‹ Return to Classroom

Creditworthiness

REVIEW | HISTORY

Meets Specifications

Dear Excellent Student,
Thank you for your submission. I enjoyed reviewing your work because it was great. This submission meets all of our expectations. The Classification Model is not an easy course to grasp, however, the task was really understood. A lot of work has been done and you should be proud of yourself. Continue practicing on these projects and other projects of yours and you will become the best in your domain.

Keep up the good work and good luck in future projects!!

Business and Data Understanding

✓ **The section is written clearly and is concise. The section is written in less than 250 words.**

✓ **All following questions have been answered:**

1. **What decisions need to be made?**
2. **What data is needed to inform those decisions?**
3. **What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?**

Building the Training Set

✓ **The section is written clearly and is concise. The section is written in less than 100 words.**

✓ **The following question has been answered:**

1.In your cleanup process, which field(s) did you impute or remove?

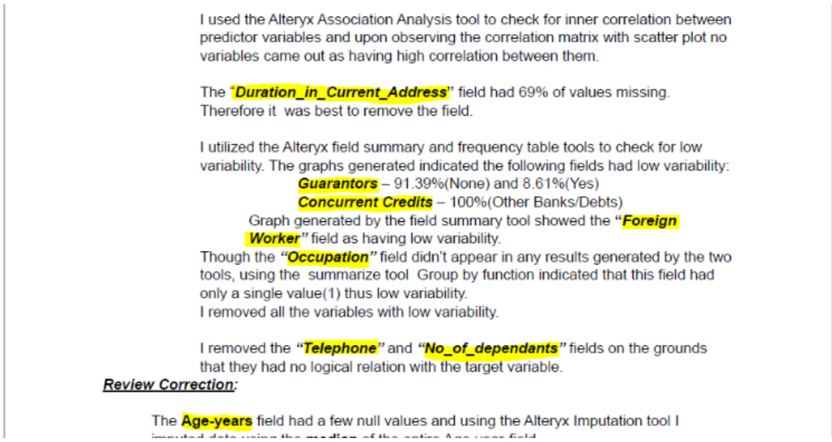
Please justify why you imputed or removed these fields. Visualizations are encouraged.

The correct fields are removed or imputed.

✅*Awesome!!*
All the variables which need to be removed and imputed are correctly identified with a reasonable justification and the visualizations .

I like the way you have justified each of the variables which need to be removed/imputed.

Tips: If you would like to better understand - in a very intuitive way - why the median is a good value to impute the missing values in Age field, check this site out: measure central -tendency



I used the Alteryx Association Analysis tool to check for inner correlation between predictor variables and upon observing the correlation matrix with scatter plot no variables came out as having high correlation between them.

The "***Duration_in_Current_Address***" field had 69% of values missing. Therefore it was best to remove the field.

I utilized the Alteryx field summary and frequency table tools to check for low variability. The graphs generated indicated the following fields had low variability:

Guarantors – 91.39%(None) and 8.61%(Yes)
Concurrent Credits – 100%(Other Banks/Debts)

Graph generated by the field summary tool showed the "***Foreign Worker***" field as having low variability.
Though the "***Occupation***" field didn't appear in any results generated by the two tools, using the summarize tool Group by function indicated that this field had only a single value(1) thus low variability.
I removed all the variables with low variability.

I removed the "***Telephone***" and "***No_of_dependants***" fields on the grounds that they had no logical relation with the target variable.

Review Correction:

The **Age-years** field had a few null values and using the Alteryx Imputation tool I imputed data using the **median** of the entire Age-year field imputed data using the **median** of the entire Age-year field.

Train your Classification Models

✓ The section is written clearly and is concise. The section is written in less than 500 words.

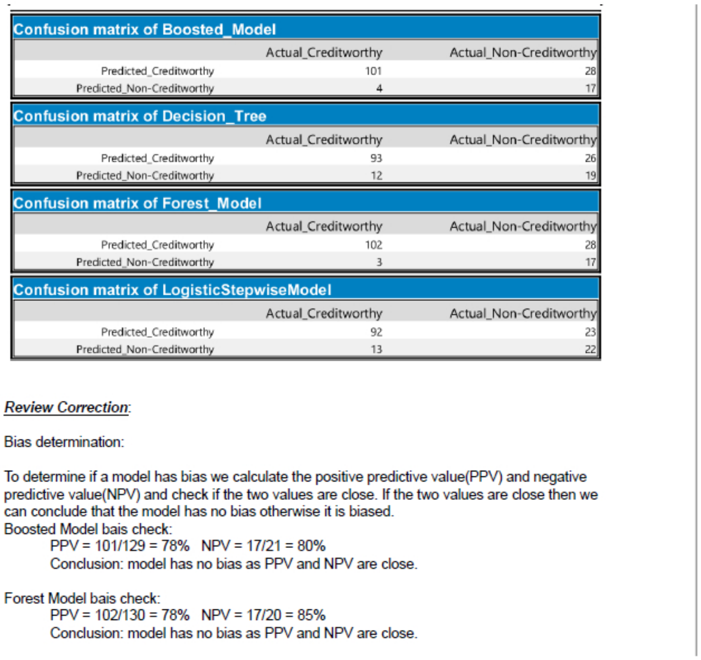
✓ All questions have been answered for each of the four models built: Logistic, Decision Tree, Forest Model, Boosted Model

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.
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?

There should be 4 sets of questions answered.

✅*Awesome work done!!*

- All the significant predictor variables are provided with the variable importance charts for all the models.
- The values in the confusion matrix and the overall percent accuracies are within range.
- The bias seen in each model is very well calculated and justified.



Confusion matrix of Boosted_Model

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

Confusion matrix of Decision_Tree

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	26
Predicted_Non-Creditworthy	12	19

Confusion matrix of Forest_Model

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

Confusion matrix of LogisticStepwiseModel

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

Review Correction:

Bias determination:

To determine if a model has bias we calculate the positive predictive value(PPV) and negative predictive value(NPV) and check if the two values are close. If the two values are close then we can conclude that the model has no bias otherwise it is biased.
Boosted Model bais check:
PPV = 101/129 = 78%   NPV = 17/21 = 80%
Conclusion: model has no bias as PPV and NPV are close.

Forest Model bais check:
PPV = 102/130 = 78%   NPV = 17/20 = 85%
Conclusion: model has no bias as PPV and NPV are close.

Writeup

✓ The section is written clearly and is concise. The section is written in less than 250 words.

✓ All questions have been answered:

1. 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: Your manager only cares about how accurate you can identify people who qualify and do not qualify for loans for this problem.

1. How many individuals are creditworthy?

DOWNLOAD PROJECT

RETURN TO PATH