Machine Learning for Credit Default Prediction

Suci Rahma Nura

June 2023

24 out of **102** Financial Technology

have an NPL (Non-Performing Loan)

percentage **above 5%**

There are even Fintech companies with an NPL (Non-Performing Loan) ratio reaching 66.27%

www.cnbcindonesia.com

Table of Contents

Objective

- What kind of debtors have a tendency to default?

- What machine learning models are suitable for predicting debtors defaults?

Data Understanding

Data Understanding

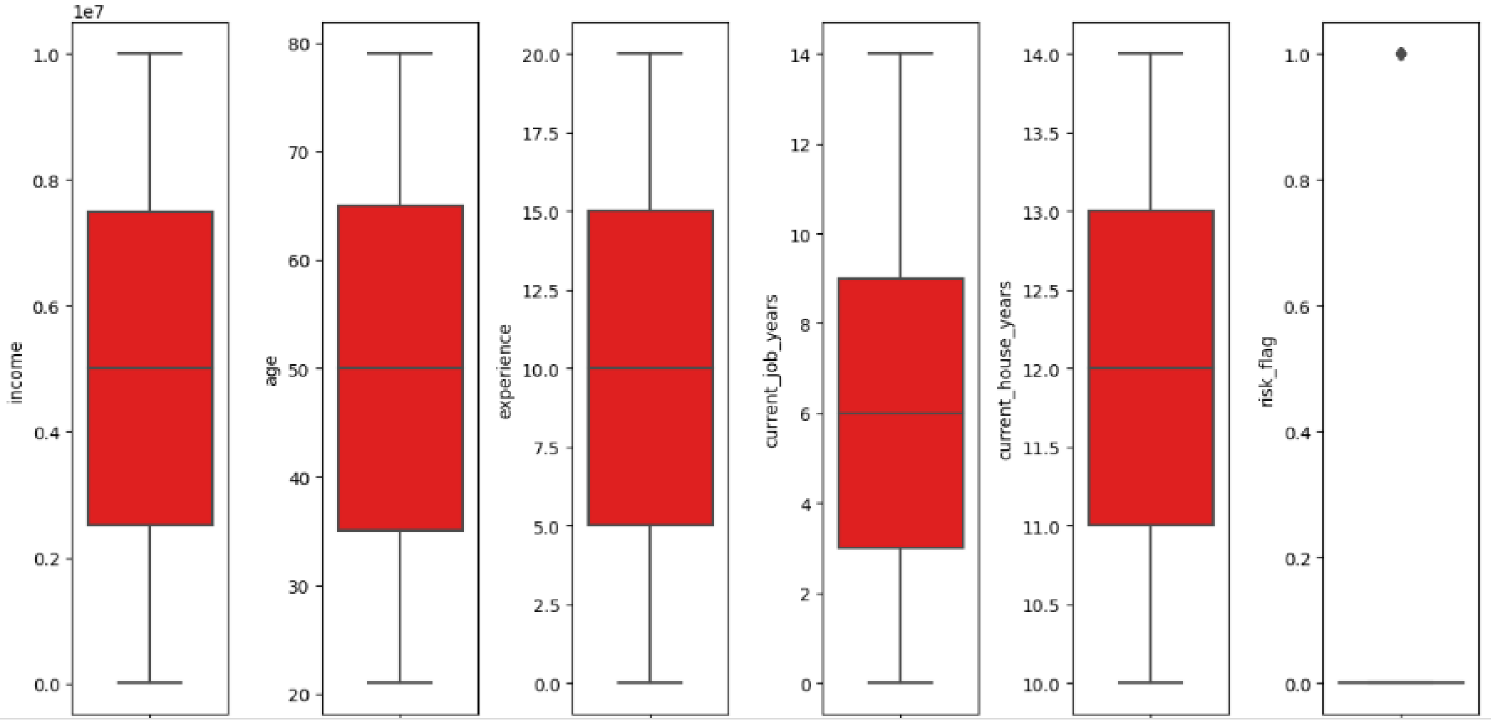
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 252000 entries, 0 to 251999
Data columns (total 13 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   Id                 252000 non-null  int64
