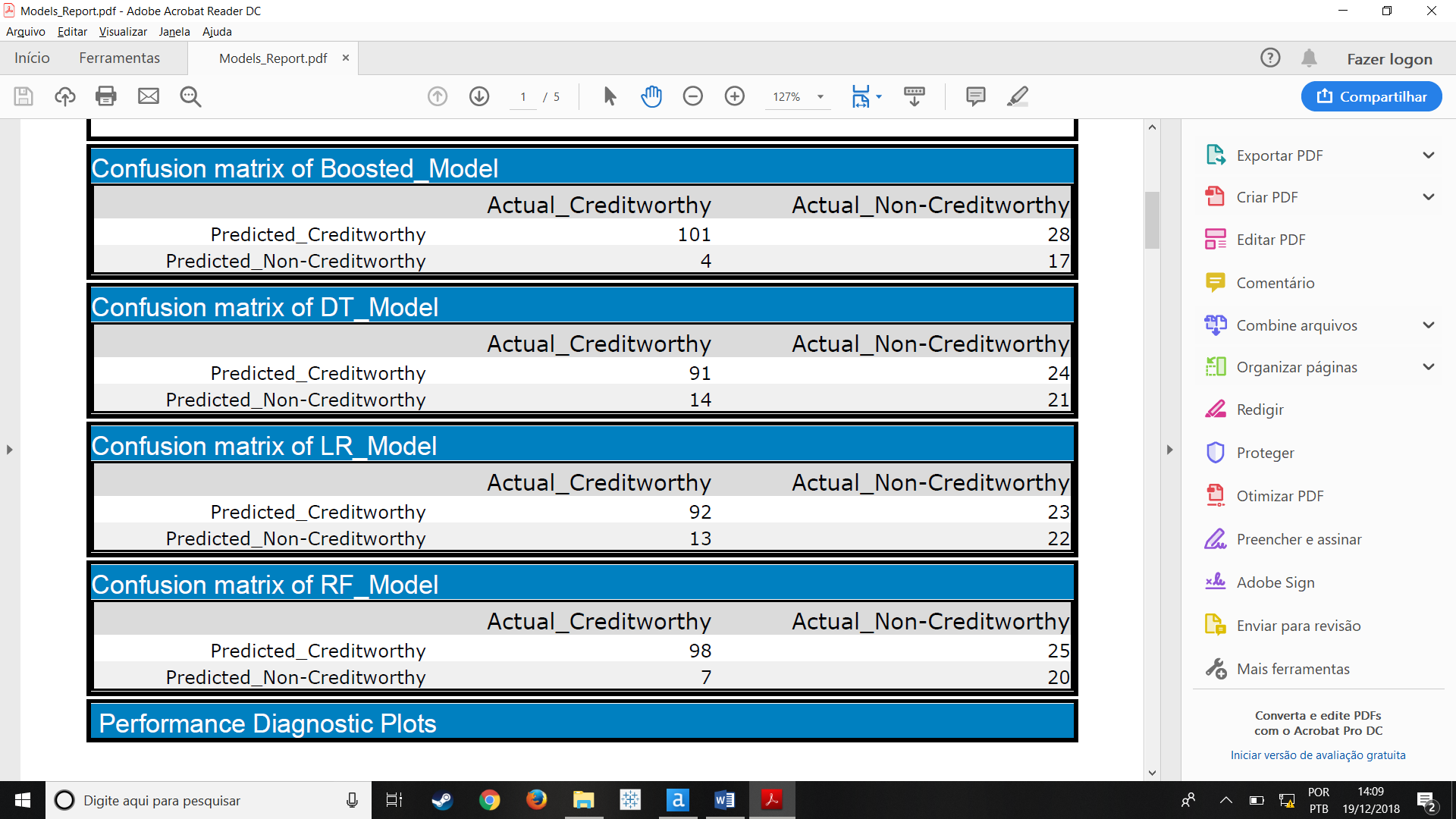
* Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.



The Boosted model approach reached 5 most important variables that are *Account Balance, Credit Amouont, Payment Status of Previous Credit, Payment Status of Previous Credit, Duration of Credit Month* and *Purpose.*

* 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?

The Boosted Model achieved an overall accuracy of 78.67% on the validation set, its Accuracy to predict Creditworthy customers is the highest achieved by all 4 models, however its Accuracy to predict Non Creditworthy customers is the worst, it means the model is biased to predict that the customer is creditworthy.



# 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)*

I chose to use the Random Forest Model because it achieved the highest overall accuracy along with the Boosted Model, between these two, the Random Forest Model outperformed the Boosted Model in the Non-Creditworthy accuracy. With high values in the False Positive Rate in the ROC curve, the Random Forest performed slightly better than the Boosted, even though with the True Positive Rate being less than 0.5 it shows the worst results. As discussed before the Boosted model is biased to predict that the customer is Creditworthy, so analyzing these insights I opted for choosing the Random Forest Model.

*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.

* How many individuals are creditworthy?

According to the Random Forest Model, there are 403 out of the 500 customers that are Creditworthy.