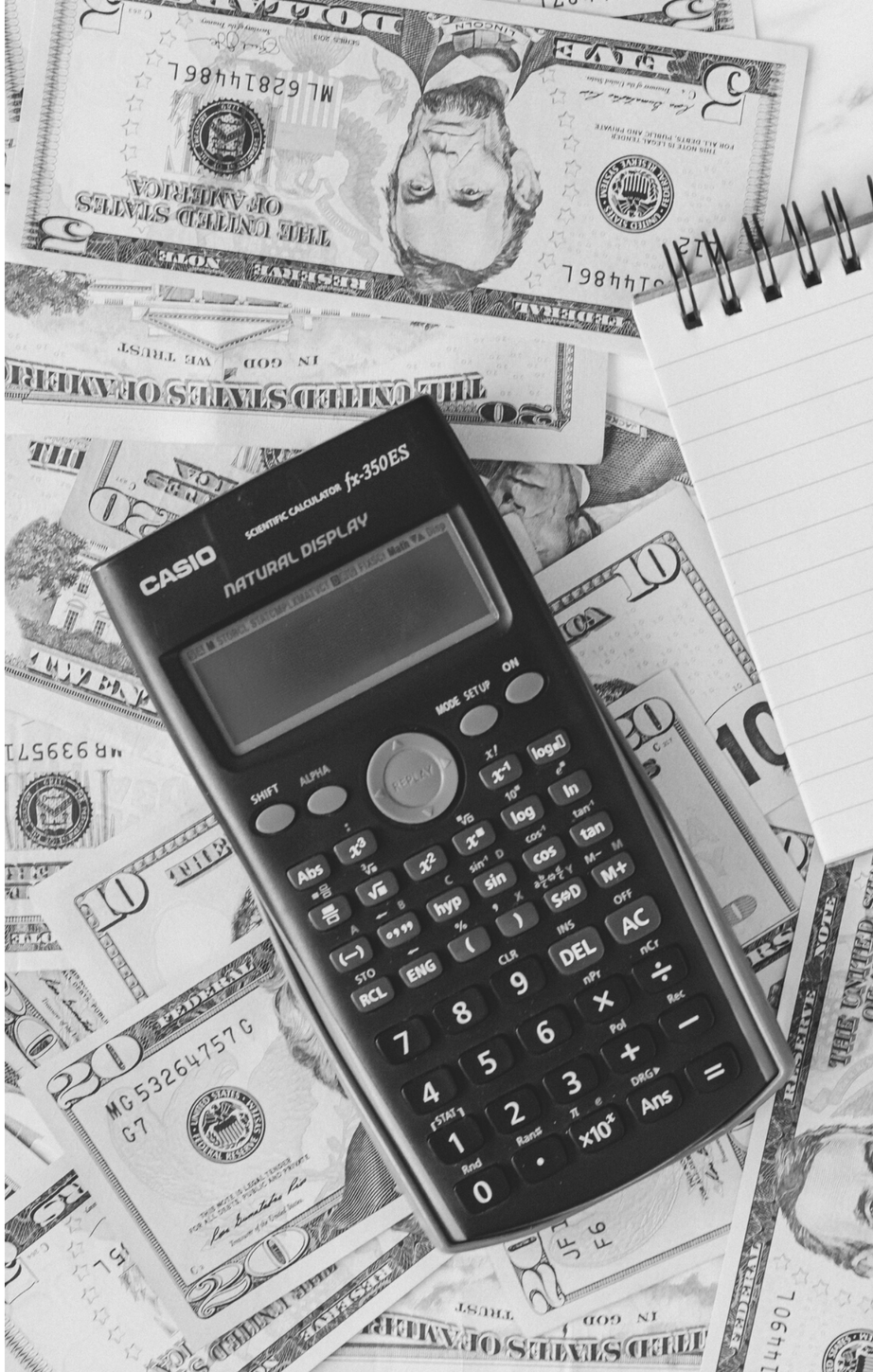
CREDIT CARD FRAUD DETECTION

Demi, Kam, Hania

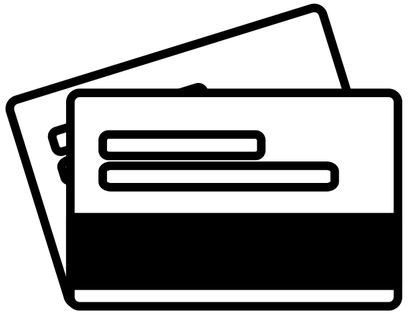
BOOTCAMP PROJECT

TODAY'S AGENDA

- Problem Statement
- Data Exploration
- Important Features
- Promising Models
- Conclusions
- Recommendations for Business



