# KNIME Machine Learning Challenge 2025

# Università degli Studi di Milano-Bicocca

# Master of Science in Data Science

Eduardo Mosca 925279
Elisa Princic 886476
Sasha Risoluti 870667

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

TEAM 13

# Access and partition dataset

TEAM 13

**Exploratory Data Analysis**

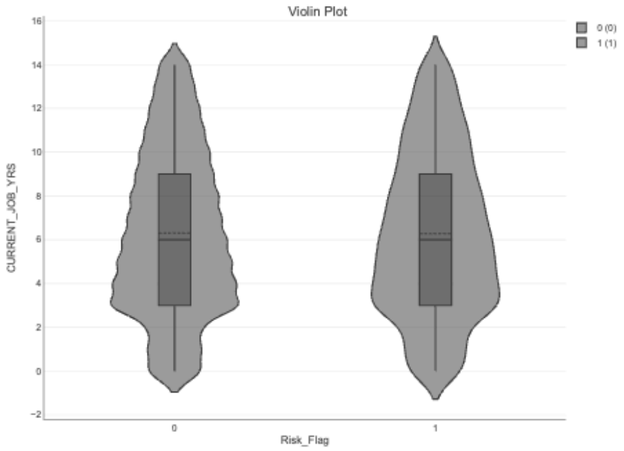
**Duplicate Row Filtering**

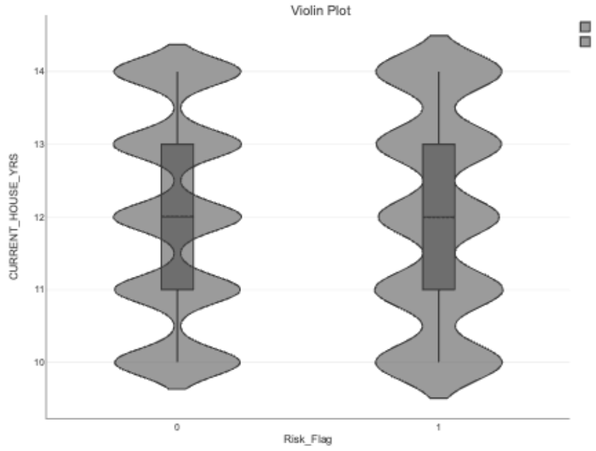
**Partitioning Train/Test**

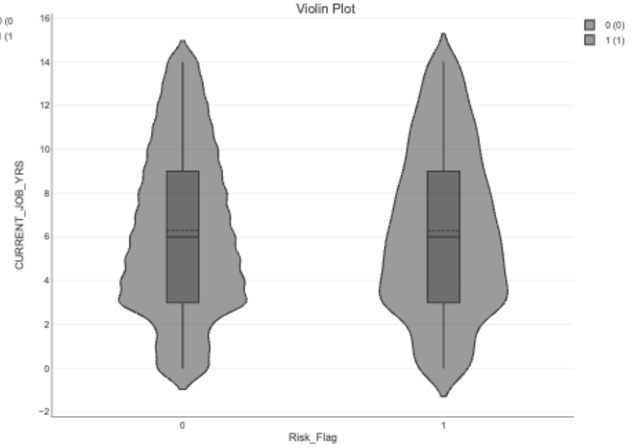
# Exploratory Data Analysis (part 1)

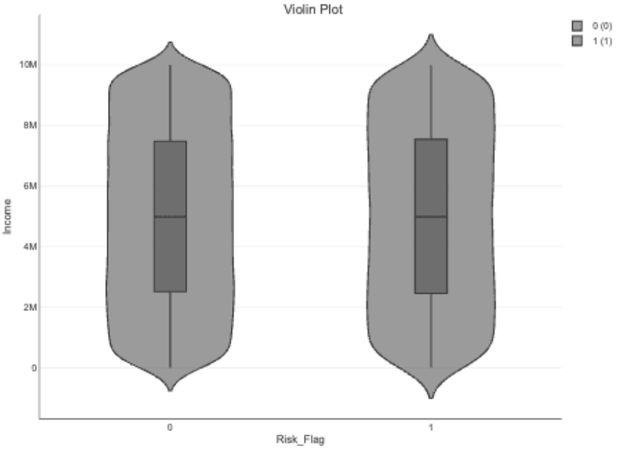
TEAM 13

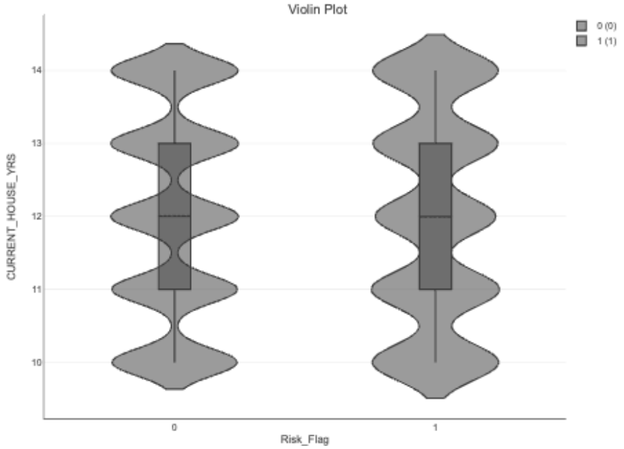
Similar distribution of target classes among features

