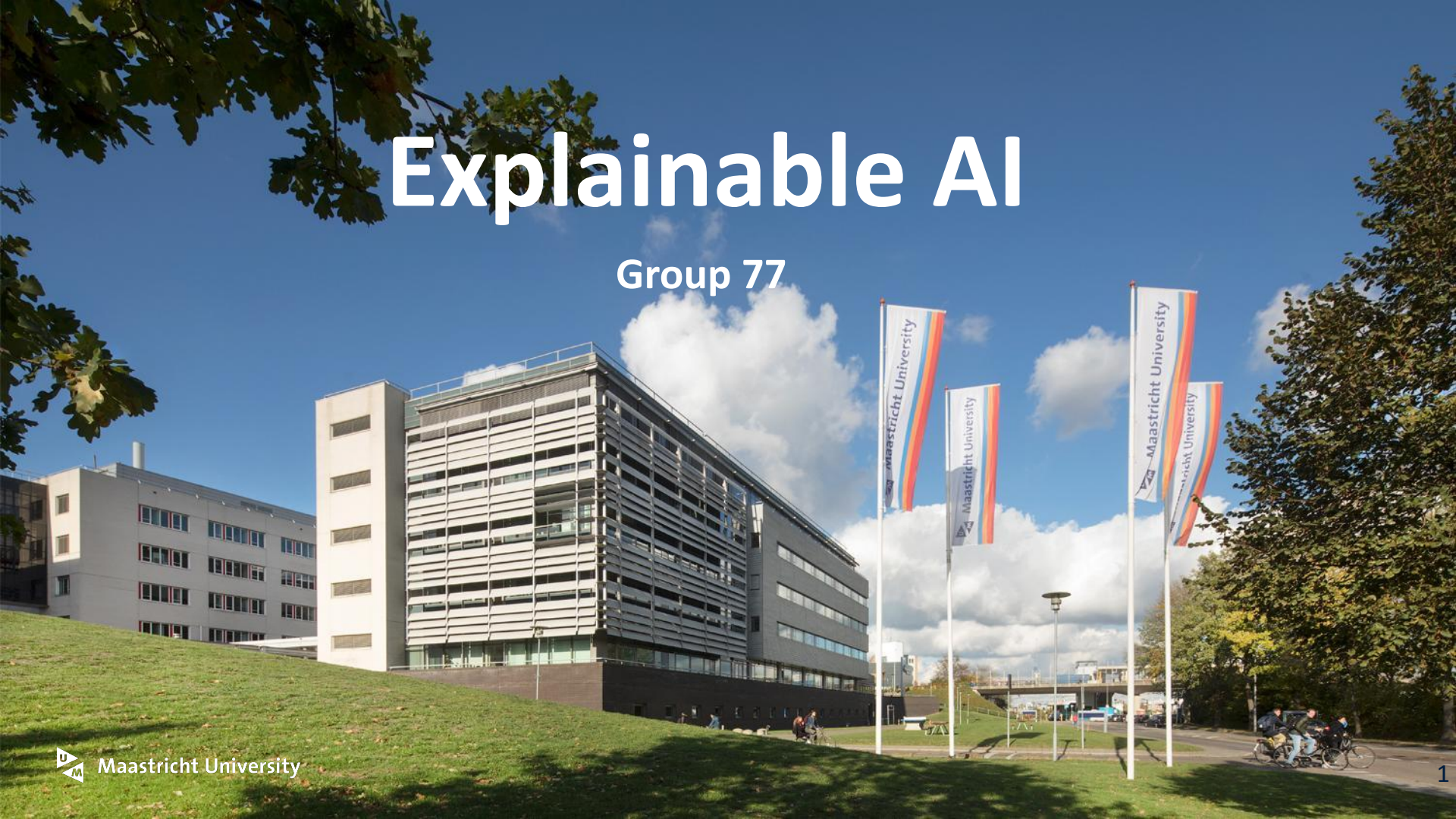
Explainable AI

Group 77



Group Members

Eleftherios Tetteris
Economist

Nikolaos Athanasopoulos
Mechanical Engineer

Chrysanthi Foti
Mathematician

Outline

Outline

- Problem Statement
- Scenarios
- EDA
- Models
- Explainability
- Evaluation

Problem Statement

Introduction-Data Structure

- German Credit Dataset
- Has multiple variants
- This variation has 9 covariates for classification
- 1 binary output (Good Risk vs Bad Risk for loan)
- Good Risk = you can probably repay the requested loan
- Bad Risk = you probably cannot repay the requested loan

Introduction-Data Structure

	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

Scenarios

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.

White Mirror Scenario

- In case of a predicted Bad Risk, one can exploit the targeted advices given by the lenders based on the XAI methods and improve any factors (that can be modified) and negatively affected the predicted score.

- Increase saving account, checking account, credit amount by either transferring money from other bank accounts that may belong to the same person or by being helped from parents/friends.

Black Mirror 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.

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.

Exploratory Data Analysis

Introduction-Numeric Data

- Age ranges 19-75 years
- Credit Amount ranges 250-18,000 credits
- Account duration ranges 4-72 months
- Job has 4 values: “unemployed”, “unskilled and non resident”, “skilled/official” & “management/self_employed/highly_qualified

