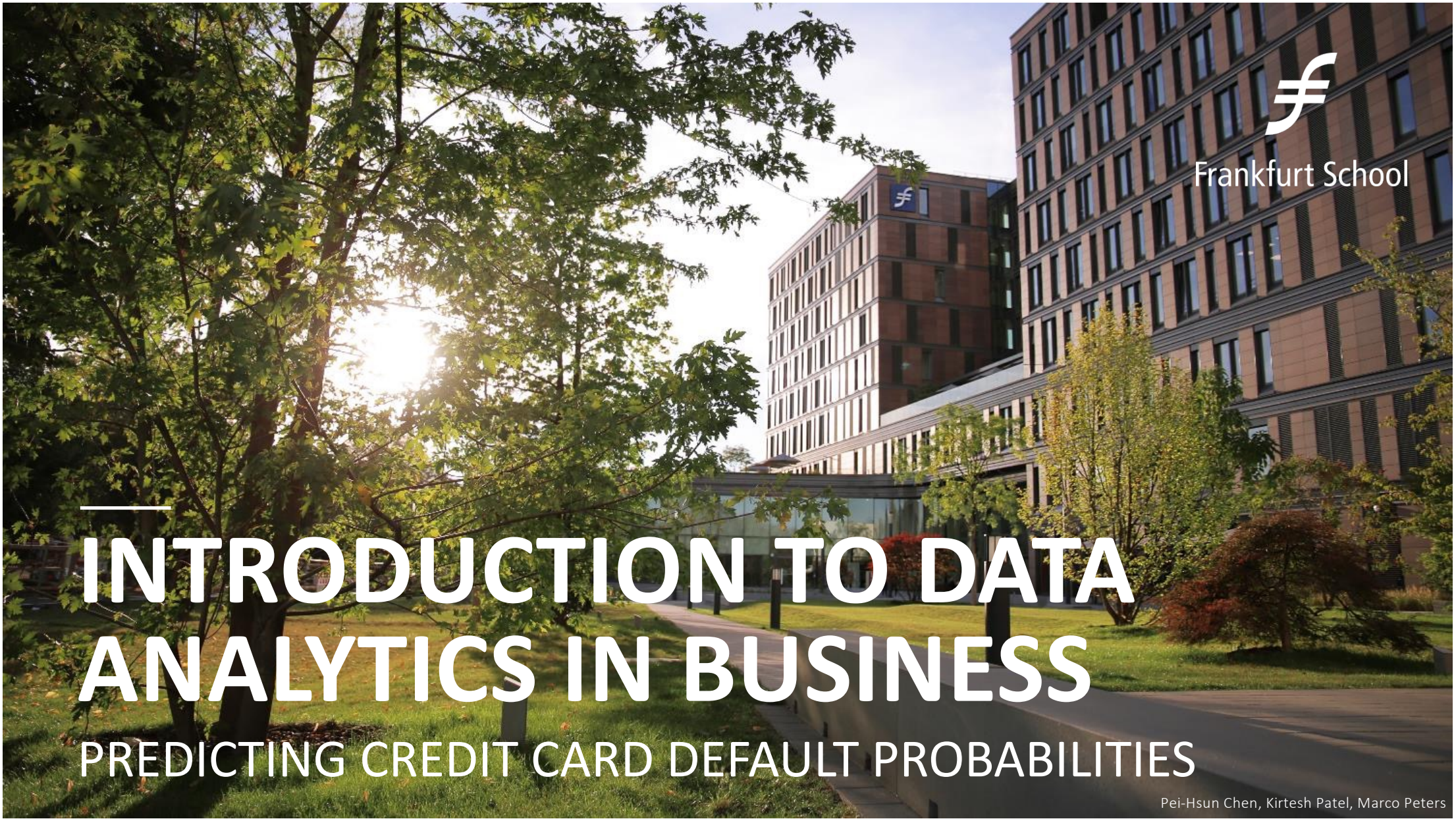
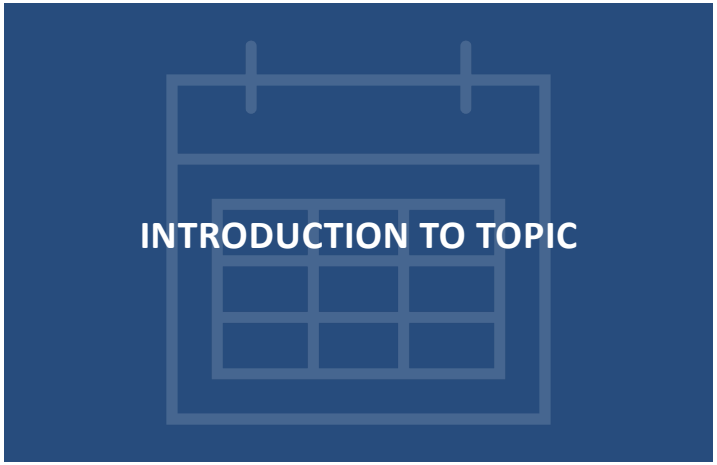