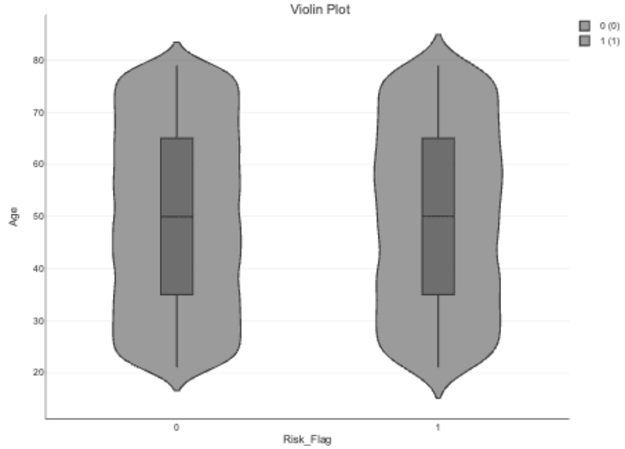


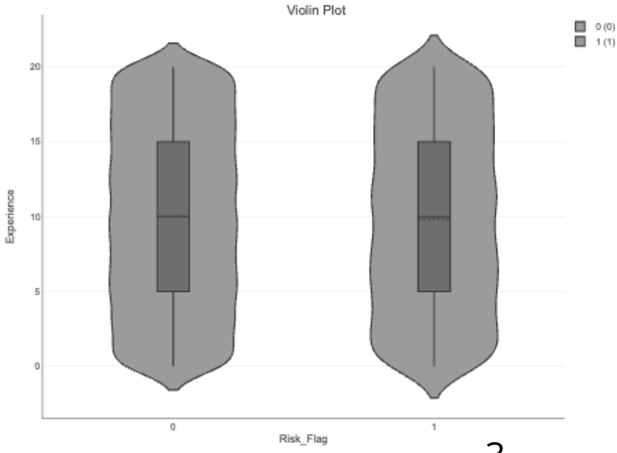



# Exploratory Data Analysis (part 2)

TEAM 13

- Unbalanced dataset 80-20
- Data seemed uniformally distributed in each feature
- Similar percentage of risky and non-risky in each categorical variable value

**Cross Tabulation of Age by Risk_Flag**

| Frequency<br>Row Percent | 0 | 1 | Total |
|---|---|---|---|
| 21 | 633 | 134 | 767 |
| | 82,5293% | 17,4707% | |
| 22 | 581 | 151 | 732 |
| | 79,3716% | 20,6284% | |
| 23 | 588 | 128 | 716 |
| | 82,1229% | 17,8771% | |
| 24 | 609 | 155 | 764 |
| | 79,712% | 20,288% | |
| 25 | 583 | 150 | 733 |
| | 79,5362% | 20,4638% | |
| 26 | 591 | 146 | 737 |
| | 80,19% | 19,81% | |

**Cross Tabulation of Profession by Risk_Flag**

| Frequency<br>Row Percent | 0 | 1 | Total |
|---|---|---|---|
| Air_traffic_controller | 725 | 176 | 901 |
| | 80,4661% | 19,5339% | |
| Analyst | 687 | 156 | 843 |
| | 81,4947% | 18,5053% | |
| Architect | 644 | 171 | 815 |
| | 79,0184% | 20,9816% | |
| Army_officer | 629 | 172 | 801 |
| | 78,5268% | 21,4732% | |
| Artist | 667 | 181 | 848 |
| | 78,6557% | 21,3443% | |
| Aviator | 681 | 163 | 844 |
| | 80,6872% | 19,3128% | |

