CREDIT DEFAULT

USING MACHINE LEARNING CLASSIFICATION ON TAIWANESE DATA

By Daniel Baumann

TEXT

WHAT IS CREDIT DEFAULT?

- Inability to pay a debt in due time
- Typically a 6/9 month time frame

Why is default bad?

- It represents a large sunk cost to lending societies
- Liability contracts often won’t recover all costs!

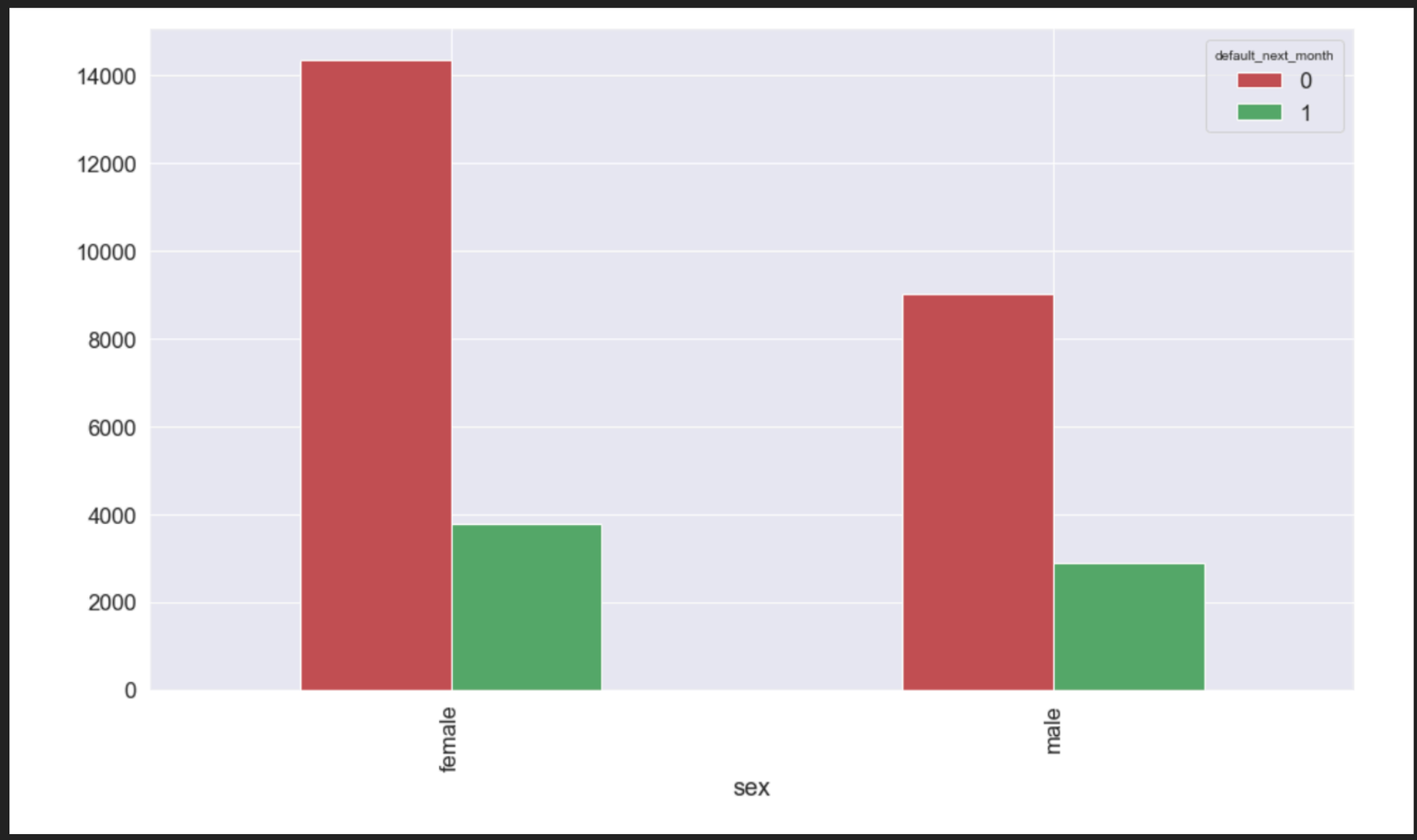
THE IMPORTANCE OF CLASSIFYING DEFAULTERS

- Classifying defaulters correctly can indirectly improve profitability
  - You can refuse credit to those you deem risky
  - Raise interest rates to riskier individuals
  - Aversion to offering high credit limits to risky individuals