 1   income             252000 non-null  int64
 2   age                252000 non-null  int64
 3   experience         252000 non-null  int64
 4   married            252000 non-null  object
 5   house_ownership    252000 non-null  object
 6   car_ownership      252000 non-null  object
 7   profession         252000 non-null  object
 8   city               252000 non-null  object
 9   state              252000 non-null  object
 10  current_job_years  252000 non-null  int64
 11  current_house_years  252000 non-null  int64
 12  risk_flag          252000 non-null  int64
dtypes: int64(7), object(6)
memory usage: 25.0+ MB
```

- **This dataset is obtained from a Hackathon. https://www.kaggle.com/datasets/gargvg/univai-dataset?select=univ.ai_Training+Data.csv**
- **This dataset has 12 columns and 252000 rows**
- **Target feature : 'risk_flag'**
- **No missing and duplicate values**

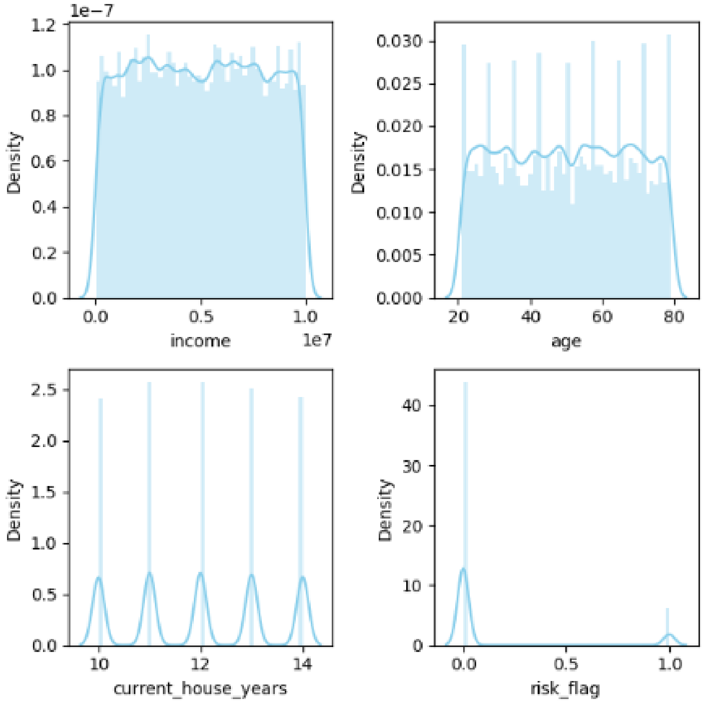
Exploratory Data Analysis

EDA



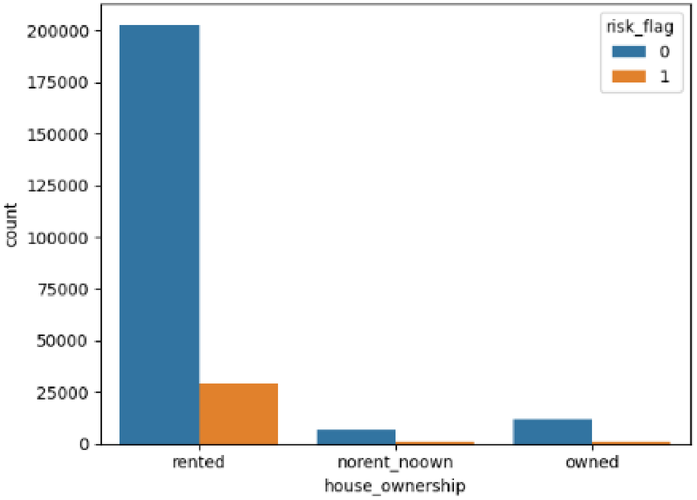
We can see, that risk-flag is the only column that has outliers and it reasonable, since it only has 2 unique value (0 and 1). So we can say that there are no outliers in this dataset.

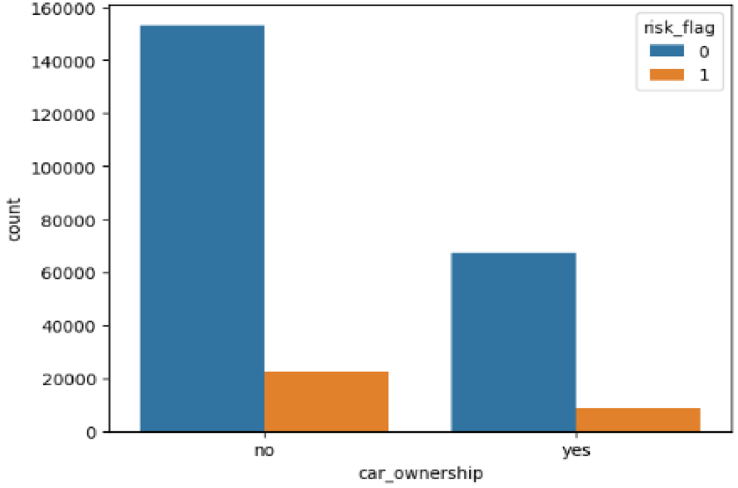
EDA



Somehow, the columns are not simetrical, but not skew either

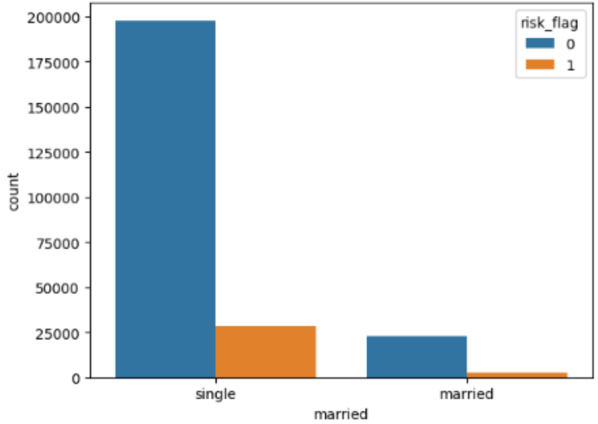
EDA





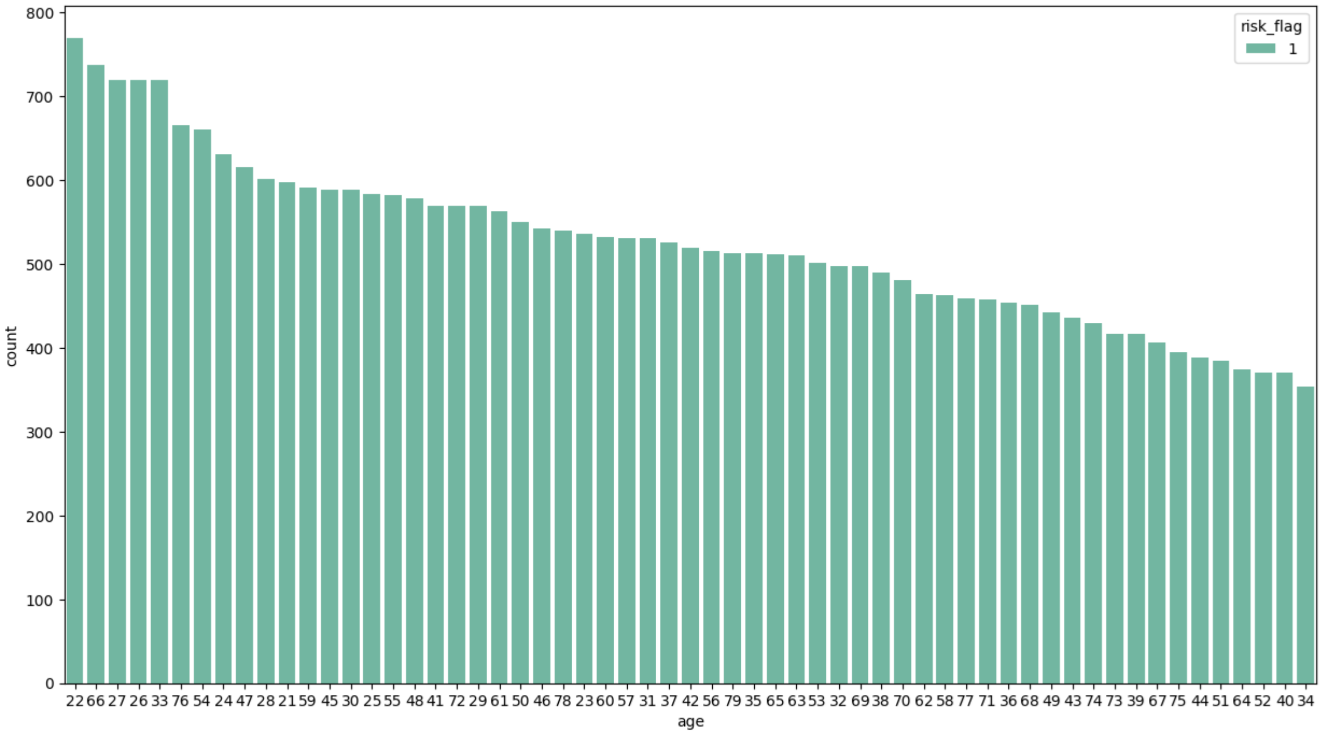
notes : blue bars indicate **non-risky** debtors and orange bars indicate **high-risk** debtors.