Introduction-Data Structure-Categorical data

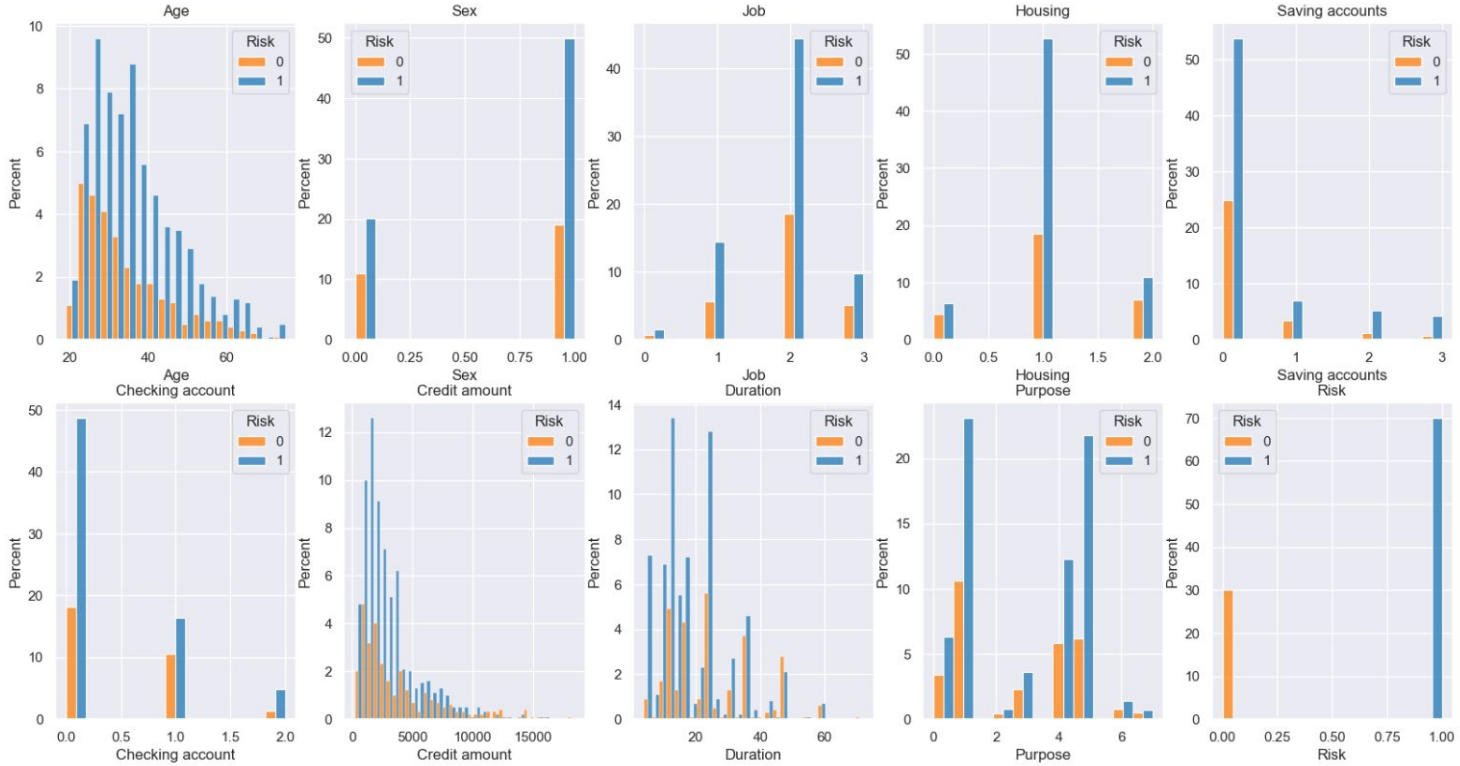
```
Sex : ['male' 'female']

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']
```

Categorical Data Cleaning

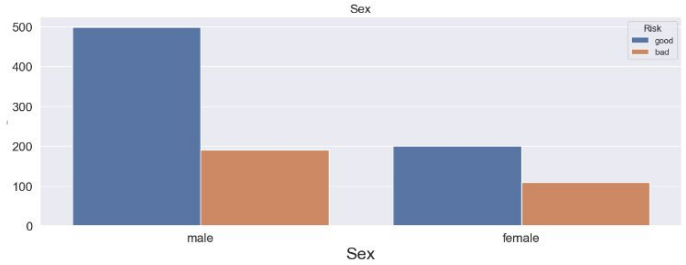
- 3 possibilities for NaN values: drop columns, drop rows and replace with mode
- Drop rows and drop columns do not lead to good results, so mode was used in the end
- Afterwards, the categorical features were encoded into integers

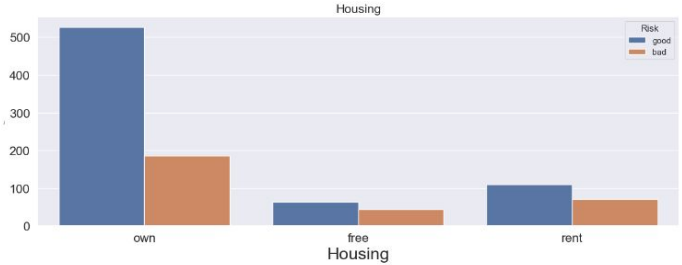
```
Sex
{'female': 0, 'male': 1}
--------------------------------------------------
Housing
{'free': 0, 'own': 1, 'rent': 2}
--------------------------------------------------
Saving accounts
{'little': 0, 'moderate': 1, 'quite rich': 2, 'rich': 3}
--------------------------------------------------
Checking 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}
```

