UNIVERSITY OF ZAGREB

Faculty of Electrical Engineering and Computing

Laboratory for Financial and Risk Analytics

STAREBEI funded research project

Prediction of Cashflow Timing and Patterns in International Bank Accounts

BRUNO GASPEROV | 09 DECEMBER 2021

PRESENTATION BREAKDOWN

I. Introduction - the research project
II. Data and exploratory analysis (EA)
III. Modeling
IV. Results and discussion
V. Conclusion

I. Introduction - the research project

I. I. The underlying problem

- EIB manages multiple bank accounts in different countries and currencies, for dealing with counterparties scattered throughout the globe
- Each account typically receives and sends out hundreds of transactions on a daily basis (cash inflows and outflows)
- For liquidity management purposes, i.e., to be able to meet its obligations on time, the EIB would like to able to **predict intraday cash inflow timings** and patterns with satisfactory accuracy

"Liquidity risk is *the risk to an institution's financial condition or safety and soundness arising from its inability (whether real or perceived) to meet its contractual obligations*" (federalreserve.gov)

- The focus is on cash inflows since outflows are under direct control of the EIB
- Main idea: utilize historical data to extract patterns (regularities) in cashflows and leverage them to make (as accurate as possible) predictions about the future (timing of the incoming cash inflows)
- It would be beneficial to the EIB to know not only the expected (predicted) timings but also the uncertainties of the predictions

I. I. The underlying problem (cont.)

Specific example:

- A certain EIB account has a balance of 10 mil. EUR on a certain day
- The EIB needs to make a payment of 20 mil. EUR (from this account) to counterparty CP1 by noon (Luxembourg time)
- The EIB is expecting to receive a payment (on the same account) of 15 mil. EUR from counterparty CP2 at *some time* that day
- If the payment by CP2 arrives prior to noon, no problem arises
- However, the EIB is not sure about the timing of the incoming payment (it could happen at some time in the afternoon as well)
- Should the EIB pre-fund the account with an additional 10 mil. EUR or wait for the CP2 payment and risk being late?
- If only we could know the predicted timing of the expected cashflow and the prediction uncertainty

I. II. Research goals and questions

Research goals

- Design and develop an AI-based solution for predicting intraday cash inflow timing and patterns in international bank accounts

Research questions

- Can historical cashflow data be leveraged (via machine learning techniques) to make accurate predictions on timings of incoming cashflows?
- What are the key features (variables) that conduce to (or detract from) predictability?

I. III. Team

Tutor EIB Group:
Aghzinnay Omar
Head of Liquidity Planning and Control Unit
Finance Directorate I Back Office Treasury

University tutor:
Assoc. Prof. Zvonko Kostanjcar, PhD
Head of the Laboratory for Financial and Risk Analytics, FER, UNIZG

STAREBEI junior researcher:
Bruno Gasperov, MSc
PhD student and research associate
Laboratory for Financial and Risk Analytics, FER, UNIZG

Researcher:
Stjepan Begusic, PhD
Post-doc researcher
Laboratory for Financial and Risk Analytics, FER, UNIZG

Laboratory for Financial and Risk Analytics (lafra.fer.hr)

Research topics:

- Risk modelling and portfolio optimization
- Reinforcement learning for market making
- Machine learning applications in finance

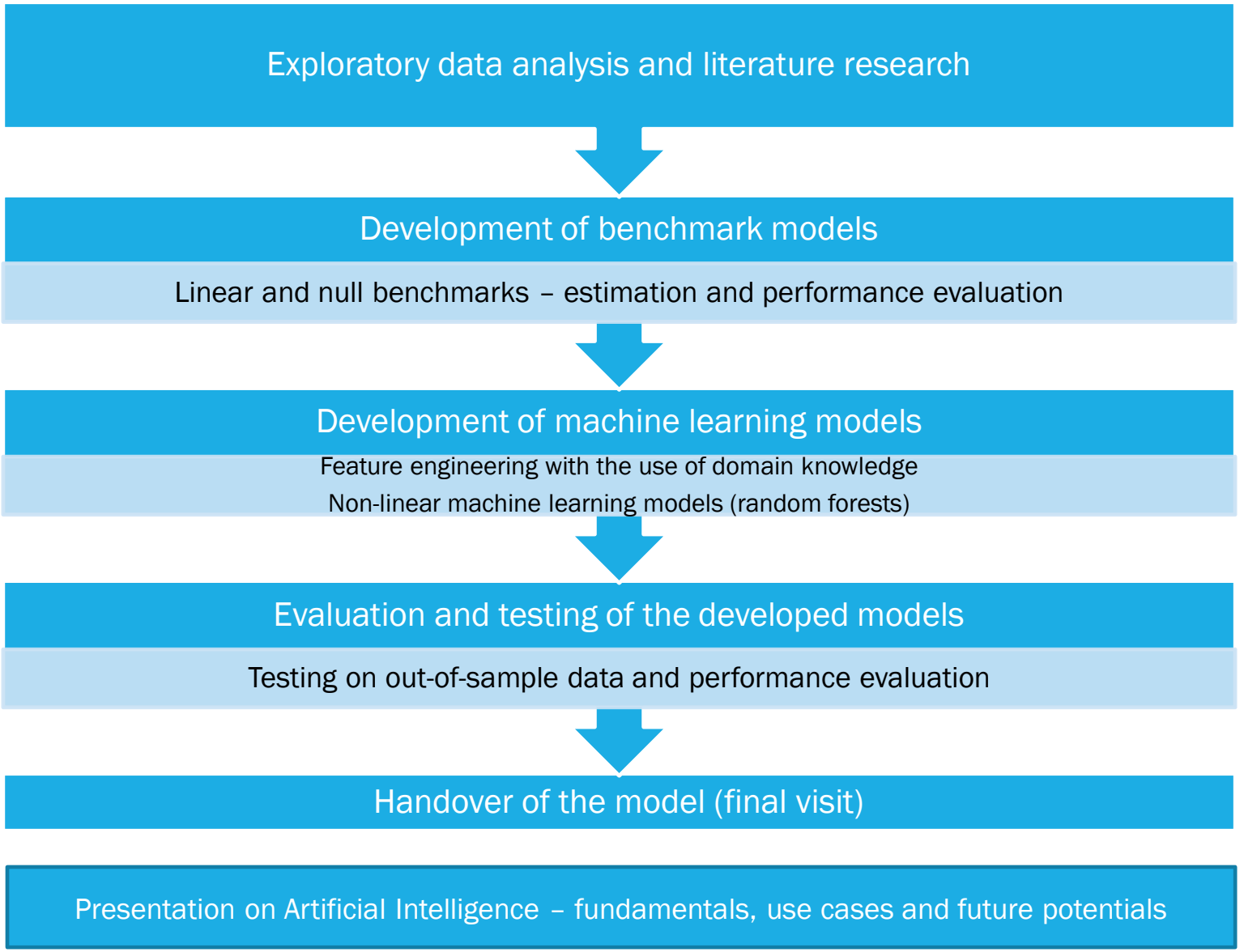
I. IV. Project workflow

Duration:

Nov 2020 – Oct 2021

Number of phases:

