Explainable AI

Group 77

Chrysanthi Foti
Eleftherios Tetteris
Athanasopoulos Nikolaos

Maastricht University

Introduction

Introduction-Data Structure

- German Credit Dataset (multiple variants)
- 9 covariates for classification
- 1 binary output (Good Risk vs Bad Risk)

	Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose	Risk
0	67	male	2	own	NaN	little	1169	6	radio/TV	good
1	22	female	2	own	little	moderate	5951	48	radio/TV	bad
2	49	male	1	own	little	NaN	2096	12	education	good
3	45	male	2	free	little	little	7882	42	furniture/equipment	good
4	53	male	2	free	little	little	4870	24	car	bad

Introduction-Data Structure-Numeric data

	Age	Job	Credit amount	Duration
count	1000.000000	1000.000000	1000.000000	1000.000000
mean	35.546000	1.904000	3271.258000	20.903000
std	11.375469	0.653614	2822.736876	12.058814
min	19.000000	0.000000	250.000000	4.000000
25%	27.000000	2.000000	1365.500000	12.000000
50%	33.000000	2.000000	2319.500000	18.000000
75%	42.000000	2.000000	3972.250000	24.000000
max	75.000000	3.000000	18424.000000	72.000000

Introduction-Data Structure-Categorical data

```
Sex : ['male' 'female']

Housing : ['own' 'free' 'rent']

Saving accounts : [nan 'little' 'quite rich' 'rich' 'moderate']

Checking account : ['little' 'moderate' nan 'rich']

Purpose : ['radio/TV' 'education' 'furniture/equipment' 'car' 'business'
 'domestic appliances' 'repairs' 'vacation/others']

Risk : ['good' 'bad']
```

Introduction-Data Structure-Categorical data (continued)

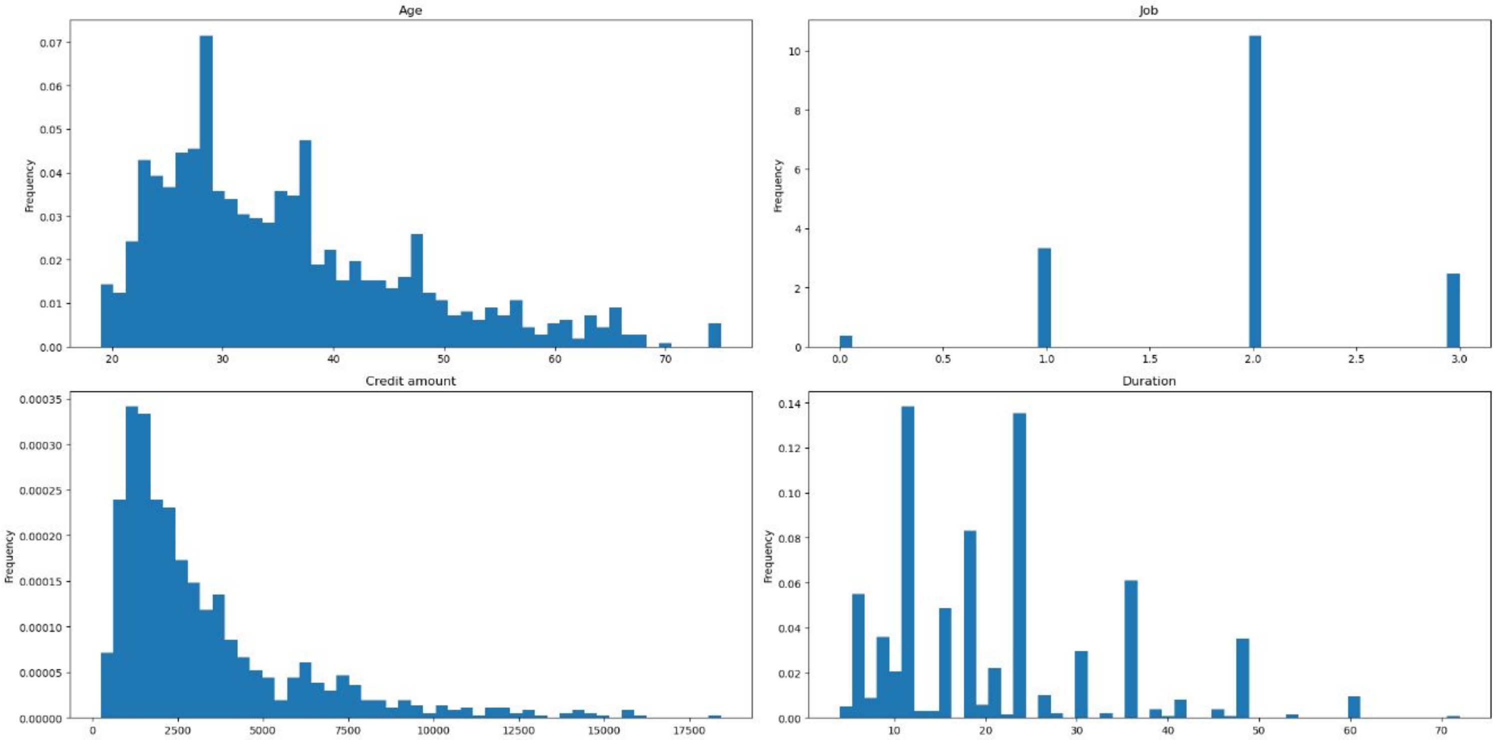
	Sex	Housing	Saving accounts	Checking account	Purpose	Risk
count	1000	1000	817	606	1000	1000
unique	2	3	4	3	8	2
top	male	own	little	little	car	good
freq	690	713	603	274	337	700