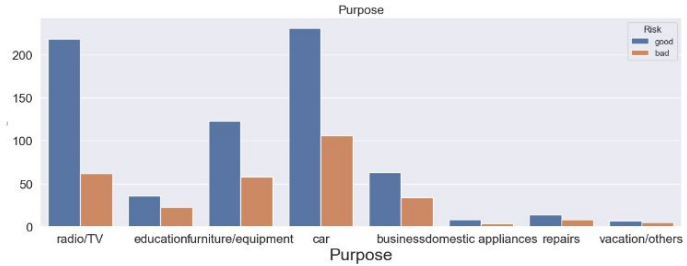
Data Distributions



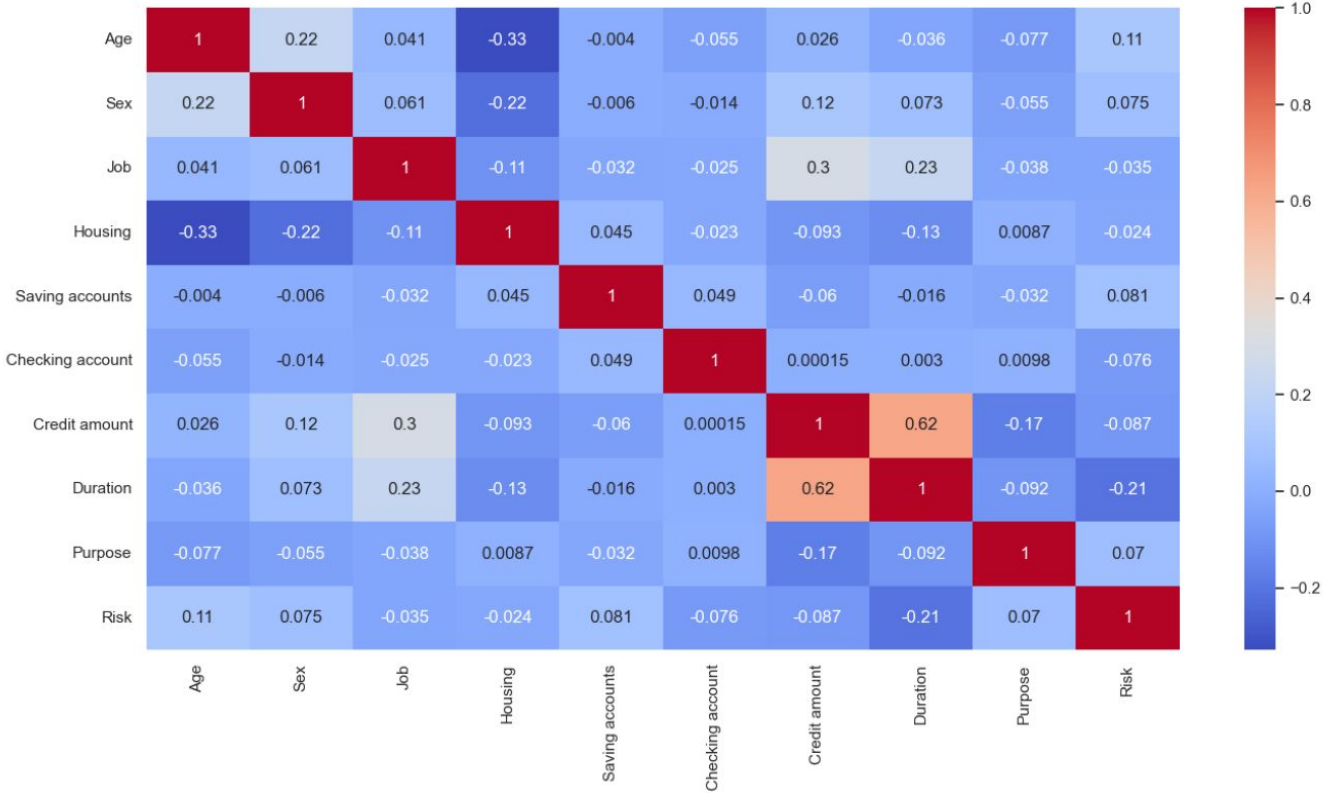
Data Distributions - Categorical



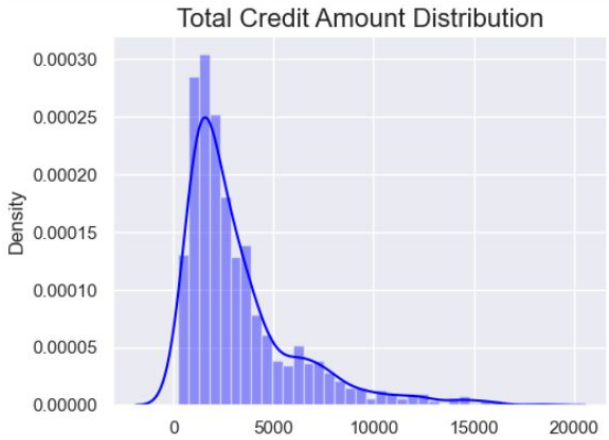


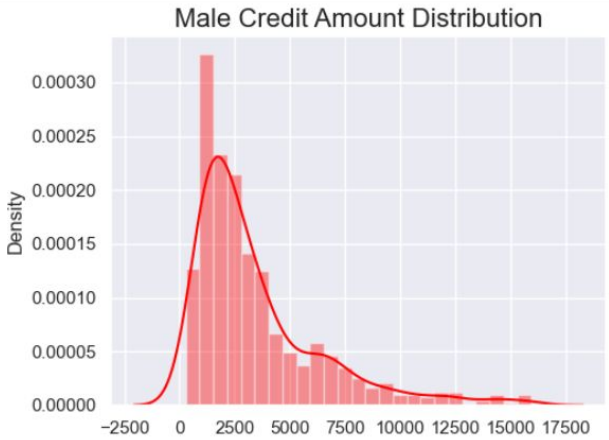


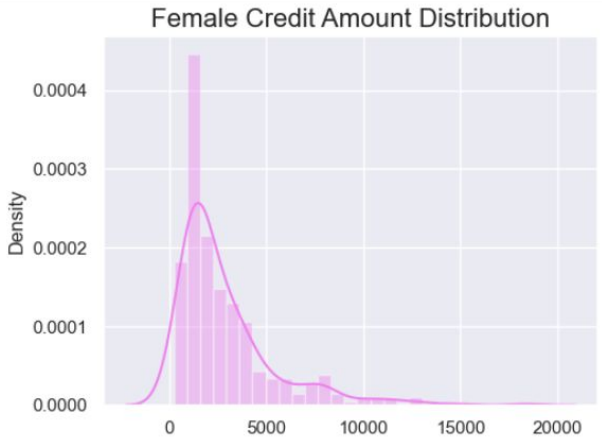
Feature Correlation Heatmap-Spearman r



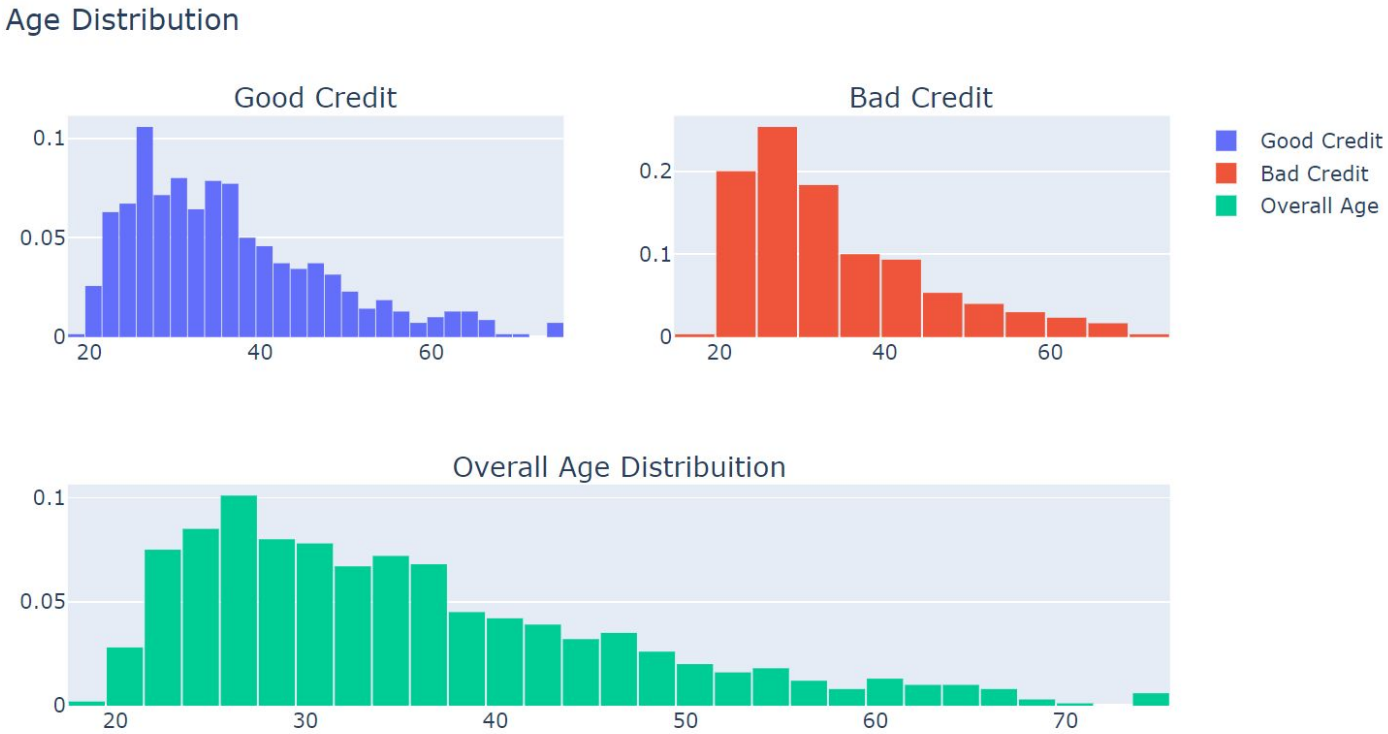
Male-Female vs Credit Amount



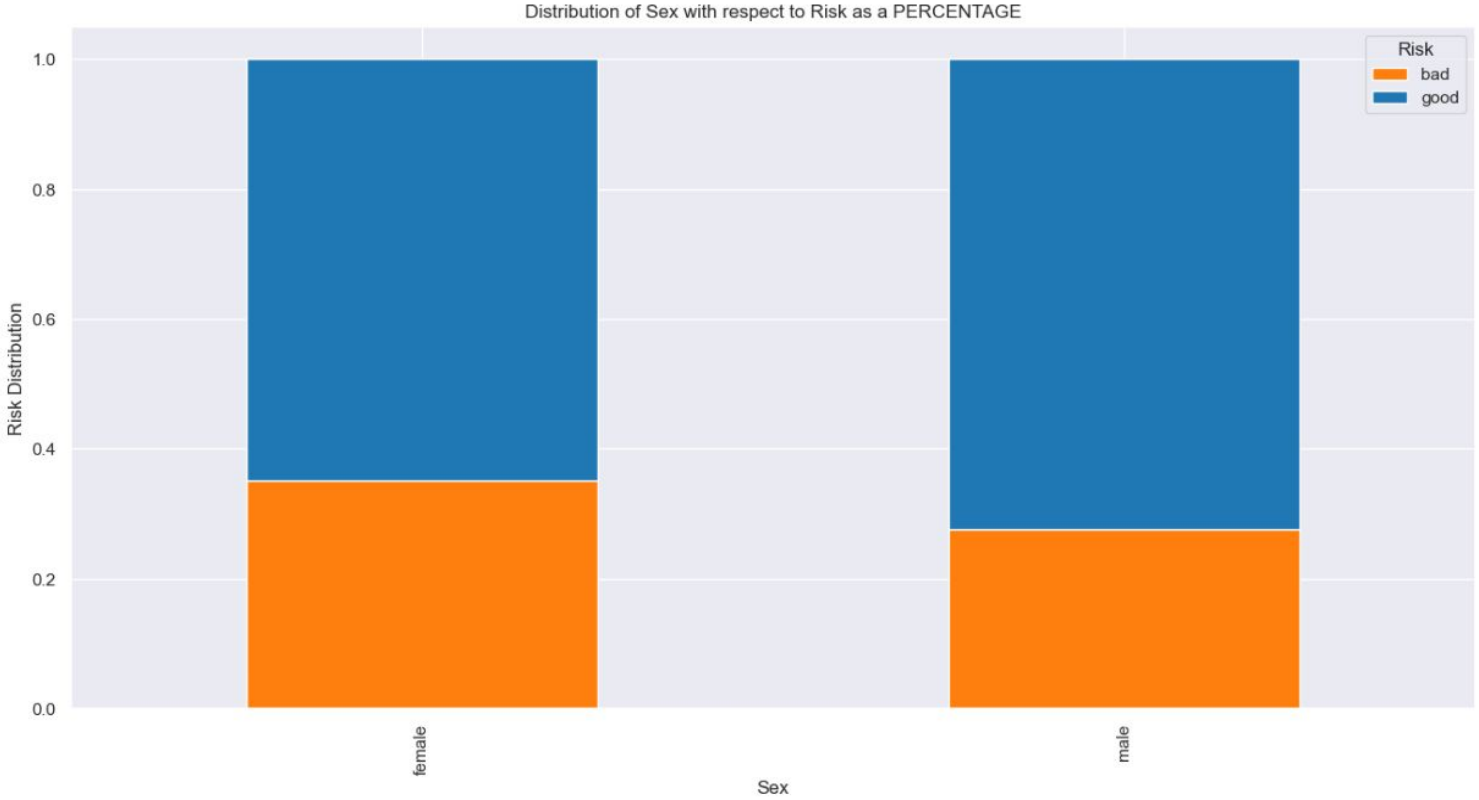




Age vs Risk


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

Sex vs Risk


Distribution of Sex with respect to Risk as a PERCENTAGE
Risk
bad
good
Risk Distribution
1.0
0.8
0.6
0.4
0.2
0.0
female
male
Sex

Models

Models-Things we learned from Midterm