<Axes: xlabel='married', ylabel='count'>



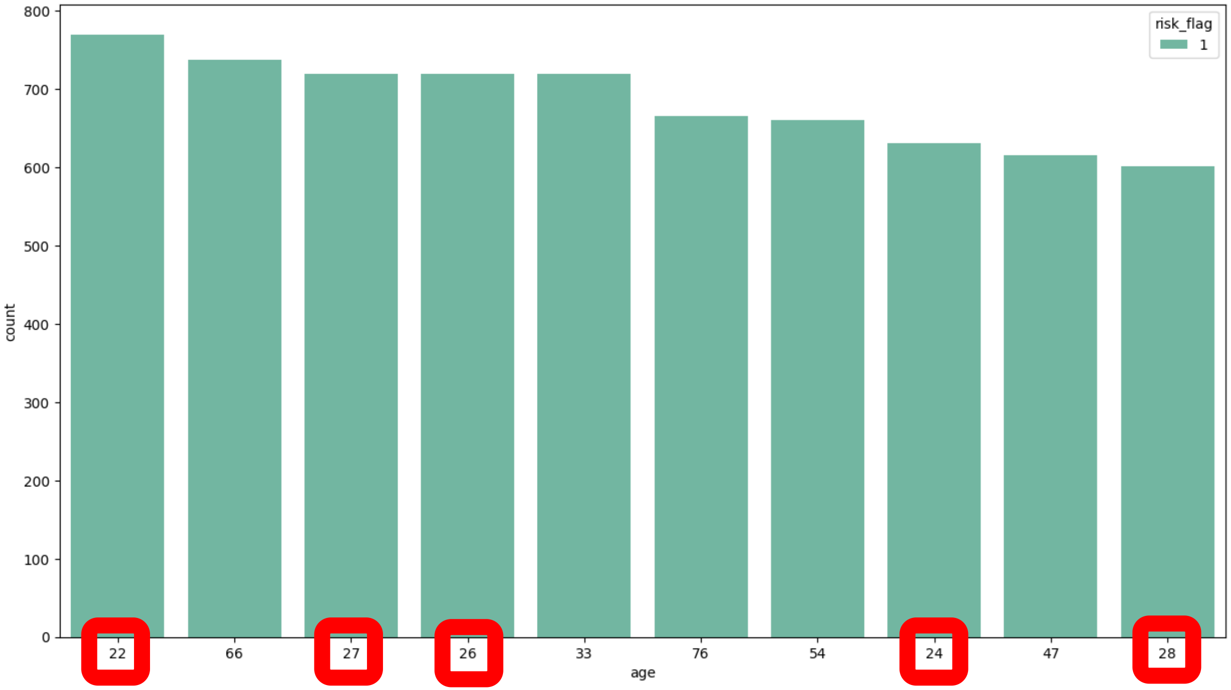
Based on this graph, debtors who are single, live in a rented house, and are not car owners have a higher risk of default than those who are married, live in their own house and own a car.

EDA



There is no visible risk in certain age groups, which means that any age can be at risk of default.

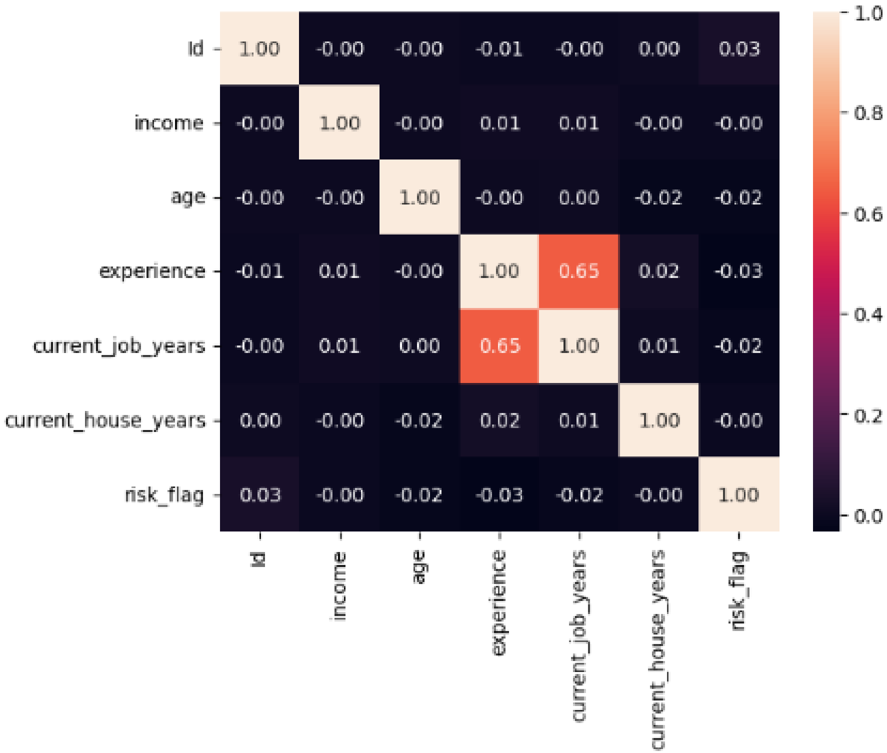
EDA



But, If we take the 10 ages with the highest risk, we can see that the 20s are more at risk than other ages.

EDA

Multivariate Analysis



Based on the heatmap above, experience and current_job_years have a stronger correlation than the others, although the correlation is only 0.65, still < 0.7.

EDA - Chi Square Test