INTRODUCTION TO DATA ANALYTICS IN BUSINESS

PREDICTING CREDIT CARD DEFAULT PROBABILITIES

Pei-Hsun Chen, Kirtesh Patel, Marco Peters

CONTENT


INTRODUCTION TO TOPIC


DATA SET


METHODOLOGY &
DATA PREPERATION


FINDINGS


CONCLUSION


LIMITATIONS &
NEXT STEPS

INTRODUCTION TO TOPIC

RESEARCH PROBLEM & RESEARCH QUESTION

Drivers	Details	Outcome
Increased number of credit card users	• Credit card ownership has increased significantly over the past decades, making it today very wide-spread • Percentage of owning a credit card in an American household is at over 70% percent (Ma, Y.H. (2020)*)	Credit card loans make up an exceptionally high share of Banks' debt ($917B in 2019, +5% over 2018)
Lowered prerequisites toward applicants	• With continuously decreasing interest rates, prerequisites applicants of credit cards have to meet have been decreasing as well	Increased risk for loans given to default

Research Question: What individuals, given a defined set of characteristics, should be admitted a 1-month loan in form of a credit card?

* Ma, Y.H. (2020) Prediction of Default Probability of Credit-Card Bills. Open Journal of Business and Management, 8, 231-244

DATA SET

DETAILS ON DATA SET 1 – CHARACTERISTICS

The first data set contains all relevant features of credit card holders.

SOURCE:
- Kaggle

SIZE:
- 438,557 rows * 18 columns
- 438,510 users (IDs), 17 characteristics/user

FEATURES:
- ID
- Gender
- Owns a car
- Owns a realty
- No. of children
- Annual income
- Income type
- Education level
- Marital status
- Housing type
- Age
- Days employed
- Has a mobile phone
- Has a work phone
- Has a phone
- Has an e-mail
- Occupation
- No. of family members

	ID	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL
0	5008804	M	Y	Y	0	427500.0
1	5008805	M	Y	Y	0	427500.0
2	5008806	M	Y	Y	0	112500.0
3	5008808	F	N	Y	0	270000.0
4	5008809	F	N	Y	0	270000.0
...	...	...	...	...	...	...
438552	6840104	M	N	Y	0	135000.0
438553	6840222	F	N	N	0	103500.0
438554	6841878	F	N	N	0	54000.0
438555	6842765	F	N	Y	0	72000.0
438556	6842885	F	N	Y	0	121500.0

DETAILS ON DATA SET 2 – PERFORMANCE

The second data set contains the performance of individual credit card holders from the first data set (how long it took them to repay credit)

SOURCE:
- Kaggle

SIZE:
- 1,048,575 rows * 3 columns
- 45,985 users (IDs), 2 performance indicators

FEATURES:
- ID
- Months balance
- Status

	ID	MONTHS_BALANCE	STATUS
0	5001711	0	X
1	5001711	-1	0
2	5001711	-2	0
3	5001711	-3	0
4	5001712	0	C
...	...	...	...
1048570	5150487	-25	C
1048571	5150487	-26	C
1048572	5150487	-27	C
1048573	5150487	-28	C
1048574	5150487	-29	C

METHODOLOGY & DATA PREPERATION

METHODOLOGY

1 DATA SELECTION & PREPERATION

- Data set identification and integration into python
- Data preparation for data set visualization and verification
- Visualization

2 PRE-PROCESSING FOR MACHINE LEARNING

- Data manipulation
- Correlation analysis and feature selection
- PCA and train-/test splitting

3 MODEL EXECUTION

- K-nearest neighbour
- Decision tree
- Naive Bayes
- Random forest

4 CONCLUSIONS & RESEARCH QUESTION ANSWERING

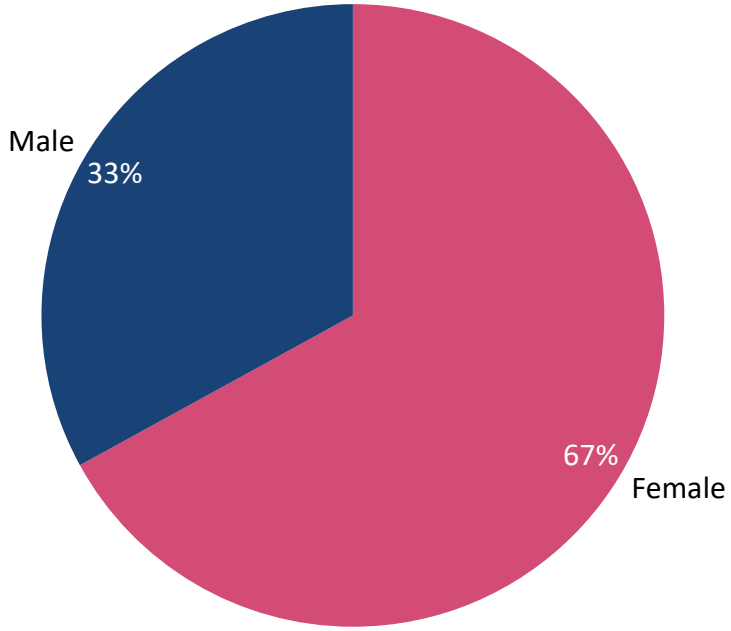
- Model performance comparison
- Feature importance
- Takeaways and drawback to initial research question

