AI Tools Lab 2020 - DBMS

Florian Betz & Astrid Büchner
Summerterm 2020
Dr. Todor Ivanov

Table of Content

The Use Case

Prediction of default payment of credit card clients.
To do so, we use classification trees as well as logistic regression.

The Dataset

- The data set contains information of Taiwanese credit card clients
- The dataset captures a timeframe from April 2005 to September 2005
- It was uploaded to kaggle.com in 2016 and there is no copyright for it
- Please find the whole dataset here

1. Introduction of the dataset
1.2 Overview of the dataset

Dataset size: 30‘000
Columns: 25
Depentent variables: 23
Rows: 30‘000

The Variables

ID	ID of each client (numbers datapoints consecutively)
LIMIT_BAL	Amount of given credit in NT dollars (includes individual & family/supplementary credit)
SEX	1 = male, 2 = female
EDUCATION	1 = graduate school, 2 = university, 3 = high school, 0,4,5,6 = others
MARRIAGE	Marital status: 1 = married, 2 = single, 3 = divorced, 0 = others
AGE	Age in years
PAY_0	Repayment status in September 2005
PAY_2	Repayment status in August 2005
PAY_3	Repayment status in July 2005
PAY_4	Repayment status in June 2005
PAY_5	Repayment status in May 2005
PAY_6	Repayment status in April 2005

NOTE: Possible values and their meaning valid for all PAY_ columns:

-2 = no consumption
-1 = pay duly
0 = the use of revolving credit
1 = payment delay for one month
2 = payment delay for two months
...
8 = payment delay for eight months
9 = payment delay for nine months and above

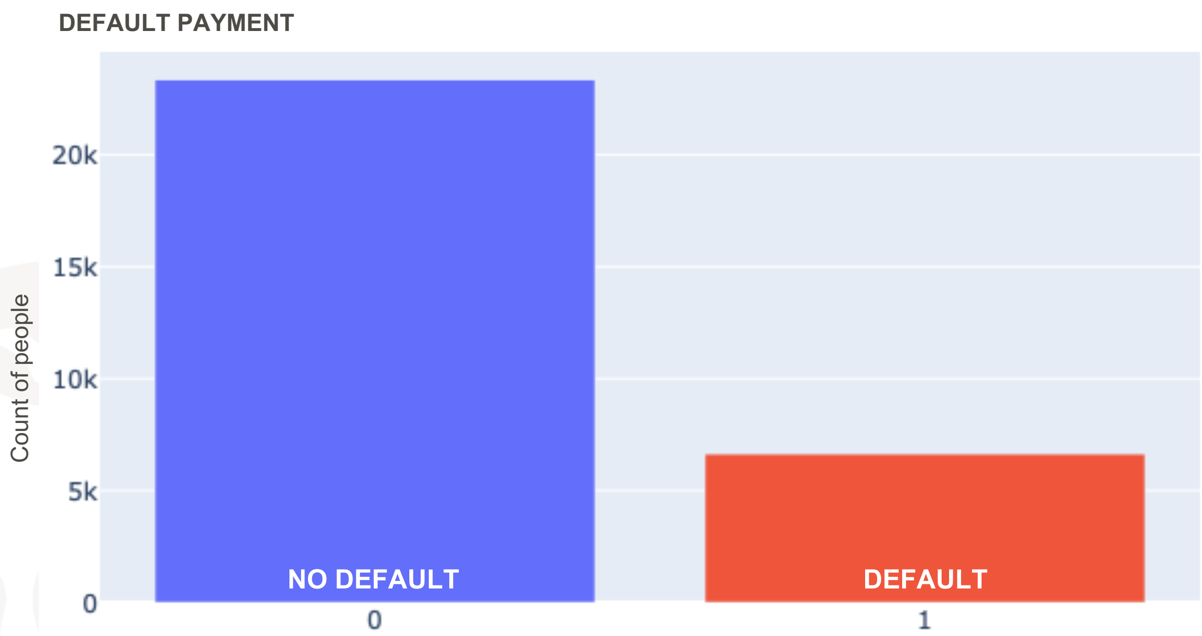
1. Introduction of the dataset
1.2 Overview of the dataset

