PAUL
BATEMAN

Predicting Loan Defaults

September 2021

Business brief

- **to understand who is likely to default**
- **who they should lend to in the future**

Dataset comprises over 40,000 historical loans, over 100 variables
Client is Lending Club, a US-based peer-to-peer lending platform

Approach Taken

Data examination and domain knowledge

Examination of the data and data dictionary. Decided to select key variables as an MVP.

LC initially operated a p-2-p lending platform.

Explored the cleaned data.

Made a logistic regression model to explain binary variable, 'defaulted'.

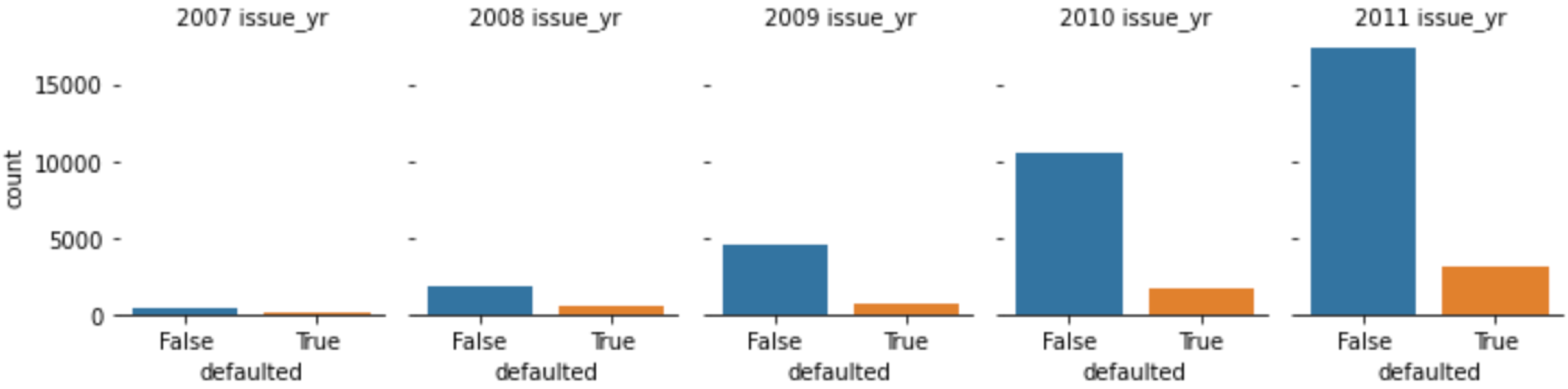
Credit Risk

To reduce default risk, LendingClub focuses on high-credit-worthy borrowers, declining approximately 90% of the loan applications it received as of 2012 and assigning higher interest rates to riskier borrowers within its credit criteria. Only borrowers with FICO score of 660 or higher can be approved for loans.
The statistics on LendingClub's website state that, as of December 31, 2016, 62.3 percent of borrowers report using their loans to refinance other loans or pay credit card debt.

Loan Performance Statistics

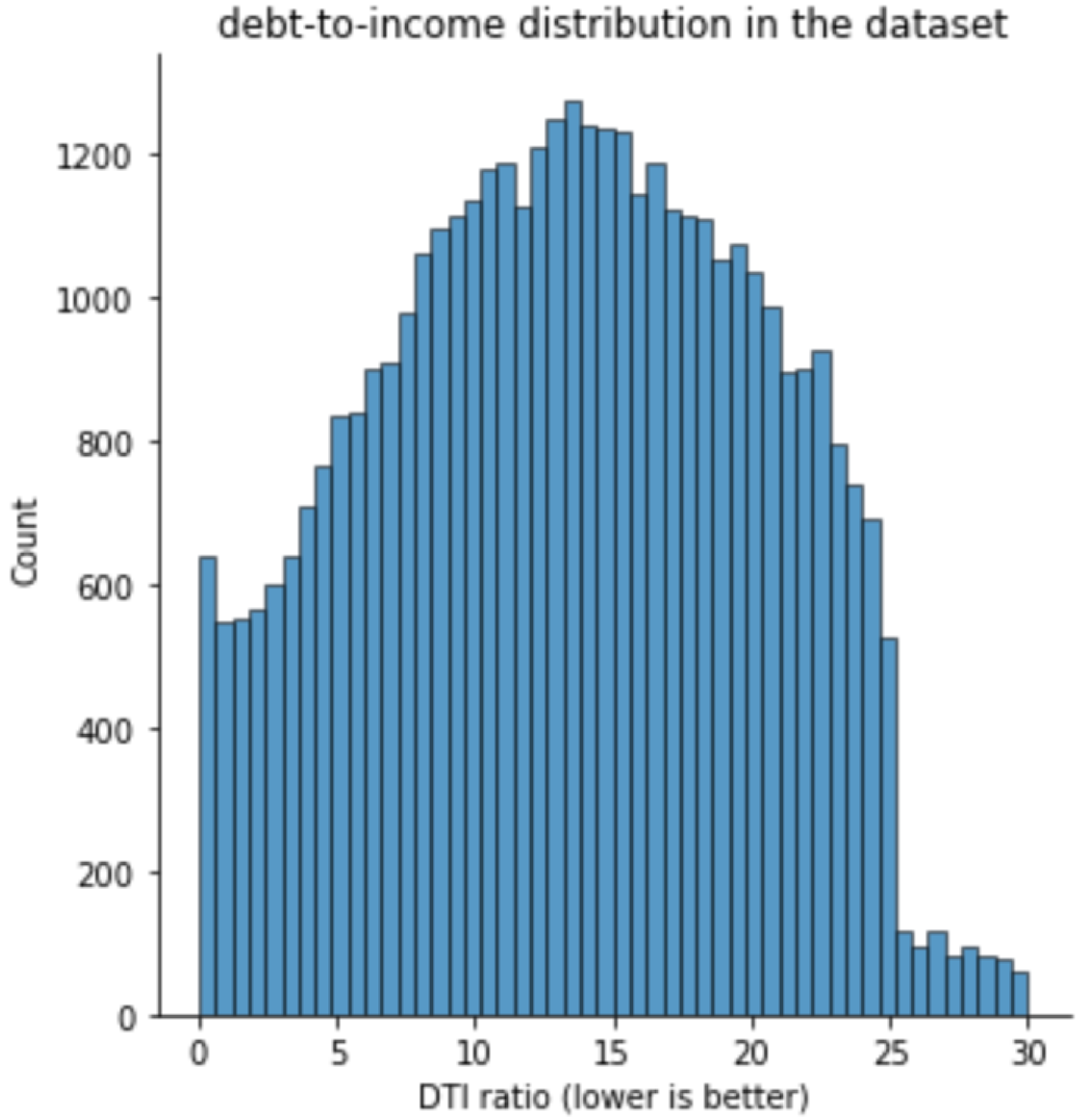
As of June 30, 2015, the average LendingClub borrower has a FICO score of 699, 17.7% debt-to-income ratio (excluding mortgage), 16.2 years of credit history, $73,945 of personal income and takes out an average loan of $14,553 that s/he uses for debt consolidation or for paying off credit card debts. The investors had funded $11,217,348,156 in loans, with $1,911,759,192 coming from Q2 2015. The nominal average interest rate is 14.08%, default rate 3.39%, and an average net annualized return (net of defaults and service fees) of 8.93%. The average returns of investment for LendingClub lenders are between 5.47% and 10.22%, with 23 straight quarters of positive returns as of the second quarter of 2013.