Mark



PROBLEM STATEMENT

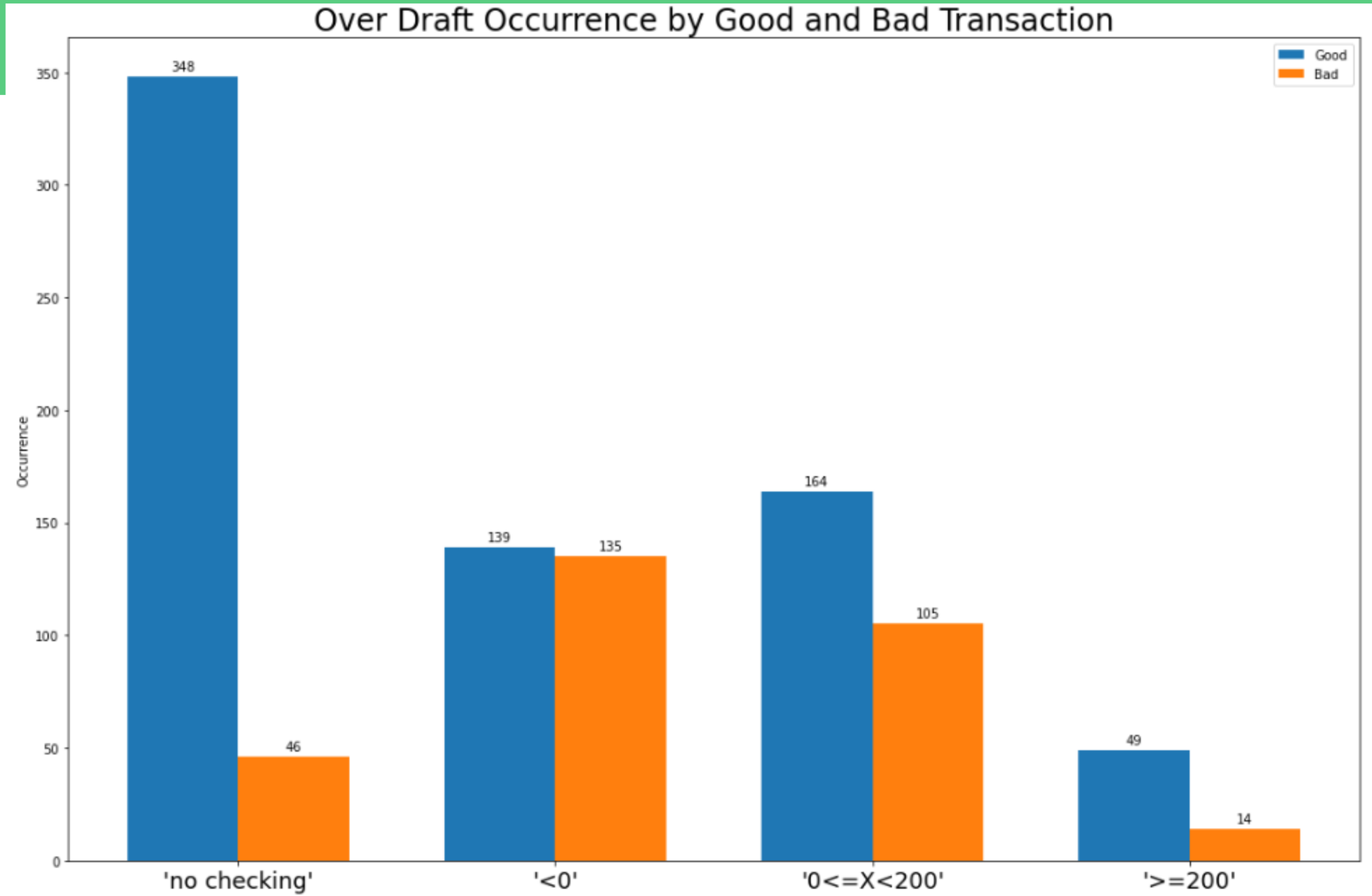
- Predict fraudulent transactions
- German Credit Risk dataset