```
Chi-square statistic: 111.89204667099783
p-value: 3.773053705715196e-26
marital status has a significant correlation with risk flag
```

```
Chi-square statistic: 182.98924138871385
p-value: 1.8381930028370595e-40
house_ownership has a significant correlation with risk flag
```

```
Chi-square statistic: 145.42374419378916
p-value: 1.7350853850183746e-33
car_ownership has a significant correlation with risk flag
```

The chi square test was conducted on categorical data such as married, house_ownership, and car_ownership. The result is that the variables married, house_ownership, and car_ownership have a fairly strong correlation with the variable risk_flag.

Data Pre Processing

Data Pre Processing

In this case the methods that will be used are :

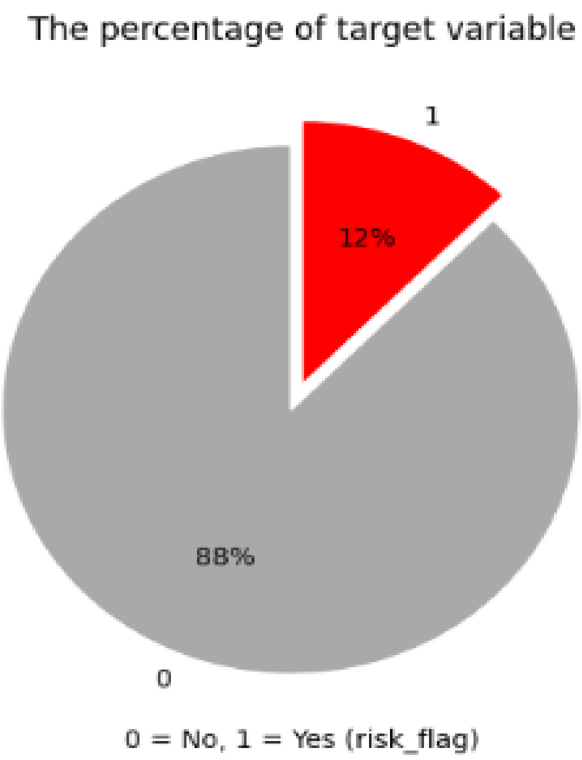
- **One hot encoding** for the variables married and car_ownership.
- **Frequency encoding** for the profession, city, and state variables.
- **Scaling** on the age variable, experience variable, current_job_years, and current_house_years.

Handling Imbalance