BILL_AMT1	Amount of bill statement in September 2005 (NT dollar)
BILL_AMT2	Amount of bill statement in August 2005 (NT dollar)
BILL_AMT3	Amount of bill statement in July 2005 (NT dollar)
BILL_AMT4	Amount of bill statement in June 2005 (NT dollar)
BILL_AMT5	Amount of bill statement in May 2005 (NT dollar)
BILL_AMT6	Amount of bill statement in April 2005 (NT dollar)
PAY_AMT1	Amount of previous payment in September 2005 (NT dollar)
PAY_AMT2	Amount of previous payment in August 2005 (NT dollar)
PAY_AMT3	Amount of previous payment in July 2005 (NT dollar)
PAY_AMT4	Amount of previous payment in June 2005 (NT dollar)
PAY_AMT5	Amount of previous payment in May 2005 (NT dollar)
PAY_AMT6	Amount of previous payment in April 2005 (NT dollar)
default.payment.next.month	Default payment (1 = yes, 0 = no)

Note: The explanation of the variables given for the dataset was incomplete. We adjusted the variable explanation in relation to a kaggle user, who contacted the responsible professor and asked for the missing explanations. You can find his post here.

1. Introduction of the dataset
1.2 Overview of the dataset

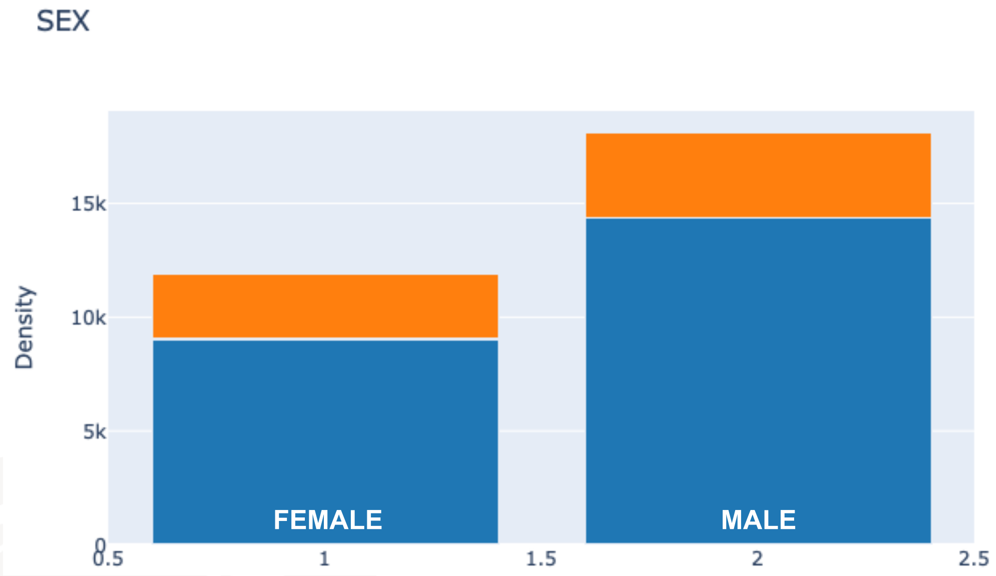
Visualization of Attributes



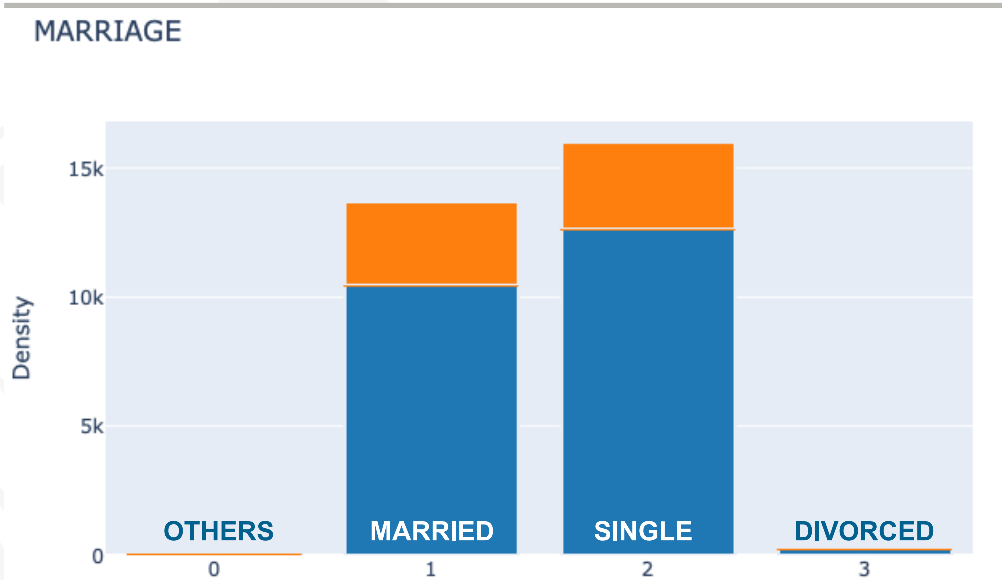
- ➢ The majority of clients pays their bills.
- ➢ There are still 6’598 default payments out of 30’000, which is about 22%.
- ➢ In the following, we will take a closer look at the distribution of default payments in terms of demographic data.

1. Introduction of the dataset

1.2 Overview of the dataset



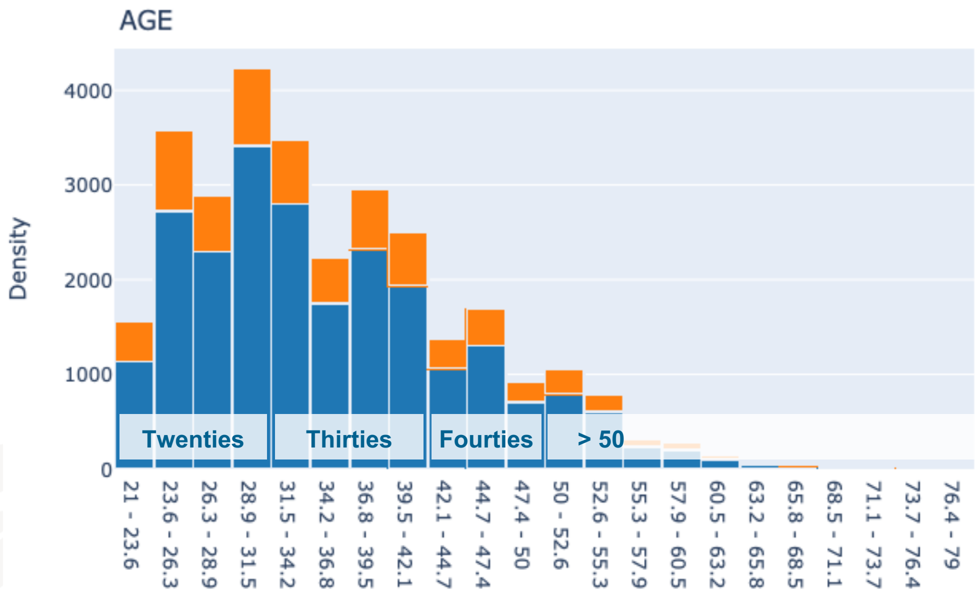
- ➢ The dataset contains more male than female subjects
- ➢ In absolute numbers it seems men rather default