DATA EXPLORATION

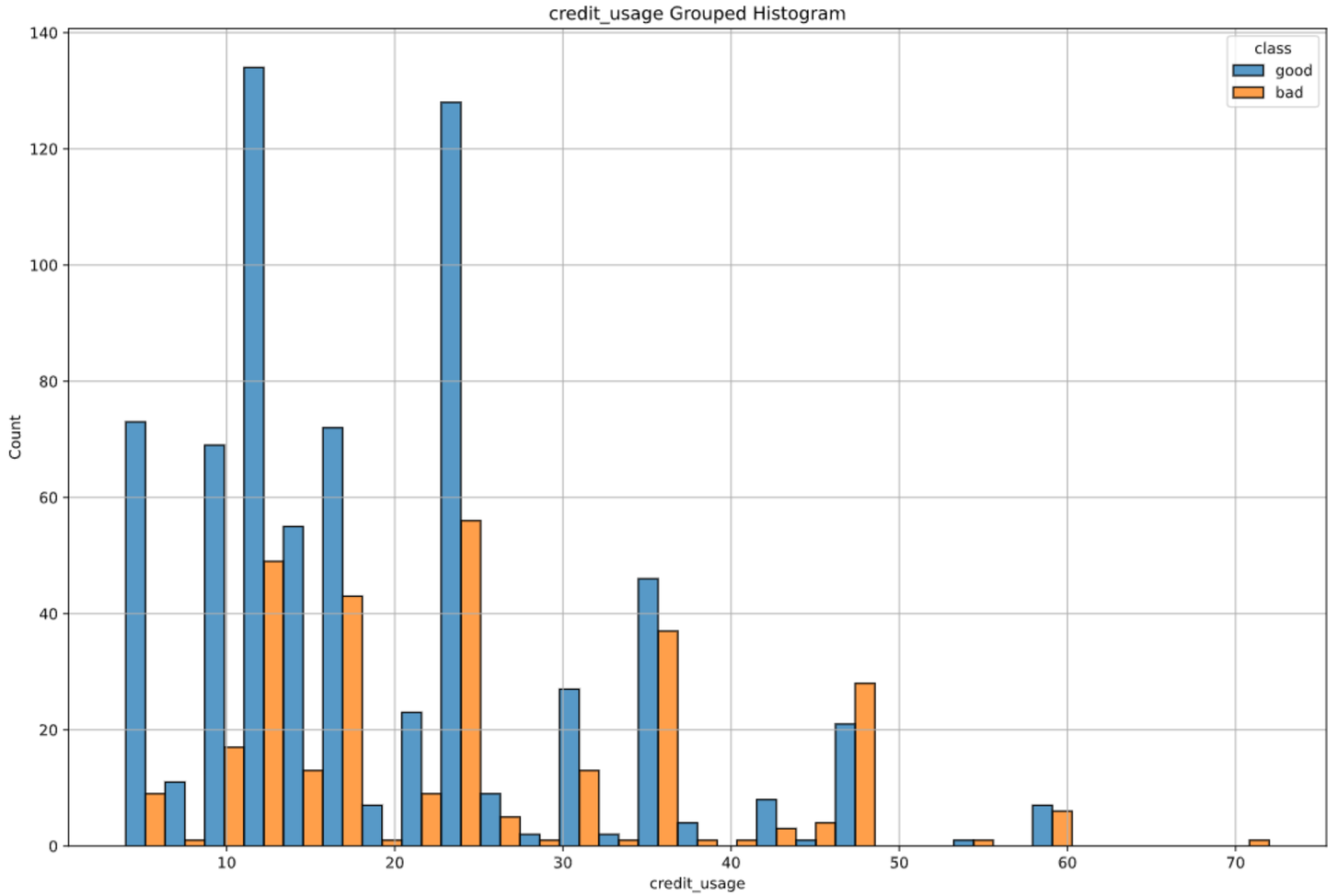
FEATURES IN THE DATA SET

	features	n_values
0	over_draft	['<0', '0<=X<200', 'no checking', '>=200']
1	credit_usage	[6, 48, 12, 42, 24, 36, 30, 15, 9, 10, 7, 60, ...
2	credit_history	['critical/other existing credit', 'existing p...
3	purpose	[radio/tv, education, furniture/equipment, 'ne...
4	current_balance	[1169, 5951, 2096, 7882, 4870, 9055, 2835, 694...
5	Average_Credit_Balance	['no known savings', '<100', '500<=X<1000', '>...
6	employment	['>=7', '1<=X<4', '4<=X<7', unemployed, '<1']
7	location	[4, 2, 3, 1]
8	personal_status	['male single', 'female div/dep/mar', 'male di...
9	other_parties	[none, guarantor, 'co applicant']
10	residence_since	[4, 2, 3, 1]
11	property_magnitude	['real estate', 'life insurance', 'no known pr...
12	cc_age	[67, 22, 49, 45, 53, 35, 61, 28, 25, 24, 60, 3...
13	other_payment_plans	[none, bank, stores]
14	housing	[own, 'for free', rent]
15	existing_credits	[2, 1, 3, 4]
16	job	[skilled, 'unskilled resident', 'high qualif/s...
17	num_dependents	[1, 2]
18	own_telephone	[yes, none]
19	foreign_worker	[yes, no]
20	class	[good, bad]

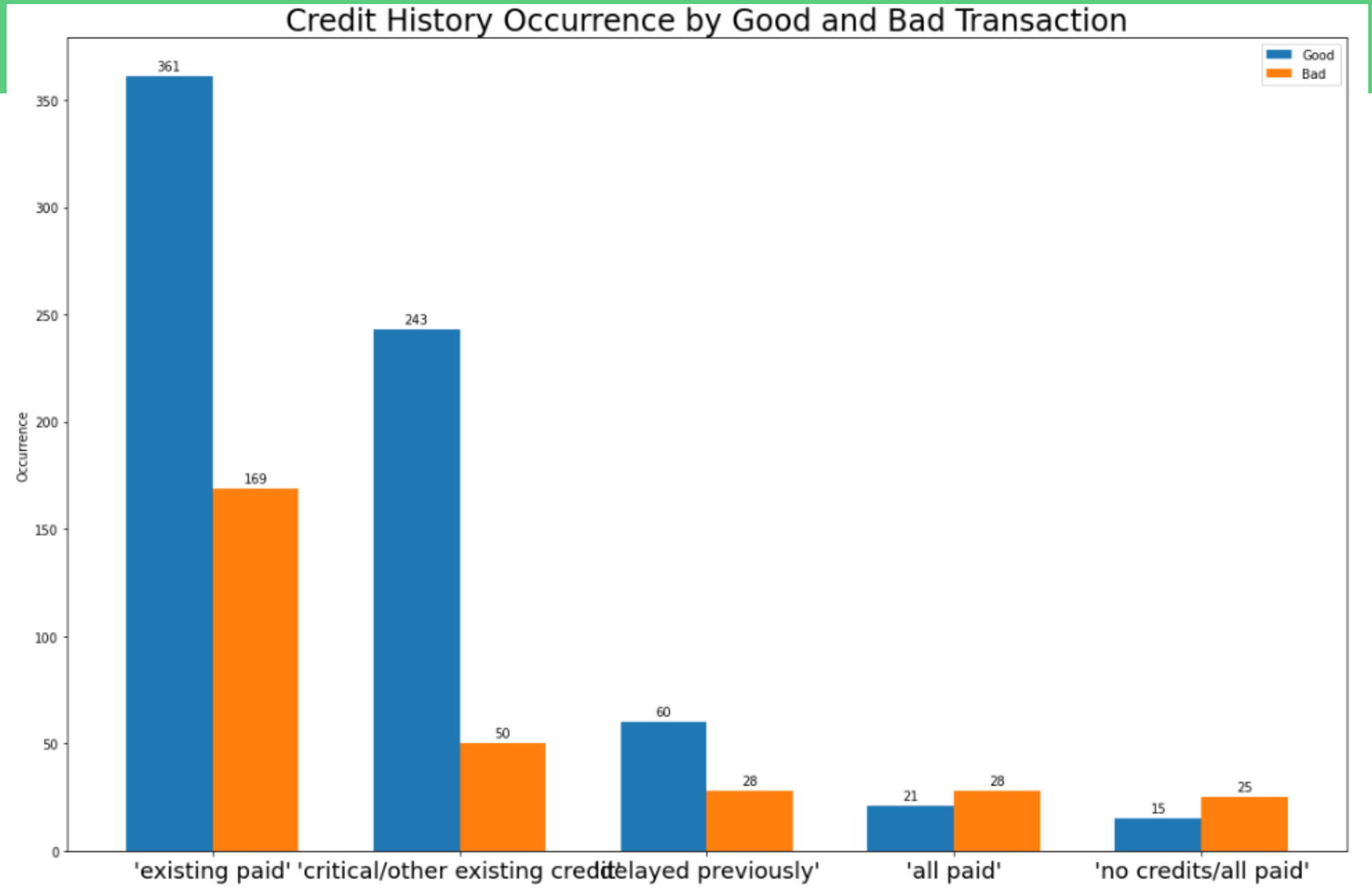
OVER DRAFT


Over Draft Occurrence by Good and Bad Transaction
Good
Bad
Occurrence
0
50
100
150
200
250
300
350
348
46
139
135
164
105
49
14
'no checking'
'<0'
'0<=X<200'
'>=200'