4



II. Data and exploratory analysis (EA)

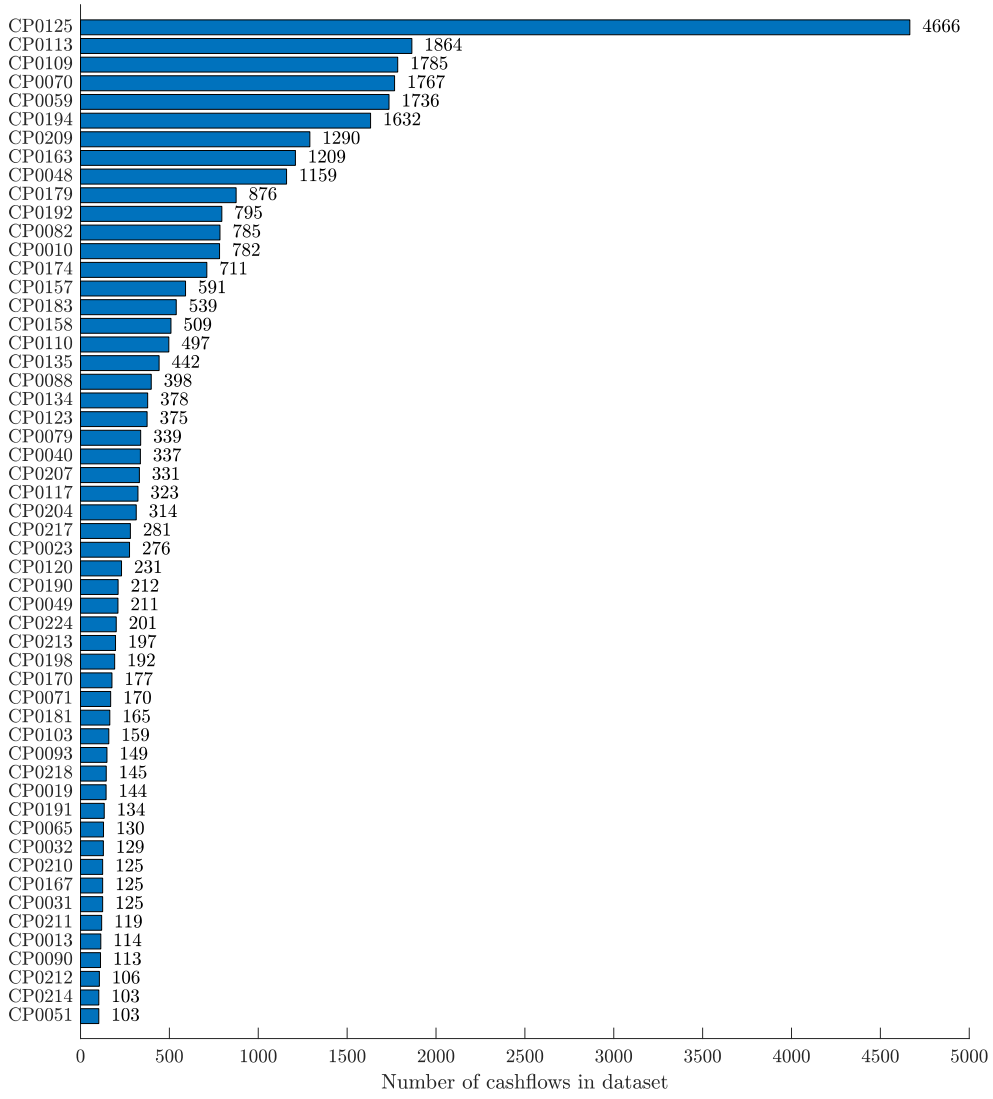
II. I. Datasets

The EIB's bank/account | The target | Anonymized CP | Other ratings available

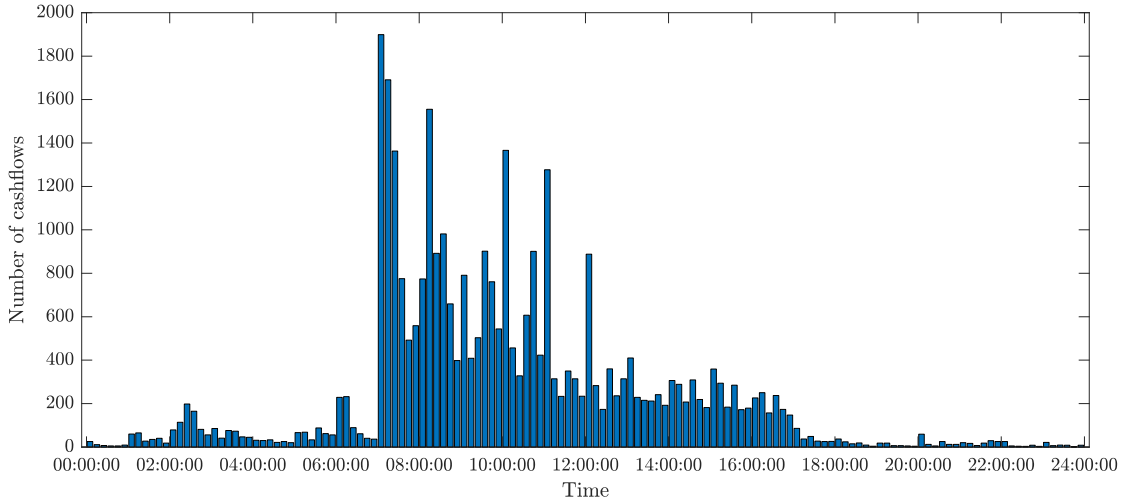
	Our Bank	Our Account	CF Payment Date	Time Stamp	CF Payment Amount EUR	CF Payment Currency	Counterparty	Client Country Code	Client Country Name	New EIB Internal Rating	Instrument Group	Portfolio	Account Opening Time	Account Closing Time	Account Cutoff Time
0	BKHANDLOWY WAW	PL66103015080000000300751067	2019-01-02	0 days 12:12:00	3.812833e+06	PLN	CP0049	PL	Poland	A1	MM-DEPO	RSI2-S	0 days 01:30:00	0 days 18:00:00	0.666667
1	BKNYMELLON NYC	890-0545-747	2019-01-02	0 days 15:03:00	8.232272e+03	USD	CP0174	NL	Netherlands	Aa3	FX	3PM-SSMED-FX	0 days 02:00:00	1 days 01:00:00	0.937500
2	BKNYMELLON NYC	890-0545-747	2019-01-02	0 days 15:03:00	5.337479e+04	USD	CP0174	NL	Netherlands	Aa3	FX	3PM-SSMED-FX	0 days 02:00:00	1 days 01:00:00	0.937500
3	BKNYMELLON NYC	890-0545-747	2019-01-02	0 days 15:19:00	1.516379e+05	USD	CP0070	GB	United Kingdom	A1	SWAP-IR	BLT-F	0 days 02:00:00	1 days 01:00:00	0.937500
4	BKNYMELLON NYC	890-0545-747	2019-01-02	0 days 03:09:00	1.318081e+06	USD	CP0123	GB	United Kingdom	A1	FX-SWAP	TA1FX-SWAP-OUTR	0 days 02:00:00	1 days 01:00:00	0.937500

- Historical data provided by the EIB
  - Cashflows datasets + datasets containing opening/closing/cut-off times for different accounts
- **46,780** cashflows, spanning the period from Jan 2019 to the first part of Oct 2021
  - Generally, a "large enough" number of observations needed for "data-hungry" ML methods
- Cashflow timing is the target variable (intraday time, the date is known in advance)
- Key variables are shown in the figure above (in total **32** variables)