```
[83] #check distribution of target variable
     new_df1['risk_flag'].value_counts()
```

```
0    221004
1     30996
Name: risk_flag, dtype: int64
```

This dataset has a fairly imbalanced amount between risk_flag that is worth 1 and risk_flag that is worth 0, for that we need to do imbalance handling using smote.



Modelling with Balance Data

Model with Balance Data

Machine learning model to be tested are :

1. Logistic Regression
2. Decision Tree
3. KNN
4. Random Forest
5. Gaussian Naive Bayes
6. Gradient Boosted Tree

Model with Balance Data

Because this data set is a **credit risk classification case** and the data is balanced, we will focus on the **recall** and **accuracy** evaluation matrix. The following is a recap of the model values:

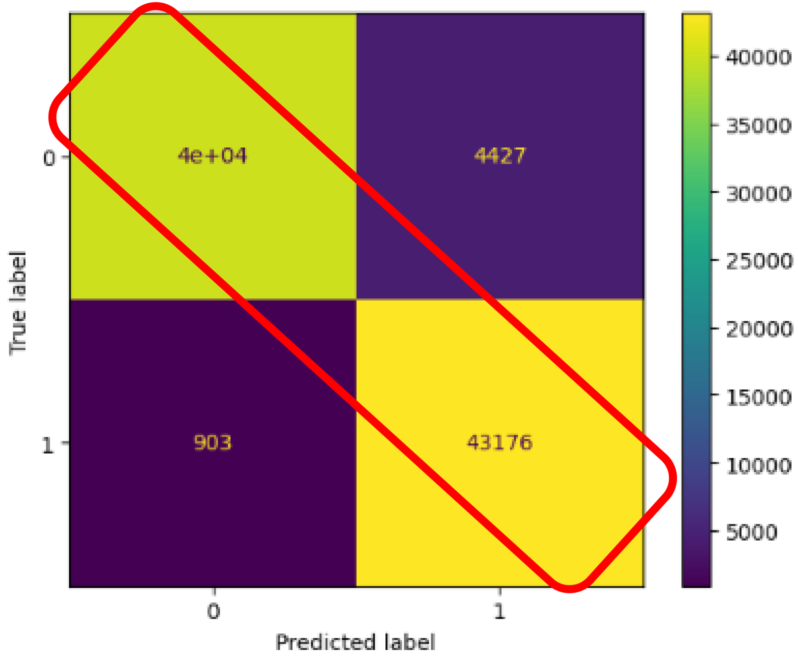
Jenis Model	Accuracy	F1 Score	Recall	Precision
Logistic Regression	0.5071	0.2245	0.5751	0.1395
Decision Tree	0.8713	0.6157	0.8306	0.4891
KNN	0.8610	0.5002	0.5605	0.4516
Random Forest	**0.8869**	0.6339	0.7892	0.5296
Gaussian Naive Bayes	0.2662	0.2249	**0.8581**	0.1294
Gradient Boosted Tree	0.6444	0.2634	0.5126	0.1773

The accuracy value of Random Forest is higher than the recall of Gaussian Naive Bayes, so we will use **Random Forest** as the model because it has the highest evaluation matrix value.

88.69%

Based on the test results of several methods, the best model is Random Forest using matrix evaluation Accuracy with an evaluation rate is 88.69%.

Confusion Matrix



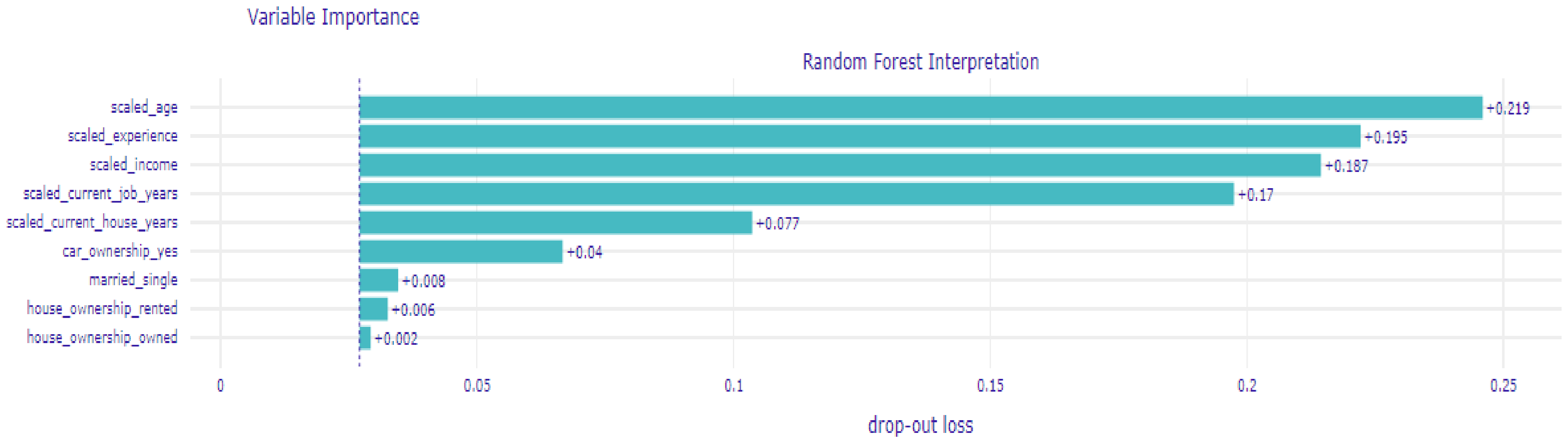
- 4427 means 4427 not high risk that predicted as high risk.
- 903 means 903 have high risk that predicted as not high risk.
- 4e+04 means 4e+04 that correctly predicted as not high risk
- 43176 means 43176 that correctly predicted as high risk.

Dalex

Dalex

Next, we will use dalex to see which variables are most highly correlated with the target variable after data processing.

Dalex

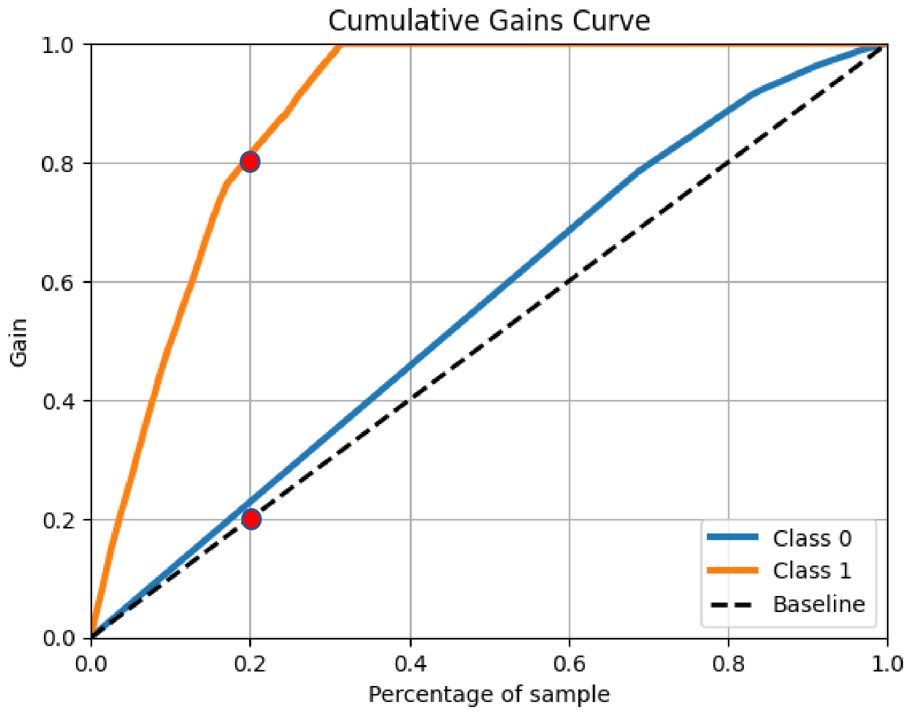


