MSTW'22 Hackathon

S3S2'19

Dexter, Dongheng, Jason, Ze Li, Ming Roong

Team Biography

Dexter Woo Teng Koon
- The Chinese University of Hong Kong
- BSc in Quantitative Finance and Risk Management Science, Minor in Mathematics and Statistics

Lee Dongheng
- Nanyang Technological University, Singapore
- BSc in Mathematical Sciences with a Minor in Finance

Lee Jason
- Asia Pacific University of Technology & Innovation (APU)
- BSc (Hons.) in Computer Science with specialism in data analytics

Liew Ze Li
- Monash University
- BEng (Hons.) in Chemical Engineering

Ku Ming Roong
- UOW Malaysia KDU University College
- BSc (Hons.) in Computer Science

Presentation Outline

1. Our Business Case and Use Cases
2. Exploratory Data Analysis
3. Data Cleaning
4. Data Pre-Processing
5. Machine Learning: Clustering and Classification Models
6. Conclusion

Presentation Outline

1. Our Business Case and Use Cases
2. Exploratory Data Analysis
3. Data Cleaning
4. Data Pre-Processing
5. Machine Learning: Clustering and Classification Models
6. Conclusion

1. Our Case: Profiling Customers Through Credit Risk Assessment

Dataset

- South German Credit (UPDATE) Data Set by Prof Ulrike Grömping, from UCI Machine Learning Repository [1]
- 1000 instances with 21 attributes, all values are real integers
- For classification, regression, and clustering

Background

- Who we are: German bank
- Past dataset of credit applications
- Customer segmentation and predictive models are important in credit risk assessment

[1] https://archive.ics.uci.edu/ml/datasets/South+German+Credit+%28UPDATE%29#

1. Our Case: Profiling Customers Through Credit Risk Assessment (con't)

Data Dictionary

Column	Variable name	Description	Type
laufkont	status	status of the debtor's checking account with the bank	categorical (ordinal)
laufzeit	duration	credit duration in months	numerical (discrete)
moral	credit_history	history of compliance with previous or concurrent credit contracts	categorical (ordinal)
verw	purpose	purpose for which the credit is needed	categorical (nominal)
hoehe	amount	credit amount in DM	numerical (discrete)
sparkont	savings	debtor's savings	categorical (ordinal)
beszeit	employment_duration	duration of debtor's employment with current employer	categorical (ordinal)
rate	installment_rate	credit installments as a percentage of debtor's disposable income	categorical (ordinal)
famges	personal_status_sex	combined information on sex and marital status	categorical (nominal)
buerge	other_debtors	Is there another debtor or a guarantor for the credit?	categorical (nominal)
wohnzeit	present_residence	length of time (in years) the debtor lives in the present residence	categorical (ordinal)
verm	property	the debtor's most valuable property, i.e. the highest possible code is used	categorical (ordinal)
alter	age	age in years	numerical (discrete)
weitkred	other_installment_plans	installment plans from providers other than the credit-giving bank	categorical (ordinal)
wohn	housing	type of housing the debtor lives in	categorical (ordinal)
bishkred	number_credits	number of credits including the current one the debtor has (or had) at this bank	categorical (ordinal)
beruf	job	quality of debtor's job	categorical (ordinal)
pers	people_liable	number of persons who financially depend on the debtor	categorical (ordinal)
telef	telephone	Is there a telephone landline registered on the debtor's name?	categorical (ordinal)
gastarb	foreign_worker	Is the debtor a foreign worker?	categorical (ordinal)
kredit	credit_risk	Has the credit contract been complied with (good) or not (bad)?	categorical (ordinal)

Given The Dataset, We Have Two Use Cases:

1. Using our clustering model, we can identify clusters from input data based on some common features.

Cluster 1

Bad Credit Risk

Cluster 2

Good Credit Risk

2. Using our classification model, we can estimate the probability of a particular customer having a good credit risk.

Customer 1: Male, 35 years old, Car Loan
Customer 2: Male, 20 years old, Apartment Loan