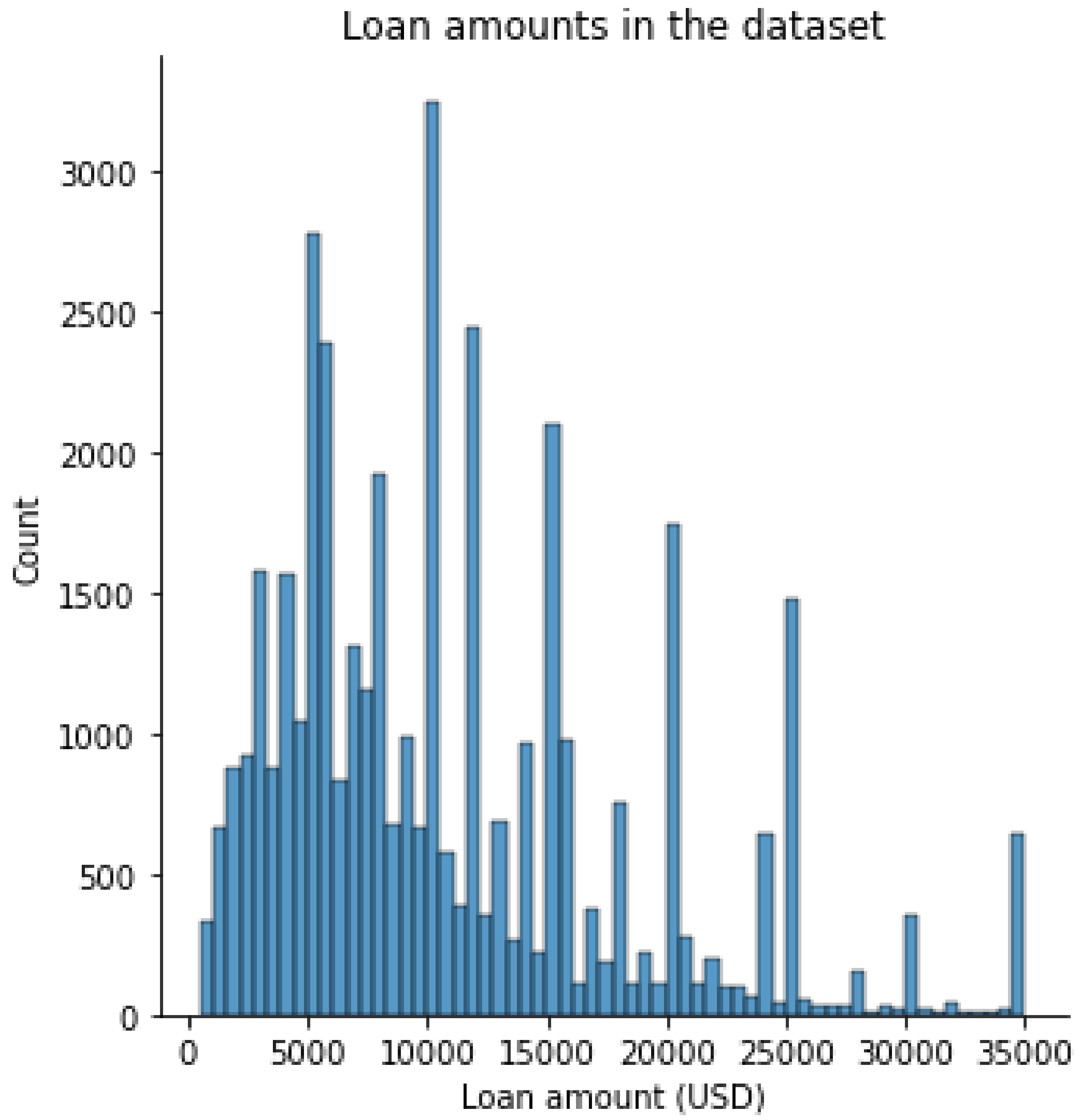
Data description



Default rate: 27% 21% 14% 14% 15%

Data description





Log model 1

01 Loan characteristics (5)

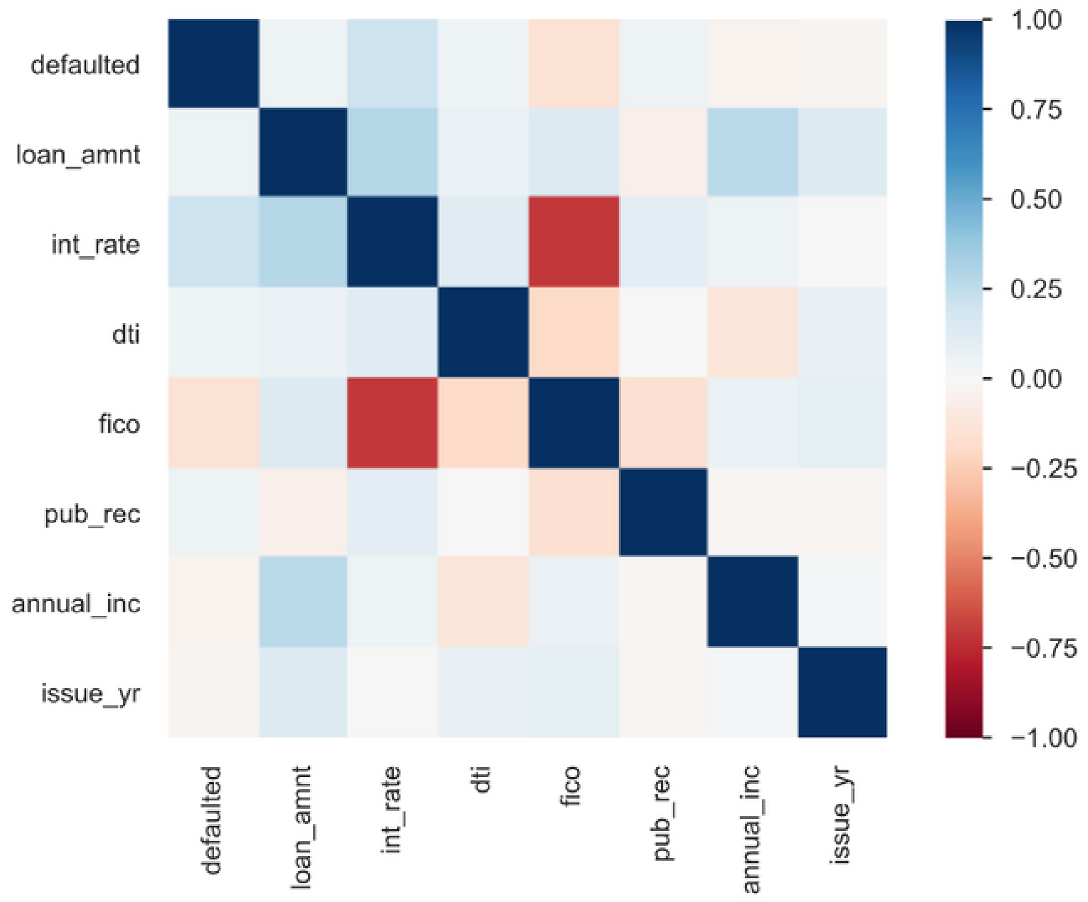
- 'defaulted' (Y/N)
- 'loan_amnt'
- 'term' (36 or 60 months)
- 'issue_d'
- 'int_rate'

02 Credit risk indicators (4)

- 'sub_grade'
- 'dti'
- 'fico'
- 'pub_rec'

03 Demographic variables (5)

- 'emp_length'
- 'home_ownership'
- 'verification_status'
- 'addr_state'
- 'annual_inc'