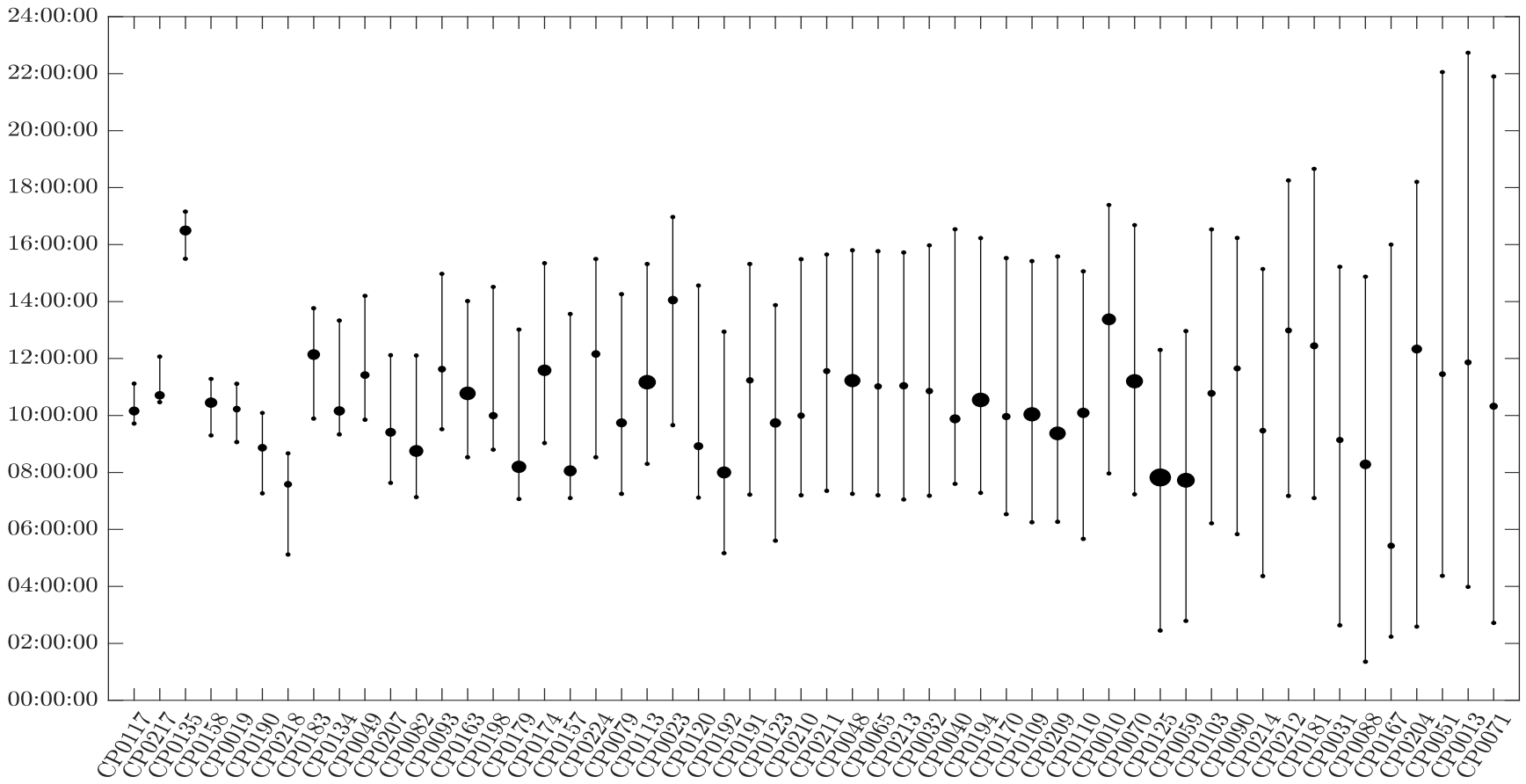
II. II. Some EA takeaways



- **Descriptive statistics and exploratory analysis**
  - Univariate analyses of all variables
  - Discovering specific patterns or groupings



II. II. Some EA takeaways (cont.)



- **Descriptive statistics and exploratory analysis**
  - Relationship between input variables (CP) and the target variable (timing)

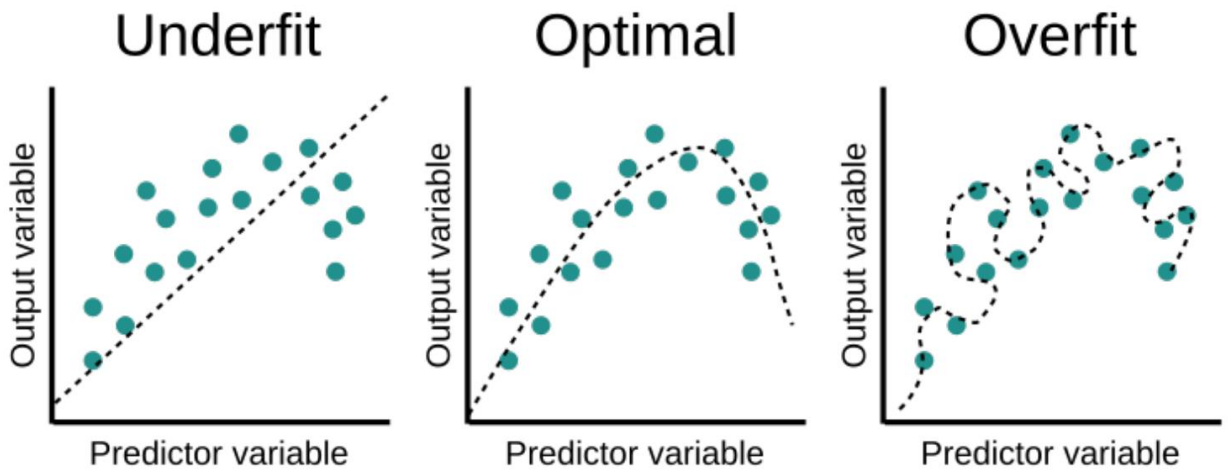
II. III. Data cleaning and preprocessing

- **Data cleaning:**
  - Cashflows with suspiciously low amounts (less than 0.01 EUR) were discarded as erroneous
  - Cashflow payment amounts were all converted to the same currency (EUR)
  - Account times (open, close, cut-off) were merged into the cashflow dataset
  - Some timestamps needed fixing
    - Cashflows that arrived before the opening time were assumed to arrive precisely at the opening time
- **Data preprocessing:**
  - Selecting informative variables
    - Collinearities – the information contained in certain variables is already contained in other variables (e.g. New EIB Internal Rating and Counterparty, Account and Bank)
  - One-hot encoding for categorical variables
  - Feature engineering (handcrafted features) with the use of domain knowledge