Categorical Data Cleaning

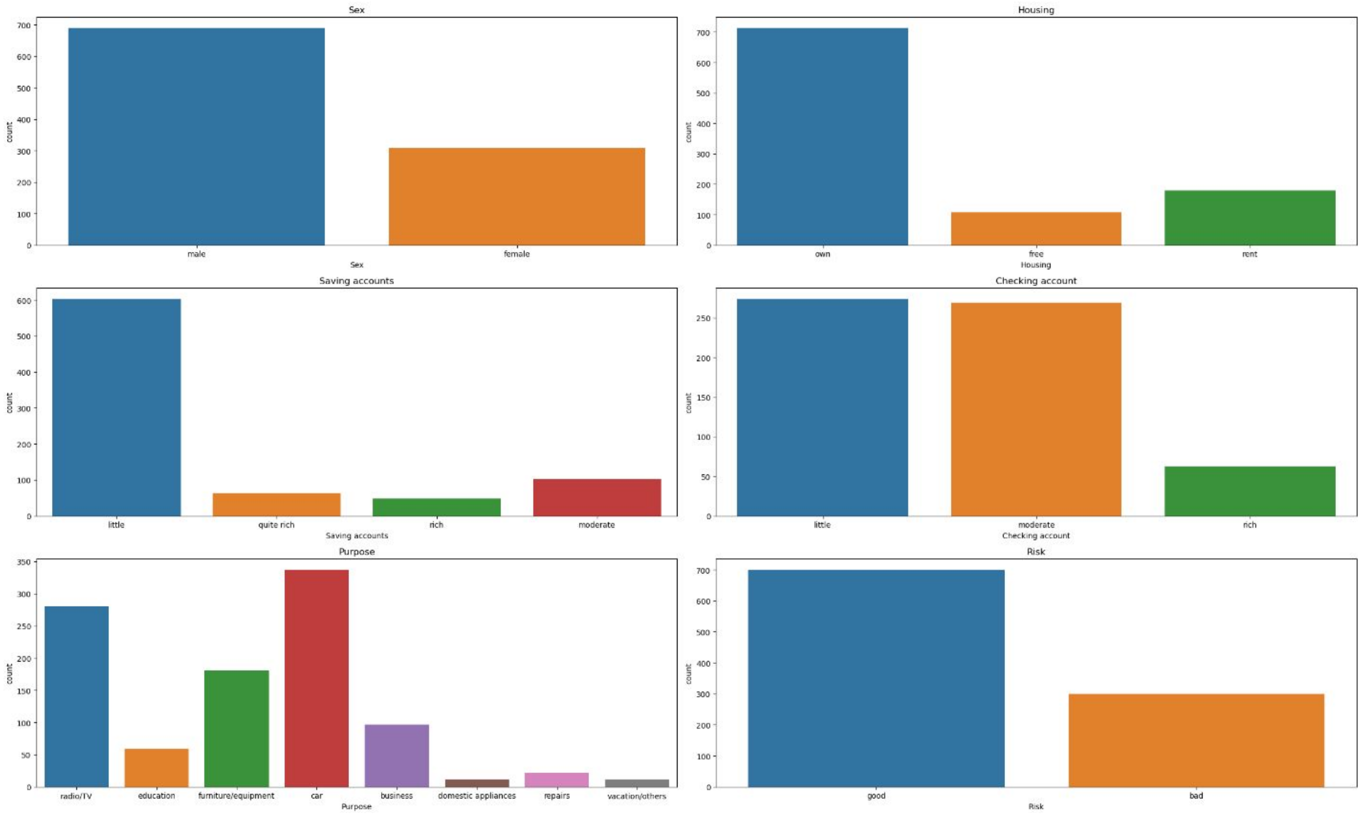
- For now, NaN values were filled with respective mode
- In the end, different methods will be examined as well (dropping columns, dropping rows)
- Afterwards, the categorical features were encoded into integers

```
Sex
{'female': 0, 'male': 1}
--------------------------------------------------
Housing
{'free': 0, 'own': 1, 'rent': 2}
--------------------------------------------------
Saving accounts
{'little': 0, 'moderate': 1, 'quite rich': 2, 'rich': 3}
--------------------------------------------------
Checking account
{'little': 0, 'moderate': 1, 'rich': 2}
--------------------------------------------------
Purpose
{'business': 0, 'car': 1, 'domestic appliances': 2, 'education': 3, 'furniture/equipment': 4, 'radio/TV': 5, 'repairs': 6, 'vac
ation/others': 7}
--------------------------------------------------
Risk
{'bad': 0, 'good': 1}
--------------------------------------------------
```