Based on the barplot above, age, income, `current job years, and experience have a high effect on the risk flag. And age has the highest effect.

Business Case

"How much benefit will this model bring to the company??"

To find out we will use Cummulative Gain Curve



We can say that the results of the plot_cumulative_gain are good, for example, if we take a sampling of 0.2, then we have got in Class 1 is 0.8, which means 4x better performance than we do not use the model (baseline).

notes : The baseline on this curve is not the model baseline, but a prediction made without the model.

Business Case

	With Model	Without Model
Saved	35.360	8.840
Failed	8.840	35.360
Total cost	132.600,00	132.600,00
Bruto	70.720.000,00	17.680.000,00
Netto	70.587.400,00	17.547.400,00

4x better performance

Save = $ 53,040,000

Final summary

Debtors who are single, live in a rented house, and do not own a car are very risky to finance.

Variables that have a high effect on credit risk are customer age, married status, home ownership and car ownership.

In this case, the model using Random Forest with Accuracy evaluation matrix has the best performance, with an evaluation of 88.69%.

With a high level of evaluation, of course this prediction also has a high impact on business, at least it can help reduce losses > 88% + 4x company income.

Business Recommandations

Avoid financing debtors who are single, live in a rented house, do not own a car, aged between 20s years old.

However, if you still want to finance debtors with these categories, you may be able to submit additional data such as additional income, proof of ownership of other assets, or a guarantor from the family who is able and willing to take financial responsibility.

Appendix

Here are some of the methods I used to find the best model

1. - One Hot Encoding for Married, home ownership, car ownership
   - Frequency Encoding for Profession, city, state

2. - One Hot Encoding for Married, car ownership