- Gradient Boosting Classifier is slightly worse than Random Forest, but with same explainability
- We need an inherently interpretable model in order to compare with LIME & SHAP
- Oversampling is important

Models-Testing

- A plethora of models were tested to see results (Random Forests, Logistic Regression, SVM, GradientBoostingClassifiers, Naive Bayes etc)
- In the end, Random Forests and Logistic Regression were selected
- The selection was done with respect to AUC score, while also looking at Fb scores (b=0.5, 1, 2)
- Also did oversampling (Random, ADASYN, SMOTE)

Models Motivation - Random Forests

- Handling Non-Linear relationships
- Robust to Outliers
- Overfitting Reduction
- High Performance

Models Motivation - Logistic Regression

- Inherently Interpretable model
- Explicit Feature Importance through coefficients
- Regularization can reduce overfitting
- Somewhat robust to outliers

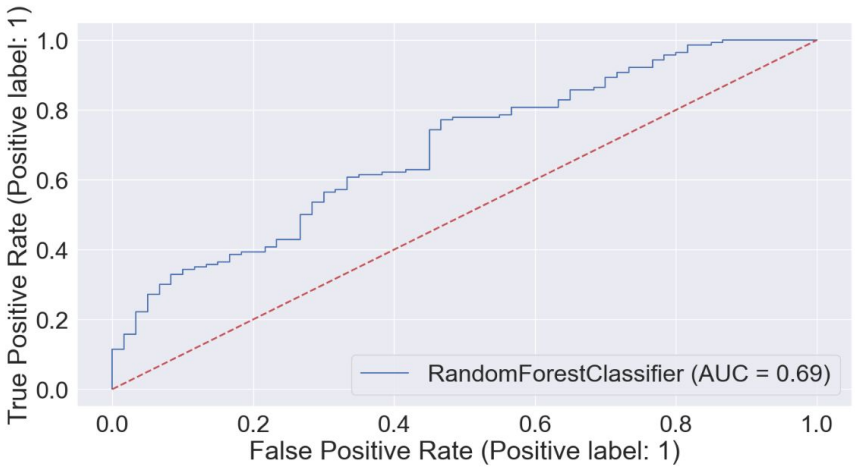
Models Results

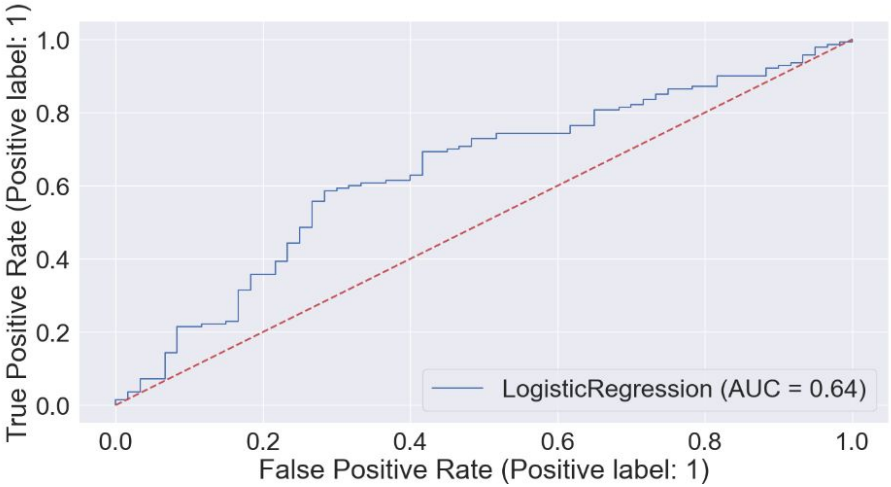
- Random Oversampling is the best resample strategy
- Grid Search for Random Forest produced the best performing model
- Also performed Grid Search for class weights on Logistic Regression

Models Results - continued

Model	F1	F0.5	F2	AUC
Best Random Forest	0.70	0.77	0.80	0.69
Best Logistic Regression	0.65	0.77	0.69	0.64