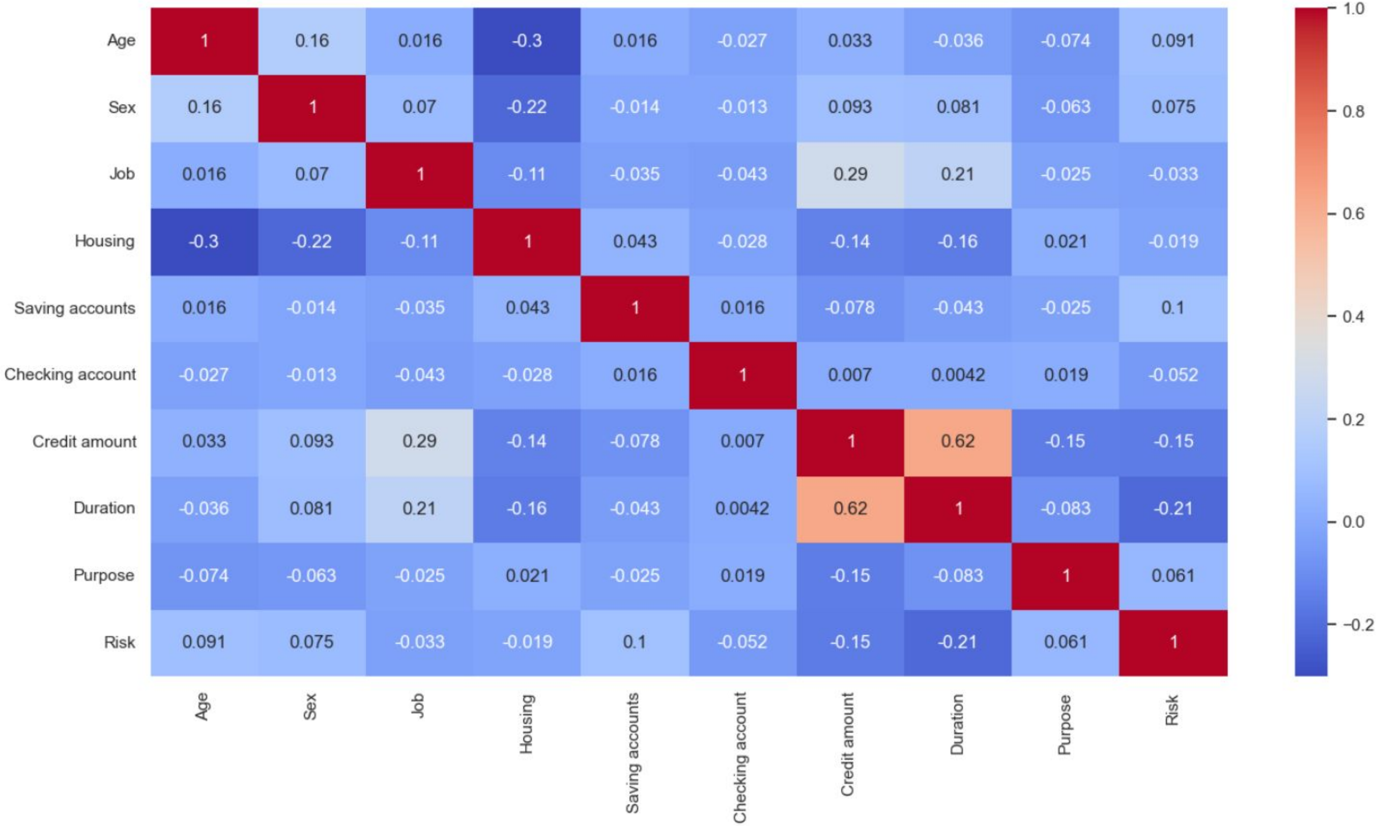
Data Distributions - Numeric



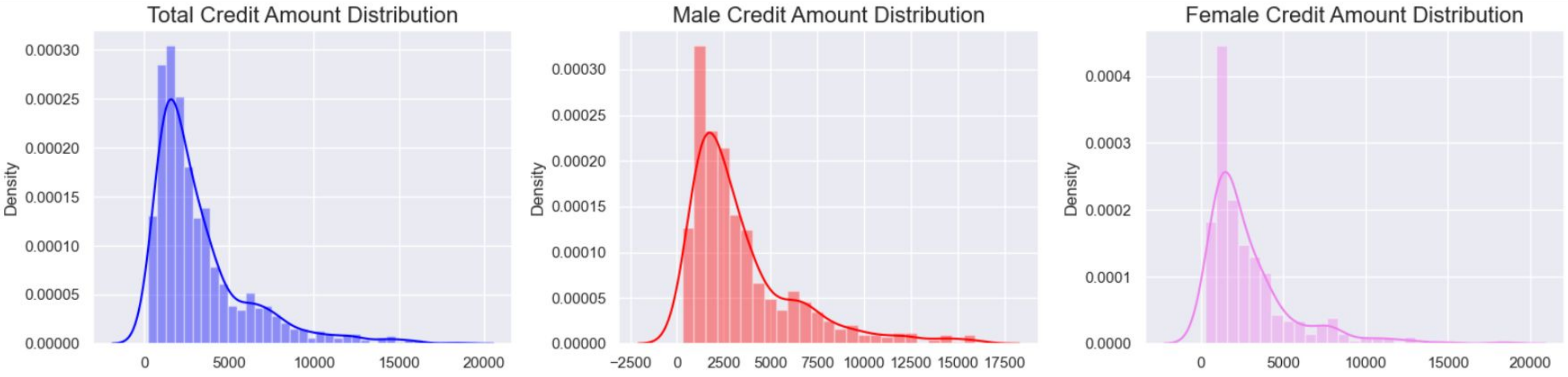
Data Distributions - Categorical



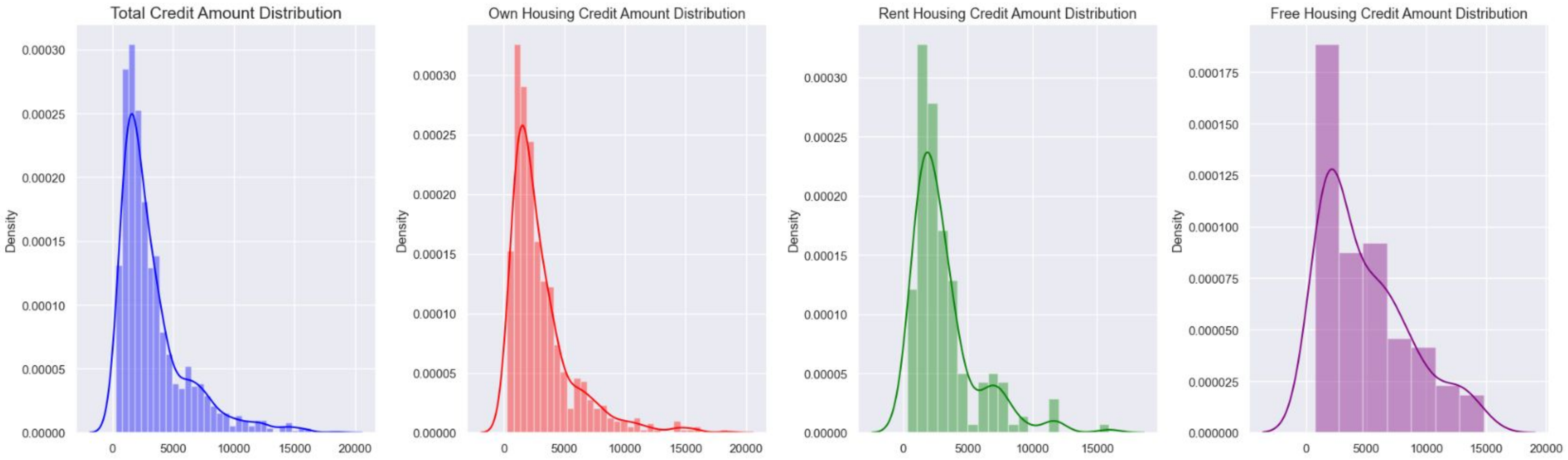
Feature Correlation Heatmap



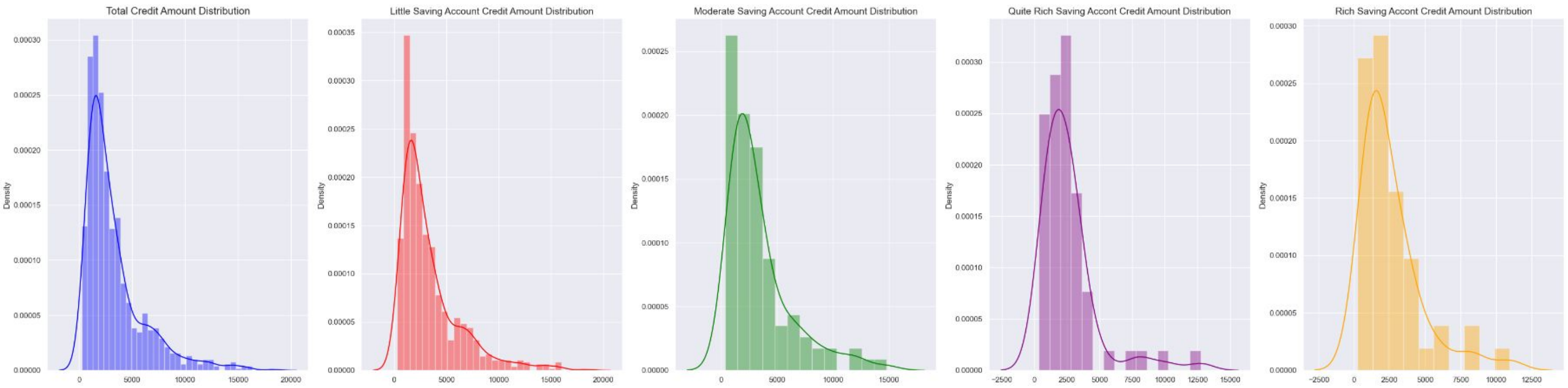
Male-Female vs Credit Amount



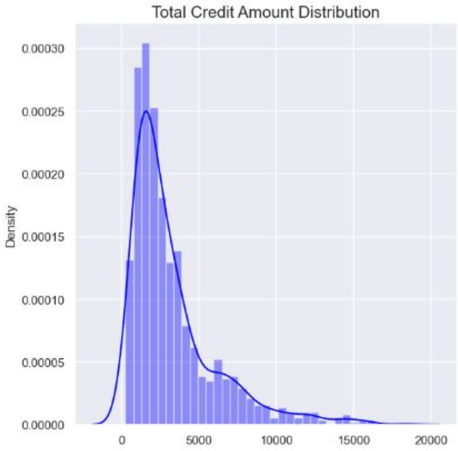
Housing vs Credit Amount

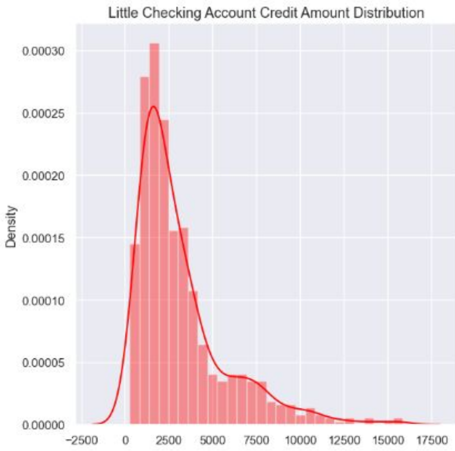


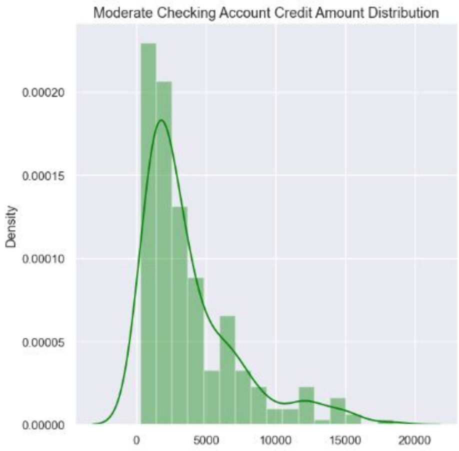
Saving Account vs Credit Amount