DATA PREPARATION

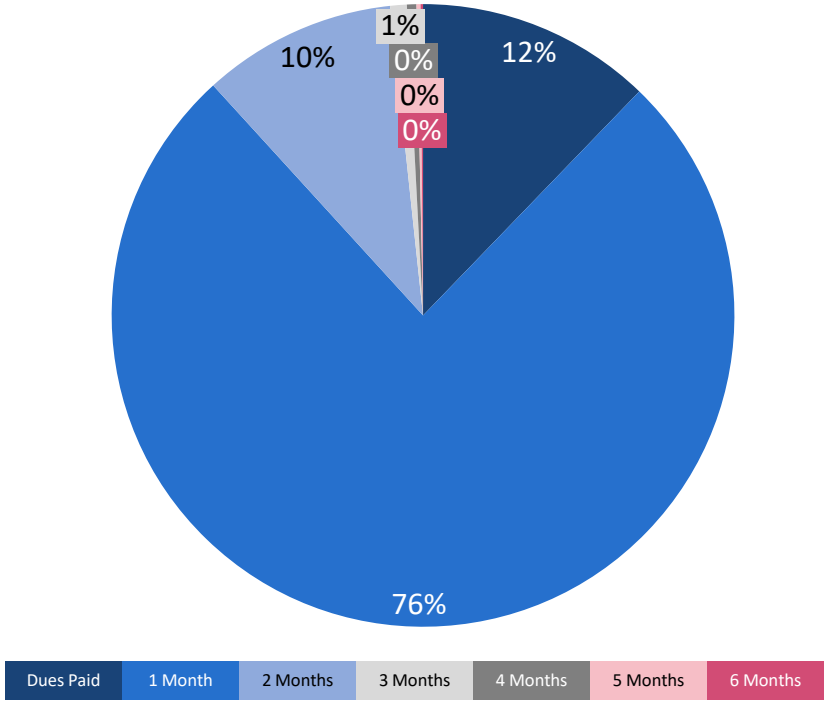
1. **Pre-processing of the performance data set**
   - 1.1 Transforming status indicators X, C to integers (-1)
   - 1.2 Adjusting data types to be integers
   - 1.3 Increasing all status indicators by 1, such that status ranges from 0 - 7
   - 1.4 Grouping entire data set by ID, scrapping all ID duplicates, keeping only the entry with highest status (i.e., worst performance)
2. **Merging of the performance data set and the characteristics data set (inner merge)**