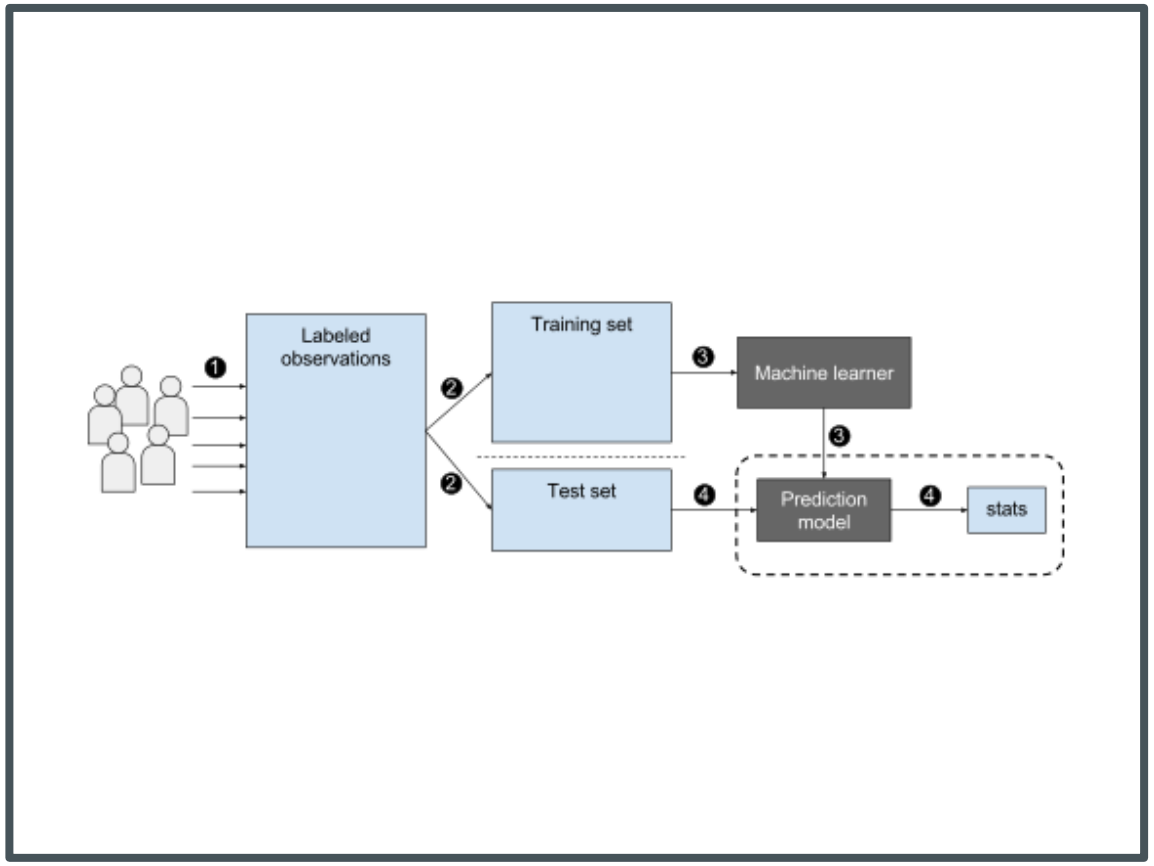
III. Modeling

III. I. OVERFITTING/ UNDERFITTING

- Among the central problems in machine learning
- Finding the right level of model complexity
  - Overly high complexity leads to overfitting (capturing the noise and not only genuine patterns)
  - Overly low complexity leads to underfitting (not capturing patterns properly)
  - Both overfitting and underfitting lead to poor generalization (performance on unseen examples)



III. II. SUPERVISED LEARNING



- The goal is to find the mapping from the input to the output
  - Given the counterparty, portfolio, cashflow amount in EUR and other variables, can we predict the timing?
- Data – annotated examples: INPUT ⟶ TARGET
  - $(INPUT_1, TARGET_1)$, $(INPUT_2, TARGET_2)$, …, $(INPUT_N, TARGET_N)$
- Output – prediction of target variable

III. III. BENCHMARK MODELS

- Simple models used to gauge the performance of our ML models
- Clearly underfitting
  - **Null model**
    - the output (cashflow timing prediction prediction) is given by the mean CF timing
  - **Generalized linear model (GLM)** where f is the logistic function

$$y = f(z) = f(a + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p)$$

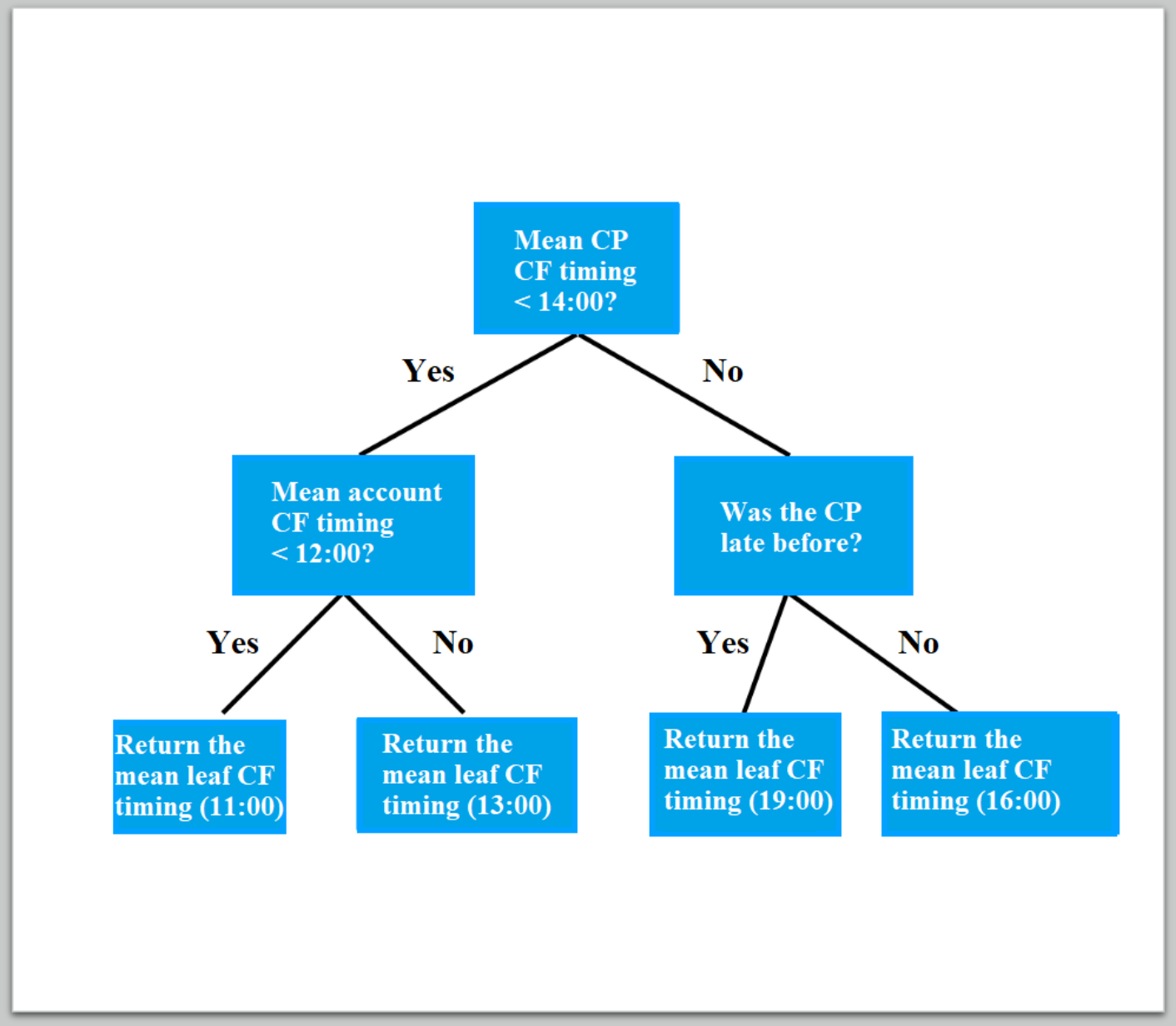
$$f(z) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-(a + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p)}}$$

    - x_i are the features (CP, portfolio, etc.), y the target variable (intraday cashflow timing)
  - **Weighted generalized linear model (WGLM)**
    - as above but more importance is given to cashflows with larger payment amounts