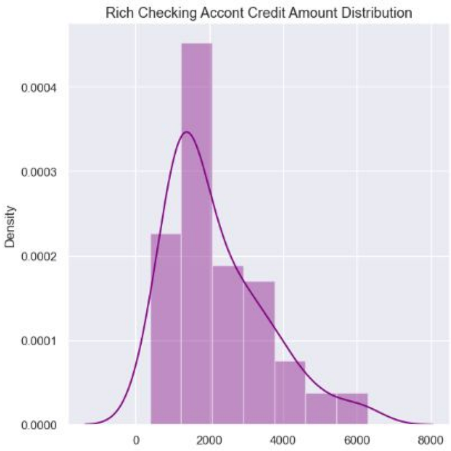


Checking Account vs Credit Amount

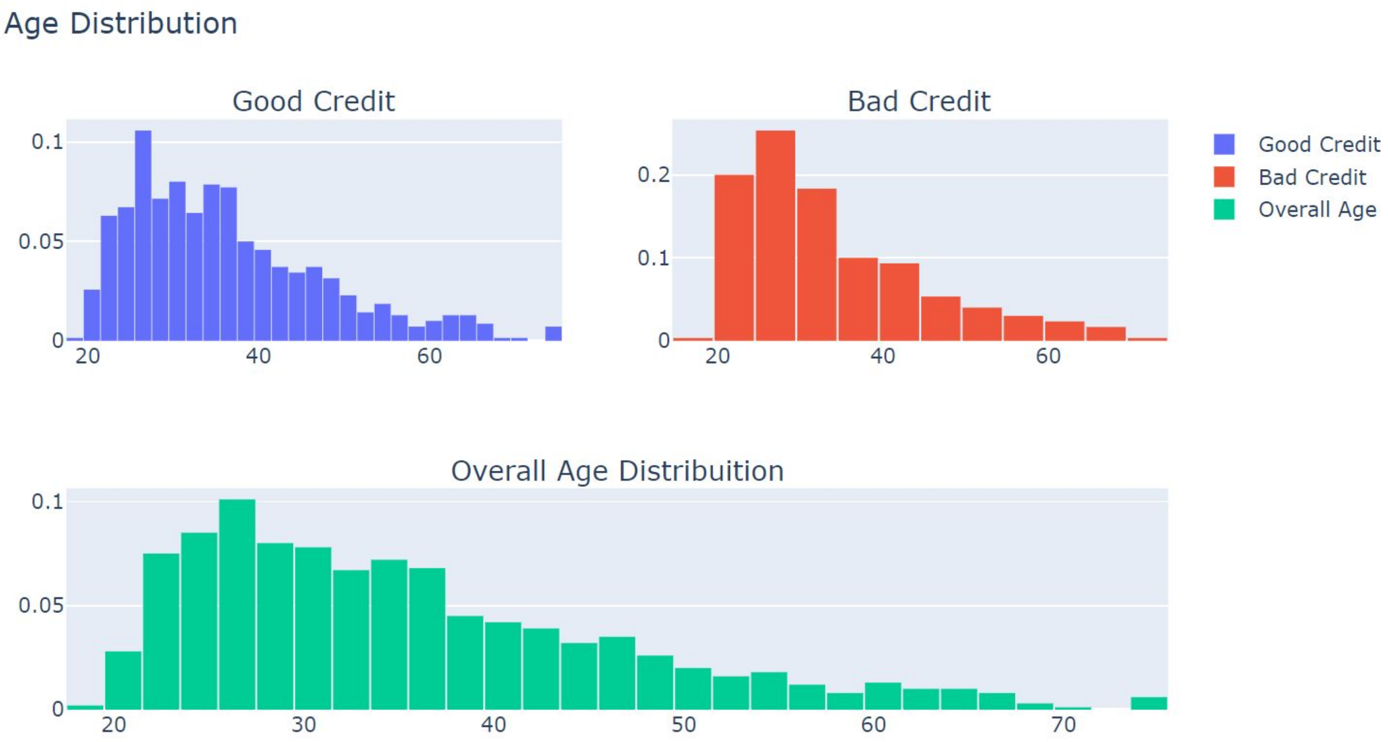








Age vs Risk


Age Distribution
Good Credit
Bad Credit
Overall Age Distribuition
0.1
0.05
0
0.2
0.1
0
20
40
60
20
30
40
50
60
70
Good Credit
Bad Credit
Overall Age

Scenario

Black Mirror Scenario

- In a more dystopian scenario, the use of XAI can perpetuate existing **biases** and **discrimination** in the lending process.