DATA VISUALIZATION I

Gender Distribution

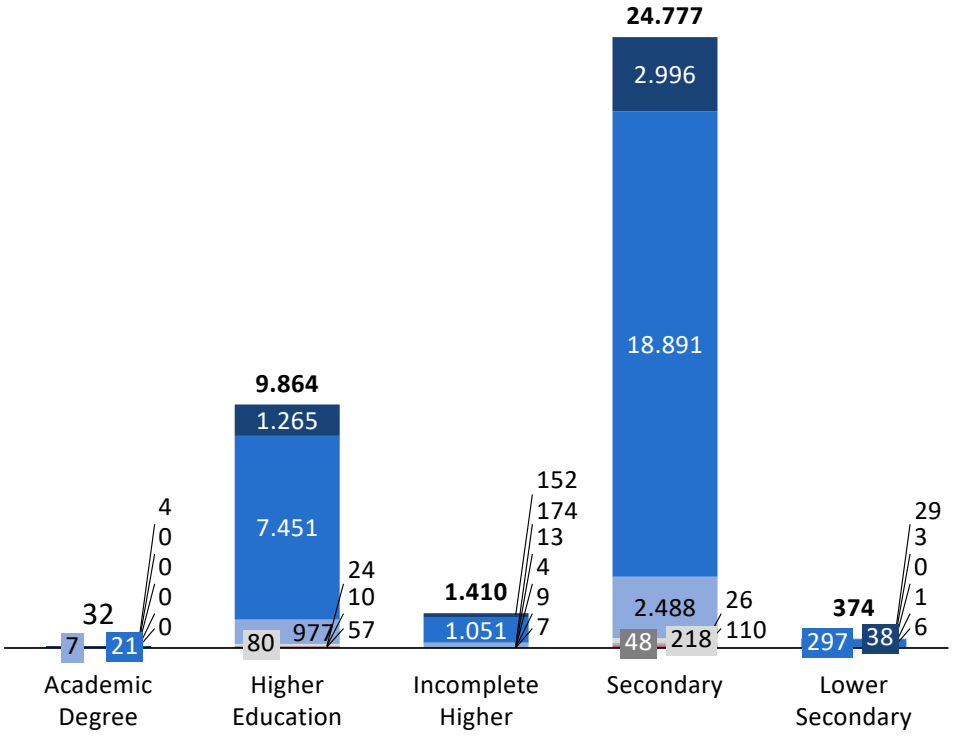


Status Distribution

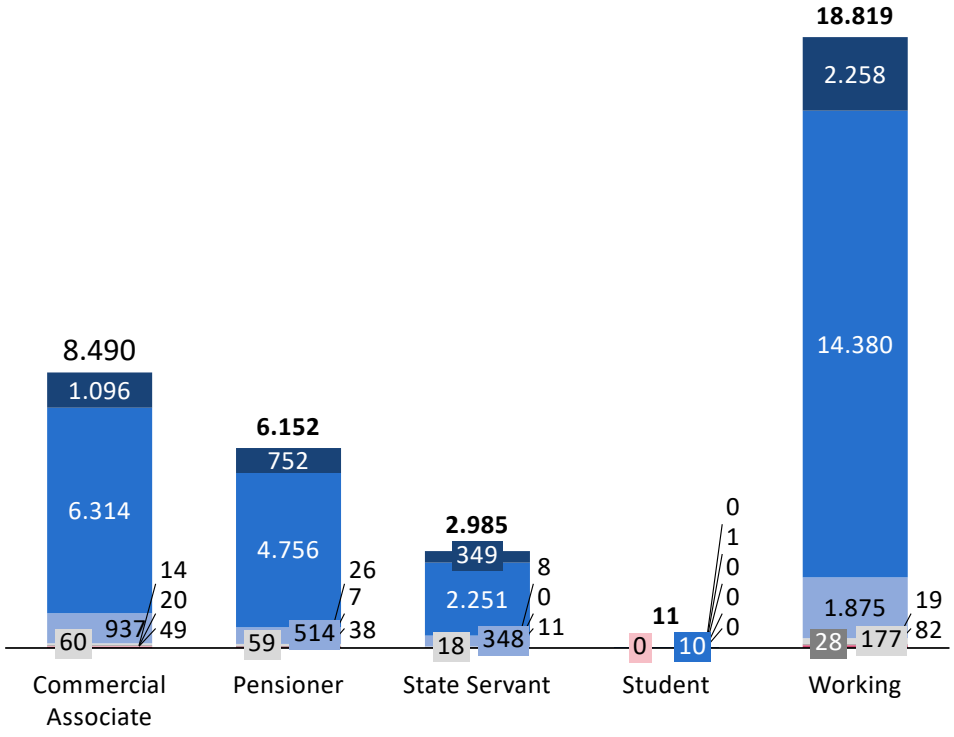


DATA VISUALIZATION II

Education / Due Paid



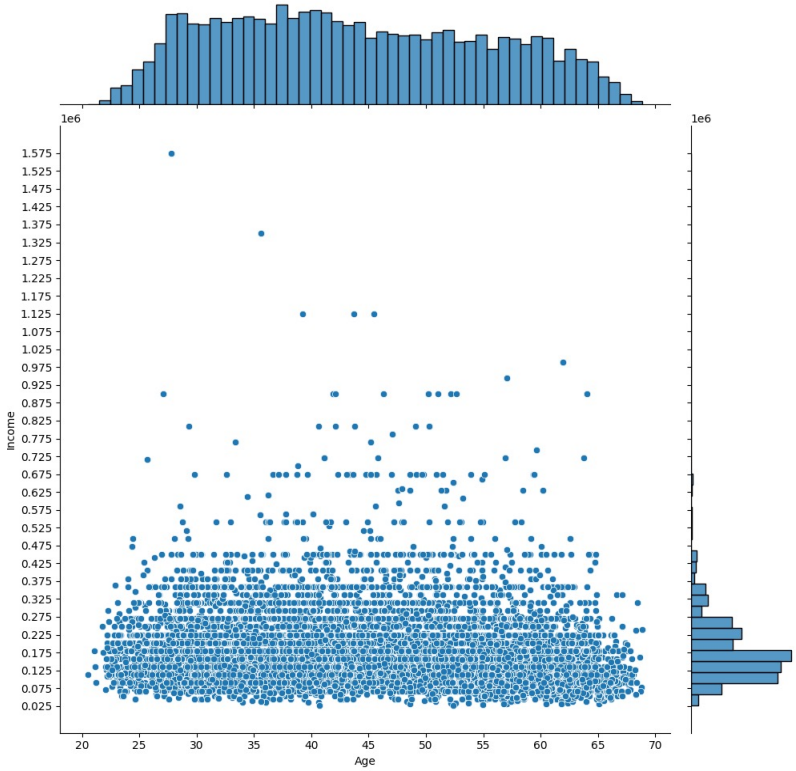
Income Type / Due Paid



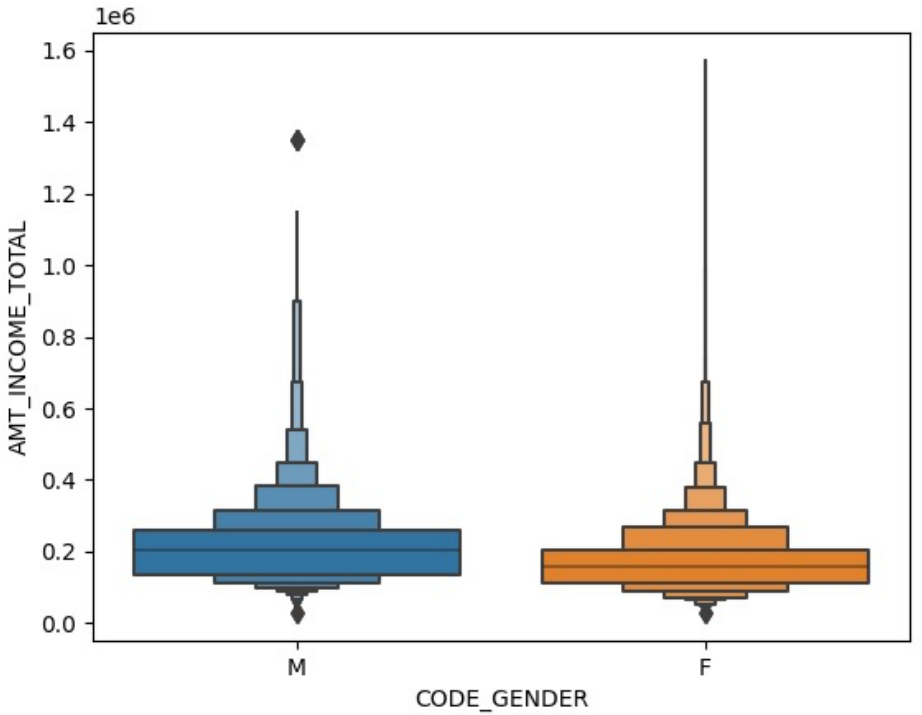
Dues Paid | 1 Month | 2 Months | 3 Months | 4 Months | 5 Months | 6 Months