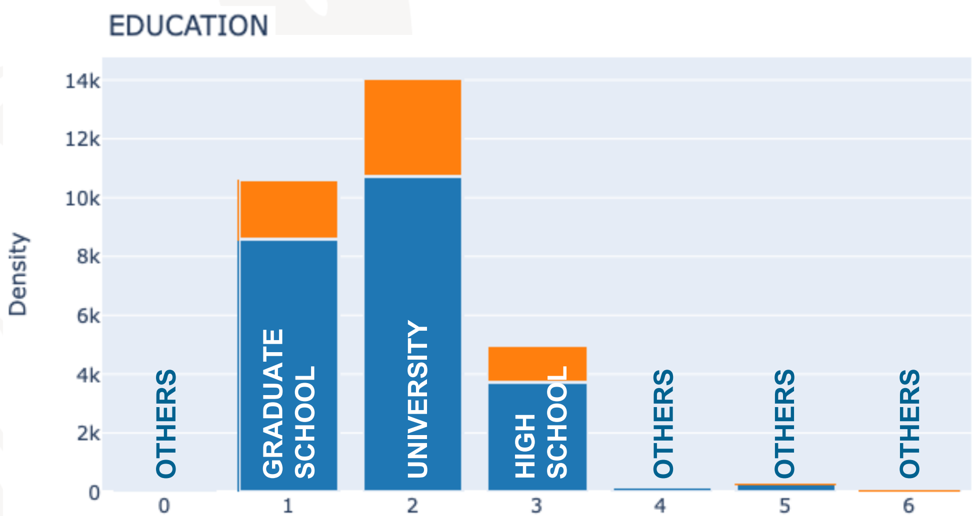
- ➢ Most people in the dataset are either single or married
- ➢ The minority is divorced
- ➢ The status “others” (= 0) is neglectable in this case

1. Introduction of the dataset

1.2 Overview of the dataset



- ➢ The majority of the subjects is in their twenties or thirties
- ➢ The higher the age (starting at 30), the fewer the count of people



- ➢ Most people in the dataset hold an university degree
- ➢ overall we can say the educational background of test persons is quite high
- ➢ We can also see that 4, 5, 6 & 0 (others) are a minority and do not seem to play a big role

1. Introduction of the dataset
1.3 Classification Tree & Logistic Regression

- To adress the use case of predicting default payments we use **Classification Tree** and **Logistic Regression**.

- Further, we used the **LIME Tabular Explainer** to get a better explanation of the results.

- While using the MS InterpretML Toolkit, we made use of the Classification Tree and Logistic Regression, that are already implemented in InterpretML.
- Whereas for the AIX360 approach, we had to use the models by scikit learn and applied AIX360 tools afterwards.

2. Data Cleaning

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

To prepare and clean the dataset in order to apply the models, we made some modifications:

1. **Check for null values**
   The dataset has no null values
2. **Rename columns**
   Change of name of the independent variable to "default_pay", for convenience. Change of the column "PAY_0" to "PAY_1" for consistency
3. **Convert currency**
   To get a better reference New Taiwan Dollar is changed to Euro (Exchange rate: Euro ≈ 0.03 * Taiwan-Dollar 9. Juni, 18:11 UTC)
4. **Change "SEX" 2 to 0**
   Change of numerical representation for male clients from 2 to 0, to get a dummy variable.
5. **Drop columns containing "other/unknown"**
   The columns "EDUCATION" and "MARRIAGE" have other/unknown values. These are relatively rare, so these rows are dropped. They don't add value to the model, and cannot be interpreted
6. **Delete ID & rearrange index**
   Deletion of column "ID" (it is just a random consecutively numbering of the datapoints, no impact)
7. **Categorize data**
   Categorization of ordinal and nominal data, to change them to dummy variables
8. **Correlation matrix**
   Correlation between cardinal columns:
   - "LIMIT_BAL" has by far the biggest correlation with default payment
   - "BILL_AMTX" and "PAY_AMTX" are highly correlated among themselves, but its declining dependent on time
   - "BILL_AMT1" is more correlated with default_payment than BILL_AMT2" and so on...
   - "PAY_AMT1" is more correlated with default_payment than "PAY_AMT2" and so on...
9. **Crosstabs**
   Analysis of dependencies of ordinal and nominal data:
   - There is a big gap in defaults between single and divorced clients
   - Highly educated people default less
   - Male clients default less than female clients
   - Bigger payment delay results in higher chance of default
   - The default rate is rising depending on time (comparing "Pay_1" with "Pay_2" and so on..)
10. **Determine dependent and independent variables**
11. **Get dummies for independent variables**