- If the dataset used to train the model is **biased against certain groups** of people (e.g., based on race or gender), the model will learn to associate those characteristics with a high-risk score and may **unfairly reject** loan to qualified borrowers that belong in those groups.

- This lead to a **false sense of objectivity** and **accuracy**, enabling lenders to justify decisions that rely on flawed and biased data and hence, making it harder to detect and address instances of discrimination or unfair treatment.

White Mirror Scenario

- Providing explanations of how the risk score was predicted can **promote fairness** and **transparency** in the lending process

- Understanding the crucial features and their significance in the contribution to the resulting risk, lenders and borrowers can **build trust**, leading to more responsible borrowing and lending practices and thus making more **reliable decisions.**

- Lenders could provide **targeted advice** to help potential borrowers enhance their creditworthiness and explain in detail how factors such as their job, savings account balance, or credit score influenced their loan approval decision, increasing their chances of future approval.

Sources of Bias

Sources of Bias

- It is crucial to understand that using gender and age (the are more than 200 forms of human cognitive bias) as a basis for data-driven decisions is typically prohibited by anti-discrimination laws in numerous nations.

- Individuals or institutions that provides decisions based on attributes to other individuals or organizations, are anticipated to make unbiased decisions based on objective factors such us credit history, income, etc. rather than focusing on individual features like gender, age or religion.

Models

Models

- Random Forests (sklearn)
- GradientBoostingClassifier (sklearn)
- Raw models give very low F1 score (almost 30%) for class 0 (bad risk)
- So oversampling was used to balance classes
- Another decision was to drop “Sex” and “Age” and see resulting F1 score

Results

Model	Resampling	Dropped Columns	Weighted F1 Score
Random Forest	No	None	0.69
Random Forest	Yes	None	0.88
Gradient Boosting Classifier	No	None	0.64
Gradient Boosting Classifier	Yes	None	0.76
Random Forest	Yes	“Sex”	0.86
Gradient Boosting Classifier	Yes	“Sex”	0.75
Random Forest	Yes	“Sex”, “Age”	0.84
Gradient Boosting Classifier	Yes	“Sex”, “Age”	0.73

Results-Discussion

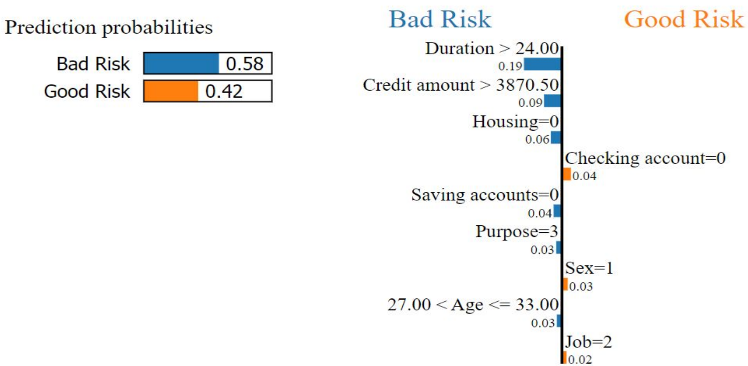
- Removing variables “Sex” and “Age” does not lead to significantly worse results
- This means that we get the same final accuracy without taking into account gender and age of applicants
- This means that the final model is less biased

Explainability

Explainability

- 2 main methods for midterm: LIME and SHAP
- Comparison between the 2 for same instances
- Both explain why the model chooses final label

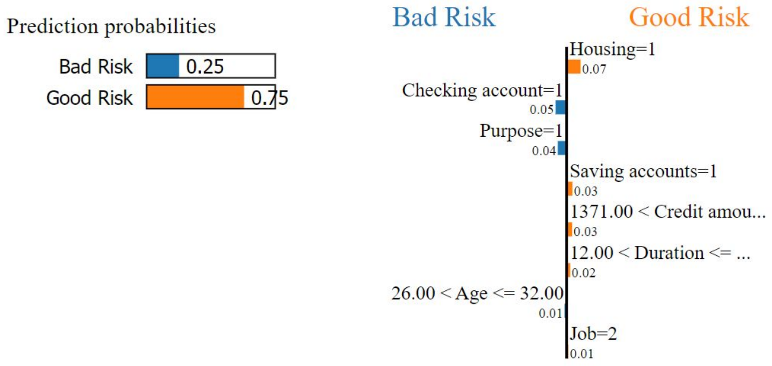
LIME-Comparison RF vs GBC, no resample, all variables



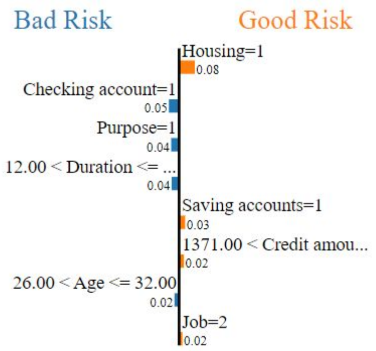
Feature	Value
Duration	48.00
Credit amount	6110.00
Housing=0	True
Checking account=0	True
Saving accounts=0	True
Purpose=3	True
Sex=1	True
Age	31.00
Job=2	True

Feature	Value
Duration	48.00
Housing=0	True
Credit amount	6110.00
Saving accounts=0	True
Checking account=0	True
Age	31.00
Job=2	True
Sex=1	True
Purpose=3	True