# Access and partition dataset

TEAM 13

Exploratory Data Analysis

Duplicate Row Filtering

Partitioning Train/Test

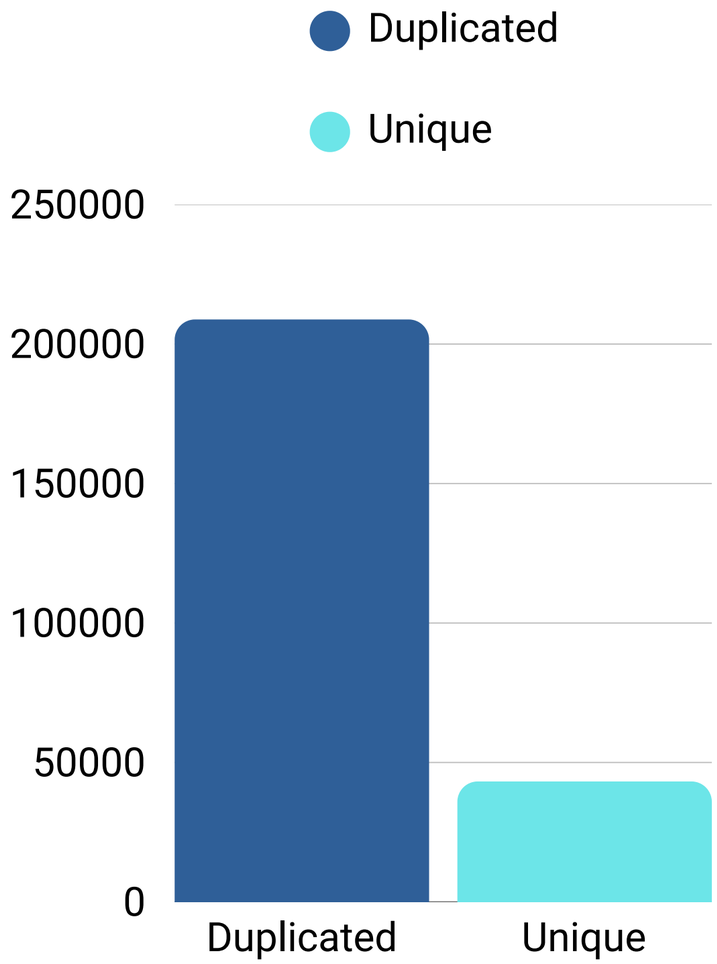
# Deduplication

TEAM 13

Affects independence between train and test set

Not strictly duplicates as ID was unique for each row

From 252,000 to 43,190 observations




Plain model with 0.9 AUC seemed to overfit

# Access and partition dataset

TEAM 13

Exploratory Data Analysis

Duplicate Row Filtering

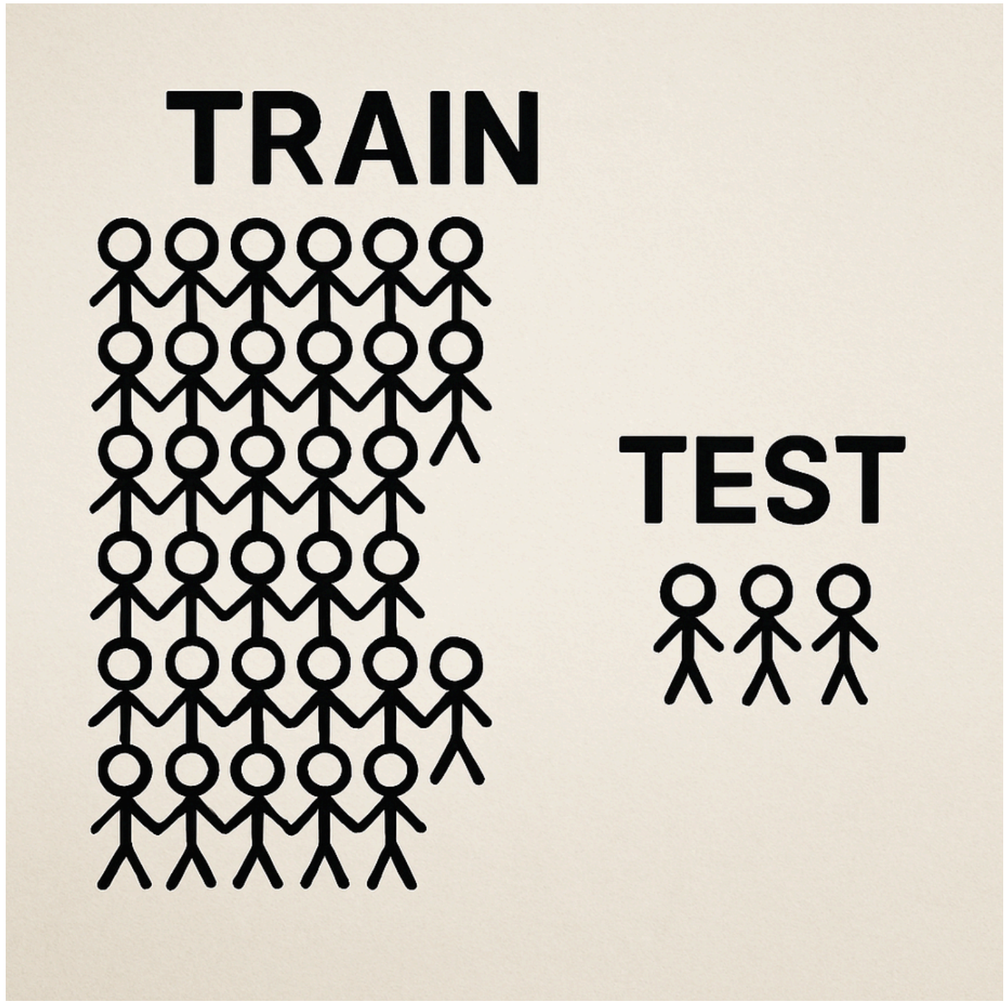
Partitioning Train/Test

# Partitioning Train/Test

TEAM 13

Train 80%

Test 20%



# Loan Approval Model - First Approach

TEAM 13

**Dealing with city and state variables**

**Filtering irrelevant and redundant variables**

**Model-specific feature selection to maximize AUC**

# Dealing with city and state variables (part 1)

TEAM 13

- Real world domain knowledge on grouping of Indian cities
- OneHotEncoding, new columns and avoiding perfect multicollinearity

| CITY |
| --- |
| Rewa |
| Parbhani |
| ... |

“Tier1"
“Tier2"
“Tier3"
“Tier4"
“Other”

| Tier1 | Tier2 | Tier3 | Tier4 |
| --- | --- | --- | --- |
| 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 0 |
| ... | ... | ... | ... |

# Dealing with city and state variables (part 2)

TEAM 13

- Delegating domain knowledge to group states based on econ/fin/dev factors

Here's a classification of Indian states into **Advanced, Medium, and Less Advanced** categories based on socio-economic indicators such as per capita income, industrialization, human development, and infrastructure. While this categorization is not official, it is widely used in economic and developmental discussions.

**States Categorized by Advancement Levels - JSON Format**

json Copia Modifica

```
{
  "AdvancedStates": [
    "Maharashtra",
    "Gujarat",
    "Karnataka",
    "Tamil Nadu",
```