Models Results - ROC Curves





Explainability

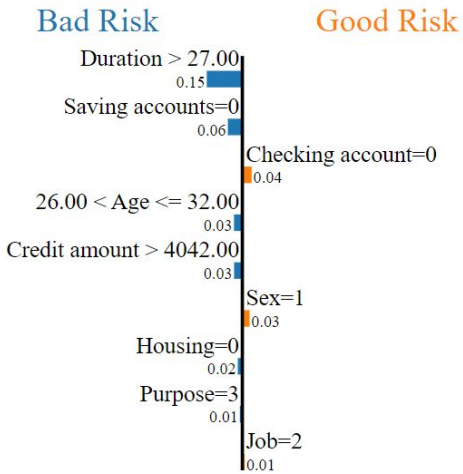
Explainability

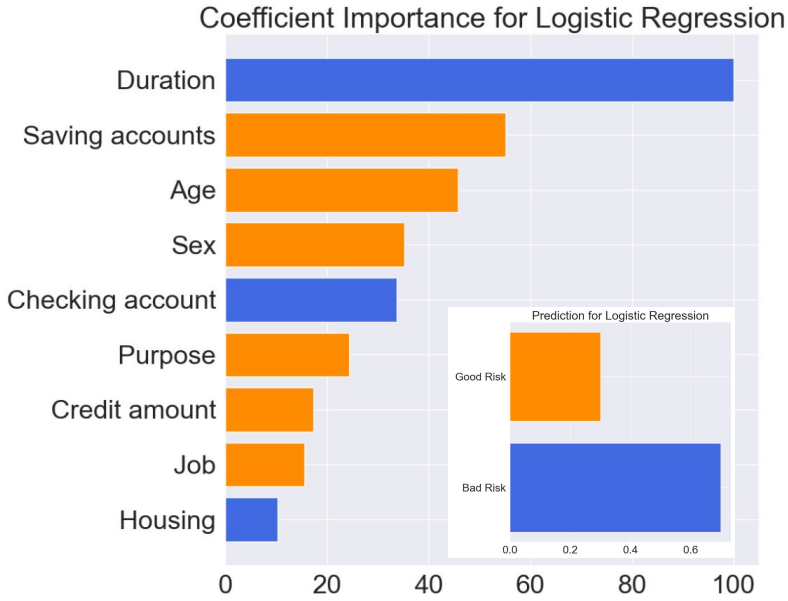
- 3 main methods: LIME and SHAP for RandomForest & Logistic Regression
- Comparison between the 3 for same instances
- All explain why each model chooses final label
- Interesting instances: where model disagrees with explainer, or both disagree with true label

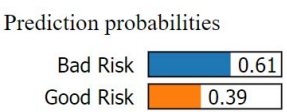
LIME & SHAP-Comparison RF with LogisticRegression

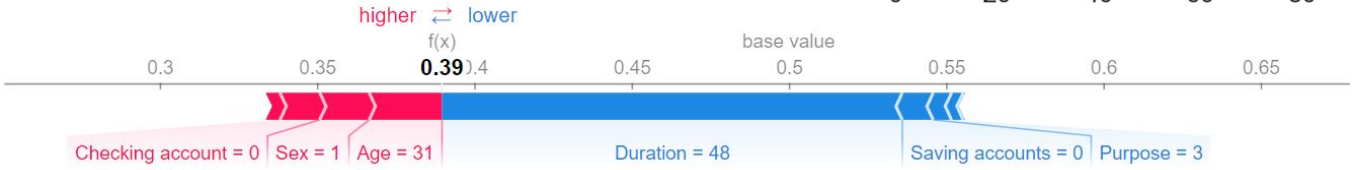
Feature	Value
Duration	48.00
Saving accounts=0	True
Checking account=0	True
Age	31.00
Credit amount	6110.00
Sex=1	True
Housing=0	True
Purpose=3	True
Job=2	True

```
Model prediction:  Good Risk
True label:  Bad Risk
LIME predicted label:  Bad Risk
```







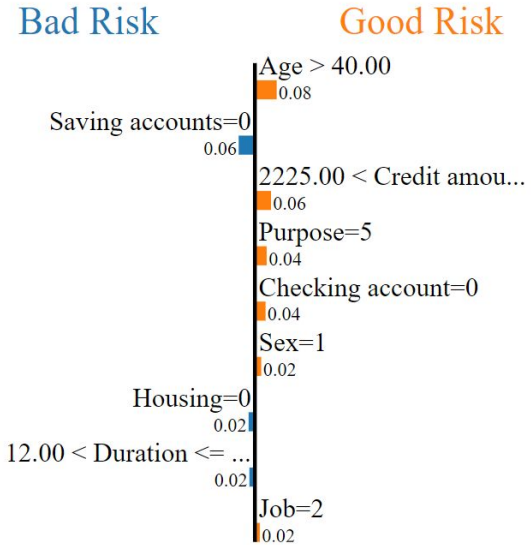


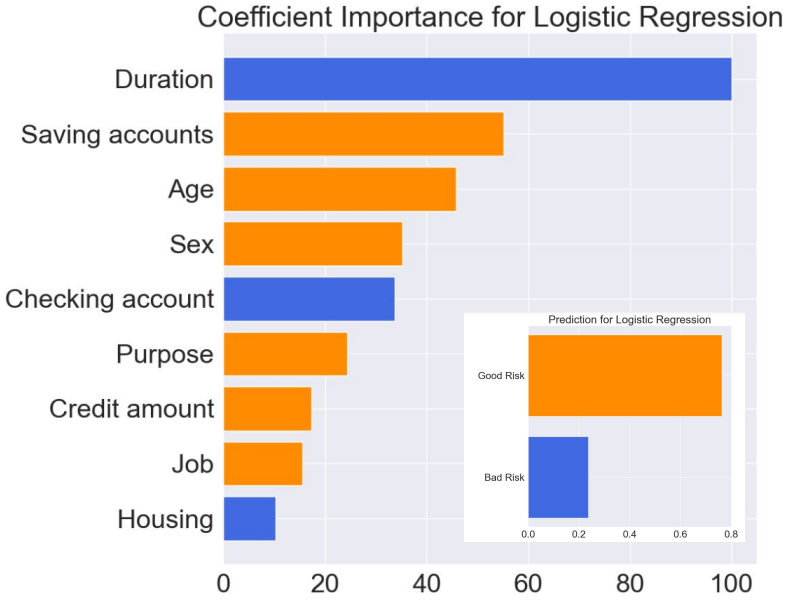
LIME & SHAP-Comparison RF with Logistic Regression (wrong pred)

