HADDAD Fayssal
MSc Finance & Big Data
NEOMA Business School

Machine Learning : Individual project

SUMMARY :

Comparing the results between RF and NN: 2
Average metrics: 2
Confusion Matrixes: 2
Interpretation of the Feature Importances results: 4
Robustness of the two models: 5
Overfitting or not ? 6
Gender bias interpretation for both NN and RF: 6
Gender Bias for NN 6
Gender Bias for RF 7

Comparing the results between RF and NN:

Average metrics:

```
Comparing Random Forest and Neural Network:

Average Precision:
Random Forest: 0.3004
Neural Network: 0.8324

Average Recall:
Random Forest: 0.4338
Neural Network: 0.1274

Average F1 Score:
Random Forest: 0.3549
Neural Network: 0.2203
```

When comparing the Random Forest and Neural Network models, we can see some differences in their performance:

In terms of average precision, the Neural Network model performs significantly better than the Random Forest model, with a score of 0.8324 compared to 0.3004. This means that, overall, the Neural Network model is better at correctly identifying positive cases among all the cases it predicted as positive.

However, when it comes to average recall, the Random Forest model outperforms the Neural Network model with a score of 0.4338 compared to 0.1274. This indicates that the Random Forest model is better at identifying the positive cases from all the actual positive cases in the dataset.

Finally, when considering the average F1 Score, which is a balance between precision and recall, the Random Forest model has a higher score of 0.3549 compared to the Neural Network's score of 0.2203. This suggests that, overall, the Random Forest model provides a better balance between precision and recall for this specific problem.

In a nutshell, while the Neural Network model has better precision, the Random Forest model has better recall and a higher F1 Score, so it's a better choice for this particular task considering the balance between the performance metrics.

Confusion Matrixes:

```
Random Forest Confusion Matrix:
[[106653    489]
 [   215    177]]

Neural Network Confusion Matrix:
[[107134      8]
 [   330     62]]
```

Random Forest:

The model correctly classified 106,653 customers who did not have repayment issues (true negatives).The model incorrectly classified 489 customers as having repayment issues when they did not (false positives).

The model correctly classified 177 customers who had repayment issues (true positives).

The model incorrectly classified 215 customers as not having repayment issues when they did (false negatives).

Neural Network:

The model correctly classified 107,134 customers who did not have repayment issues (true negatives).

The model incorrectly classified 8 customers as having repayment issues when they did not (false positives).

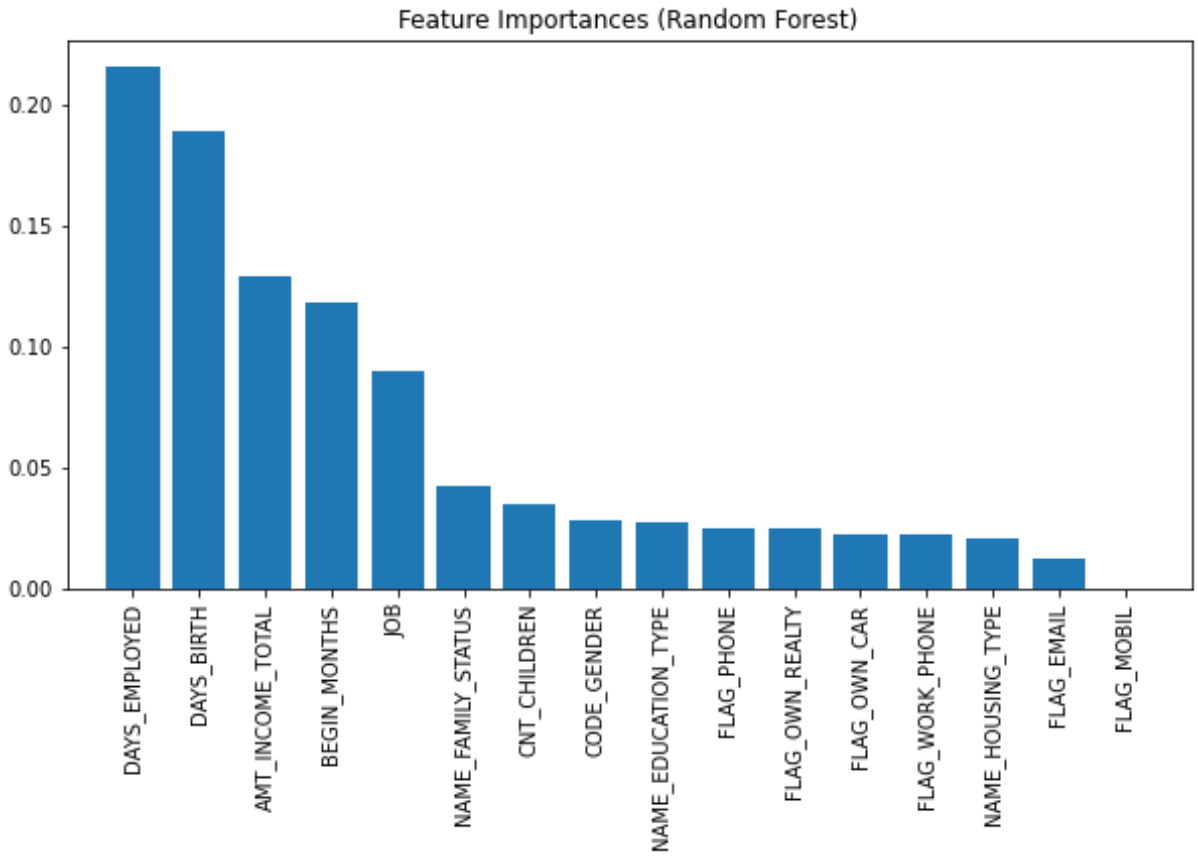
The model correctly classified 62 customers who had repayment issues (true positives).