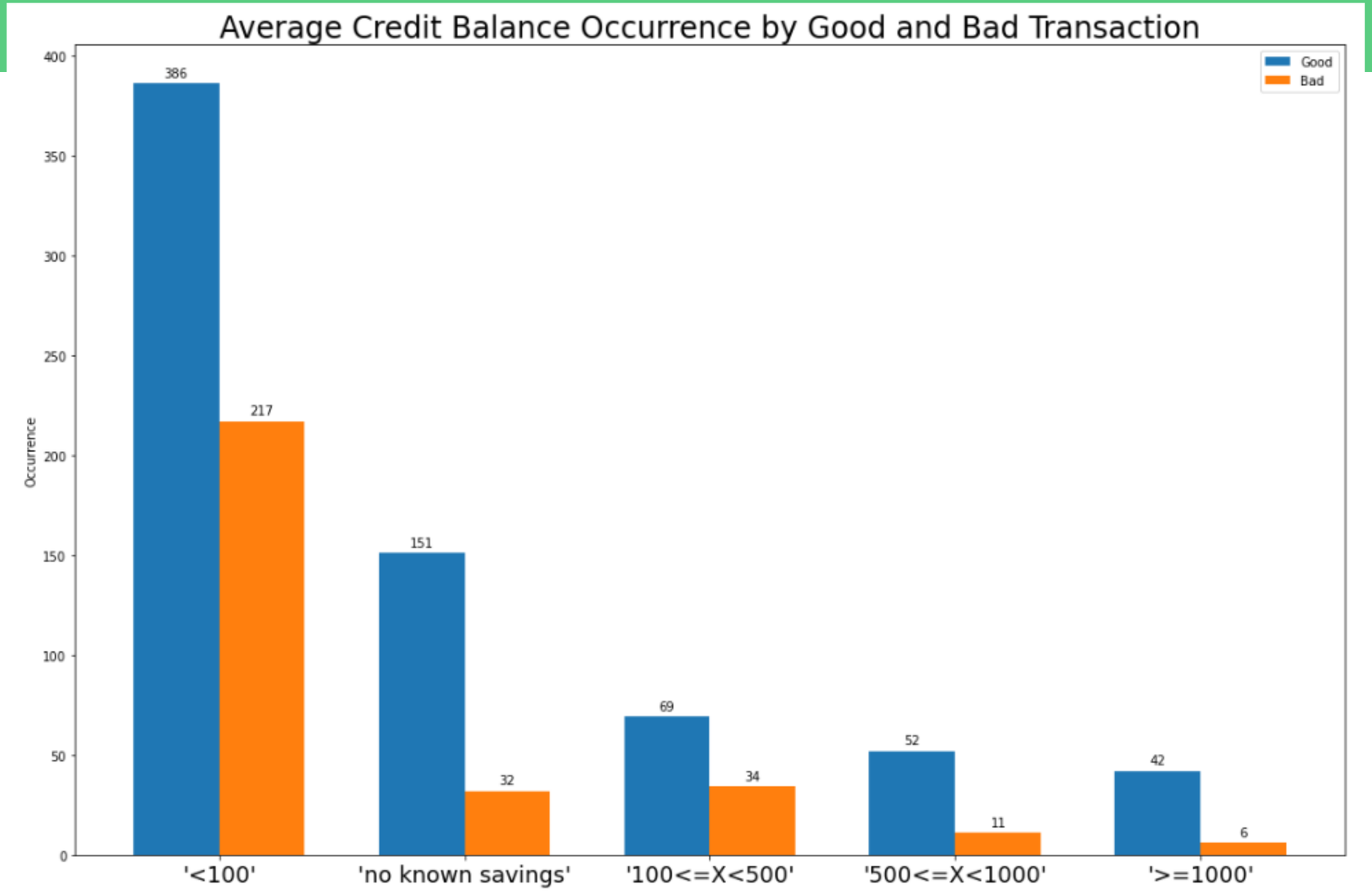
CREDIT USAGE



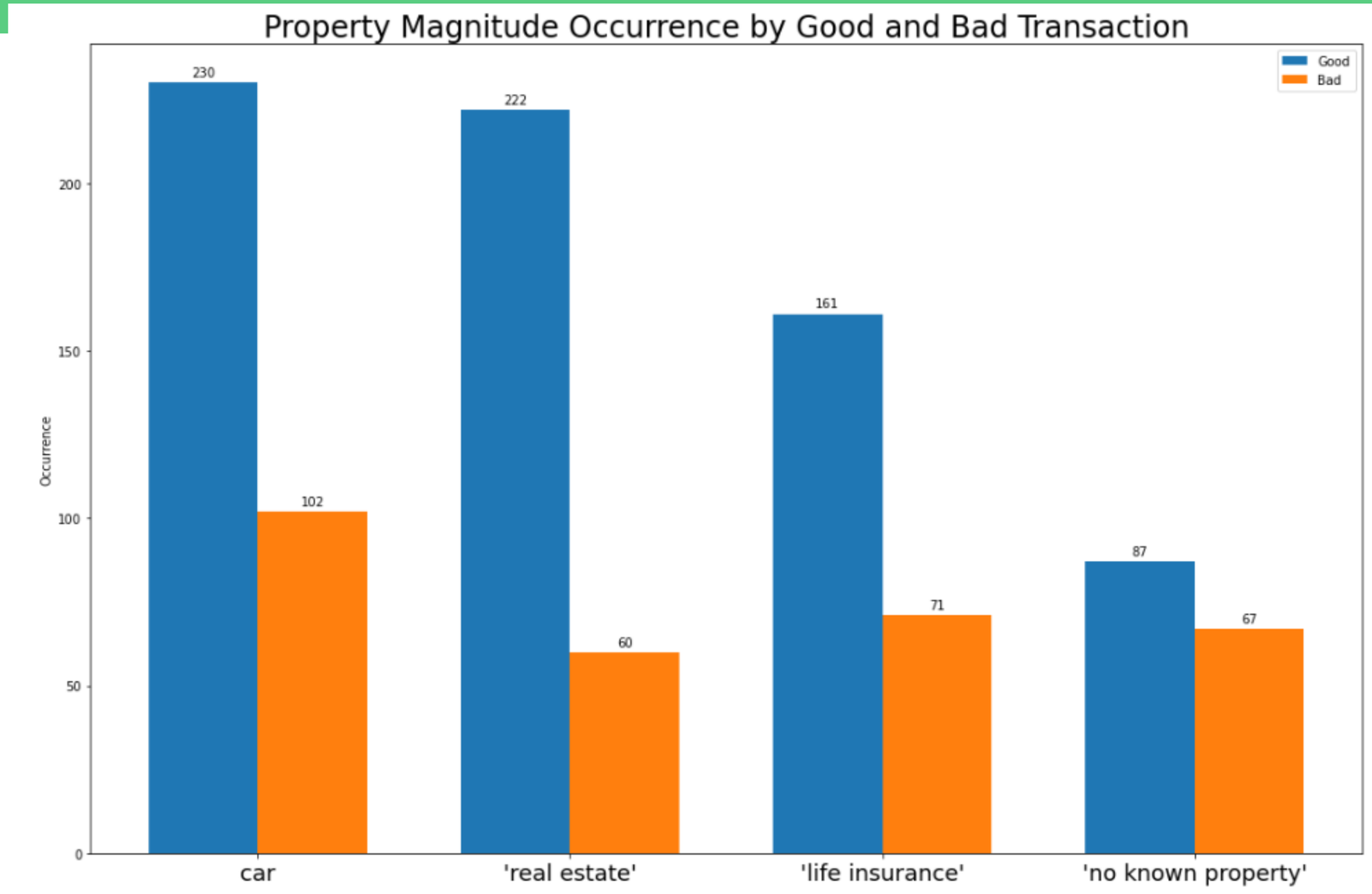
CREDIT HISTORY