III. IV. ML MODELS

Decision trees

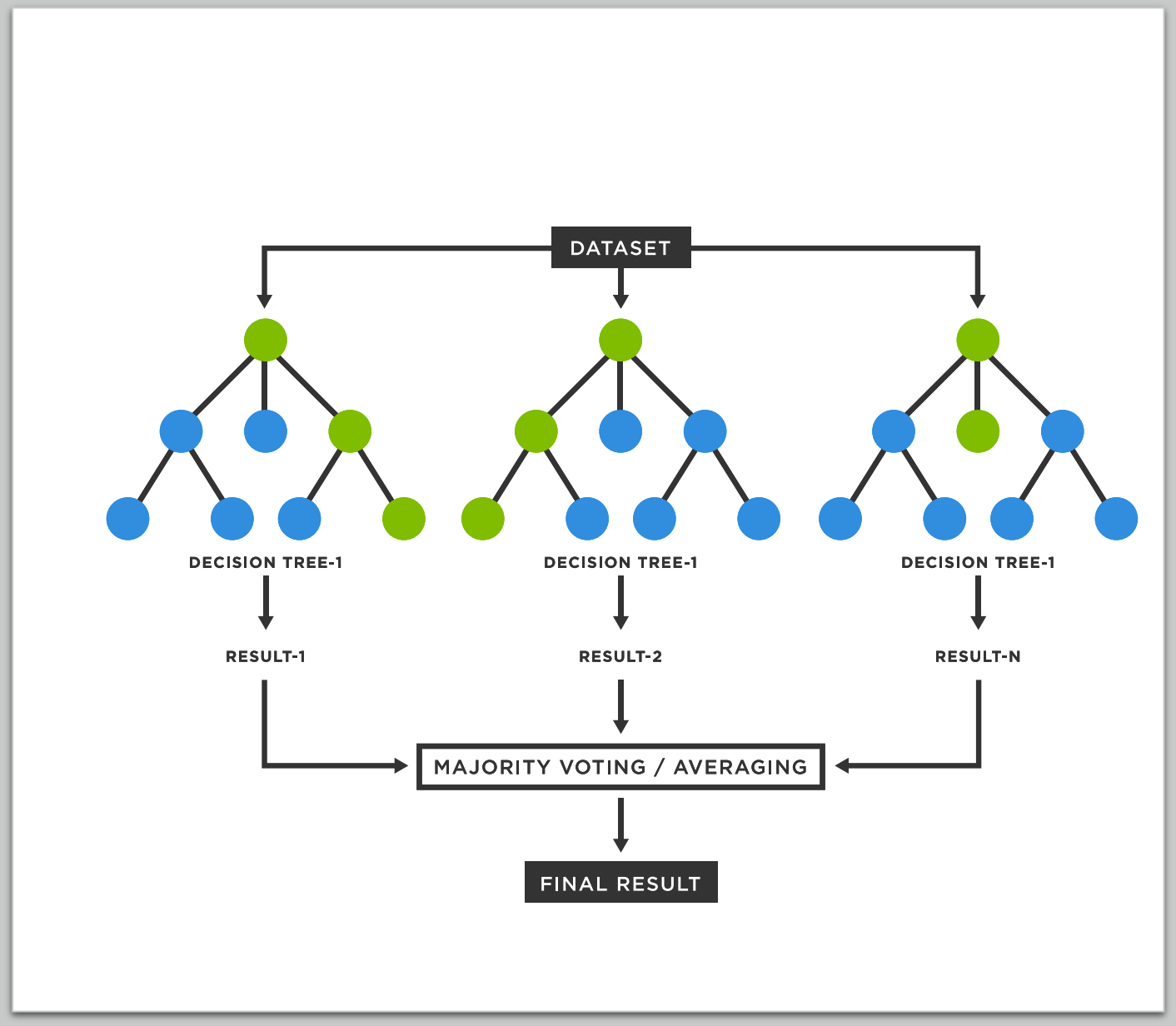
- Among the most commonly used ML models
- Advantages: simplicity, interpretability
- A tree-like graph with nodes representing certain conditions, edges representing truth/falsity of the conditions and leading to lower nodes, and the bottom nodes representing the outputs (predictions)
- Splits are selected automatically by the algorithm
- In reality, decision trees tend to overfit



III. IV. ML MODELS (CONT.)

Random forests

- Ensemble learning method that relies on multiple decision trees
- Main idea: noise cancels out with many uncorrelated trees → prevention of overfitting
- Advantages: robust to irrelevant features, invariant under scaling, versatile
- Widely used in practice for a plethora of different problems
- Key hyperparameters: number of trees, maximum depth
- Optimal model complexity

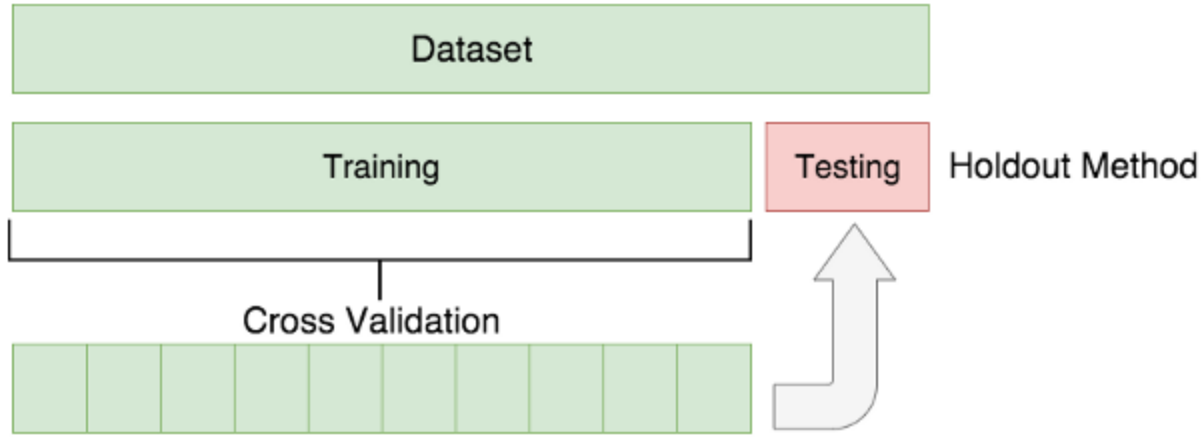


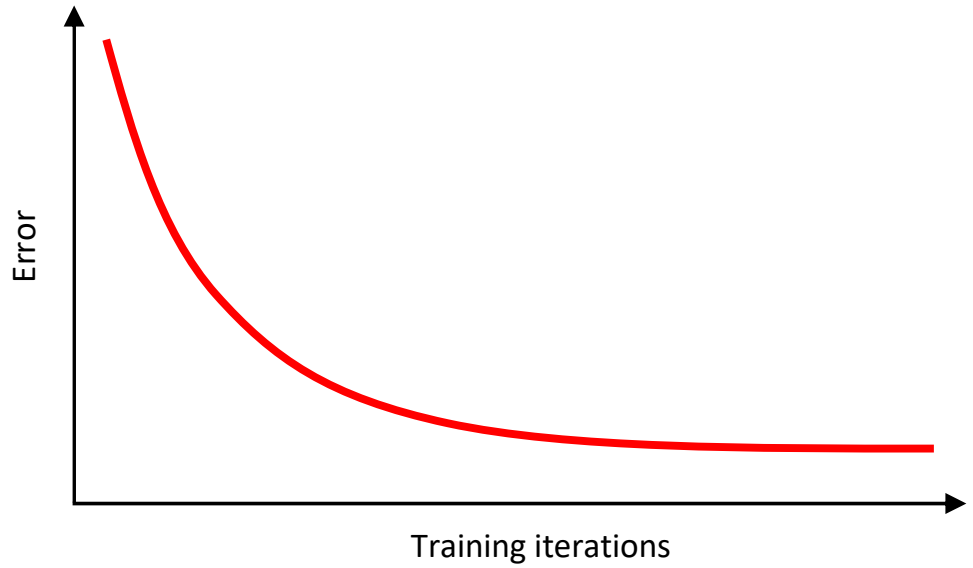
III. V. FEATURE ENGINEERING

- The process of using domain knowledge to extract features (variables) from raw data
- Used to improve the performance of the model
- Example:
  - *The average of previous N daily mean cashflow (CF) timings, by account/counterparty (CP)/portfolio/ instrument group* ⟶ **capturing short- and long- term patterns specific to a certain account, CP, etc.**
  - *The average of previous N daily mean CF timings, by CP and account/portfolio/instrument group pairs* ⟶ **capturing short- and long- term patterns specific to a certain CP and account/portfolio/instrument group pair**
  - *The average of previous N daily mean CF timings* ⟶ **capturing general short- and long- term patterns**
  - The mean timing of all cashflows (CFs) N days before for multiple values of N, determined by use of the autocorrelation function (ACF) – lags with largest autocorrelation values (in abs. value) ⟶ **capturing general short- and long- term trends**
  - **The day of the week (MON-FRI), the day of the month (1-31), the month of the year (1-12)**