WHAT VARIABLES PREDICT DEFAULT RISK? (1)

GENDER

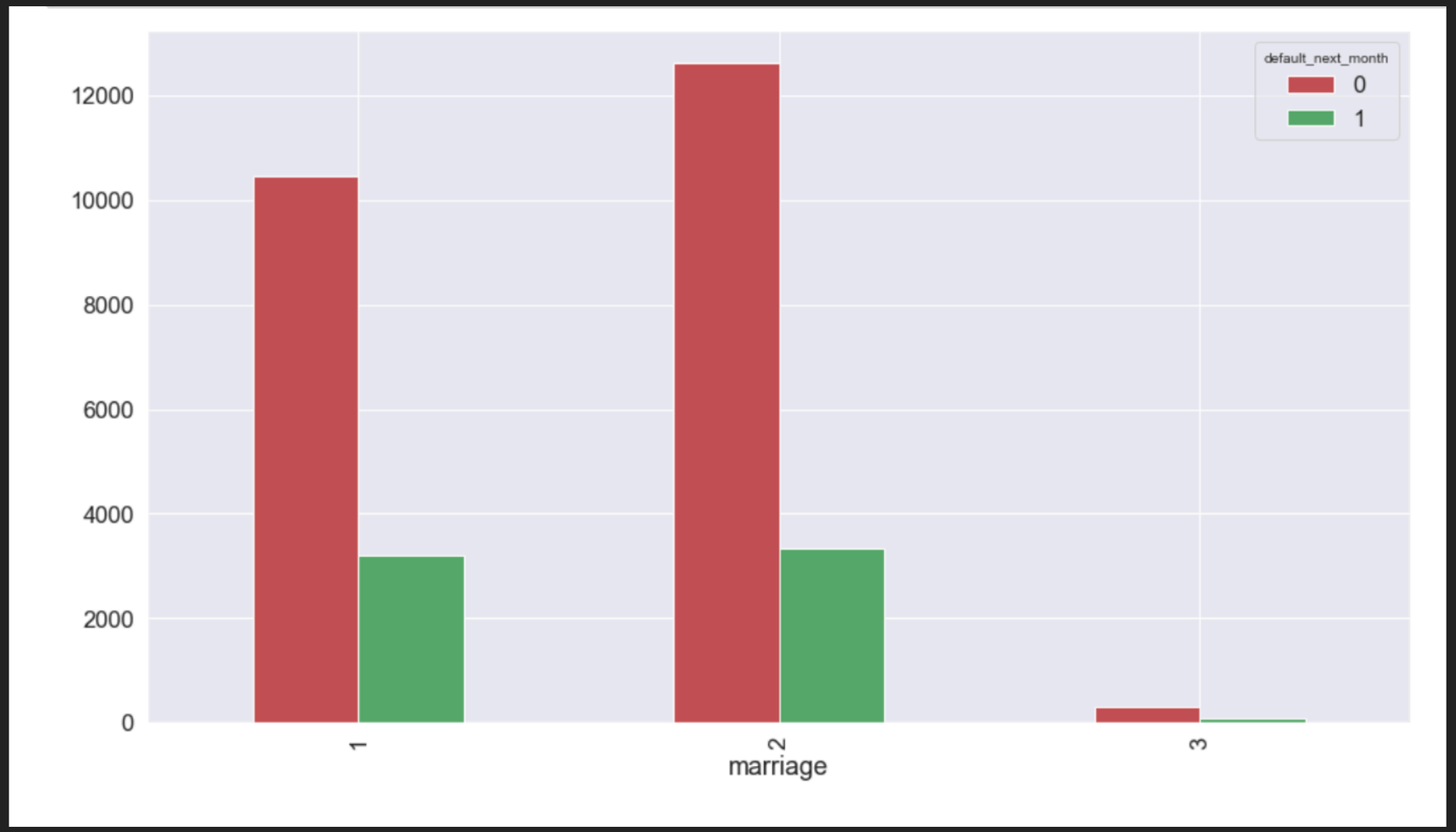
- Male default rate is 24%
- Female default rate is 21%