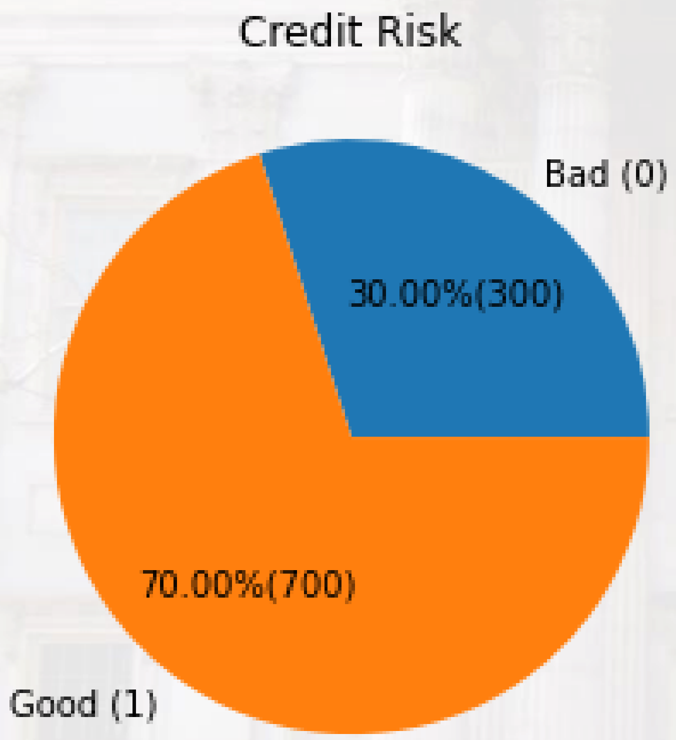
→

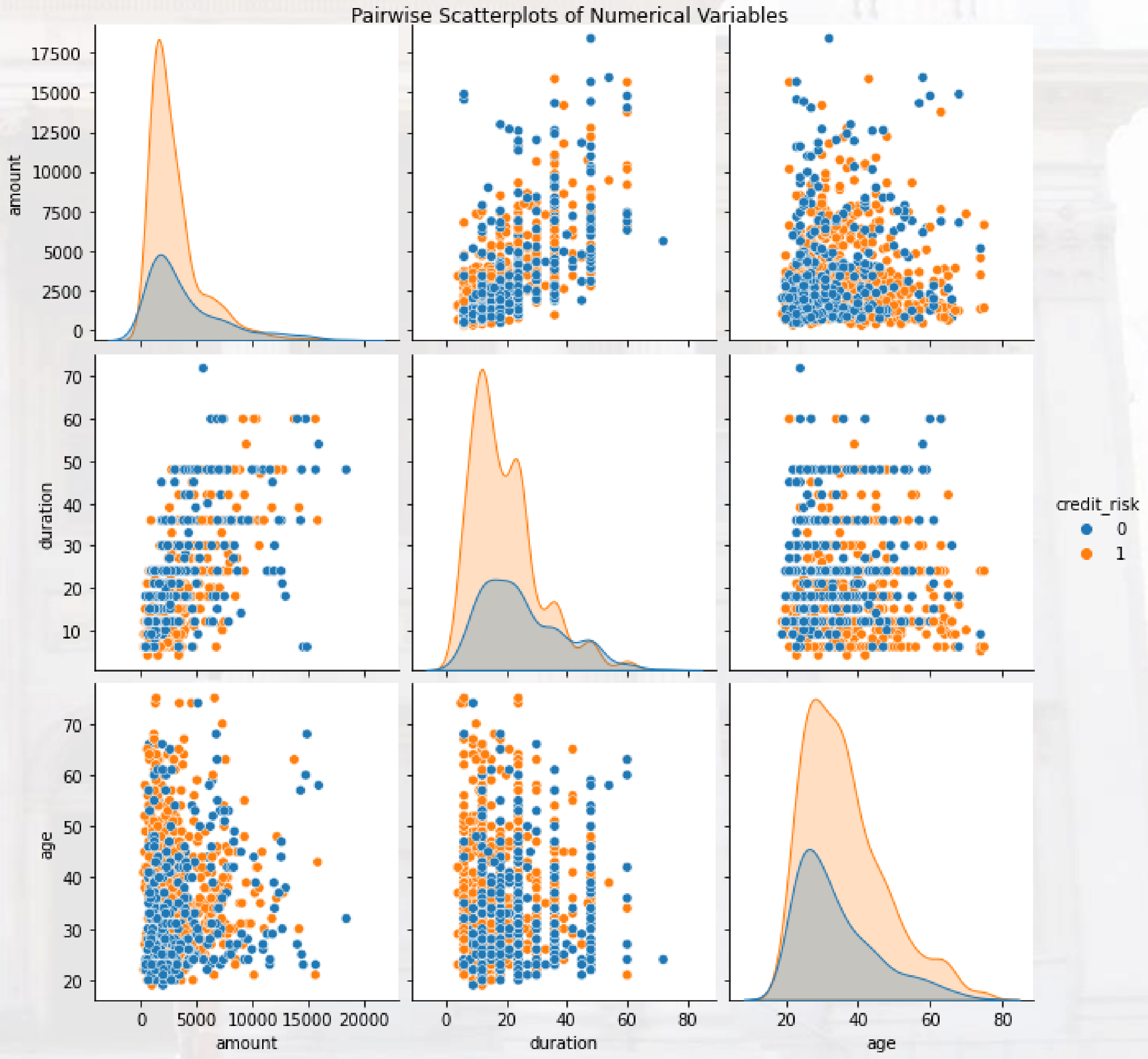
Customer 1: 80% good credit risk
Customer 2: 30% good credit risk