3. Approach to the models

	MS InterpretML	AIX360
1	Split data into training and test sets ⇢ A test size of 0.2 delivers best results	
2	Build and implement Classification Tree with the respective interpretML model ⇢ depth = 7 provides best results	Build and implement Classification Tree with the respective scikit model ⇢ depth = 7 provides best results
3	Define and implement Logistic Regression with the respective interpretML model	Define and implement Logistic Regression with the respective scikit model
4	For both models apply prediction function and check the accuracy	
5	Get a classification report	
6	Create ROC curve	
7	Import LimeTabular from interpret.blackbox	Import LimeTabularExplainer from aix360.algorithms.lime
8	Interpret local explanations	
9	Compare findings	

4. MS Interpret ML – Results

Classification Tree

Accuracy Score

```
Training accuracy: 0.8273259265541515
Test accuracy: 0.8151289009497965
```

Confusion Matrix

	Actual=True	Actual=False
Predicted = True:	434	227
Predicted = False:	863	4372

Classification Report

	precision	recall	f1-score	support
False	0.84	0.95	0.89	4599
True	0.66	0.33	0.44	1297
accuracy			0.82	5896
macro avg	0.75	0.64	0.67	5896
weighted avg	0.80	0.82	0.79	5896

Logistic Regression

Accuracy Score

```
Training accuracy: 0.8079043338139259
Test accuracy: 0.814280868385346
```

Confusion Matrix

	Actual=True	Actual=False
Predicted = True:	397	195
Predicted = False:	900	4404

Classification Report

	precision	recall	f1-score	support
False	0.83	0.96	0.89	4599
True	0.67	0.31	0.42	1297
accuracy			0.81	5896
macro avg	0.75	0.63	0.65	5896
weighted avg	0.80	0.81	0.79	5896