WHAT VARIABLES PREDICT DEFAULT RISK? (2)

MARRIAGE

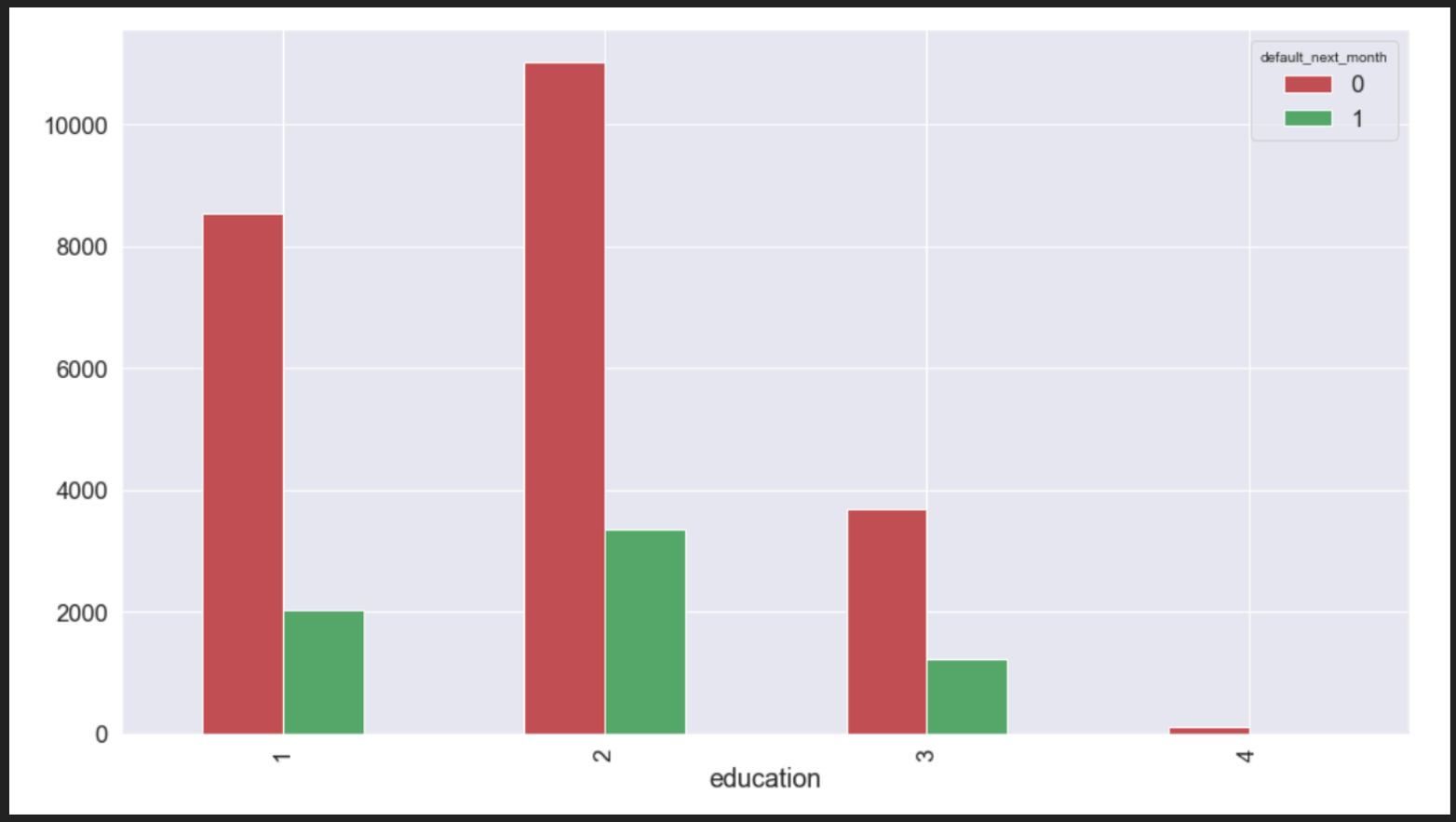
- Singles default rate is 23%
- Married individuals default rate is 21%
- Married people have more security?