LIME-Comparison RF vs GBC, oversample, no Sex

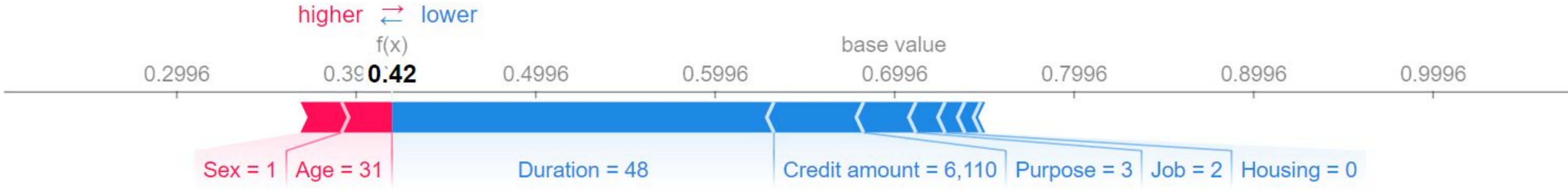


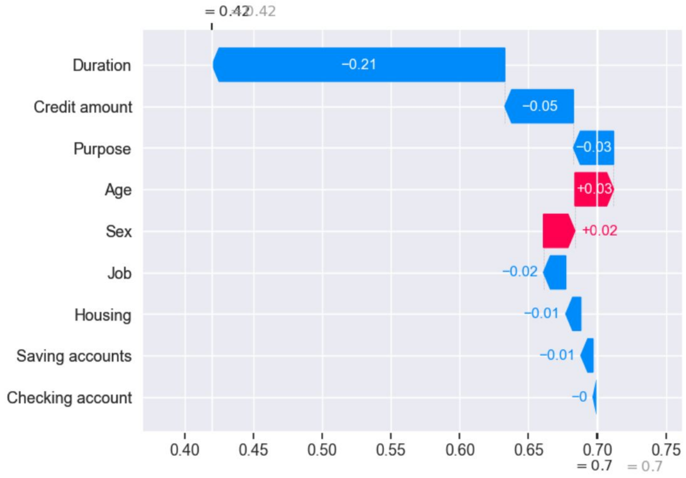
Feature	Value
Housing=1	True
Checking account=1	True
Purpose=1	True
Saving accounts=1	True
Credit amount	2278.00
Duration	18.00
Age	28.00
Job=2	True

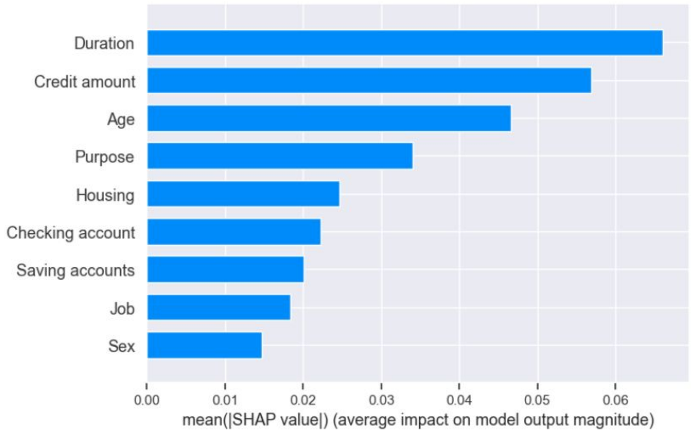


Feature	Value
Housing=1	True
Checking account=1	True
Purpose=1	True
Duration	18.00
Saving accounts=1	True
Credit amount	2278.00
Age	28.00
Job=2	True

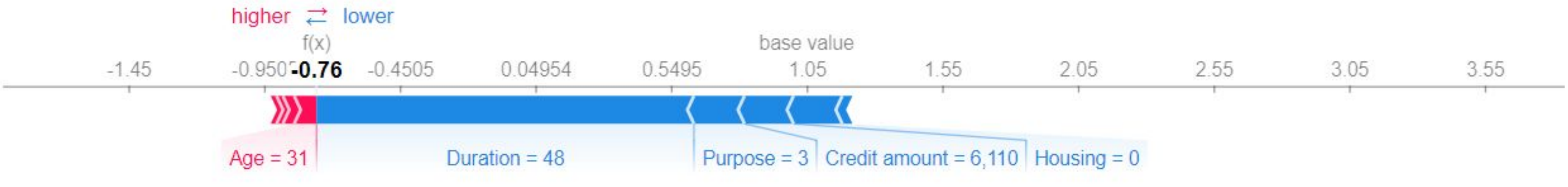
SHAP-RandomForest, no resample, all variables