III. VI. TRAINING PROCEDURE

- 80% of the dataset is used for training and validation
- The out-of-sample performance (generalization) is evaluated on the remaining 20% of the dataset (the testing set)
- MSE (Mean Squared Error) and WMSE (Weighted Mean Squared Error) are used as the objective function
  - MSE is much faster to train than MAE (Mean Absolute Error)
- $MSE = {}^{1}/_{N} \sum_i (\hat{y}_i - y_i)^2$
  - Weights w are again set to CF payment amounts – model fits larger CFs better
- $WMSE = \frac{1}{\sum_i w_i} \sum_i w_i (\hat{y}_i - y_i)^2$





IV. Results and discussion

IV. I. PERFORMANCE COMPARISON

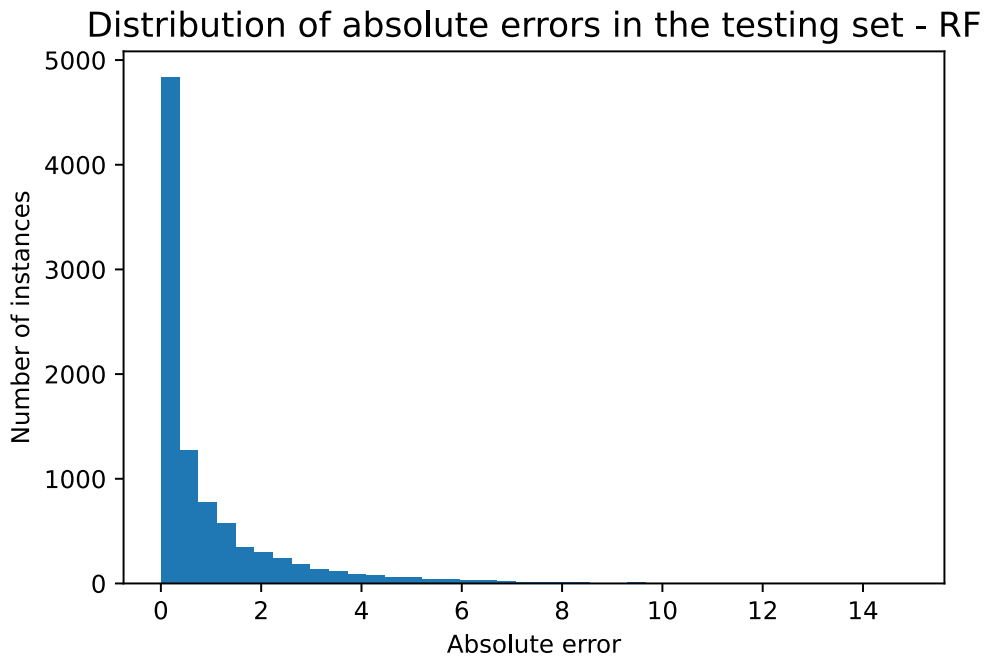
- Always important to compare against benchmark (baseline) models!
- $RMSE = \sqrt{{}^1/_N \sum_i (\hat{y}_i - y_i)^2}$,
- $MAE = {}^1/_N \sum_i |(\hat{y}_i - y_i)|$
- $RWMSE = \sqrt{\frac{1}{\sum_i w_i} \sum_i w_i (\hat{y}_i - y_i)^2}$,
- $WMAE = \frac{1}{\sum_i w_i} \sum_i w_i |(\hat{y}_i - y_i)|$
- MAE is around 0.95 ⇔ on average the predictions are 57 minutes off

	RMSE [hrs]	MAE [hrs]	*RWMSE [hrs]*	*WMAE [hrs]*
Null	3.38	2.61	4.36	3.26
GLM	2.50	1.77	4.65	3.27
WGLM	3.25	2.37	3.27	2.27
RF (MSE)	1.77	0.95	2.62	1.55
WRF (MSE)	1.76	0.96	2.61	1.54

IV. II. ERROR ANALYSIS

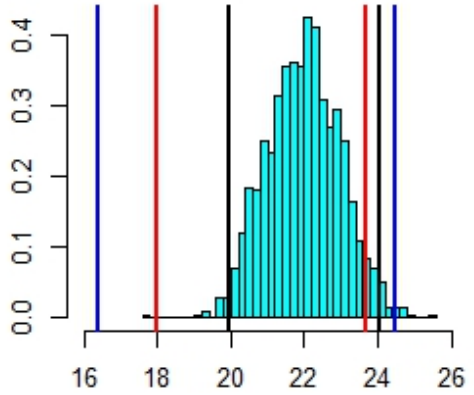
- The median error is only 0.35 (around 21 minutes) ⇔ For 50% of cashflows the model is less than 21 minutes off
- Exponential looking

- Prediction intervals:
  - Prediction: 16:39
  - 90% prediction interval: [15:50-16:59]



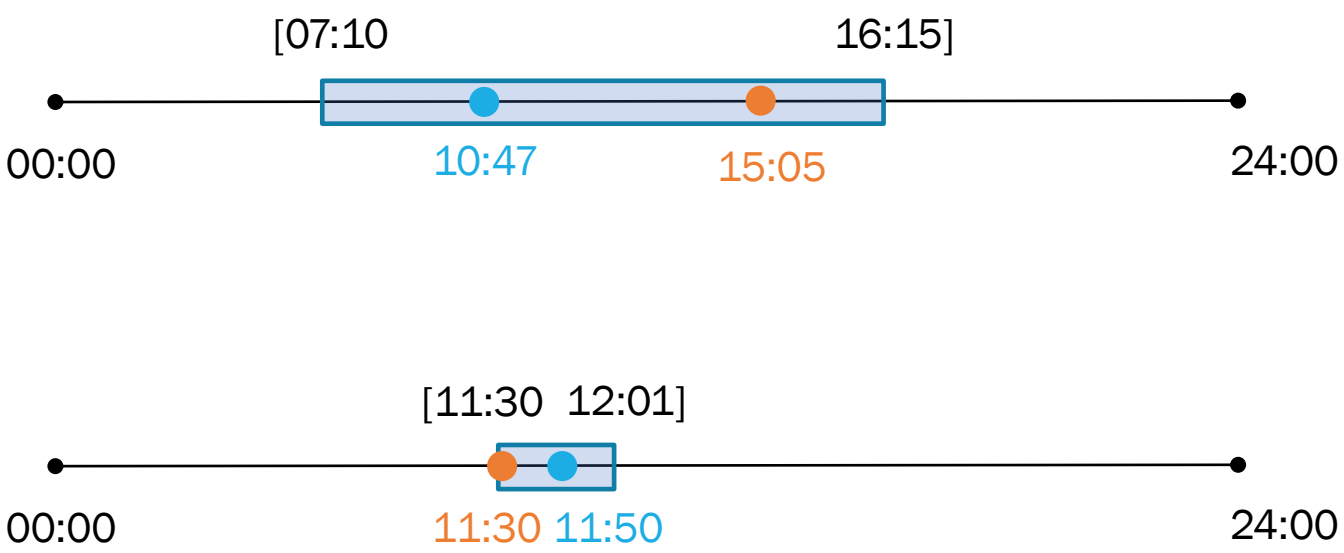
IV. III. PREDICTION INVERVALS (CONT.)