DATA VISUALIZATION III

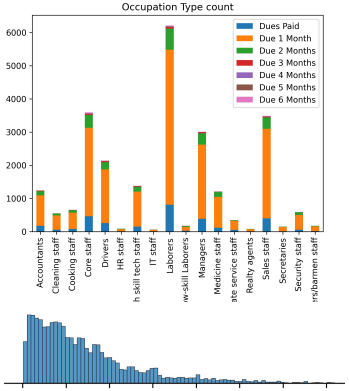
Age / Income Distribution

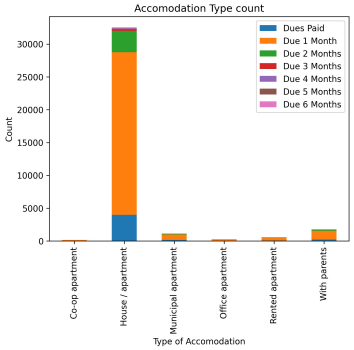


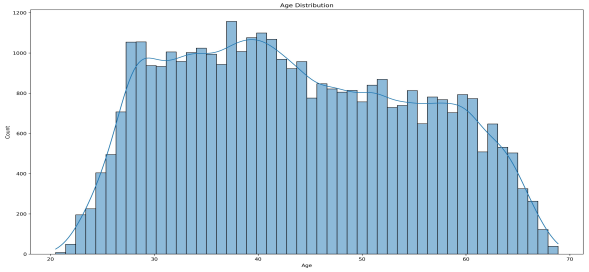
Gender / Income Distribution

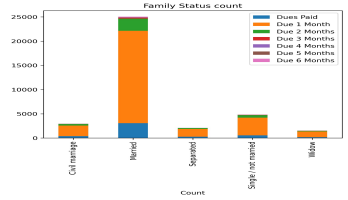


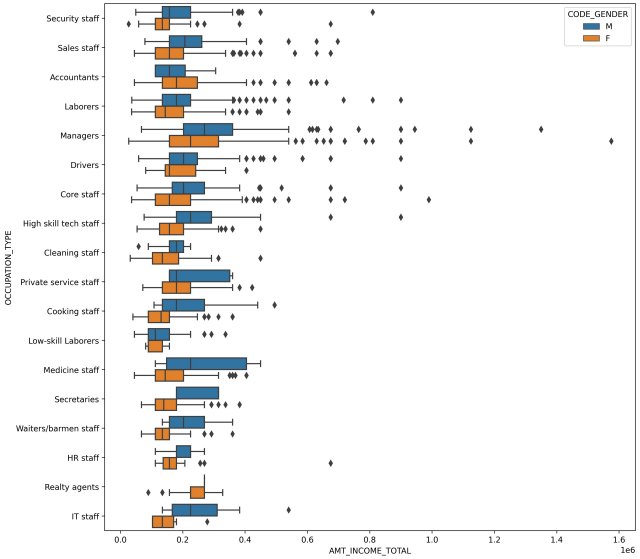
DATA VISUALIZATION – FURTHER DETAILS

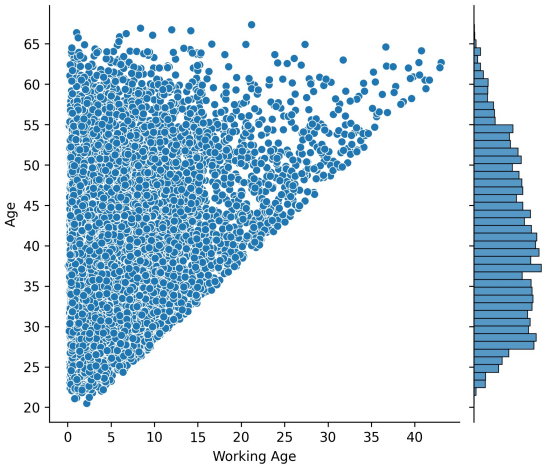


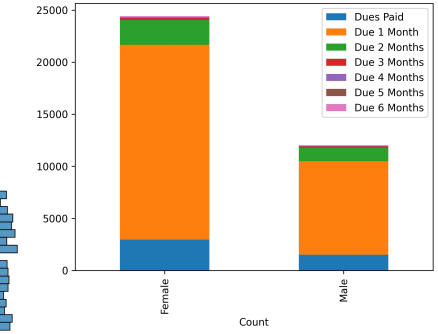


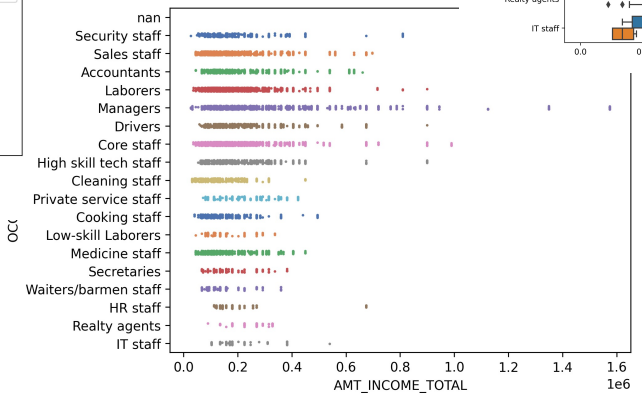


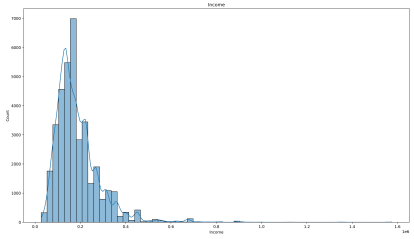