| Advanced State | Medium State |
|---|---|
| 1 | 0 |
| 0 | 0 |
| ... | ... |

# Loan Approval Model - First Approach

TEAM 13

Dealing with city and state variables

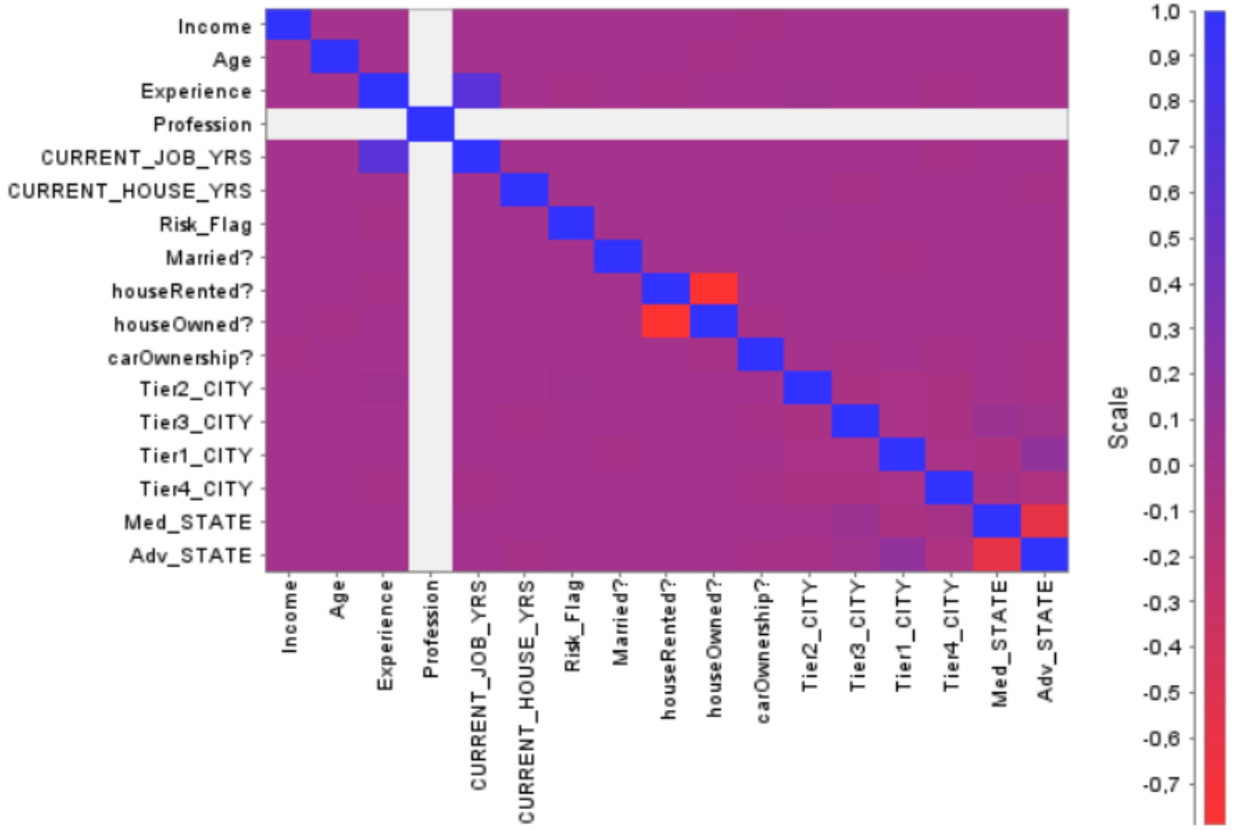
Filtering irrelevant and redundant variables

Model-specific feature selection to maximize AUC

# Filtering irrelevant and redundant variables

TEAM 13

- Correlational analysis of explanatory variables
- Removing some columns and association with the response



# Loan Approval Model - First Approach

TEAM 13

Dealing with city and state variables

Filtering irrelevant and redundant variables

Model-specific feature selection to maximize AUC

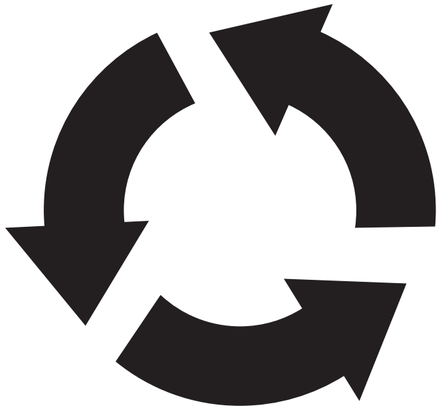
# Model-specific feature selection for max AUC

TEAM 13

- Setting up feature selection loops with different strategies
- Standardization and equal size sampling for model training
- Train/validation split and results

Standardize quantitative variables...
Equal size sampling...

Iterate through subsets...



Backwards, forwards, random...

# Loan Approval Model - Second Approach

TEAM 13

K-means on City, State and Profession

Feature Selection

XGBoost Tree Ensemble

# K-Means on City, State and Profession

TEAM 13

- Aggregated variables for each categorical value
- Evaluated different clustering algorithms to find underlying groups in data
- Picked K-Means as better algrorithm by looking at Silhouette coefficient

| City | Cluster |
|---|---|
| Adoni | 0 |
| Agartala | 1 |
| Agra | 2 |
| Ahmedabad | 3 |
| Ahmednagar | 4 |
| ... | ... |

| State | Cluster |
|---|---|
| Assam | 1 |
| Bihar | 0 |
| ... | ... |

| Profession | Cluster |
|---|---|
| Air_traffic_controller | 0 |
| Analyst | 1 |
| Architect | 2 |
| Army_officer | 3 |
| Artist | 4 |
| ... | ... |

# Loan Approval Model - Second Approach

TEAM 13

K-means on City, State and Profession

Feature Selection

XGBoost Tree Ensemble

# Feature Selection

TEAM 13

Wanted to automatically select the highest performing feature subset for an XGBoost model, by evaluating on a test set.

Highest AUC Subset:

- Prof_cluster,
- House_ownership,
- Experience,
- Age,
- Income.