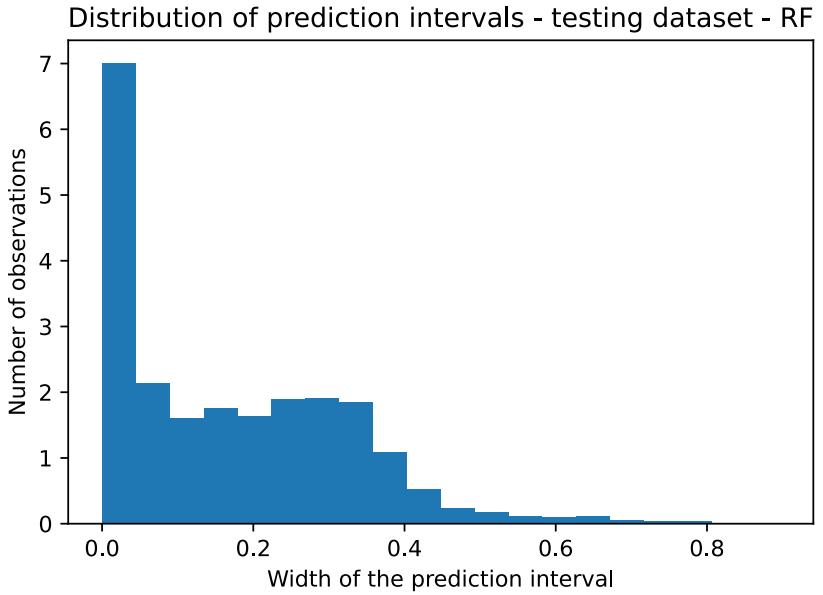
- It is desirable to quantify our confidence in each of the generated predictions (for a new instance)?
- Expectedly, certain CPs/accounts/etc. are more “recalcitrant” i.e. more difficult to predict → quantified by prediction intervals
- Example: 90% prediction interval $[12{:}00,13{:}00] \Leftrightarrow$ the probability of the CF timing falling into the interval is 90%
- The following procedure is employed:
  - We fully expand each of the N decision trees such that each leaf has only one observation.
  - The resulting N individual predictions are used to form a distribution
  - The percentiles of the distribution are used to determine the prediction intervals
    - (90% prediction interval lies between 5 and 95 percentiles of the distribution)
- This enables us to return not only the conditional mean (point estimates) but also conditional distributions
- Not to be confused with *confidence intervals* (the latter are related to estimates of the unknown true population parameter)
- Our testing confirms the validity of the approach (around 91% of the predictions lie within the 90% confidence interval)



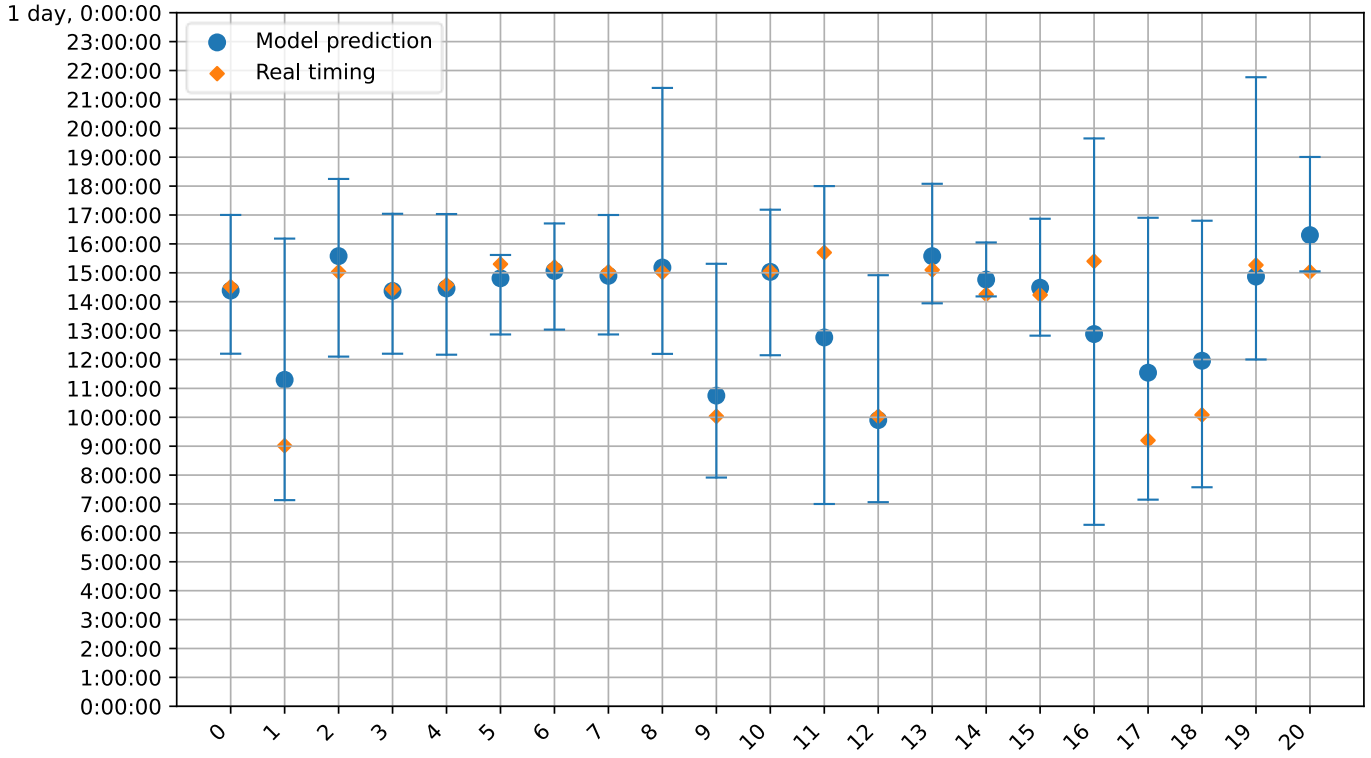
IV. III. PREDICTION INTERVALS (CONT.)

Prediction interval examples:



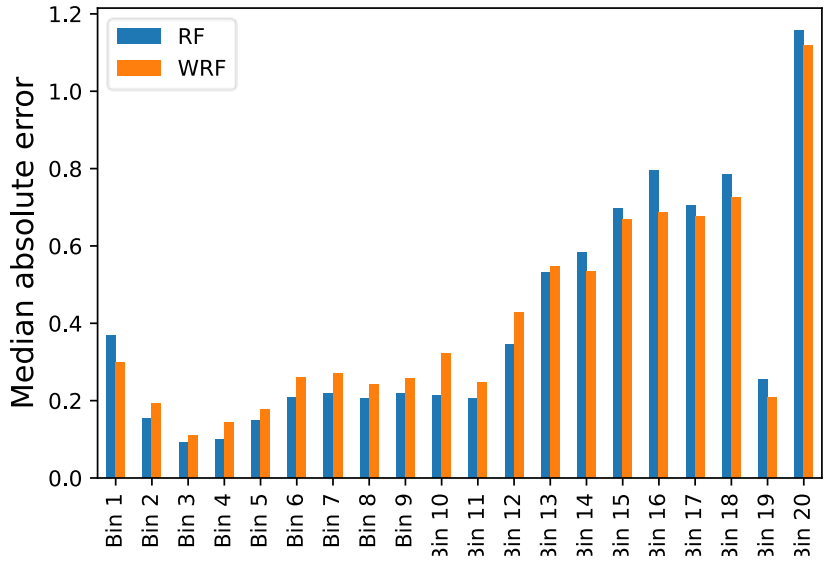


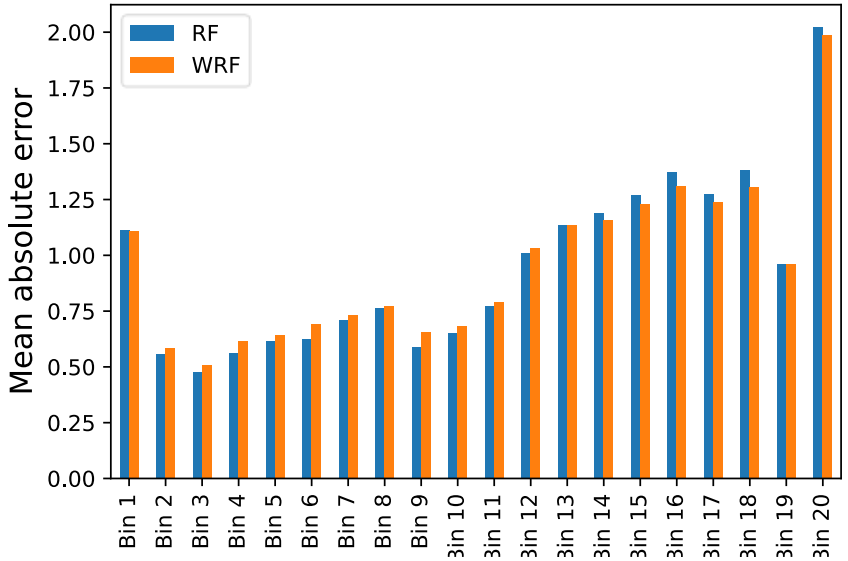
IV. III. PREDICTION INTERVALS (CONT.)