PRE-PROCESSING FOR CORRELATION ANALYSIS

2 Merging of the performance data set and the characteristics data set (inner merge)

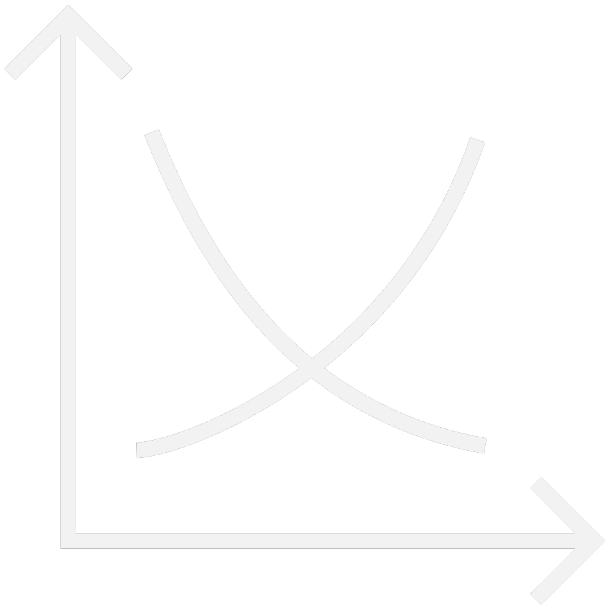
3 Pre-processing of merged data set for correlation analysis

3.1 Features: ‘gender’, ‘owns car’, ‘owns realty’ from str (yes/no) to binary (0/1)

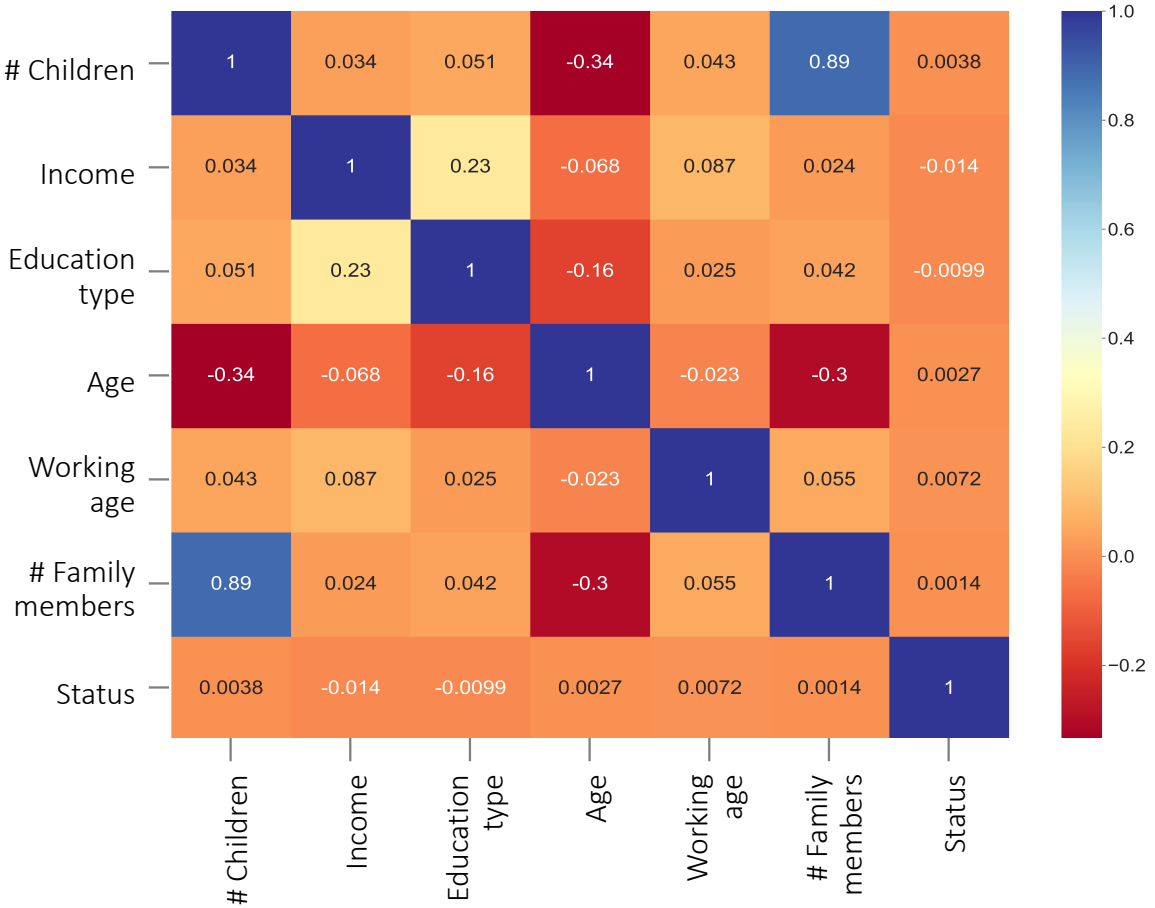
3.2 Features: ‘education type’, ‘income type’, ‘family status’, ‘housing type’, ‘occupation type’ from str to int