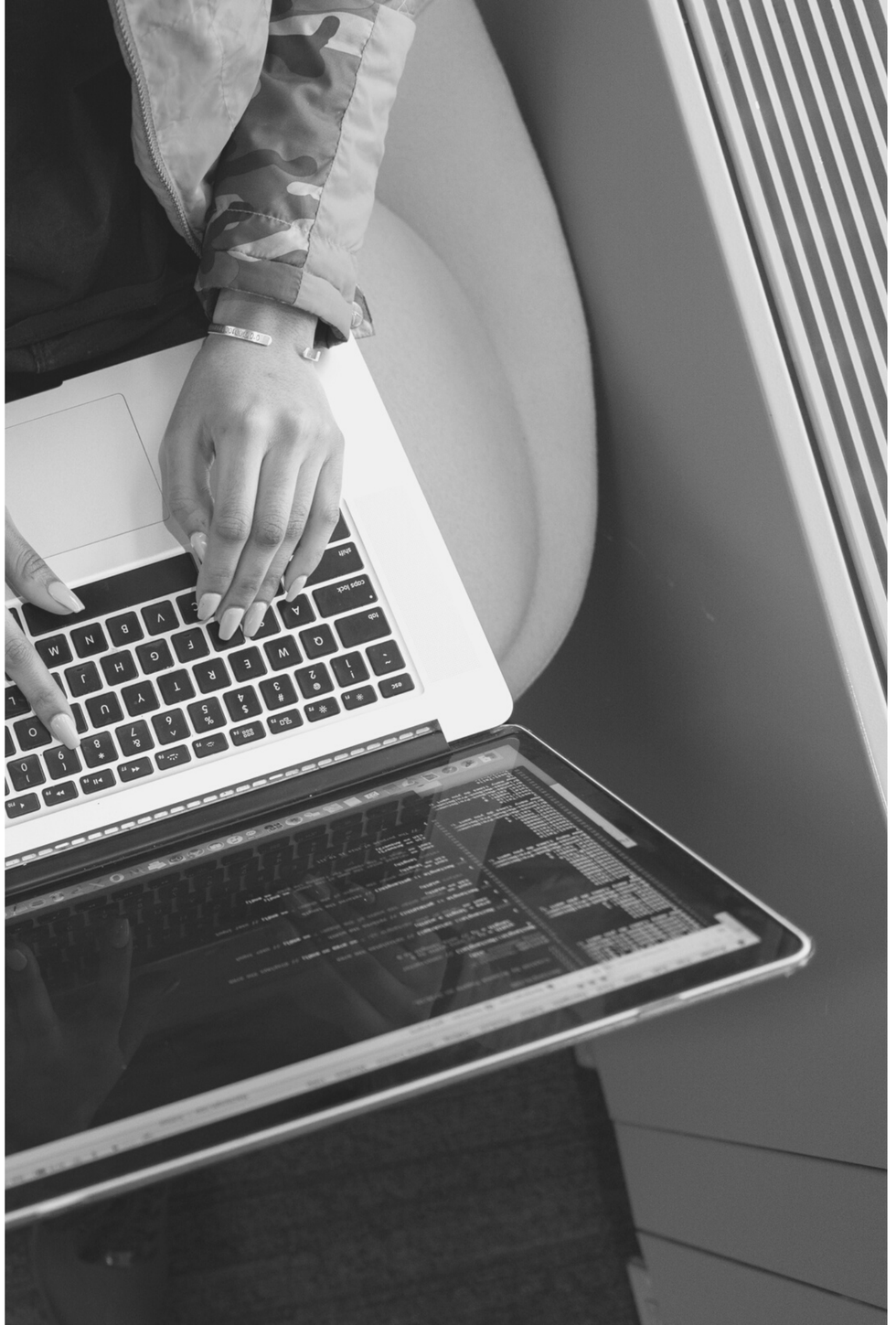
AVERAGE CREDIT BALANCE



PROPERTY MAGNITUDE



WHICH FEATURES ARE IMPORTANT?