train: 0.60
test: 0.56
AUC: 0.64
Recall score: 0.67

Log model 2

01 Loan characteristics (4)

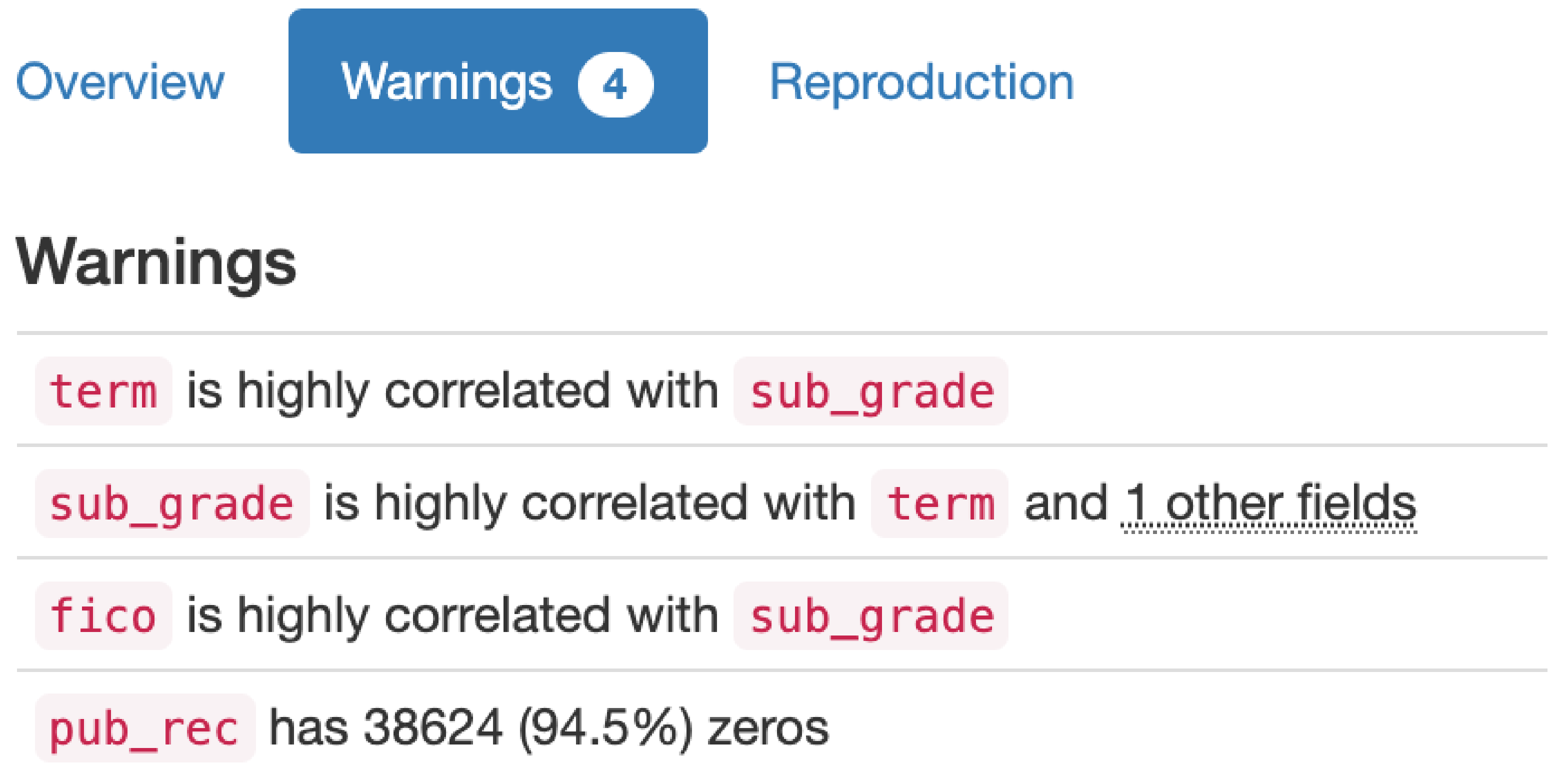
- 'defaulted' (Y/N)
- 'loan_amnt'
- 'term' (36 or 60 months)
- 'issue_d'
- ~~'int_rate'~~

02 Credit risk indicators (4)

- 'sub_grade'
- 'dti'
- 'fico'
- 'pub_rec'

03 Demographic variables (5)

- 'emp_length'
- 'home_ownership'
- 'verification_status'
- 'addr_state'
- 'annual_inc'



Warnings

term is highly correlated with sub_grade