3.3 Loan settlement performance indicator from int range (-1 to 6) to binary (0/1)

4 Correlation analysis



CORRELATION ANALYSIS



PRE-PROCESSING FOR MACHINE LEARNING

4 Correlation analysis

5 **Pre-processing for machine learning**

5.1 Scrapping of features: ‘months balance’, ’owns a mobile’, ‘status’

6 Resampling

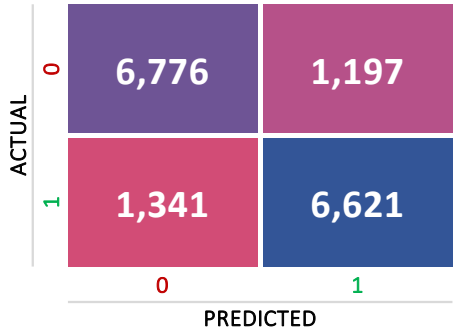
7 Principle component analysis

8 **Train-/test-splitting (75/25)**

FINDINGS

ALGORITHM PERFORMANCE COMPARISON

K-nearest neighbor



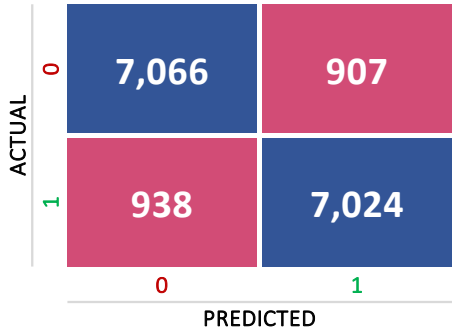
Algorithm accuracy:

84.1%

Score ratio – test/training
(overfitting)

93.3%

Decision tree



Algorithm accuracy:

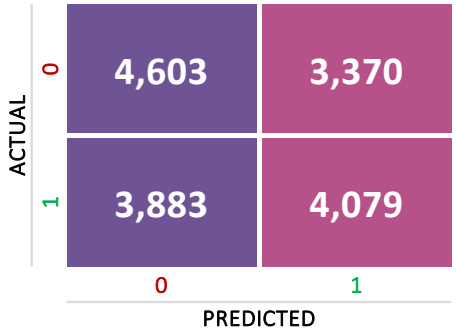
88.4%

Score ratio – test/training
(overfitting)

94.0%

+ BEST PERFORMANCE +

Naive Bayes



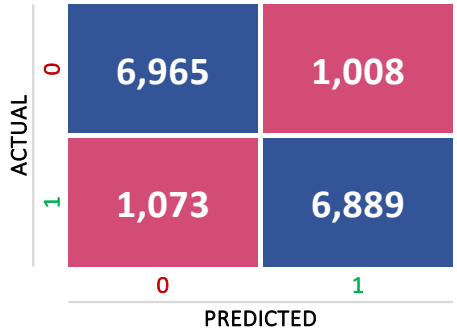
Algorithm accuracy:

54.5%

Score ratio – test/training
(overfitting)