   - Weight of Evidence (WoE) Encoding for Profession, city, state, home ownership

Here are some of the methods I used to find the best model

3. - One Hot Encoding for Married, home ownership, car ownership
   - Frequency Encoding for Profession, city, state
   - Do standard scaller for all numerical data

(this methode has the best model)

4. - One Hot Encoding for Married Married, car ownership
   - Weight of Evidence (WoE) Encoding for Profession, city, state, home ownership
   - Do standard scaller for all numerical data

Do the method repeatedly for balance and imbalance data, and also use other matrix evaluations such as Accuracy, F1 score, Recall and Precision on each method to see which evaluation matrix is better.

Comparison between baseline and model after data processing

Jenis Model	Baseline				After Processing Data			
	Accuracy	F1 Score	Recall	Precision	Accuracy	F1 Score	Recall	Precision
Logistic Regression	0.8759	0.0	0.0	0.0	0.5071	0.2245	0.5751	0.1395
Decision Tree	0.8814	0.5441	**0.569**	0.5219	0.8713	0.6157	0.8306	0.4891
KNN	0.8897	0.5338	0.5089	0.5612	0.8610	0.5002	0.5605	0.4516
Random Forest	0.8960	0.5642	0.5421	0.5881	**0.8869**	0.6339	0.7892	0.5296
Gaussian Naive Bayes	0.8759	0.0	0.0	0.0	0.2662	0.2249	**0.8581**	0.1294
Gradient Boosted Tree	0.876	0.0019	0.0009	0.6667	0.6444	0.2634	0.5126	0.1773

Refferences

1. https://medium.com/responsibleml/visualize-ml-model-bias-with-dalex-b63f182cd649
2. https://medium.com/@pararawendy19/memahami-metrik-pada-pemodelan-klasifikasi-29cd5b738ee7
3. https://ksnugroho.medium.com/confusion-matrix-untuk-evaluasi-model-pada-unsupervised-machine-learning-bc4b1ae9ae3f
4. https://www.ceicdata.com/id/indicator/indonesia/non-performing-loans-ratio
5. https://www.idscore.id/faq/detail/apa-itu-non-performing-loan-npl-serta-dampak-negatif-bagi-lembaga-keuangan

Thank you

https://wa.me/628117071712

https://medium.com/@sucirahma.srn

https://www.linkedin.com/in/sucisrn/

https://github.com/eseren

sucirahma.srn@gmail.com