# Loan Approval Model - Second Approach

TEAM 13

K-means on City, State and Profession

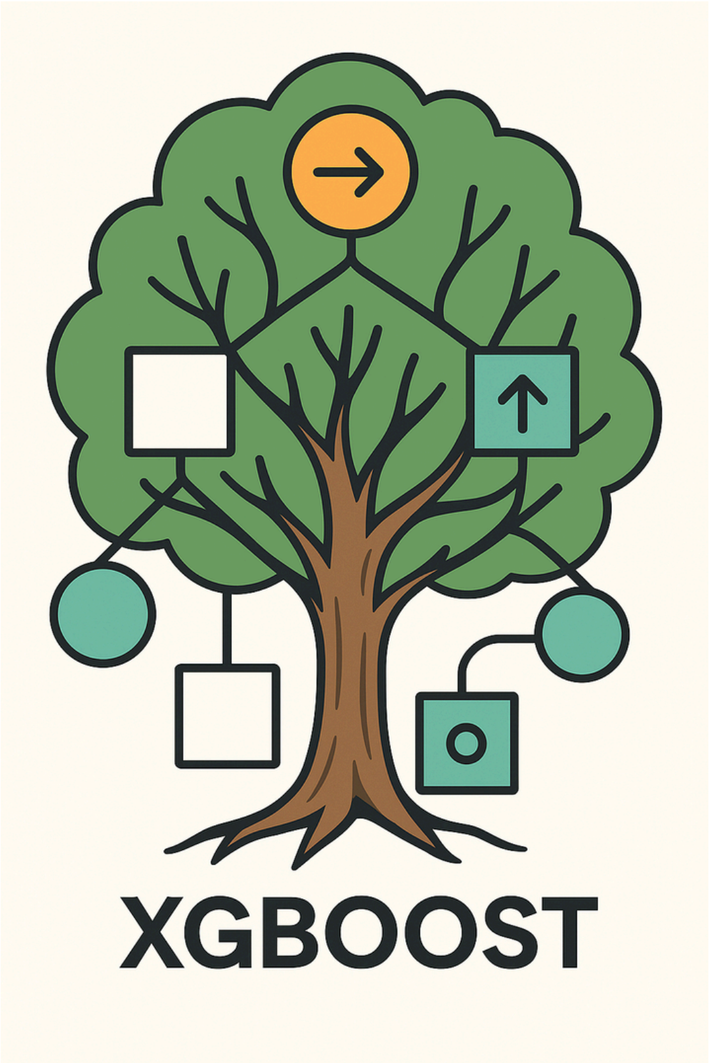
Feature Selection

XGBoost Tree Ensemble

# XGBoost Tree Ensemble

TEAM 13

- Evaluated XGBoost with selected features on a test set
- Final wrap-up of initial approaches results



# Improvements - Trying out more models

TEAM 13

**The new models**

**New models on first approach data**

**Using feature selection loops with new models**

# The new models

TEAM 13

- Wanted to cover a wider range of model complexities

Logistic Regression Learner

XGBoost Tree Ensemble Learner


Keras Network Learner

SVM Learner

H2O AutoML Learner

# Improvements - Trying out more models

TEAM 13

The new models

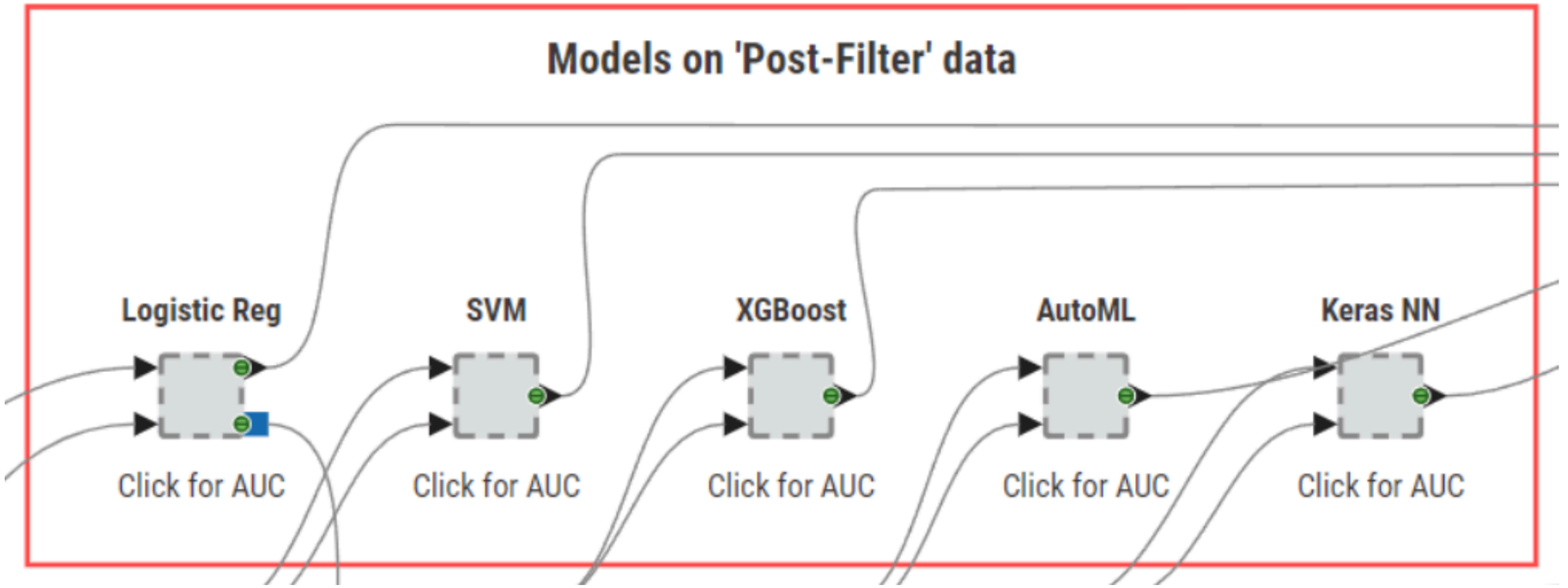
New models on first approach data

Using feature selection loops with new models

# New models on first approach data

TEAM 13

- Tried models on domain knowledge encoded data, after correlational analysis
- AutoML learner node and high performance through overfitting