sub_grade is highly correlated with term and 1 other fields

fico is highly correlated with sub_grade

pub_rec has 38624 (94.5%) zeros

train: 0.60
test: 0.56
AUC: 0.64
Recall score: 0.67

Log model 3

01 Loan characteristics (4)

- 'defaulted' (Y/N)
- 'loan_amnt'
- 'term' (36 or 60 months)
- 'issue_d'
- ~~'int_rate'~~

02 Credit risk indicators (3)

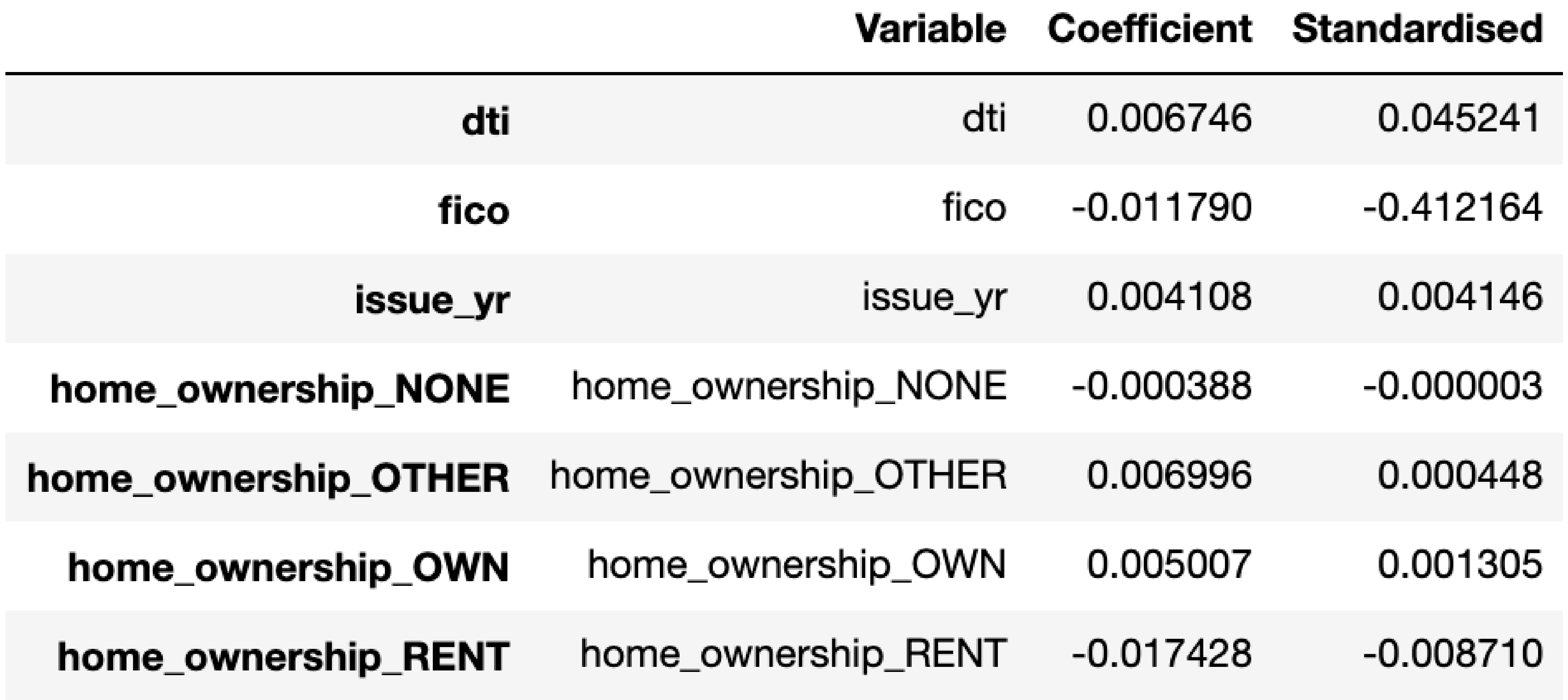
- 'sub_grade'
- 'dti'
- ~~'fico'~~
- 'pub_rec'