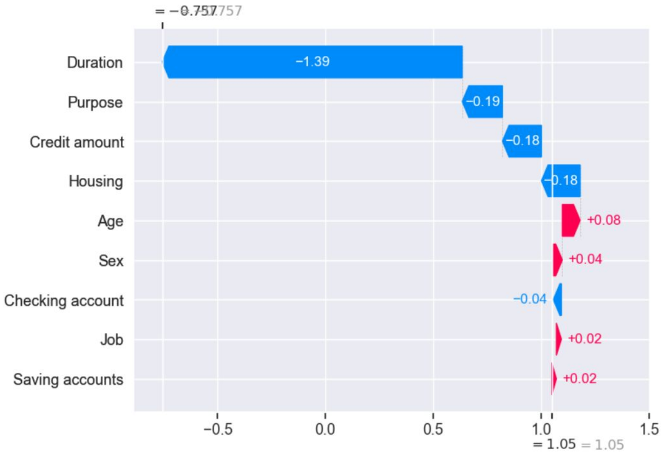


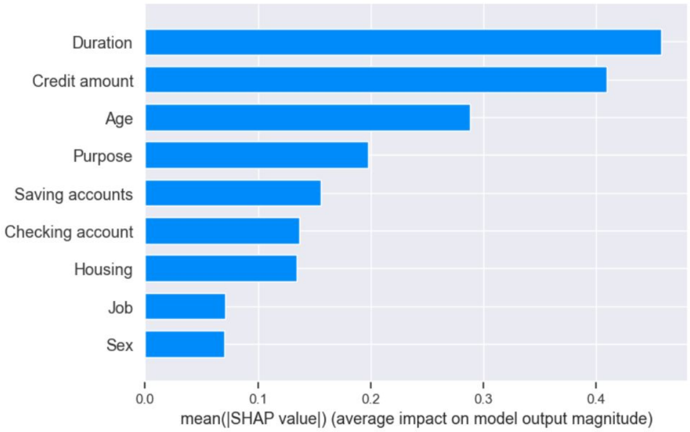




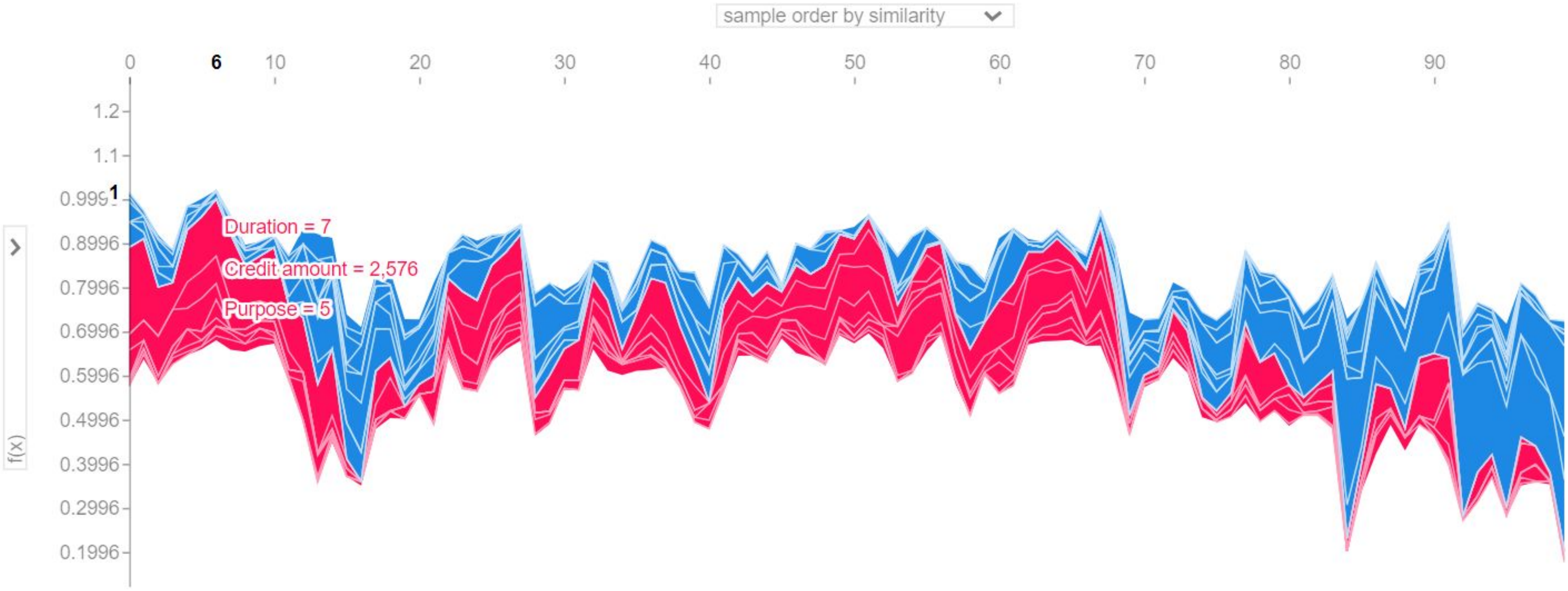
SHAP-GBC, no resample, all variables



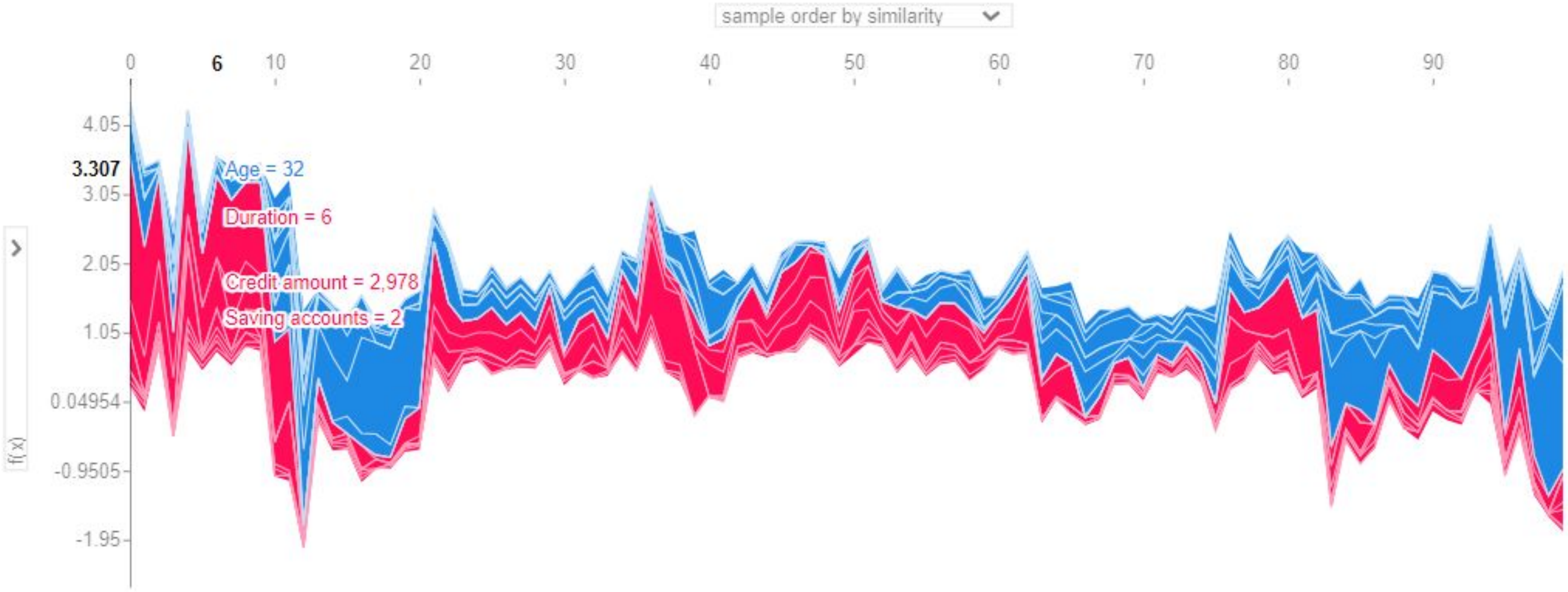




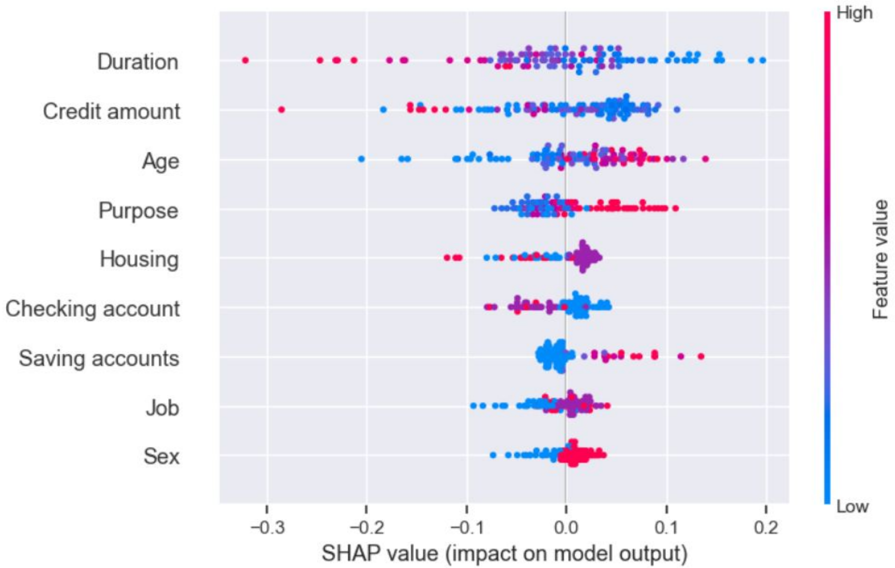
SHAP-RandomForest, no resample, all variables, waterfall

