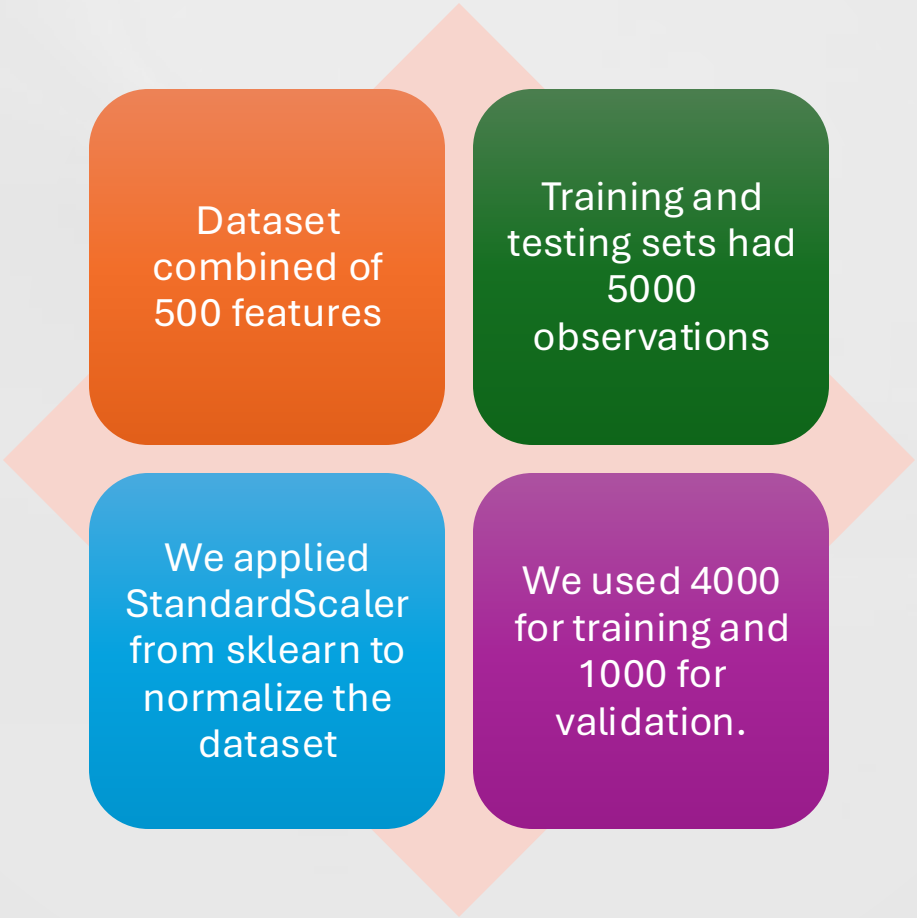
AML - Project 2

Jakub Seliga, Krzysztof Wolny

Presentation plan

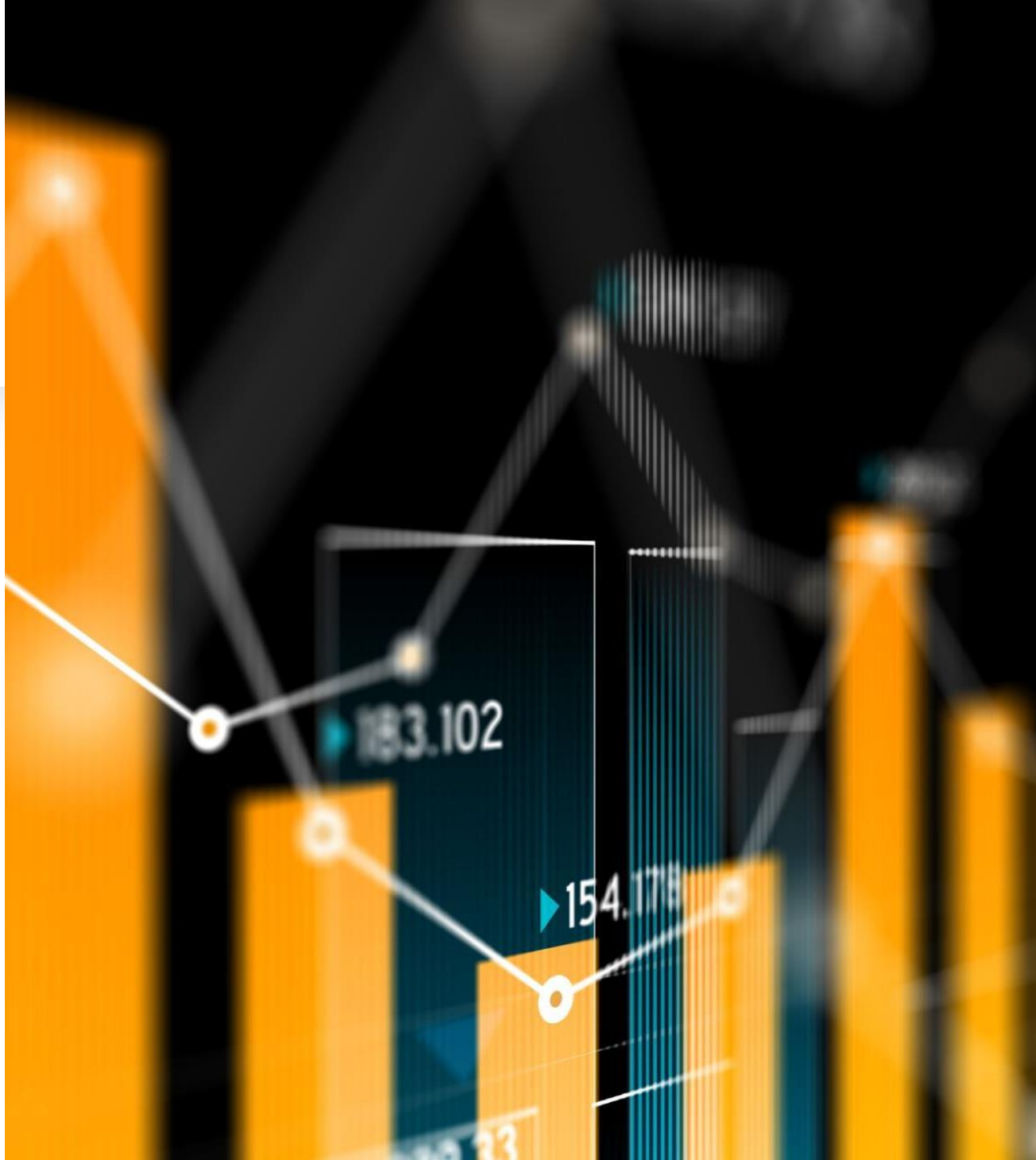
- Dataset
- Scoring
- Feature selection methods
- Classification models
- Results

Dataset


Dataset combined of 500 features
Training and testing sets had 5000 observations
We applied StandardScaler from sklearn to normalize the dataset
We used 4000 for training and 1000 for validation.

Scoring

- From the test set, we were required to select the top 20% (1000 rows) based on model performance.
- In validation, we mirrored this by selecting the top 20% (200 rows) to evaluate and compare models.
- Scoring was based on reward and cost, with a maximum possible validation score of 9800 after scaling results by a factor of 5.



Feature selction methods

- Recursive Feature Elimination (RFE) with XGBoost Classifier,
- Shap with XGBoost Classifier,
- Mean Decrease in impurity with XGBoost Classifier,
- Univariate feature selection with F-test for feature scoring

Classification models

- LOGISTIC REGRESSION,
- RANDOM FOREST CLASSIFIER,
- SUPPORT VECTOR MACHINE (SVM),
- XGBOOST CLASSIFIER.

Results

feat sel	name	sel index	nr of feat	acc	recall	precision	reward	var cost	final score
shap	Ensemble soft svm+lr	2	1	0.709	0.719836	0.695652	7600.0	200	7400.0
rfc	Ensemble soft lr+svm	2	1	0.709	0.719836	0.695652	7600.0	200	7400.0
shap	LogisticRegression	2	1	0.707	0.695297	0.702479	7600.0	200	7400.0
rfc	LogisticRegression	2	1	0.707	0.695297	0.702479	7600.0	200	7400.0
ufs	LogisticRegression	2	1	0.707	0.695297	0.702479	7600.0	200	7400.0
rfc	Ensemble soft rf+xgb+lr	2, 6	2	0.698	0.707566	0.685149	7750.0	400	7350.0

Thank you!

Jakub Seliga, Krzysztof Wolny