03 Demographic variables (5)

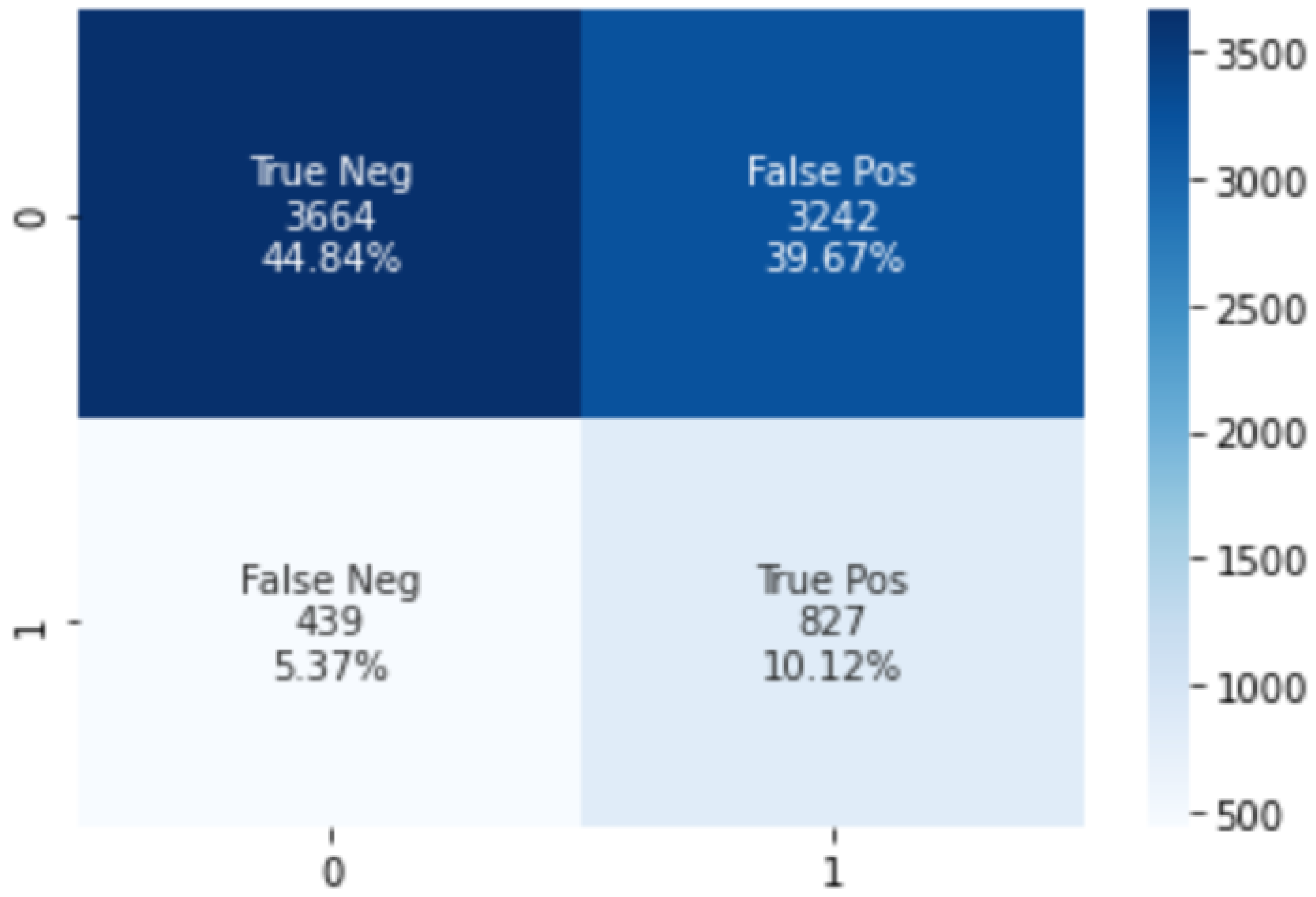
- 'emp_length'
- 'home_ownership'
- 'verification_status'
- 'addr_state'
- 'annual_inc'

train: 0.55
test: 0.56
AUC: 0.58
Recall score: 0.54

Log model final



	Variable	Coefficient	Standardised
dti	dti	0.006746	0.045241
fico	fico	-0.011790	-0.412164
issue_yr	issue_yr	0.004108	0.004146
home_ownership_NONE	home_ownership_NONE	-0.000388	-0.000003
home_ownership_OTHER	home_ownership_OTHER	0.006996	0.000448
home_ownership_OWN	home_ownership_OWN	0.005007	0.001305
home_ownership_RENT	home_ownership_RENT	-0.017428	-0.008710