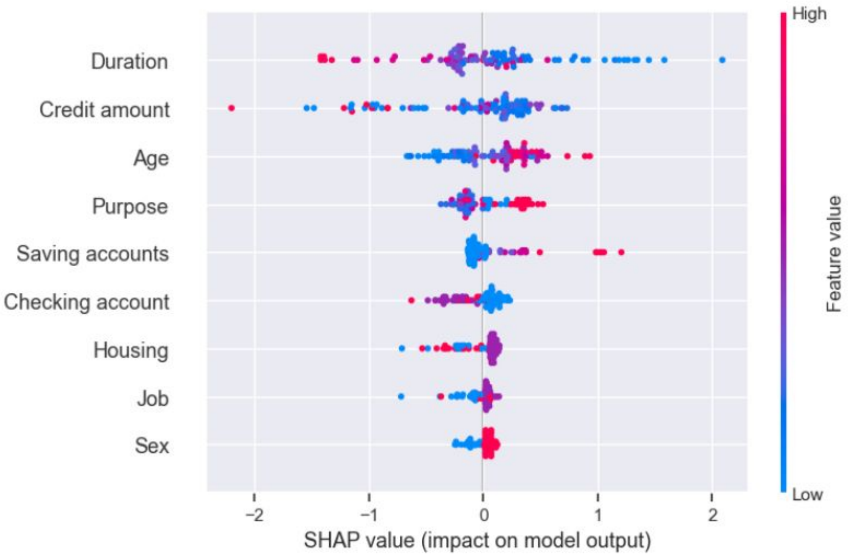


SHAP-GBC, no resample, all variables, waterfall

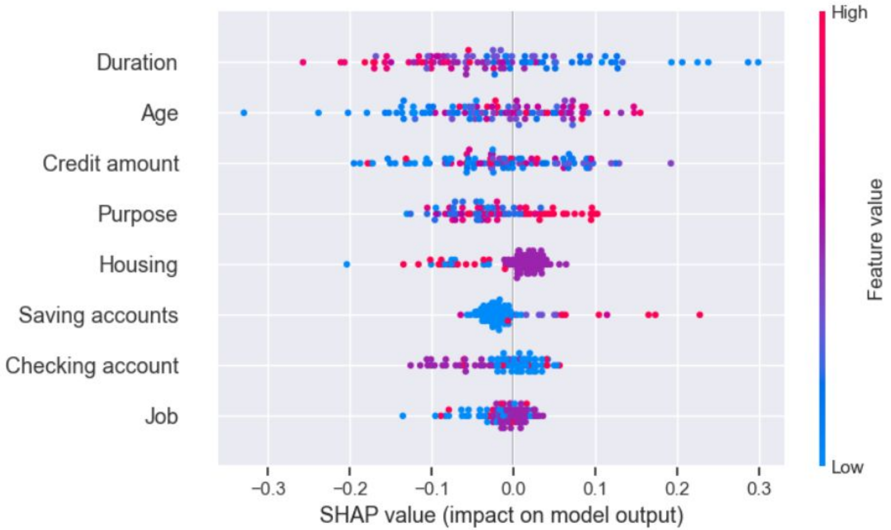


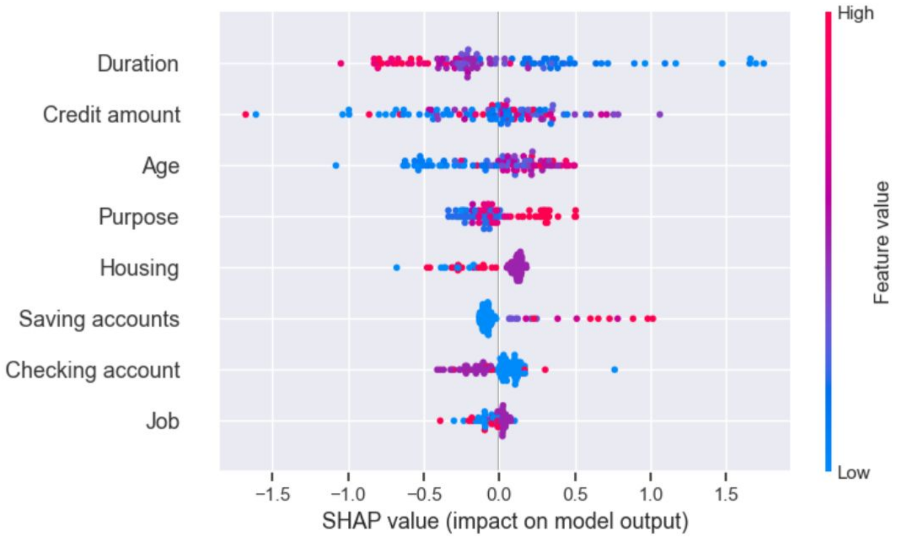
SHAP-RandomForest vs GBC, no resample, all variables, summary





SHAP-RandomForest vs GBC, oversample, dropped Sex, summary





Conclusions-Dataset

- Accept/Decline of loan could be crucial
- Declining a loan that can be repaid or accepting one that cannot be repaid has consequences both for bank and for individual
- Reasons for accept or decline should be clear

Conclusions-Models

- Basic unbalanced dataset does not allow for high F1 score on both classes
- If oversampling is performed, then both RandomForest and GradientBoostingClassifier work well
- Dropping column “Sex” and/or “Age” does not significantly decrease performance

Conclusions-Explainability

- Both LIME and SHAP provide useful insights for both models
- LIME gives only local explanations, while SHAP provides some nice global visualizations as well
- According to <u>SHAP</u>: <u>duration of account</u>, and <u>credit amount</u> are most important factors (using only these 2 covariates leads to 83% F1 score vs 88%)

Future Work

- Expand on theoretical background (scenario and sources of bias)
- Try another method such as anchors and see results
- Discover more useful insights about dataset

Thank you.
Questions?