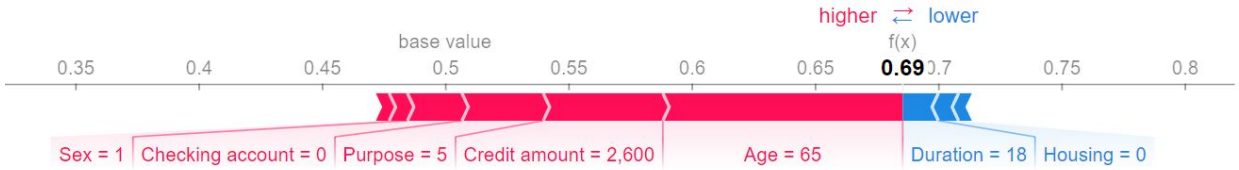
Feature	Value
Age	65.00
Saving accounts=0	True
Credit amount	2600.00
Purpose=5	True
Checking account=0	True
Sex=1	True
Housing=0	True
Duration	18.00
Job=2	True

Prediction probabilities

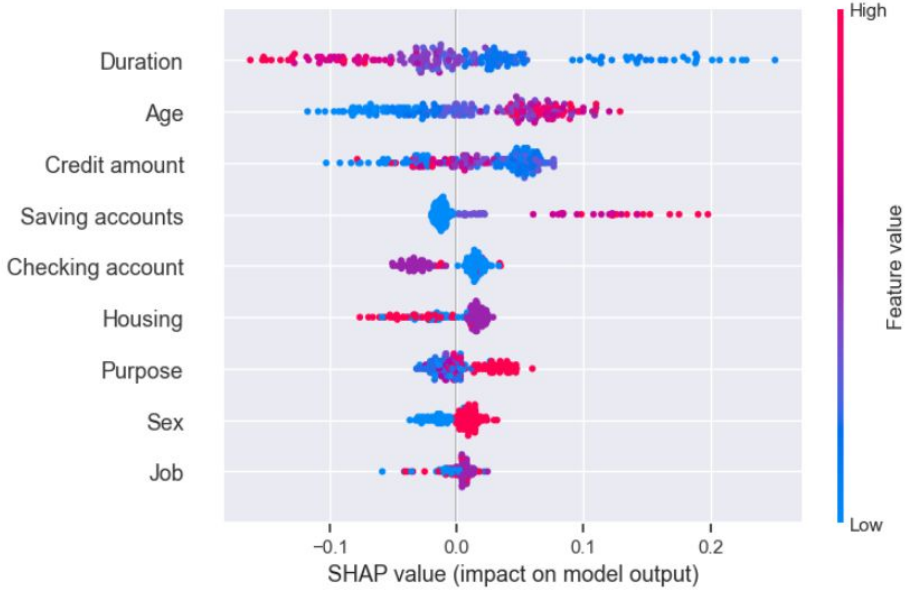

```
Model prediction:  Good Risk
True label:  Bad Risk
LIME predicted label: 1 (Good Risk)
```

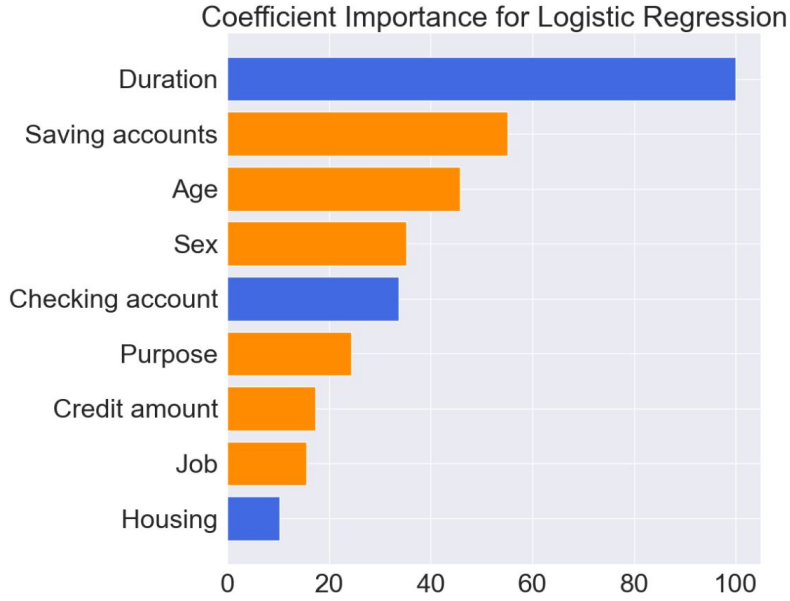






SHAP Beeswarm plot vs Logistic Regression coefficients





SHAP Waterfall Plot-How to pick instances

Conclusions-XAI

- LIME & SHAP agree in instance explainability, with almost the same feature importance
- Both tend to agree with Logistic Regression on instance level, both in feature importance and label direction
- SHAP global beeswarm plot somewhat agrees with LR coeffs, but slightly different ranking

Conclusions-XAI

- All of them give valuable insights about the final prediction
- All have certain inaccuracies
- Some coefficients and explanations do not make much sense (e.g more duration of account => good risk)
- The dataset is probably somewhat bad (original labels were not consistent)

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

Evaluation of Explanations

Evaluation of Explanations

- Independent variable: 2 different kinds of graphs (LIME & SHAP)
- Dependent variable:
  - time (in seconds) to reach a decision
- Hypothesis: The utilization of SHAP is expected to yield easier/faster decision-making outcomes than LIME