The model incorrectly classified 330 customers as not having repayment issues when they did (false negatives).

Comparison:

In comparison, the Neural Network model is better at identifying customers who do not have repayment issues (true negatives) and produces fewer false positives. However, it is not as good at identifying customers with repayment issues (true positives) and produces more false negatives than the Random Forest model.

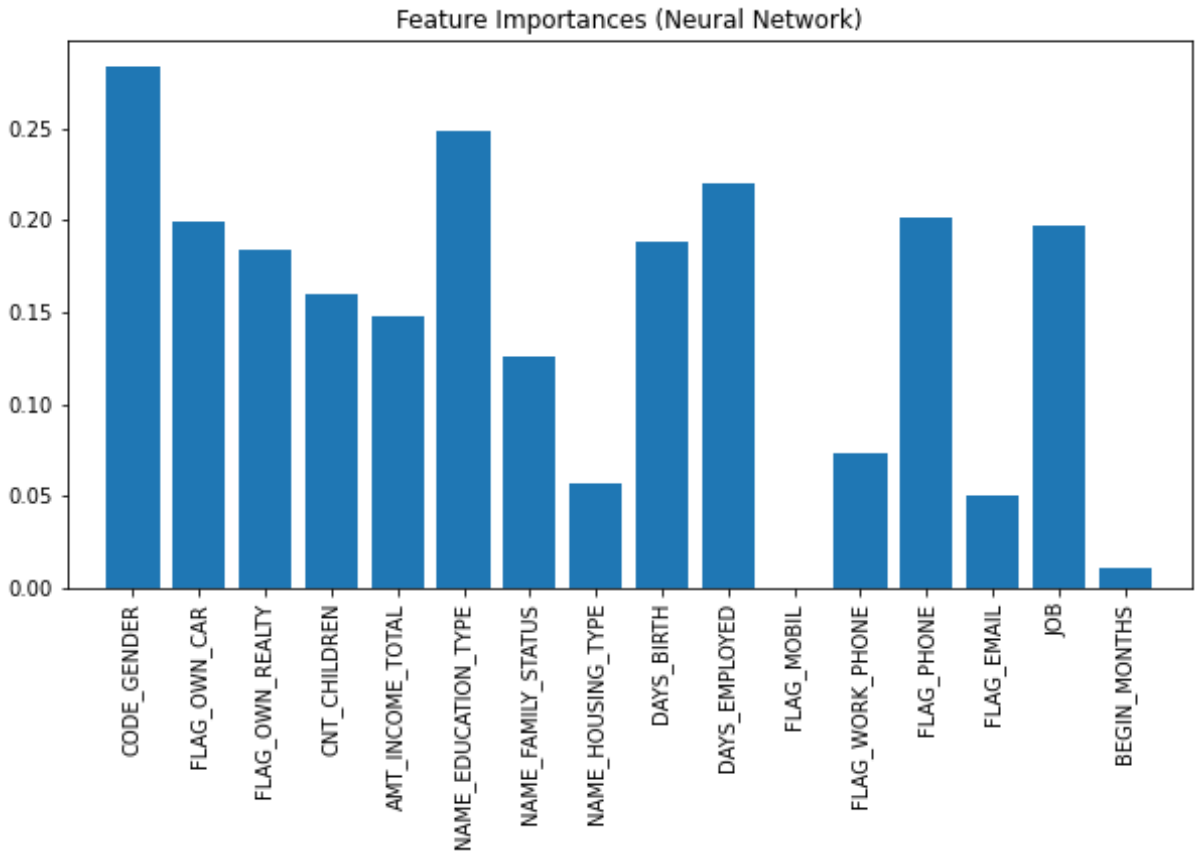
Overall, if the main objective is to minimize false positives (i.e., not bothering customers who do not have repayment issues), the Neural Network model performs better. However, if the main goal is to identify as many customers with repayment issues as possible, the Random Forest model may be a better choice.