# Improvements - Trying out more models

TEAM 13

The new models

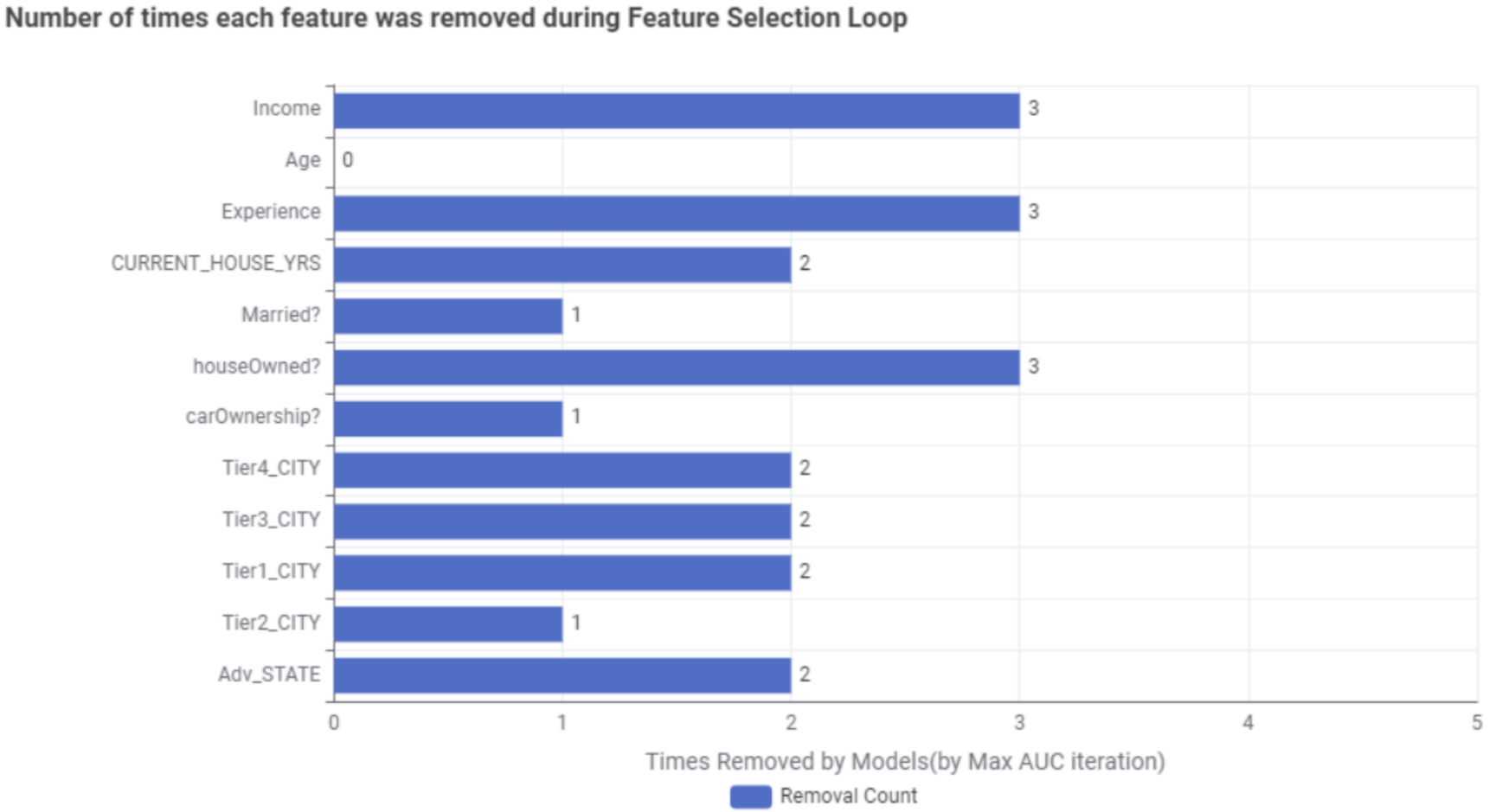
New models on first approach data

Using feature selection loops with new models

# Using feature selection loops with new models

TEAM 13

- Looked for best model specific subset in terms of AUC
- Summarized results using visualizations