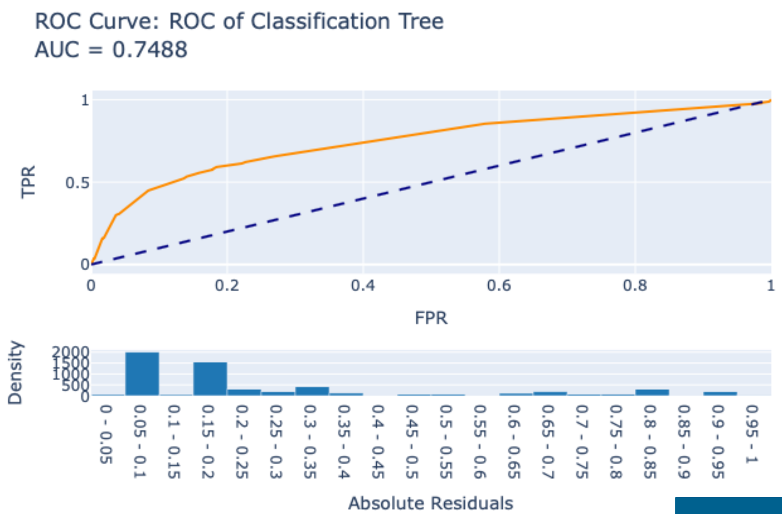
4. MS Interpret ML – Results

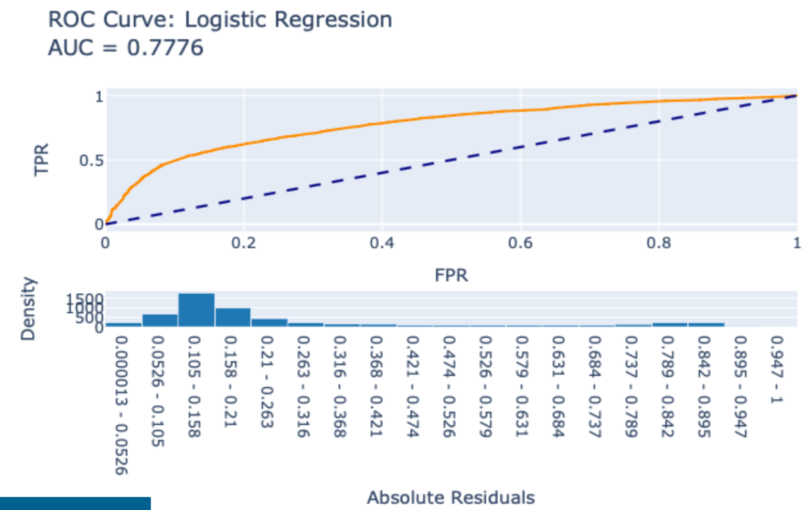
GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Classification Tree

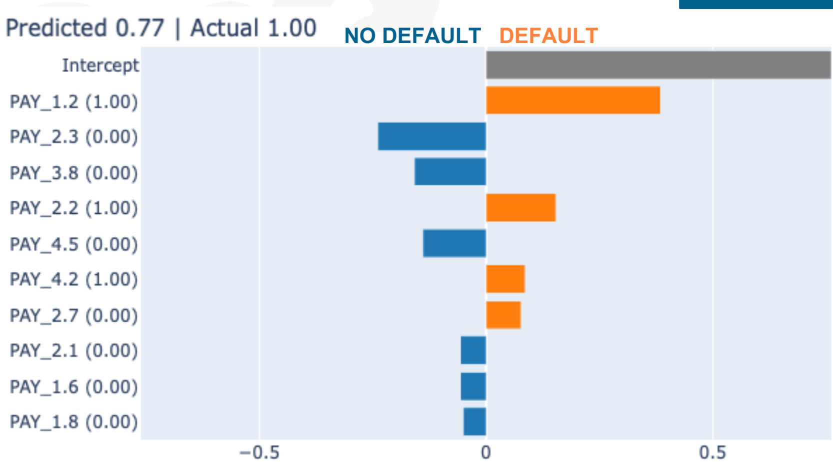
ROC Curve

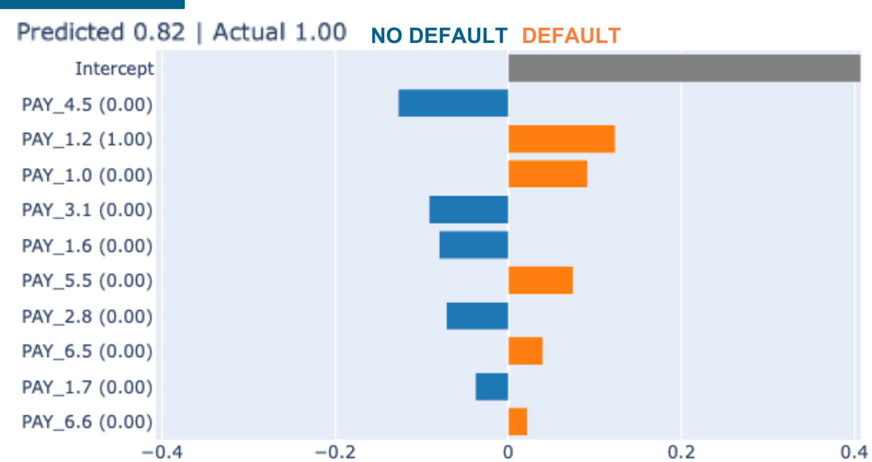
Logistic Regression





LIME Tabular Explainer





5. AIX360 – Results

Classification Tree

Logistic Regression

Accuracy Score

Classification Tree:

```
Training accuracy: 0.8273259265541515
Test accuracy: 0.8144504748982361
```