train: 0.58

test: 0.55

AUC: 0.62

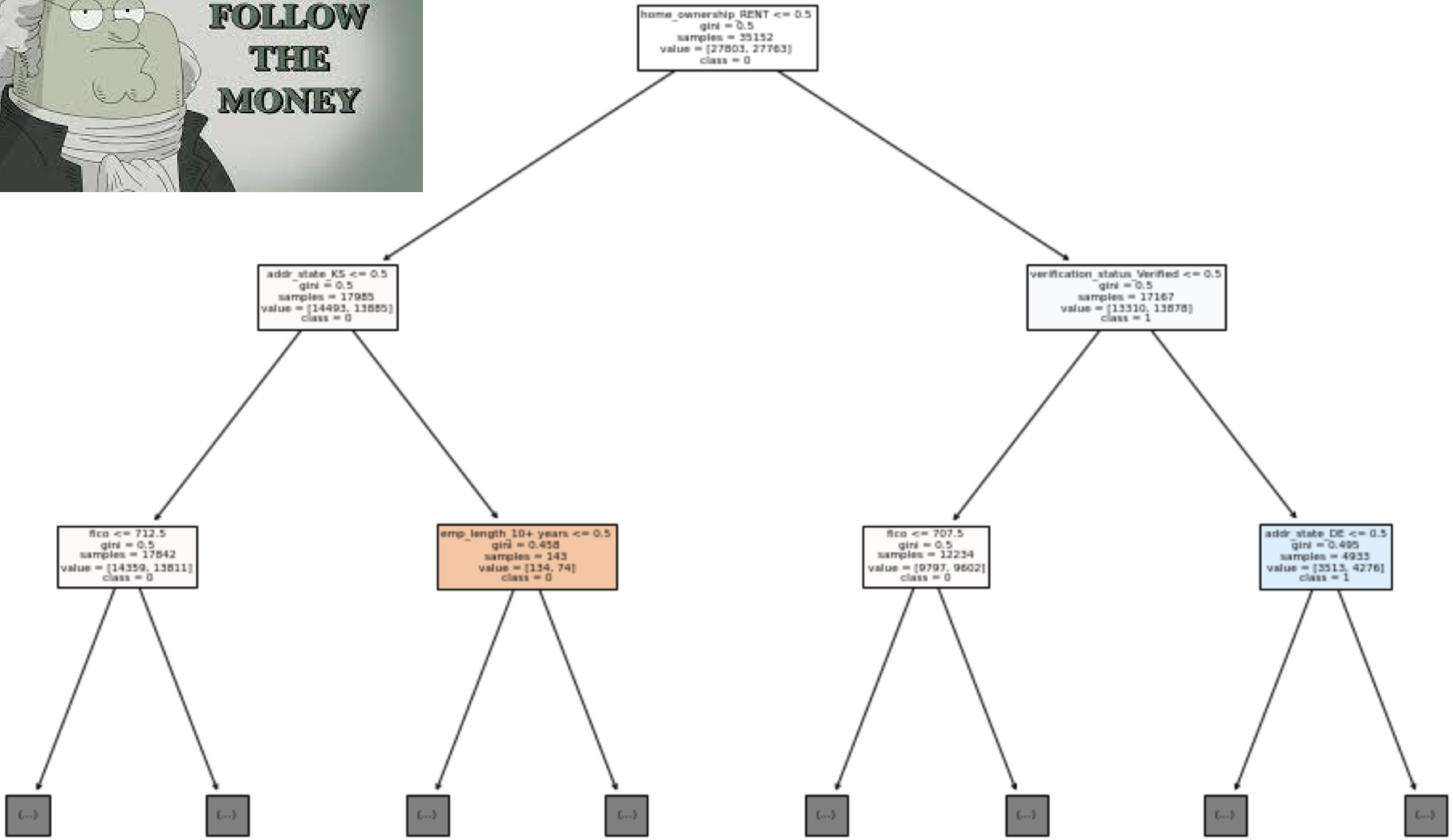
Recall score: 0.65

$$Recall = \frac{TP}{TP + FN}$$

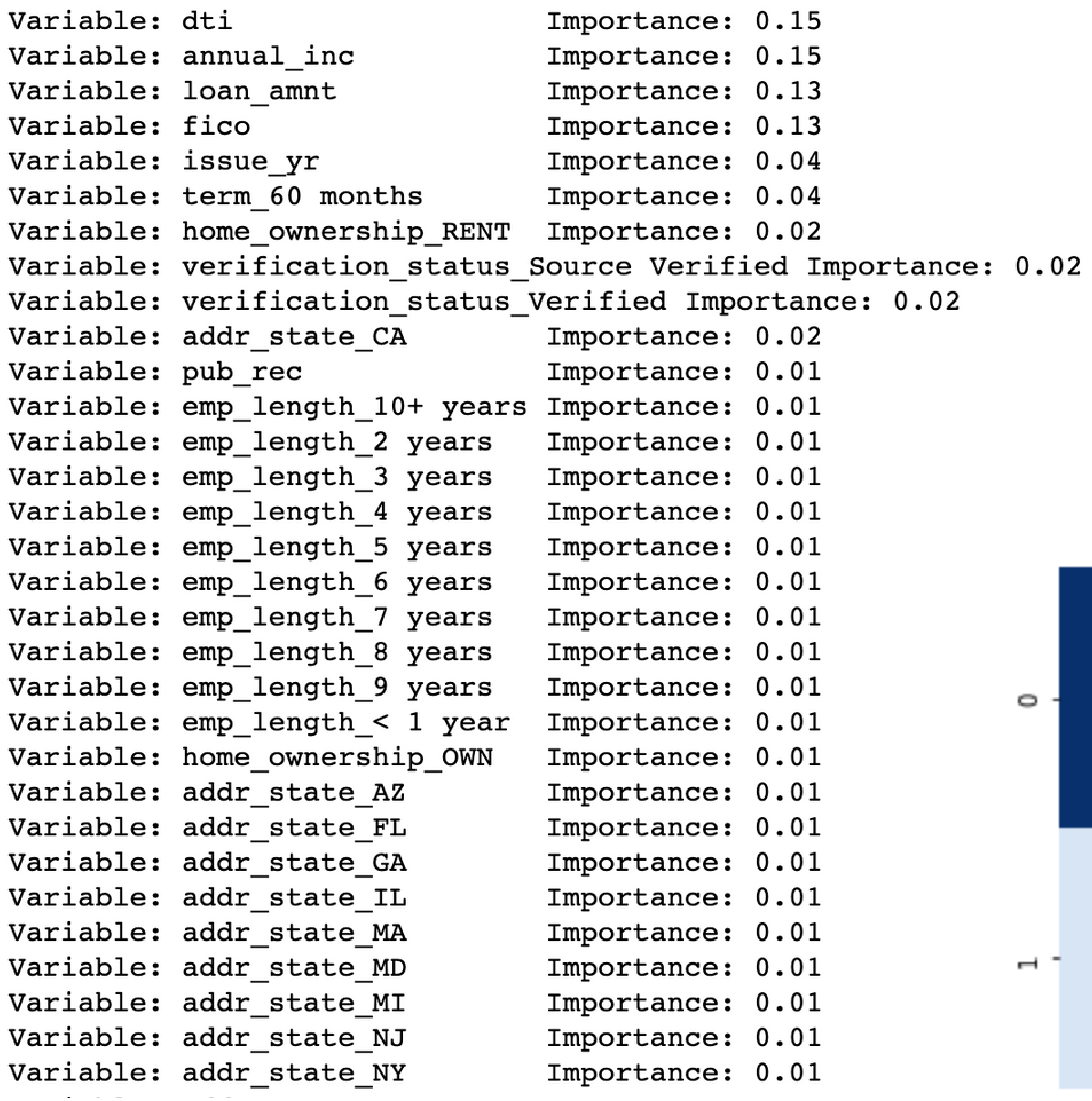
“Neither a borrower nor a lender be; for loan doth oft lose both itself and friend, and borrowing dulls the edge of husbandry.” Hamlet

Polonius gives his son Laertes advice on managing money.

Random Forest



Random Forest



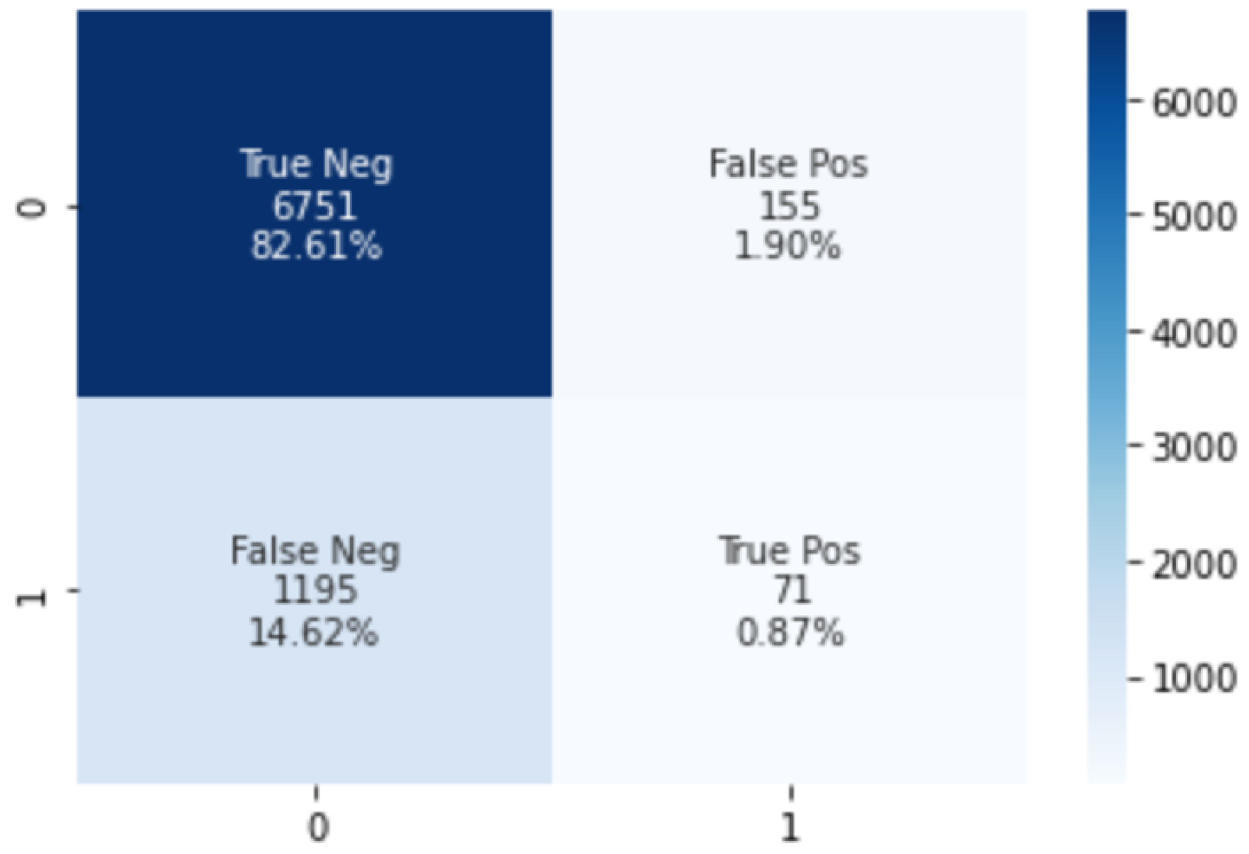
```
Variable: dti                  Importance: 0.15
Variable: annual_inc           Importance: 0.15
Variable: loan_amnt            Importance: 0.13
Variable: fico                 Importance: 0.13
Variable: issue_yr             Importance: 0.04
Variable: term_60 months       Importance: 0.04
Variable: home_ownership_RENT  Importance: 0.02
Variable: verification_status_Source Verified Importance: 0.02
Variable: verification_status_Verified Importance: 0.02
Variable: addr_state_CA        Importance: 0.02
Variable: pub_rec              Importance: 0.01
Variable: emp_length_10+ years Importance: 0.01
Variable: emp_length_2 years   Importance: 0.01
Variable: emp_length_3 years   Importance: 0.01
Variable: emp_length_4 years   Importance: 0.01
Variable: emp_length_5 years   Importance: 0.01
Variable: emp_length_6 years   Importance: 0.01
Variable: emp_length_7 years   Importance: 0.01
Variable: emp_length_8 years   Importance: 0.01
Variable: emp_length_9 years   Importance: 0.01
Variable: emp_length_< 1 year  Importance: 0.01
Variable: home_ownership_OWN   Importance: 0.01
Variable: addr_state_AZ        Importance: 0.01
Variable: addr_state_FL        Importance: 0.01
Variable: addr_state_GA        Importance: 0.01
Variable: addr_state_IL        Importance: 0.01
Variable: addr_state_MA        Importance: 0.01
Variable: addr_state_MD        Importance: 0.01
Variable: addr_state_MI        Importance: 0.01
Variable: addr_state_NJ        Importance: 0.01
Variable: addr_state_NY        Importance: 0.01
```