# Improvements - AD approach and Autoencoder

TEAM 13

Changing the problem perspective

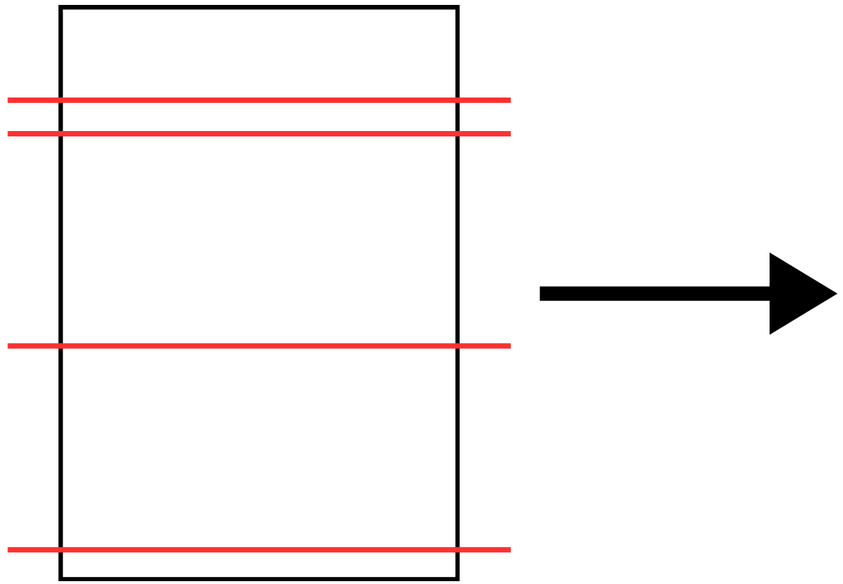
Data preparation and architecture

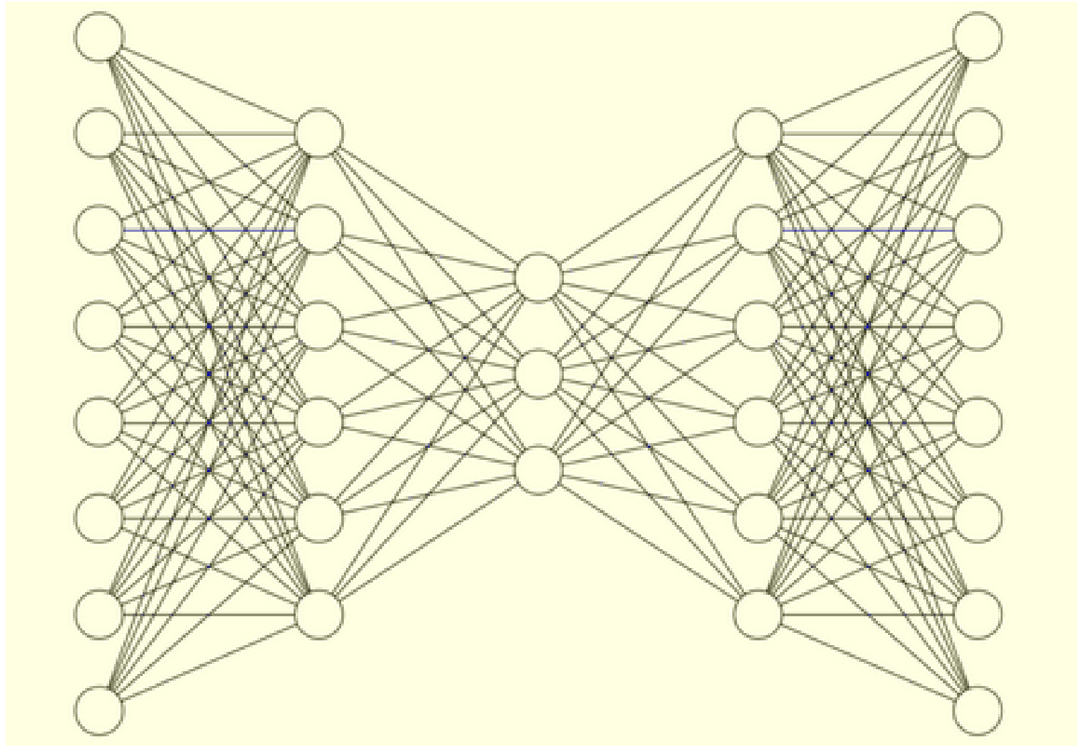
Reconstruction error and classifying anomalies

# Changing the problem perspective

TEAM 13

- Risky applicants as unusual(anomalous) entries
- How artificial neural networks can help