Presentation Outline

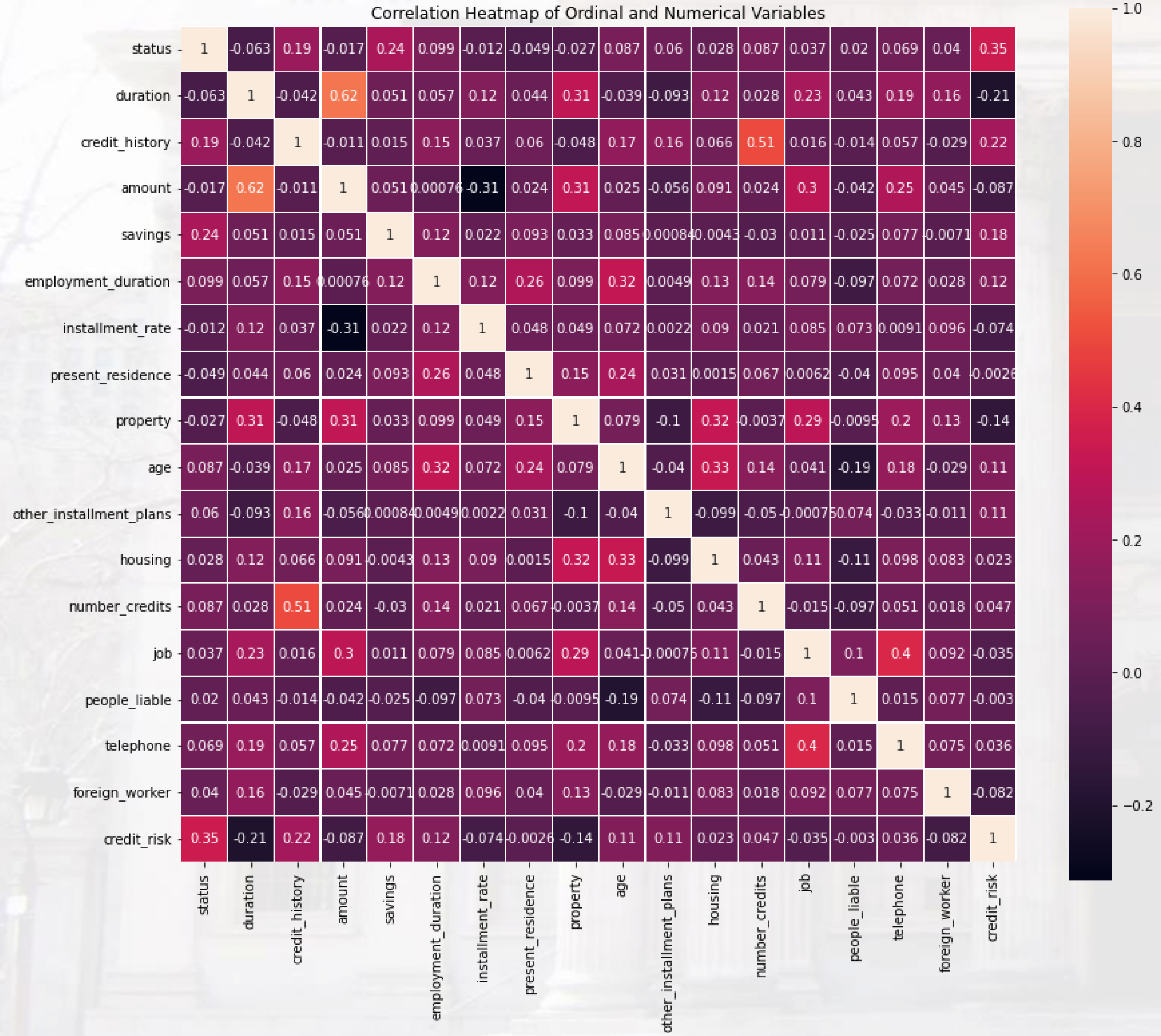
1. Our Business Case and Use Cases
2. Exploratory Data Analysis
3. Data Cleaning
4. Data Pre-Processing
5. Machine Learning: Clustering and Classification Models
6. Conclusion