WHAT VARIABLES PREDICT DEFAULT RISK? (3)

EDUCATION

- Graduate schooled (1) individuals default less than university graduates (2) and high school graduates (3)



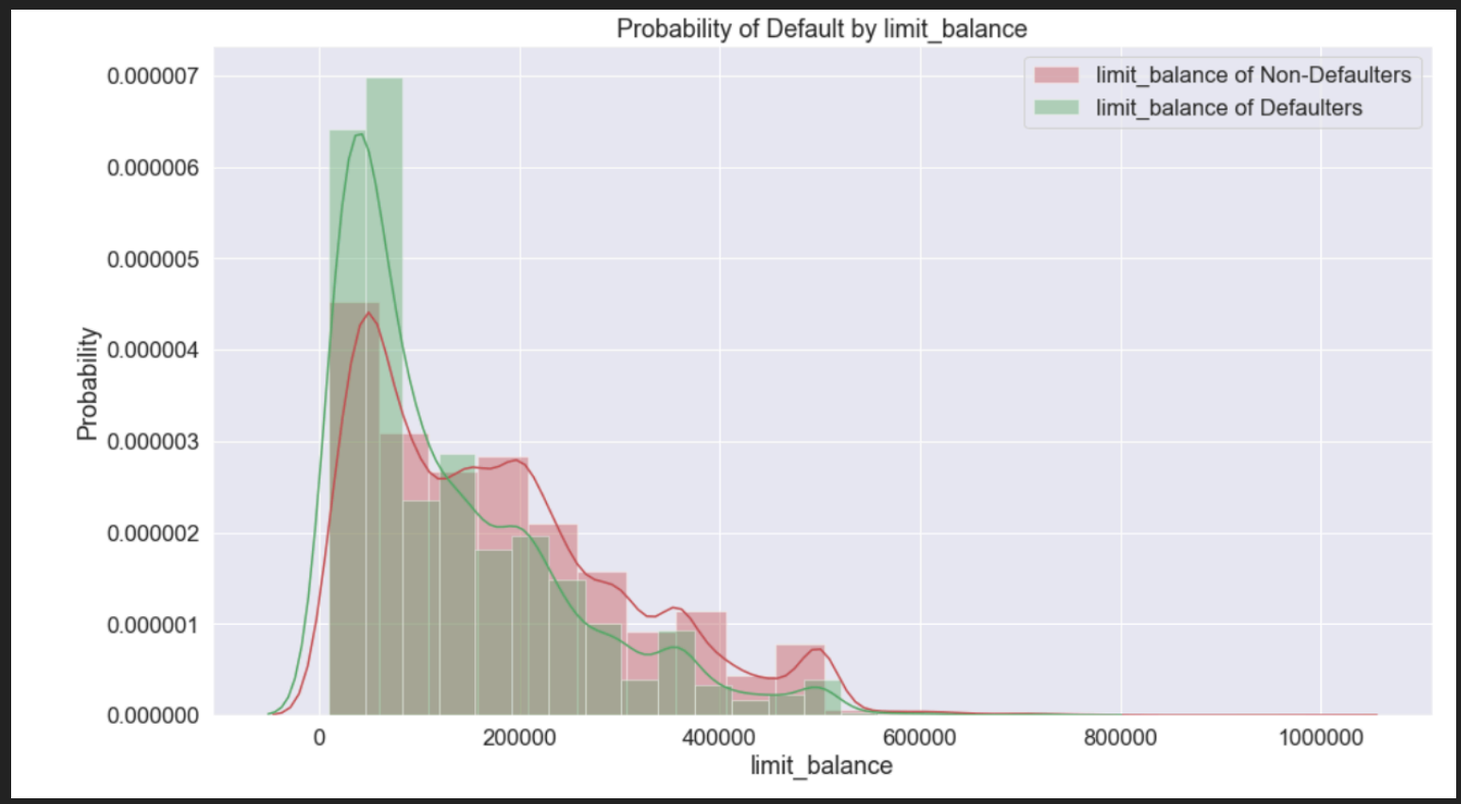
- Higher education may me less risk of default