# Improvements - AD approach and Autoencoder

TEAM 13

Changing the problem perspective

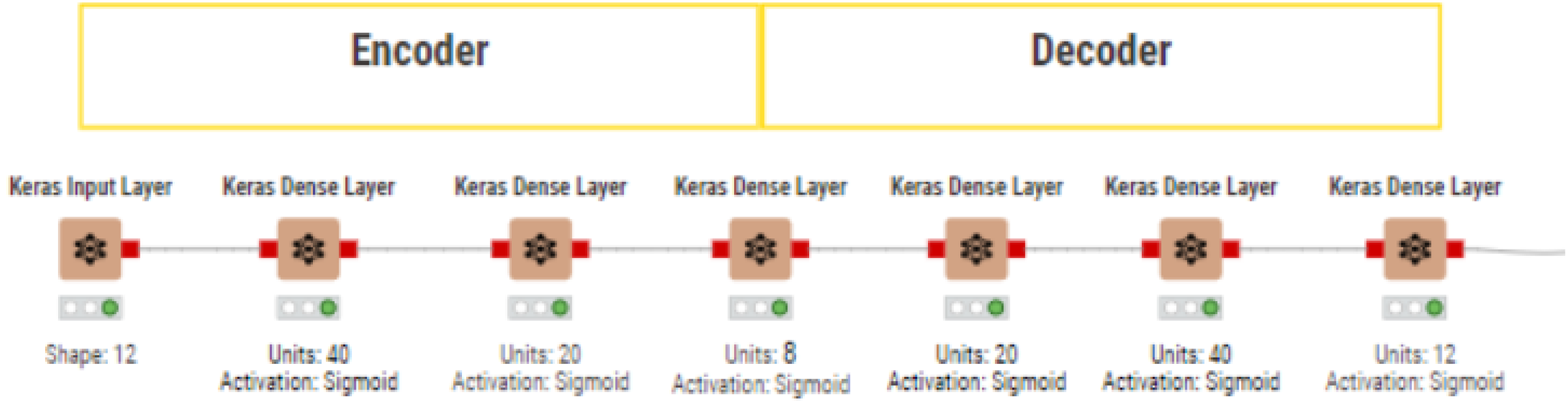
Data preparation and architecture

Reconstruction error and classifying anomalies

# Data preparation for the Autoencoder

TEAM 13

- Min-Max normalization and fully normal training data
- Neural net architecture



# Improvements - AD approach and Autoencoder

TEAM 13

Changing the problem perspective

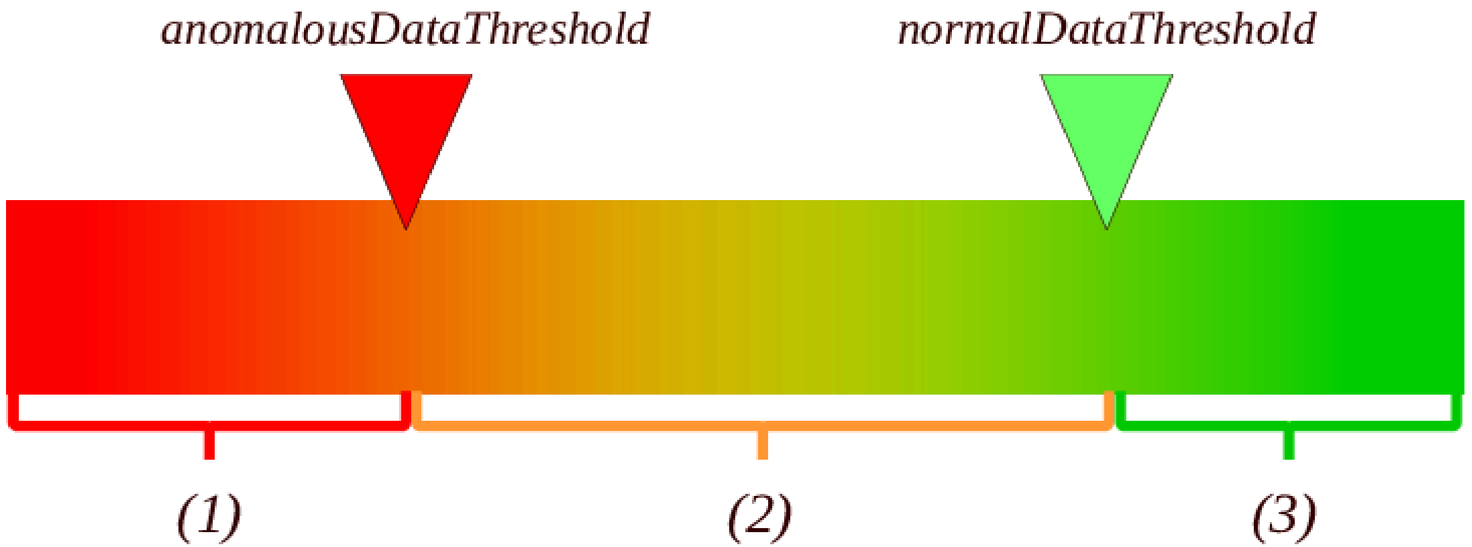
Data preparation and architecture

Reconstruction error and classifying anomalies

# Reconstruction error and classifying anomalies

TEAM 13

- How well can we reconstruct input data, having learned the normal representation?
- Computing the reconstruction error and setting a threshold for anomalies

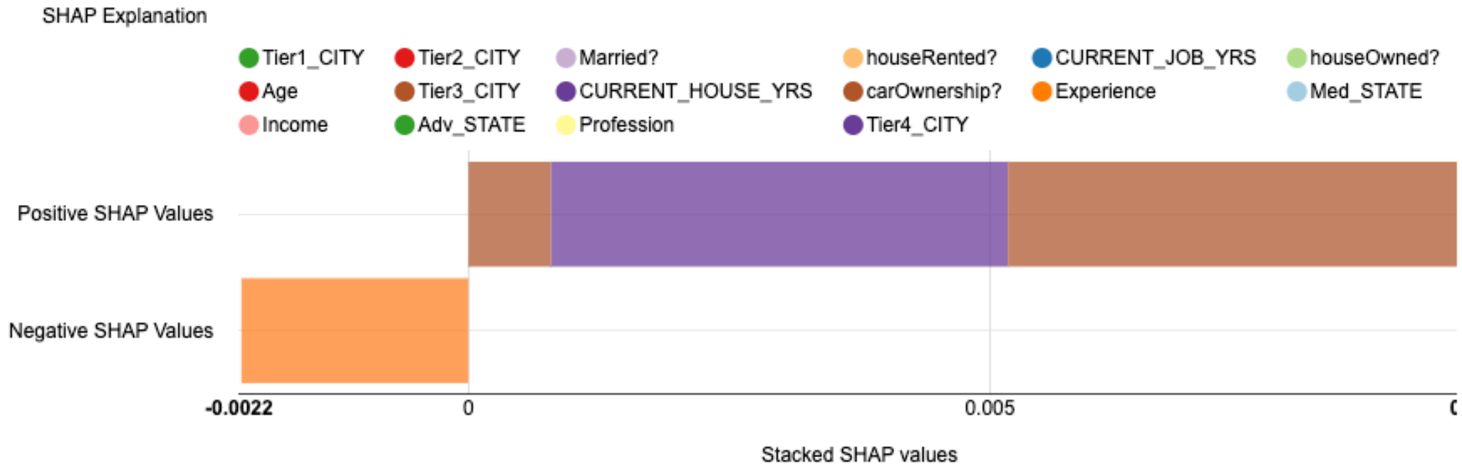



IMG reference: Explanation Method for Anomaly Detection on Mixed Numerical and Categorical Spaces - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/The-anomalous-data-range-1-the-transition-data-range-2-and-the-normal-data-range_fig2_363478906 [accessed 6 May 2025]

# Other Improvements - SHAP Values

TEAM 13

- We wanted to make the decision more transparent to the hypothetical loan applicant
- Introduction of SHAP values in the Data App




The forecast for this specific instance is 0.5. This value represents a deviation of 0.007484325650985613 from the average model forecast.

This deviation is explained by the SHAP values:

- Features with positive SHAP contributed to increase the forecast compared to the average.
- Features with negative SHAP have helped to decrease it.

The algebraic sum of all these contributions SHAP (0.01) is precisely equal to the total deviation observed (0.007484325650985613)


**TEAM 13**

# Thanks for your time!

## KNIME Machine Learning Challenge 2025

## Università degli Studi di Milano-Bicocca

## Master of Science in Data Science

**Eduardo Mosca** 925279 - **Elisa Princic** 886476 - **Sasha Risoluti** 870667