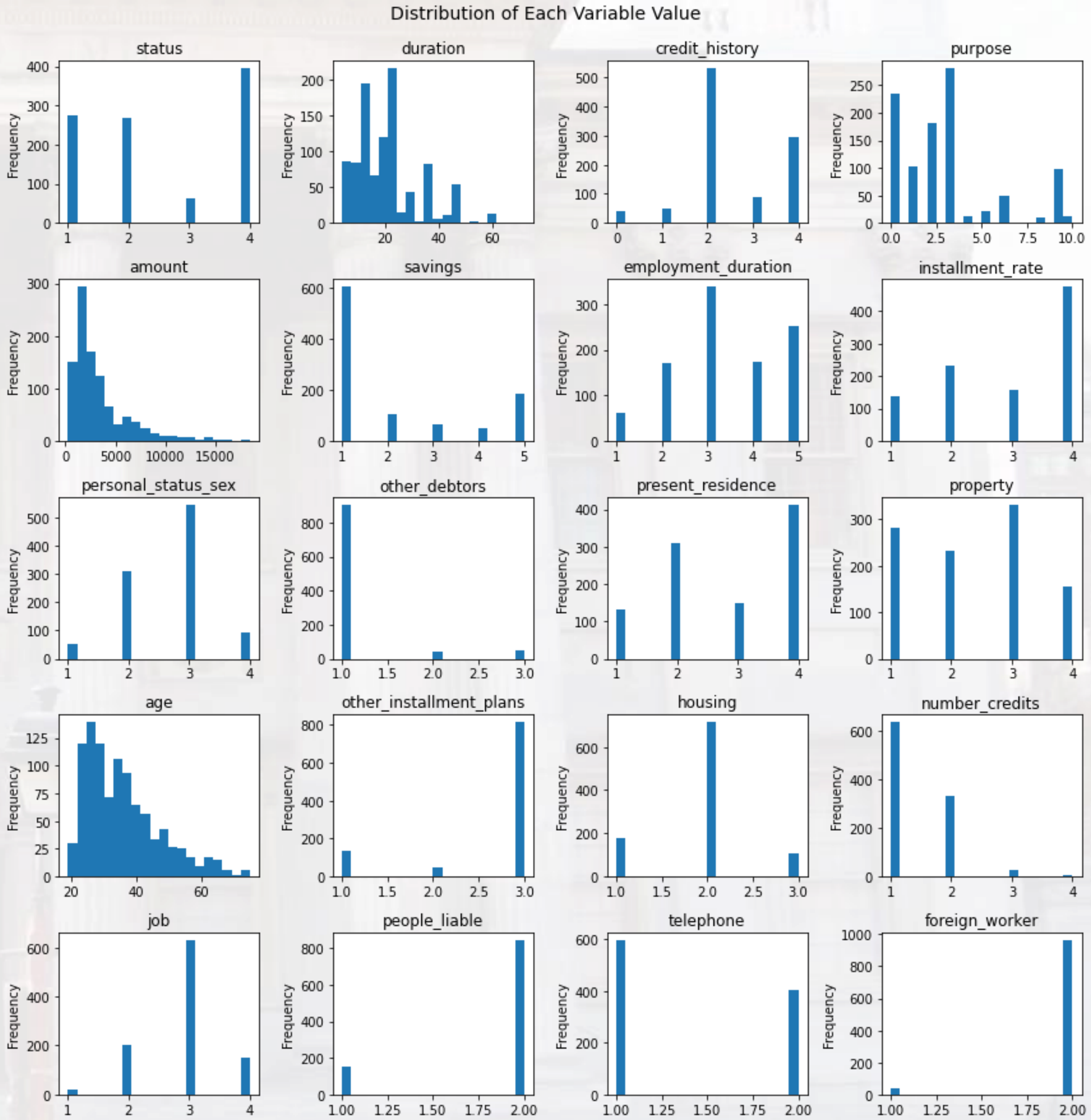
2. Exploratory Data Analysis





2. Exploratory Data Analysis (con't)



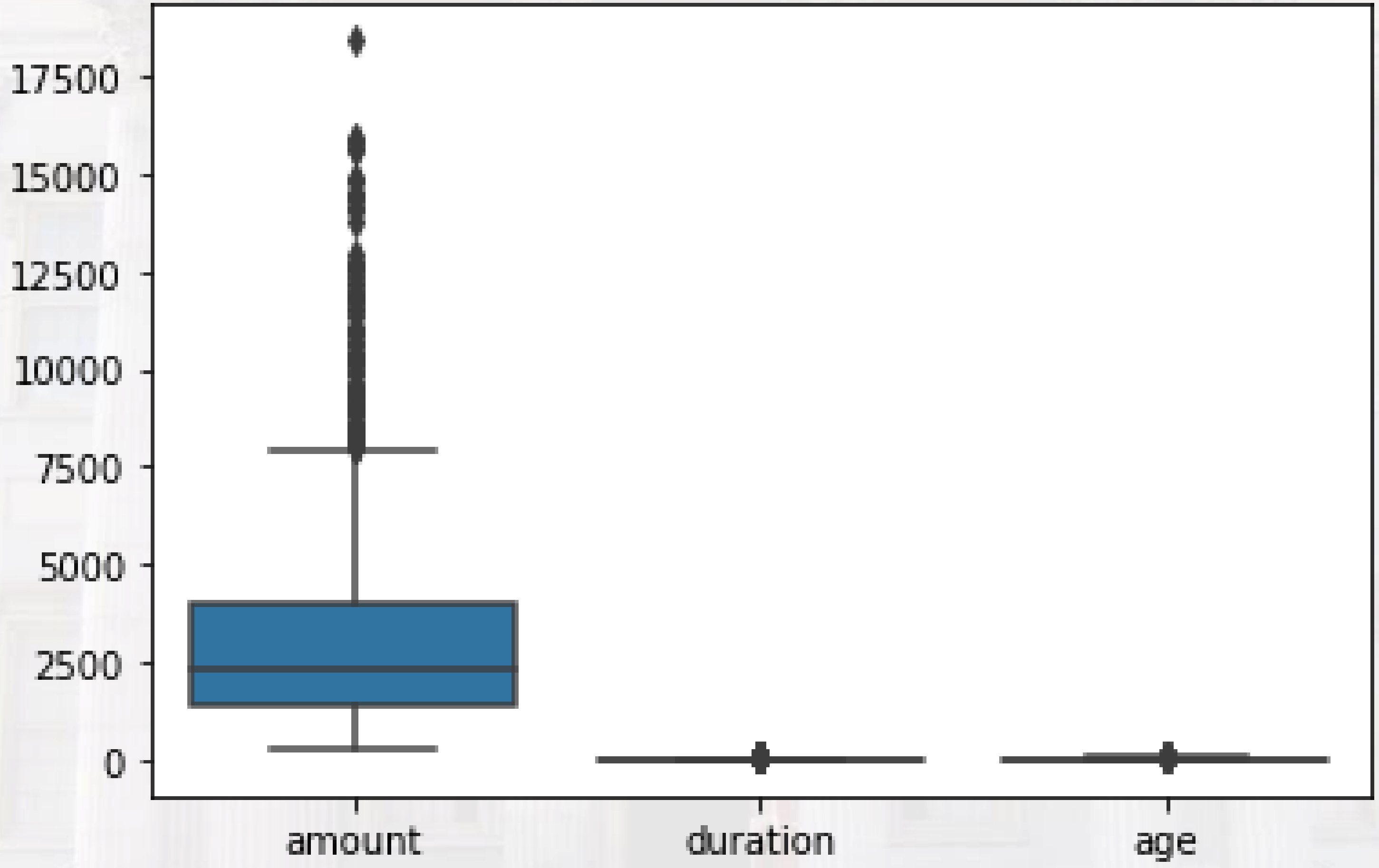


Presentation Outline

1. Our Business Case and Use Cases
2. Exploratory Data Analysis
3. Data Cleaning
4. Data Pre-Processing
5. Machine Learning: Clustering and Classification Models
6. Conclusion

3. Data Cleaning

Outliers



3. Data Cleaning (con't)

Duplicated Or Null Instances

- No duplicated or null instances found in the dataset

Multicollinearity

- None of the variables has high correlation with other variables

Abnormal Attribute

- Categorisation of *personal_status_sex* is uninterpretable

```
1 : male : divorced/separated
2 : female : non-single or male : single
3 : male : married/widowed
4 : female : single
```

Nulls

```
df.isnull().sum()
```

```
status                    0
duration                  0
credit_history            0
purpose                   0
amount                    0
savings                   0
employment_duration       0
installment_rate          0
personal_status_sex       0
other_debtors             0
present_residence         0
property                  0
age                       0
other_installment_plans   0
housing                   0
number_credits            0
job                       0
people_liable             0
telephone                 0
foreign_worker            0
credit_risk               0
dtype: int64
```

Presentation Outline

1. Our Business Case and Use Cases
2. Exploratory Data Analysis
3. Data Cleaning
4. Data Pre-Processing
5. Machine Learning: Clustering and Classification Models
6. Conclusion

4. Data Pre-Processing

Standard Scaling

- Removes the mean and scales each feature/variable to unit variance
- Avoid numerical instabilities due to large values

One Hot Encoding

- Converting categorical data variables so they can be provided to machine learning algorithms to improve predictions
- For nominal categorical data

Oversampling Biased Data

- *credit_risk* is skewed (700 good: 300 bad)
- Modify unequal data classes to create balanced data sets