Interpretation of the Feature Importances results:

The "Feature importances" charts show the importance of each input variable for the two models: Random Forest and Neural Network. The higher the bar, the more important the variable is considered for making predictions.

In the case of the Random Forest model, the most important variables are "Days_Employed", "Days_Birth", and "AMT_INCOME_TOTAL". This means that these three variables have the most significant impact on the predictions made by the random forest model.



For the Neural Network model, the most important variables are "CODE_GENDER", "NAME_EDUCATION_TYPE", and "DAYS_EMPLOYED". This means that these three variables have the most significant impact on the predictions made by the neural network model.

Robustness of the two models:

```
Random Forest:
Standard Deviation of Precision: 0.0134
Standard Deviation of Recall: 0.0151
Standard Deviation of F1 Score: 0.0133

Neural Network:
Standard Deviation of Precision: 0.1017
Standard Deviation of Recall: 0.0261
Standard Deviation of F1 Score: 0.0414
```

The Random Forest model has lower standard deviations for precision, recall, and F1 score compared to the Neural Network model. This means that the Random Forest model performs more consistently across different data splits during cross-validation. In other words, the Random Forest model is more robust than the Neural Network model, as its performance is more stable when tested on different subsets of the data.

Overfitting or not ?

```
Random Forest:
Average Training Accuracy: 0.9968
Average Validation Accuracy: 0.9942

Neural Network:
Average Training Loss: 0.0165
Average Validation Loss: 0.0150
```

The Random Forest model has an average training accuracy of 0.9968 and an average validation accuracy of 0.9942. This indicates that the model is performing well both on the training data and on the validation data, and there is a relatively small difference between the two. This suggests that the model may not be overfitting, or fitting too closely to the training data and not generalizing well to new data.

On the other hand, the Neural Network model has an average training loss of 0.0165 and an average validation loss of 0.0150. The difference between the training loss and validation loss is also small, but the magnitude of the loss values is lower compared to the accuracy values in the Random Forest model. This could indicate that the Neural Network model is not overfitting and is also performing well on the validation data.

In conclusion, based on the average accuracy and loss values, both the Random Forest and Neural Network models seem to be performing well and not overfitting.

Gender bias interpretation for both NN and RF:

Gender Bias for NN

```
Neural Network Gender Bias Evaluation:
1278/1278 [==============================] - 1s 631us/step
2084/2084 [==============================] - 1s 616us/step
Accuracy for men: 1.00
Accuracy for women: 1.00
Gender bias: 0.00
Out[251]: 0.0018370544378830678
```

The Neural Network Gender Bias Evaluation shows the performance of the neural network model in terms of accuracy for both men and women, as well as the gender bias, which is the absolute difference in accuracy between the two groups.

The results indicate that the neural network model has an accuracy of 1.00 for both men and women. This means that the model can correctly predict the target variable for all the men and women in the test set. The gender bias is 0.00, which means that there is no difference in the model's performance between men and women. This suggests that the neural network model does not exhibit any gender bias in its predictions.

As we have seen before, the “CODE_GENDER” variables have a significant impact on the Neural Network model. So, it is a good thing that there’s no gender bias, it means that the model can predict correctly the target variables for all the men and women.

Gender Bias for RF

```
Random Forest Gender Bias Evaluation:
Accuracy for men: 0.99
Accuracy for women: 0.99
Gender bias: 0.00
Out[246]: 0.0015174610980718262
```

The Random Forest Gender Bias Evaluation shows the performance of the random forest model in terms of accuracy for both men and women, as well as the gender bias, which is the absolute difference in accuracy between the two groups.

The results indicate that the random forest model has an accuracy of 0.99 for both men and women. This means that the model can correctly predict the target variable for 99% of the men and women in the test set. The gender bias is a very small number, 0.001517 for the random forest model. This value is essentially close to zero, indicating that the gender bias in the model's performance is negligible. It means that the TF model doesn’t exhibit any gender bias in its prediction.