Evaluation of Explanations - Participants

- Non-experts with basic loan knowledge
- Age range 18-55
- Independent participants
- Exclusion criterion: All individuals should not have seen the graphs before

Evaluation of Explanations - Experiment design (1)

- Online experiment - one time
- The participants will be provided with a set of graphs (8). This will include the visualised explanations given from LIME (4) and SHAP (4).
- Extrinsic evaluation: Take decision of Bad or Good risk based on the graphs

Evaluation of Explanations - Experiment design (2)

- The time will not be shown (no time-pressure)
- Textbox at the end of decision to explain the reason that reached to the specific decision
- The experiment is not safety-critical but loan decision may affect the quality of life
- Participants will not have access to the outcome (true label)

Evaluation of Explanations - Experiment design (3)

- Within subject experiment: All the participants will examine both LIME & SHAP graphs

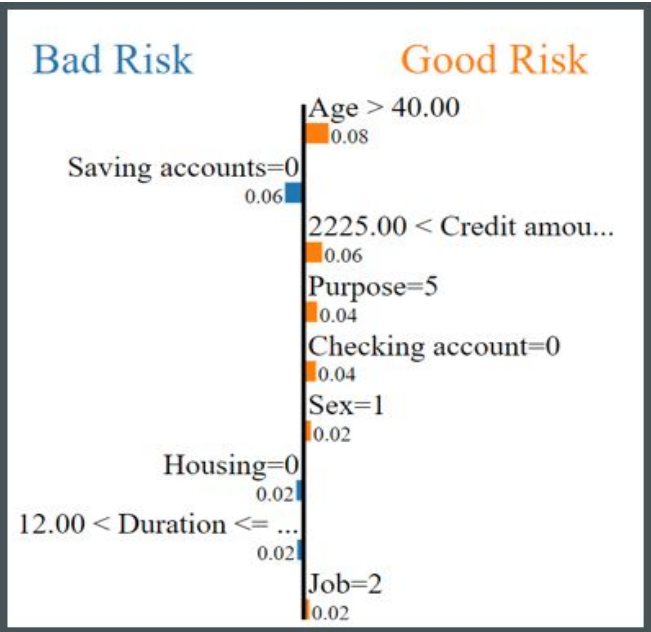
- Objective measures - the time needed to reach in a decision (bad/good risk)

You will be provided with a graph showing the predicted risk (Good/Bad) for a loan decision approval. You are called to decide based on the graph if it indicates a good or bad risk. (Check the respective box)

The predicted risk concerns of a 65 year old man, skilled employee that lives with his parents. He has little saving accounts and checking accounts while at the bank has 2600 € . His bank account is active for 18 months. The individual wants a loan to buy a radio/TV.

You will be provided with a graph showing the predicted risk (Good/Bad) for a loan decision approval. You are called to decide based on the graph if it indicates a good or bad risk. (Check the respective box)

The predicted risk concerns of a 65 year old man, skilled employee that lives with his parents. He has little saving accounts and checking accounts while at the bank has 2600 € . His bank account is active for 18 months. The individual wants a loan to buy a radio/TV.



Good Risk ☐ Bad Risk ☐

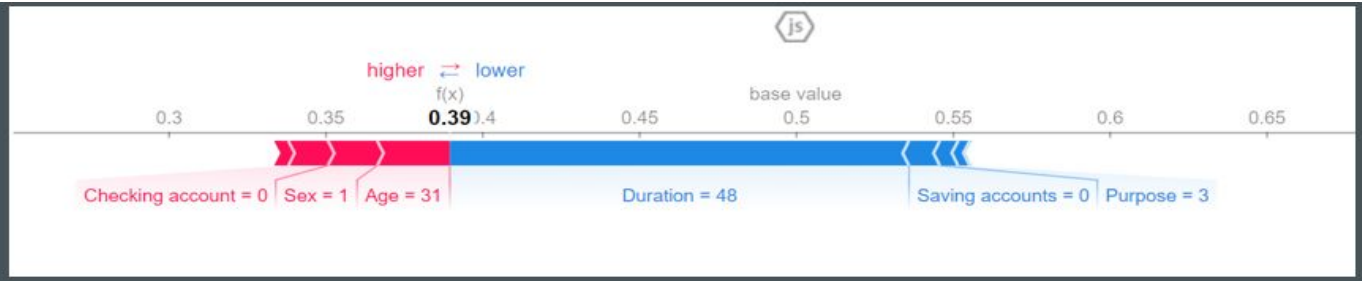
Explain why you reached to this decision:

You will be provided with a graph showing the predicted risk (Good/Bad) for a loan decision approval. You are called to decide based on the graph if it indicates a good or bad risk. (Check the respective box)

The predicted risk concerns of a 31 year old man, skilled employee that lives with his parents. He has little saving accounts and checking accounts while at the bank has 6110 € . His bank account is active for 48 months. The individual wants a loan to buy a car.

You will be provided with a graph showing the predicted risk (Good/Bad) for a loan decision approval. You are called to decide based on the graph if it indicates a good or bad risk. (Check the respective box)

The predicted risk concerns of a 31 year old man, skilled employee that lives with his parents. He has little saving accounts and checking accounts while at the bank has 6110 € . His bank account is active for 48 months. The individual wants a loan to buy a car.



Good Risk ☐ Bad Risk ☐

Explain why you reached to this decision:

Evaluation of Explanations - Analyse results

- The data will not be prepared
- Compare the results with the label of the method
- Statistically analyse the time of all graphs per method (LIME/SHAP)
- If there is no statistical difference between the times, it will still be possible to gain insights for each method from the text box

Evaluation of Explanations - Limitations

- It will be important to additionally know how much the participants trust each provided method based on the graphs.

Future Work

- Use more interpretable models and compare to logistic regression
- Try to discover how outliers are influencing results
- Expand on our evaluation of explanations

Thank you.
Questions?