Logistic Regression:

```
Training accuracy: 0.8115935883300822
Test accuracy: 0.8175033921302578
```

Confusion Matrix

Classification Tree:

	Actual=True	Actual=False
Predicted = True:	432	229
Predicted = False:	865	4370

Logistic Regression:

	Actual=True	Actual=False
Predicted = True:	468	247
Predicted = False:	829	4352

Classification Report

Classification Tree:

	precision	recall	f1-score	support
False	0.83	0.95	0.89	4599
True	0.65	0.33	0.44	1297
accuracy			0.81	5896
macro avg	0.74	0.64	0.67	5896
weighted avg	0.79	0.81	0.79	5896

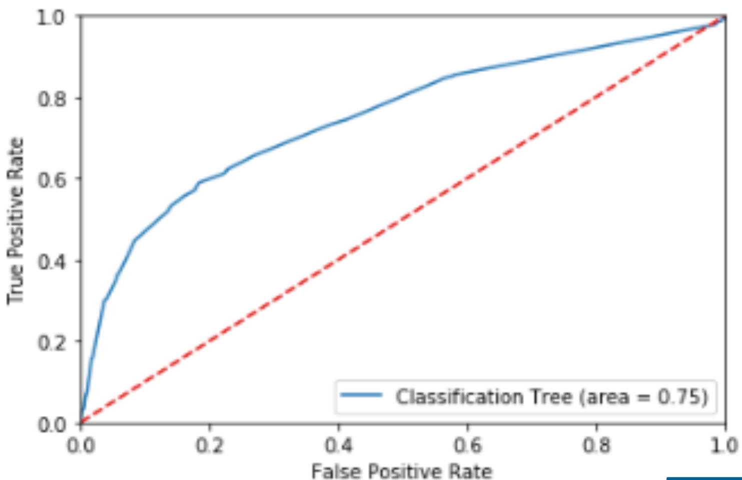
Logistic Regression:

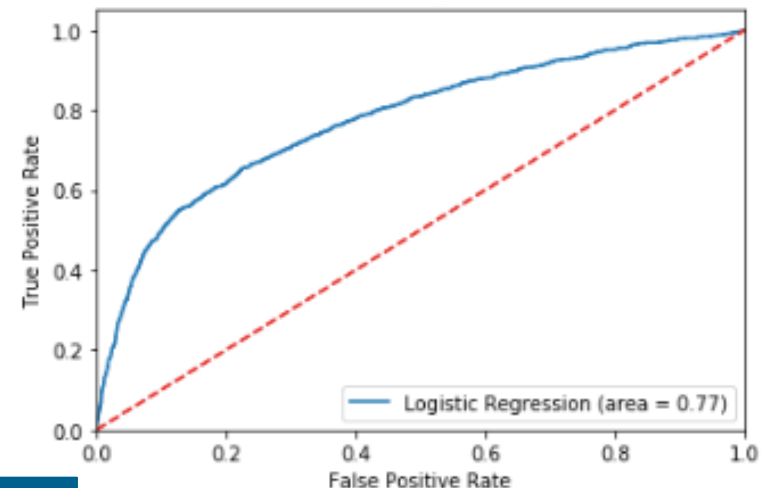
	precision	recall	f1-score	support
False	0.84	0.95	0.89	4599
True	0.65	0.36	0.47	1297
accuracy			0.82	5896
macro avg	0.75	0.65	0.68	5896
weighted avg	0.80	0.82	0.80	5896