test 0.83

AUC 0.64

Accuracy 0.83

Recall 0.06



Who is likely to default?

People with low fico scores and high dti

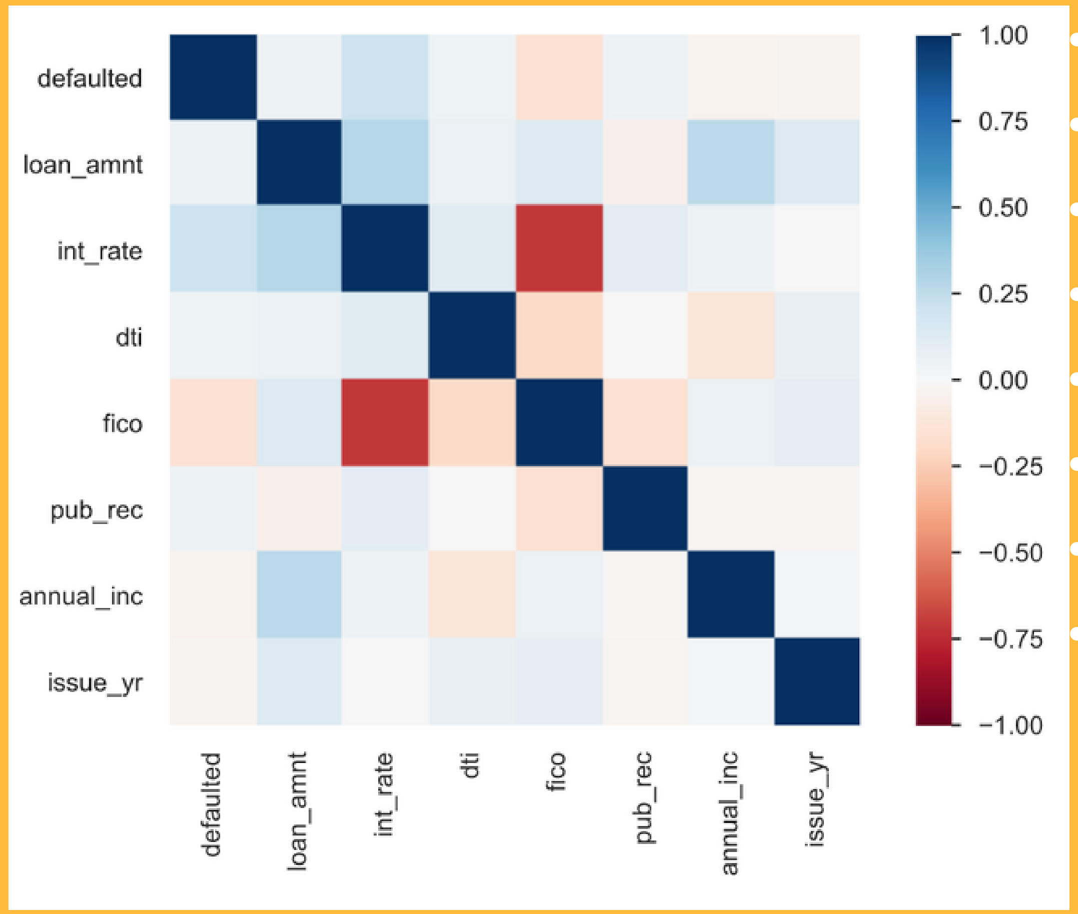
Who should we lend to? People with high fico scores and high income?

Not necessarily!

Lending considerations

expected loss =
exposure at default *
default probability *
(1-recovery rate)

- income goals
- risk appetite
- diversification



Challenges and solutions

Large data set

Comprises 114 columns and 42,325 rows.
Cleaning
Feature engineering (dates, loan_status)

Correlated variables

Some variables were correlated (e.g. int rate and fico) or were skewed (e.g. annual income) so some variables were dropped or transformed.

Biased data

Upsampling

Enhancements given more time

More demographic data

While there may be ethical considerations, the dataset is notable for missing data on age and gender. There may be further demographic variables to obtain which may explain loan default behaviour.

More diagnostic checks

Lessons for the future

Time for questions