TEXT

WHAT VARIABLES PREDICT DEFAULT RISK? (4)

LIMIT BALANCE

- Individuals with lower limit balances tend to default more



WHAT MAKES A GOOD MODEL?

THESE ARE OUR DESIRABLE PREDICTIONS:

- * True Positive*: Correctly identifying those who will default on credit
- * True Negative*: Correctly identifying those who will not default on credit

UNDESIRABLE PREDICTIONS:

- * False Positive*: Incorrectly identifying an individual who will not default, as a defaulter
- * False Negative*: Unable to identify those who will actually default

EVALUATION METRICS

- The cost of false negatives is extremely high
- We should be looking at RECALL in order to minimise false negatives

$$\text{Recall} = \frac{\text{Number of True Positives}}{\text{Number of Actual Total Positives}}$$

- i.e. "Out of all individuals we saw as actually having defaulted, what percentage of them did our model correctly identify as defaulting?"

PREDICTIVE MODELLING

Some of the models I used:

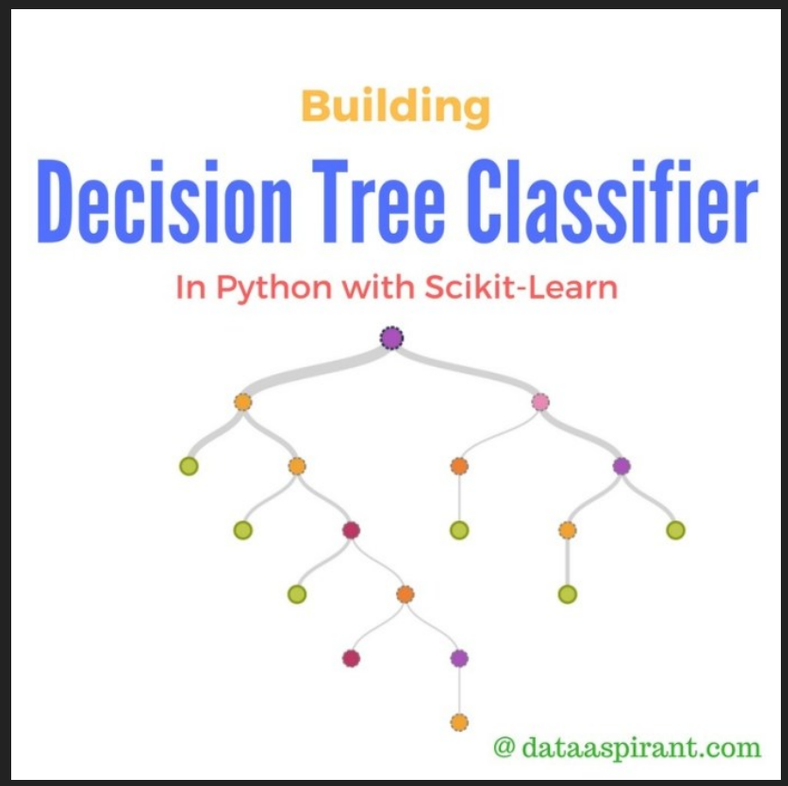
- Logistic Regression
- Decision Trees
- Support Vector Machines
- Voting Classifier
- AND THE BEST ONE:

RANDOM FOREST CLASSIFIER

TEXT

RANDOM FOREST CLASSIFIER

- Random Forest is an ensemble method which incorporates several decision trees
- Each tree will predict a certain outcome
- A decision is made by majority rule
- Recall rate of 55.7%!!!



THANK YOU

Q&A

```
AUC: 0.7713457511156865
```