Classification Tree

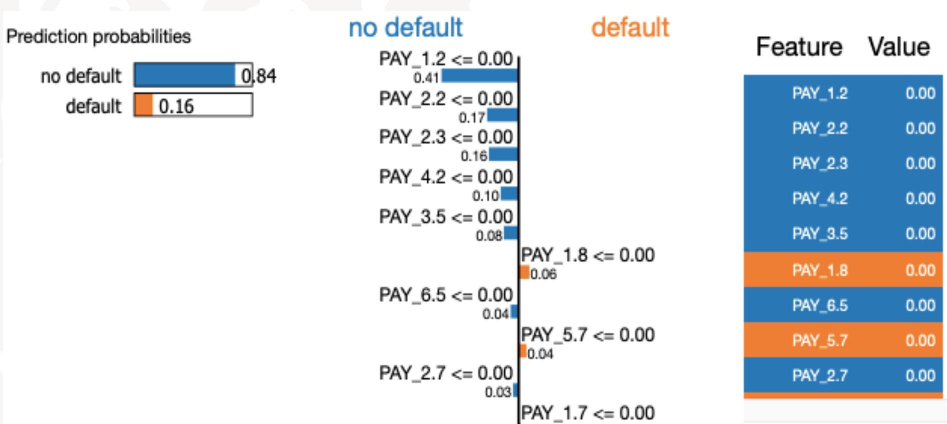
ROC Curve

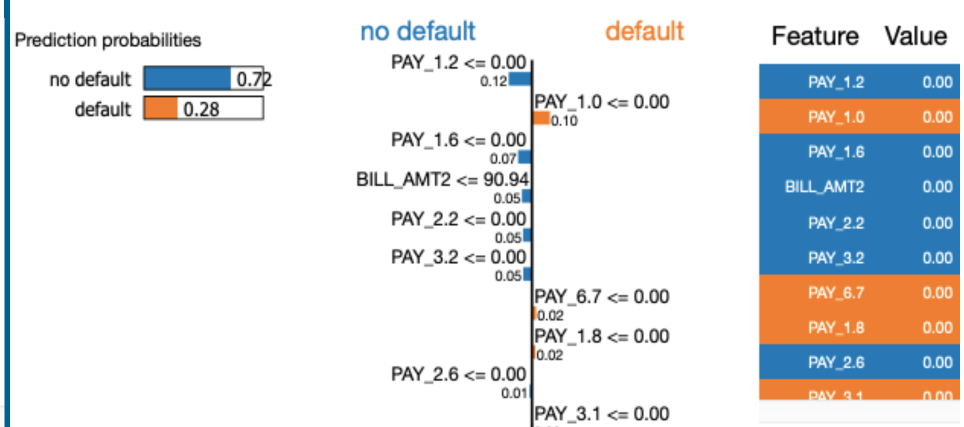
Logistic Regression





LIME Tabular Explainer





GOETHE UNIVERSITÄT FRANKFURT AM MAIN

	MS InterpretML	AIX360
Model	➢ All models needed are implemented in InterpretML	➢ Additional models (e.g. from scikit) are needed, before applying the LIME Explainer from AIX360
Visuali-zation	➢ Several visualization tools ➢ Easy to handle ➢ Clear and well readable representation	➢ Only visualization for Lime ➢ Use of other libraries to figure the data (e.g. matplotlib)
Documen-tation	➢ Good and detailed documentation ➢ Provides lots of example notebooks ➢ Great for beginners	➢ No detailed documentation ➢ Harder to find relevant information ➢ Great for intermediates to play around

Results

➢ Results in this use case are very similar, which might be due to the quite low complexity. Thus, a clear favorite cannot be stated in terms of comparison of the findings.

Difficulties

- ➢ Quite imbalanced dataset
  - ➢ Solving the issue by downsampling (decreasing dataset to 13‘196)
  - ➢ TPR increased with that change of the dataset
  - ➢ But high decrease of TNR and also of the AUC
- ➢ Some values in PAY_X have only a few counts, therefore predictions based on these values can be misleading

Further Research

- ➢ Need of a larger dataset
- ➢ More information about clients
- ➢ Larger time frame of observation