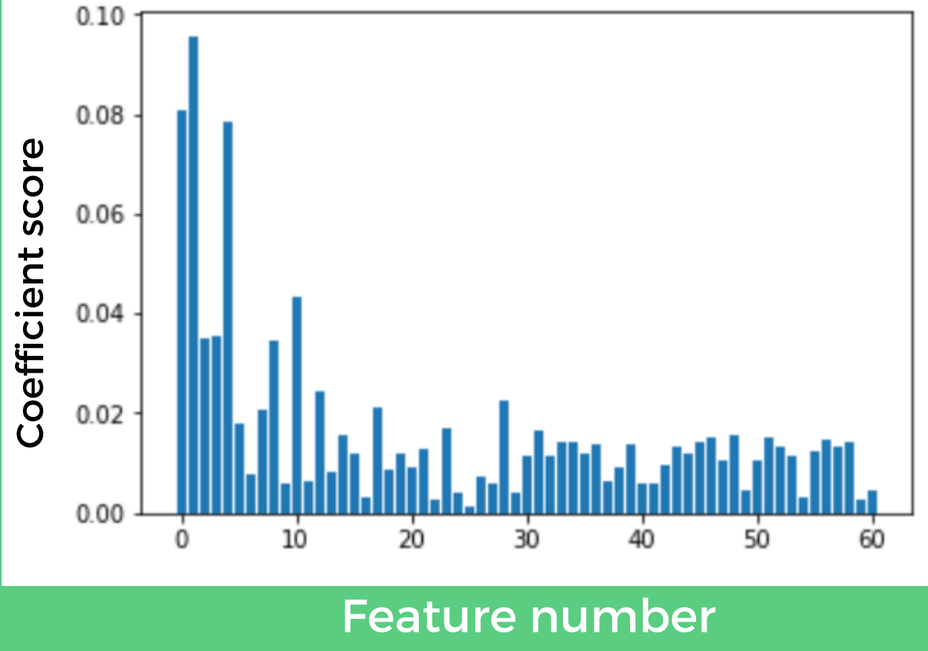


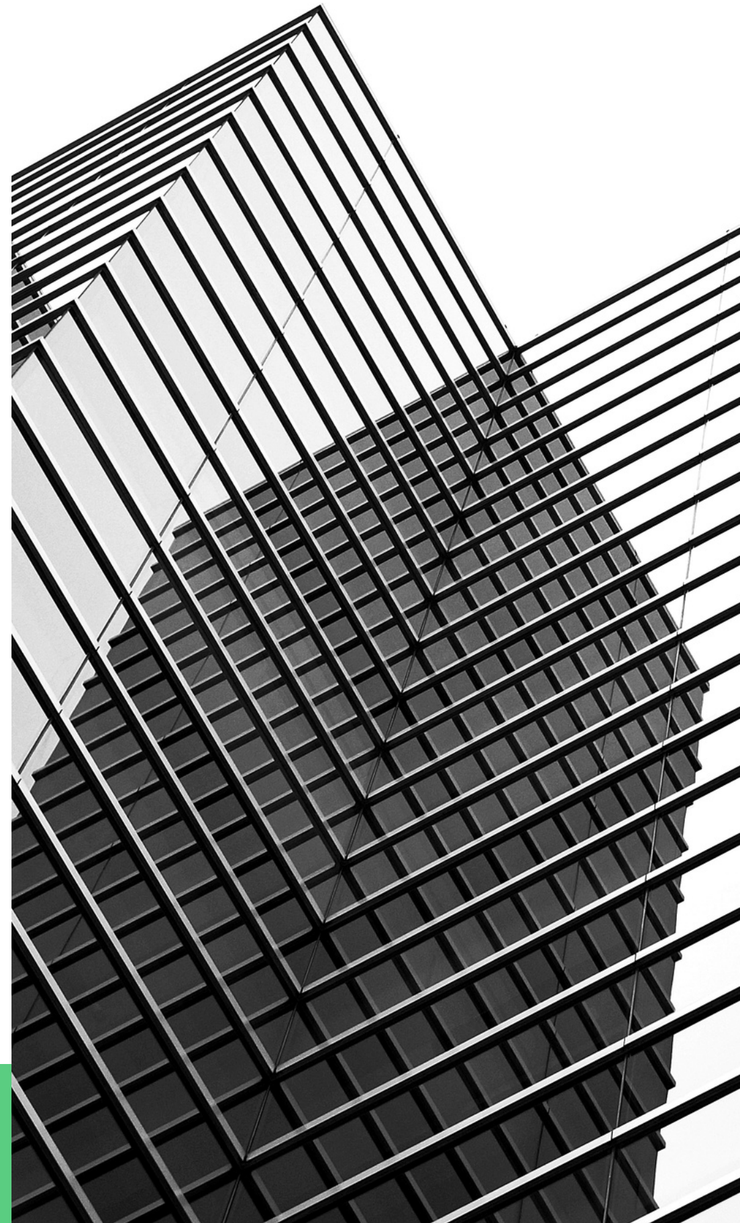
CHECKING IMPORTANCE MANUALLY

Is the feature making the model score better?

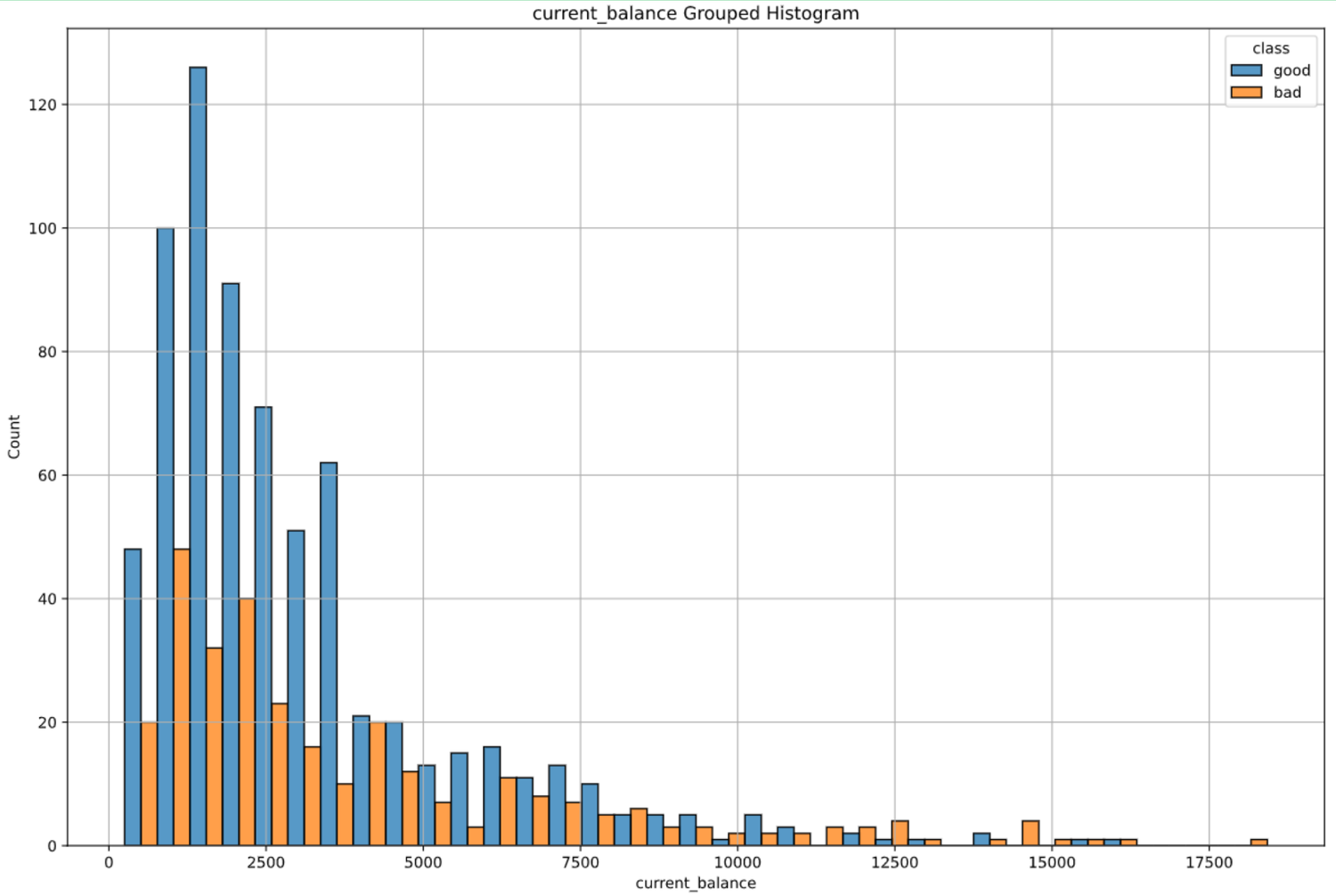
RANDOM-FOREST COEFFICIENT

Which features have the highest coefficient score?