Presentation Outline

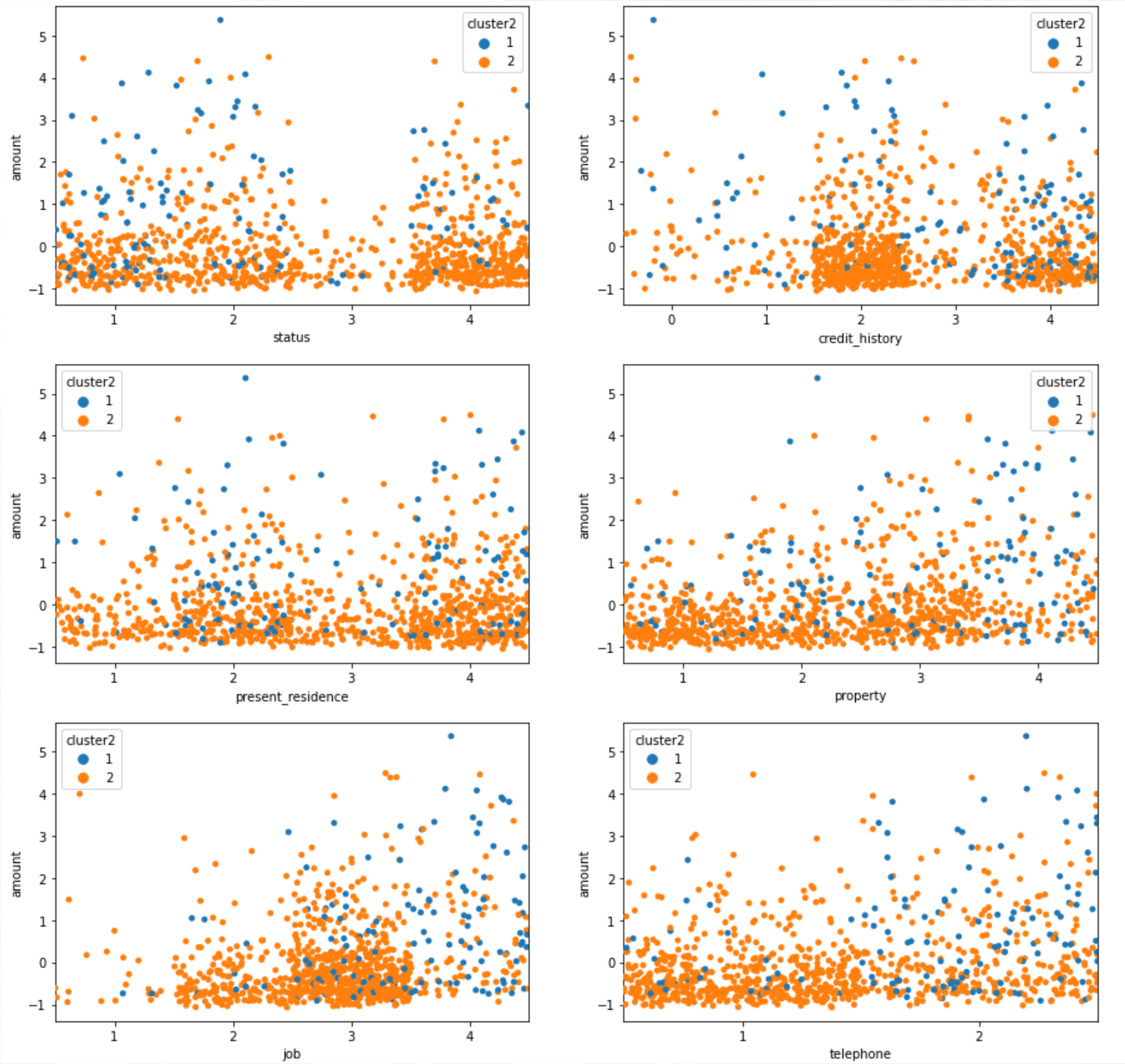
1. Our Business Case and Use Cases
2. Exploratory Data Analysis
3. Data Cleaning
4. Data Pre-Processing
5. Machine Learning: Clustering and Classification Models
6. Conclusion