99.1%

Random forest



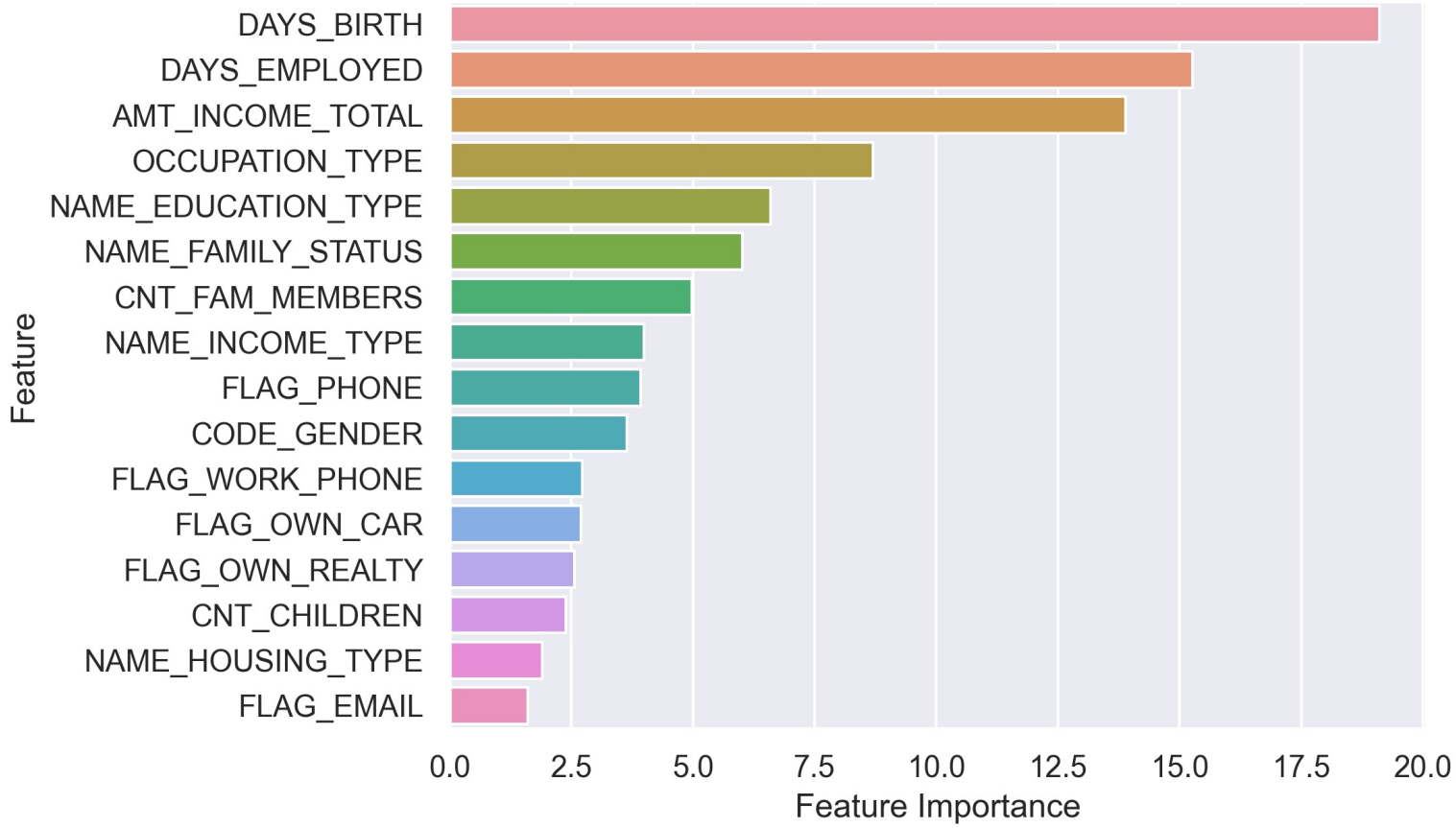
Algorithm accuracy:

86.9%

Score ratio – test/training
(overfitting)

96.0%

DEEP DIVE: RANDOM FOREST



EXEMPLARY CLASSIFICATION SHOW-CASE I

Feature	Specification
Gender	Male (M)
Owns car	Yes (Y)
Owns realty	Yes (Y)
# of children	0 (0)
Income	150,000 (150000)
Income type	Working
Education type	Academic degree
Family status	Single / not married

Feature	Specification
Housing type	House / apartment
Age	36 years (13232)
Days employed	10 years (3790)
Work phone	Yes (Y)
Land line	No (N)
Email	Yes (Y)
Occupation	High skill tech staff
# of family members	2 (2.0)

Outcome:

APPROVED

CONCLUSION

CONCLUSION

INDIVIDUALS' CREDIT DEFAULT LIKELINESS CAN BE ACCURATELY PREDICTED

1

- We have found two methods to accurately determine, if a given person should be admitted a credit card or not.
- Both decision tree and random forest offer reliable results to help answer the initially set out research question

2

- Number of age, days employed, income, and occupation type are among the most important features when it comes to determining credit default likeliness (see random forest results)

However…

EXEMPLARY CLASSIFICATION SHOW-CASE II

Feature	Specification
Gender	Male (M)
Owns car	No (N)
Owns realty	No (N)
# of children	0 (0)
Income	20,000 (20000)
Income type	Student
Education type	Higher education
Family status	Single / not married

Feature	Specification
Housing type	Rented apartment
Age	24 years (8819)
Days employed	2 years (800)
Work phone	No (N)
Land line	No (N)
Email	Yes (Y)
Occupation	IT staff
# of family members	4 (4.0)

Outcome:

APPROVED

EXEMPLARY CLASSIFICATION SHOW-CASE III

Feature	Specification
Gender	Male (M)
Owns car	Yes (Y)
Owns realty	Yes (Y)
# of children	2 (2)
Income	300,000 (300000)
Income type	Pensioner
Education type	Academic degree
Family status	Married

Feature	Specification
Housing type	House apartment
Age	71 years (25932)
Days employed	0 years - pensioner (0)
Work phone	No (N)
Land line	No (N)
Email	Yes (Y)
Occupation	Core staff
# of family members	6 (6.0)

Outcome:

DENIED

LIMITATIONS & NEXT STEPS

LIMITATIONS & NEXT STEPS

SKEWED DATA SET	• The data used was in large parts made up of a non-representative set of individuals (e.g., average income >USD 150,000) • Therefore, when classifying a randomly chosen person with a set of characteristics representative of e.g., German middle class, the algorithms may provide unexpected and unrealisitc results
DERIVATION OF TANGIBLE GUIDELINES FOR POLICY MAKERS	• A first idea would be, to use this classification in order to aid decision making regarding different service offerings • An example might be three different credit card options (e.g., silver, gold, platinum), offering different credit limits, interests and time horizons • The details of such business models do of course require a lot more further research, focusing on the topics of risk assessments, target market & business models, pricing strategy, and many more