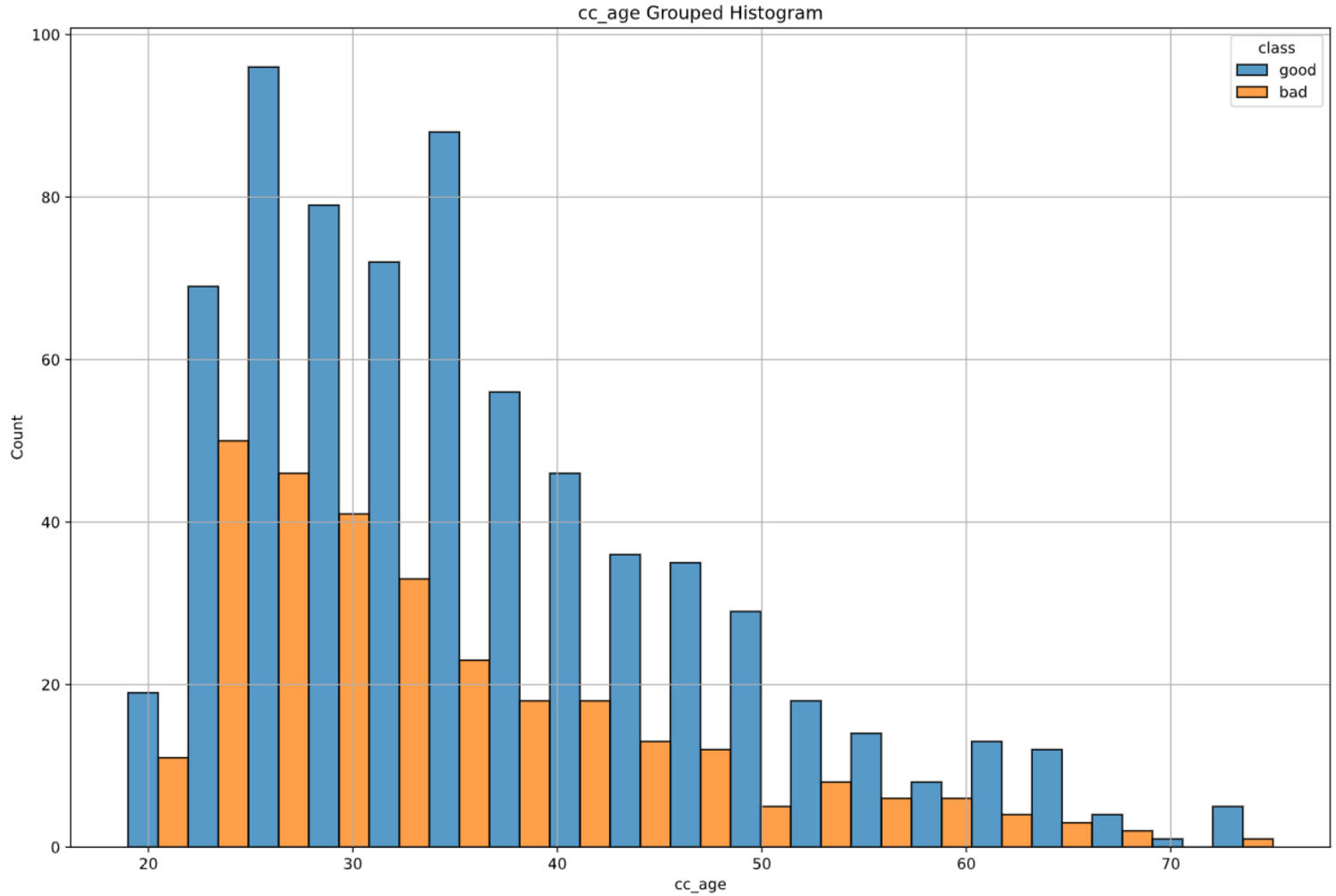


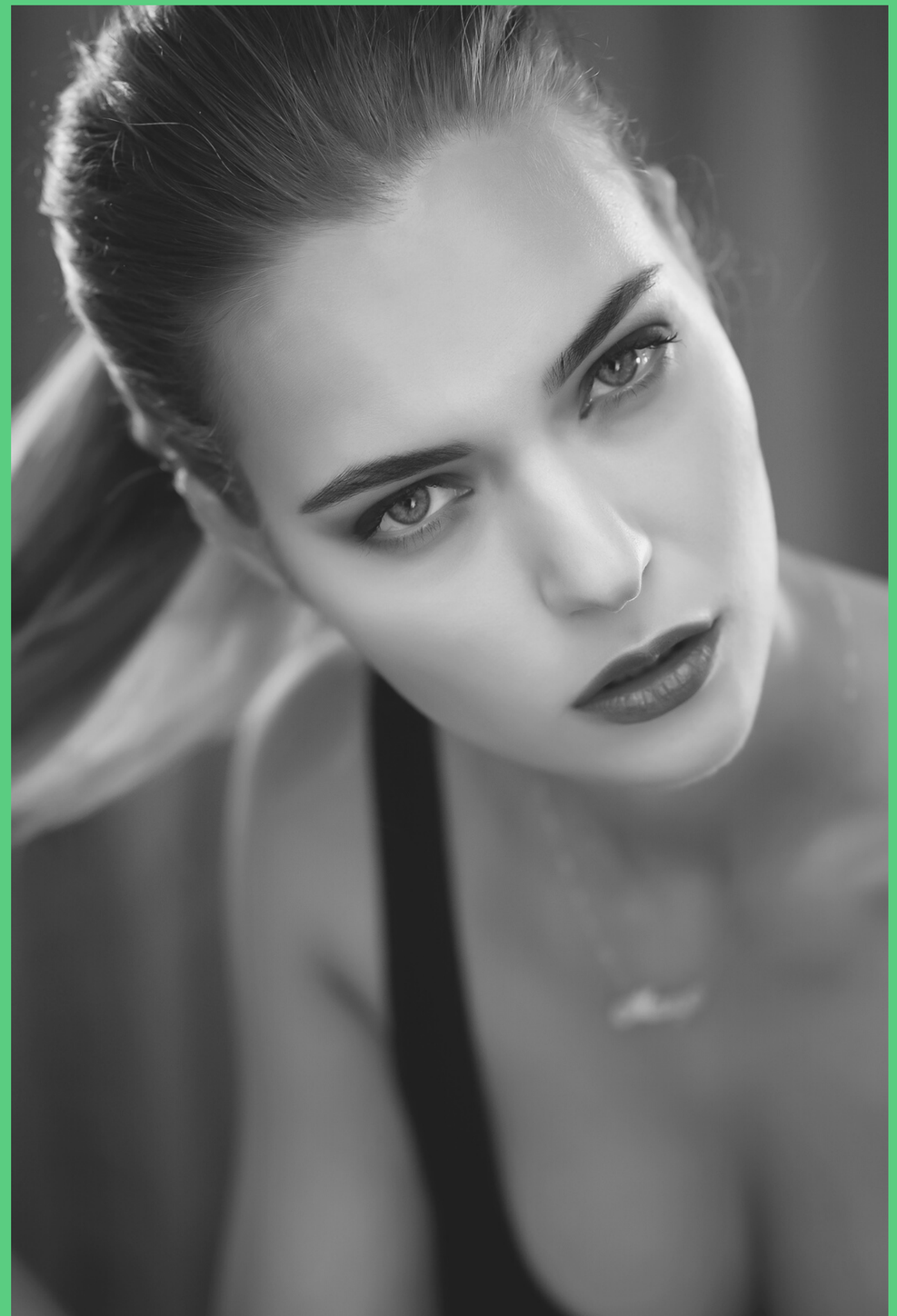


CURRENT BALANCE



AGE


cc_age Grouped Histogram
class
good
bad
Count
cc_age
0
20
40
60
80
100
20
30
40
50
60
70



PROMISING MODELS

WHAT MODELS DID WE TRY?

... we hear you asking

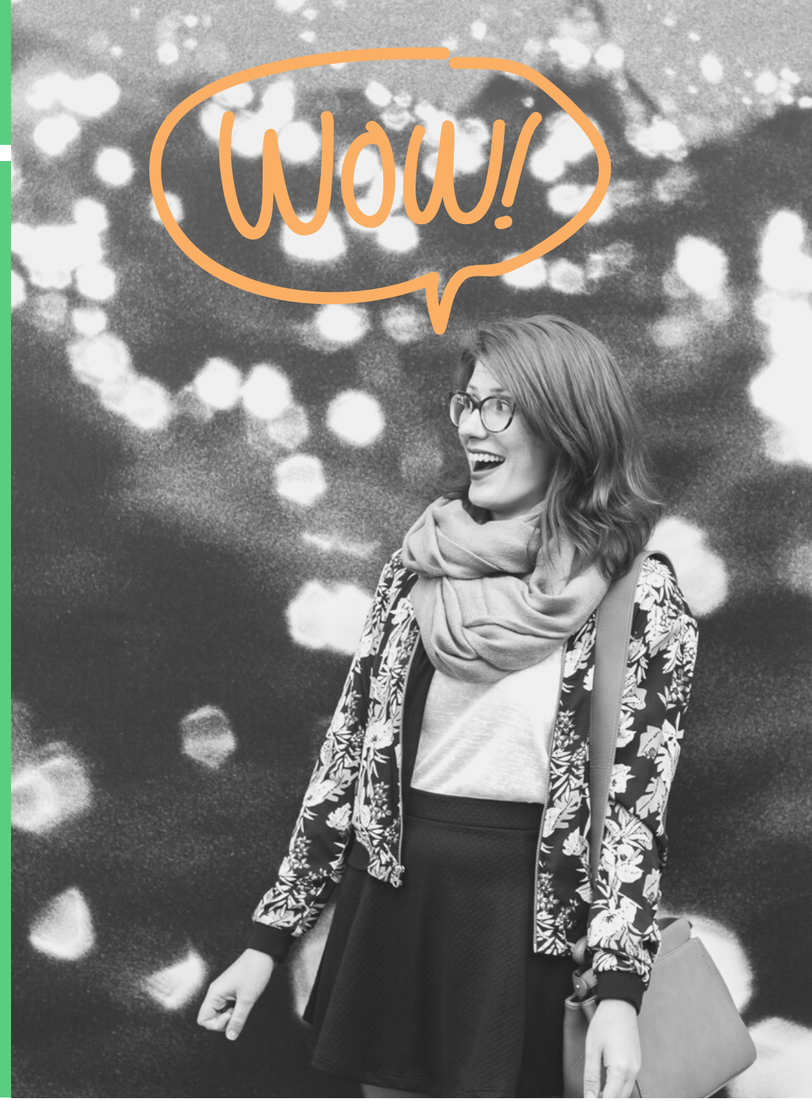
- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machines
- Unsupervised Learning

	Random Forest	Logistic Regression	SVM
Original Training Score	100%	78.3%	97%
Original Test Score	74%	76.8%	70.1%
Final Training Score	85.6%	78.3%	70%
Final Test Score	75.2%	76.8%	70%

CONCLUSIONS

THE BEST MODEL WITH THE SCORE OF

76.80%



LOGISTIC REGRESSION

Should Mark use our model?

Predicted	0	1	All
True			
0	74	76	150
1	40	310	350
All	114	386	500

1 - good transactions
0 - bad transactions

RECOMMENDATIONS

ENSEMBLING

NEURAL NETWORKS

UNSUPERVISED LEARNING ANOMALY DETECTION

UPDATING DATA

- own telephone
- location
- more data

ADDING MEANINGFUL FEATURES

- Timing of transactions.

HIRING MORE EXPERIENCED PEOPLE TO WORK AT YOUR BANK

THANKS FOR LISTENING!

QUESTIONS?

Hanna Sarnecka
Demi Kuit
Kamalen Reddy