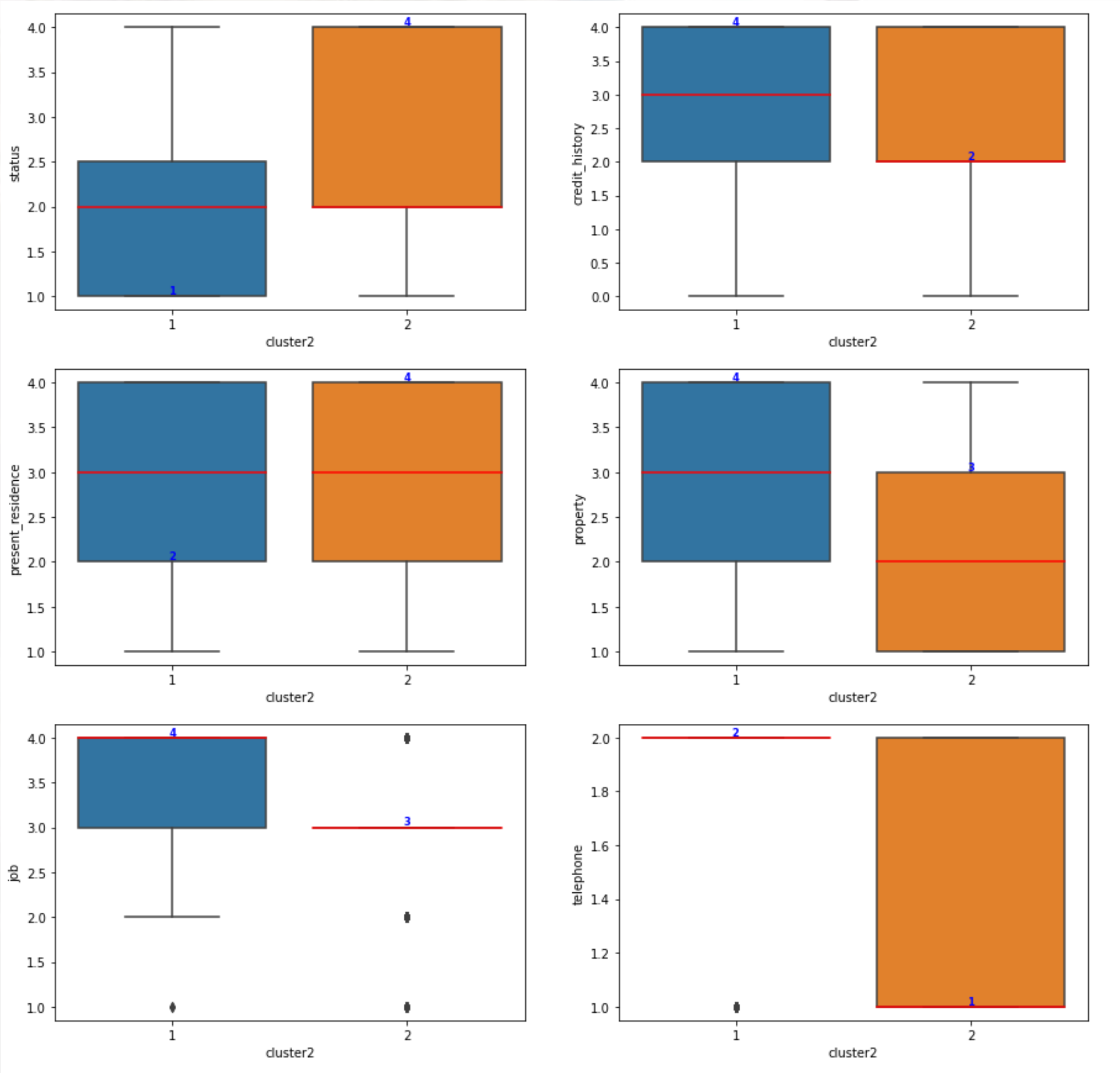
5a. Machine Learning: Clustering Model

k-Modes Clustering

Key Observations:

- Cluster 1 as “bad” credit risk and Cluster 2 as “good” credit risk.
- 6 Attributes of interest: *credit_history, status, present_residence, property, job, telephone*
- In comparison to Cluster 1, users in Cluster 2 are more able to pay off debts, have higher earning potentials, have lived in their current residence for a longer period, are less likely to own valuable property, work in more stable environments, and are more likely to be contacted.


cluster2
1
2
amount
status
credit_history
present_residence
property
job
telephone


status
credit_history
present_residence
property
job
telephone
cluster2

5b. Machine Learning: Classification Model

Logistic Regression

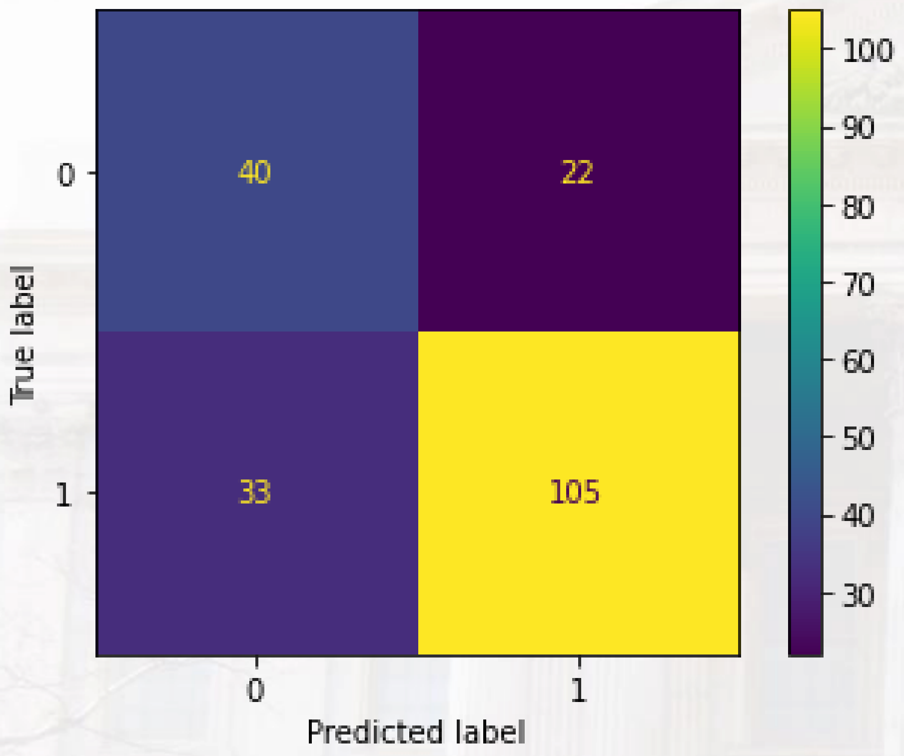
Reason

- Able to provide probabilistic classification
- Able to analyse feature importance
- Easier to explain to customers why their loan application fails

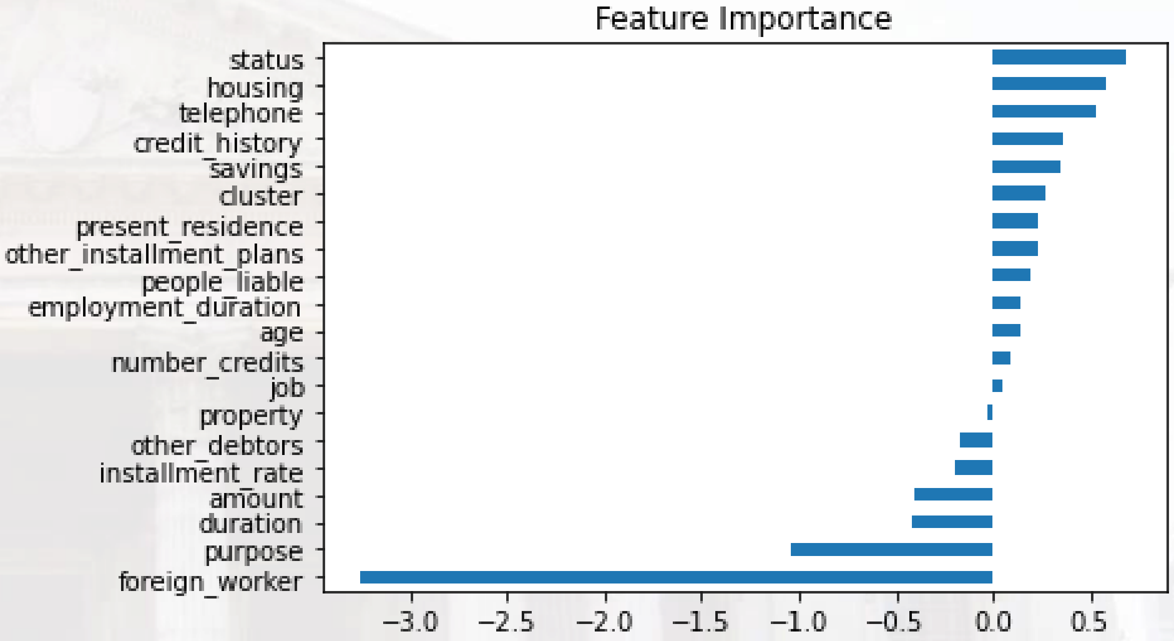
Aim

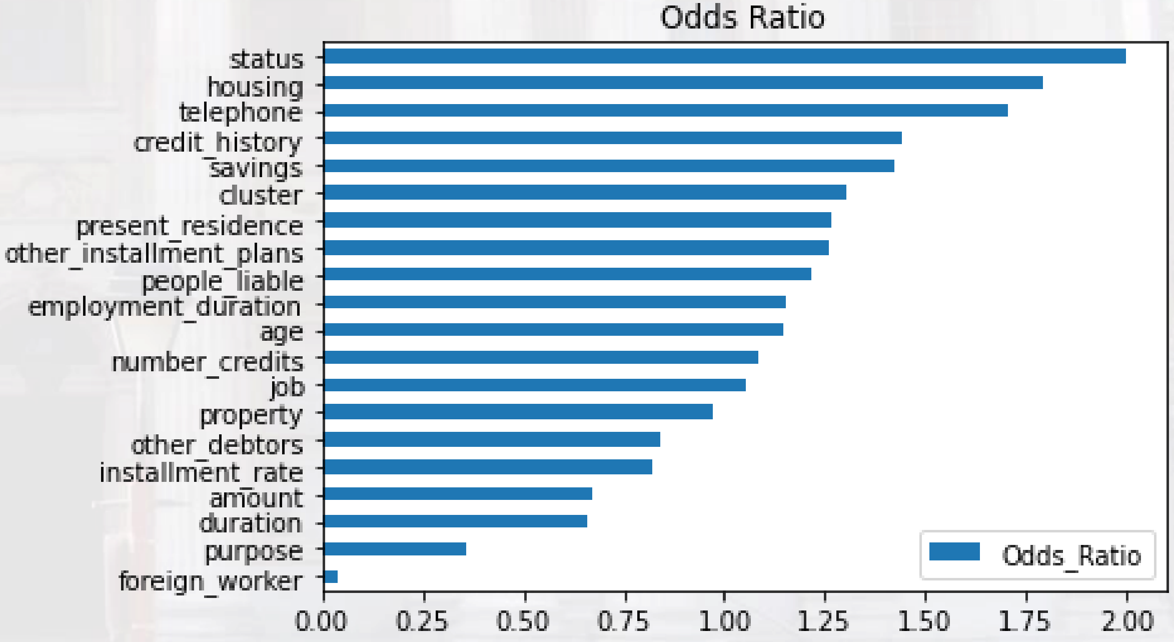
- High **precision** score: avoid **false positives** as they bring huge losses
- High accuracy score

*Note: Typos corrected at **bold areas**



	precision	recall	f1-score	support
0	0.55	0.65	0.59	62
1	0.83	0.76	0.79	138
accuracy			0.73	200
macro avg	0.69	0.70	0.69	200
weighted avg	0.74	0.72	0.73	200





Presentation Outline

1. Our Business Case and Use Cases
2. Exploratory Data Analysis
3. Data Cleaning
4. Data Pre-Processing
5. Machine Learning: Clustering and Classification Models
6. Conclusion

6. Conclusion

Using Our Model, We Can:

- Perform customer segmentation
- Estimate the probability of good credit risk

Future Work:

- Synthesise data to simulate real life credit applications
- Improve our machine learning models
- Design loan approval strategies

Dataset Problems:

- Definition of *credit_risk* is not clear
  - For example, we may define good *credit_risk* as: able to repay 90% of the loan before the end of loan period
- Confusing categorisation of sex and marital status (*personal_status_sex*)


Thank you for listening!