IV. IV. ERROR ANALYSIS (CONT.)

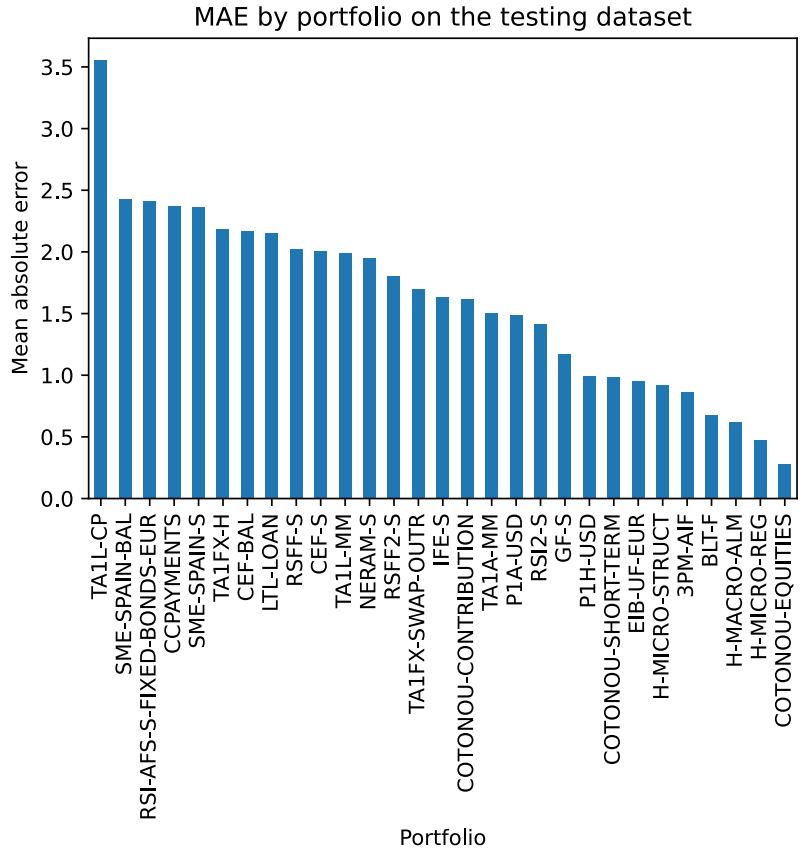
- Note: cashflows with larger payment amounts seem to be more difficult to predict (on average). WRF results in smaller error for large payment amount.

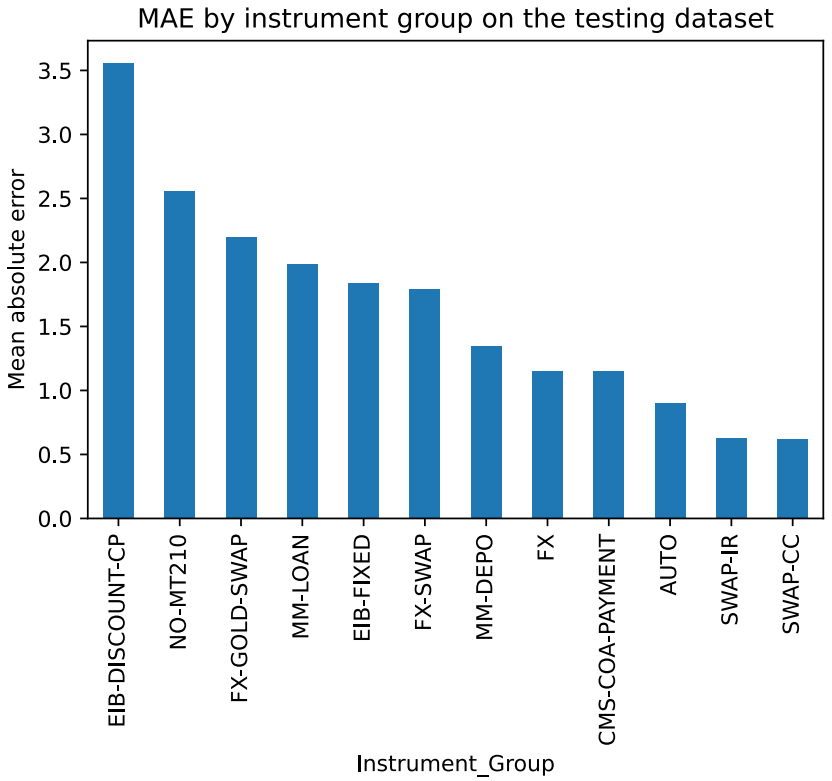




IV. IV. ERROR ANALYSIS (CONT.)

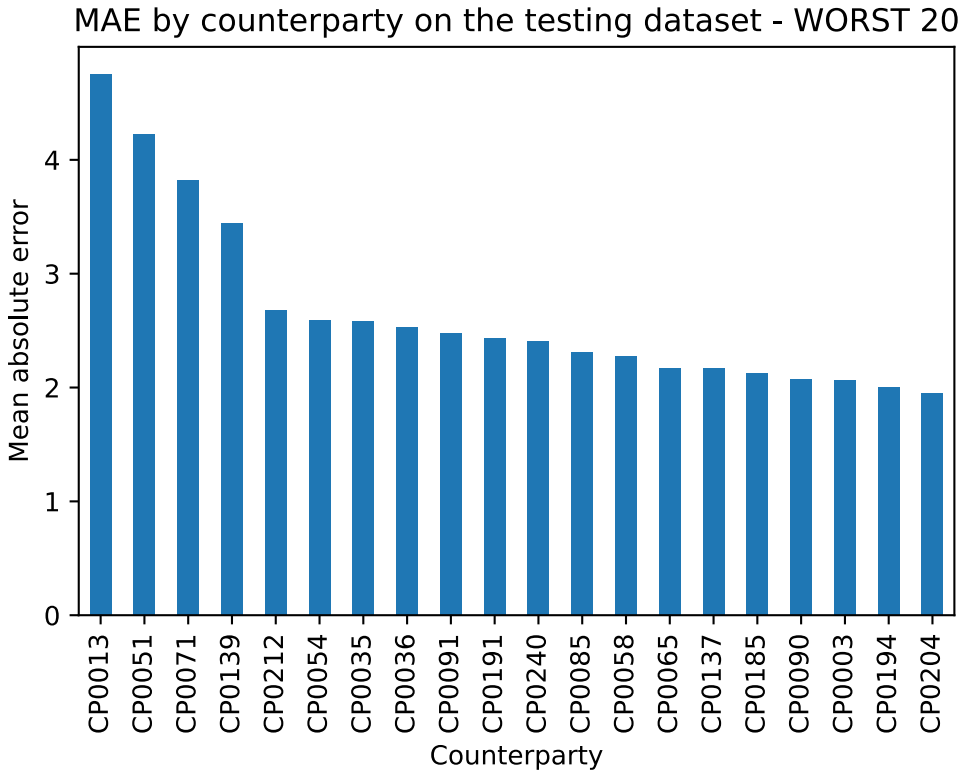
- Note the great differences in predictability between various portfolios / instrument groups.

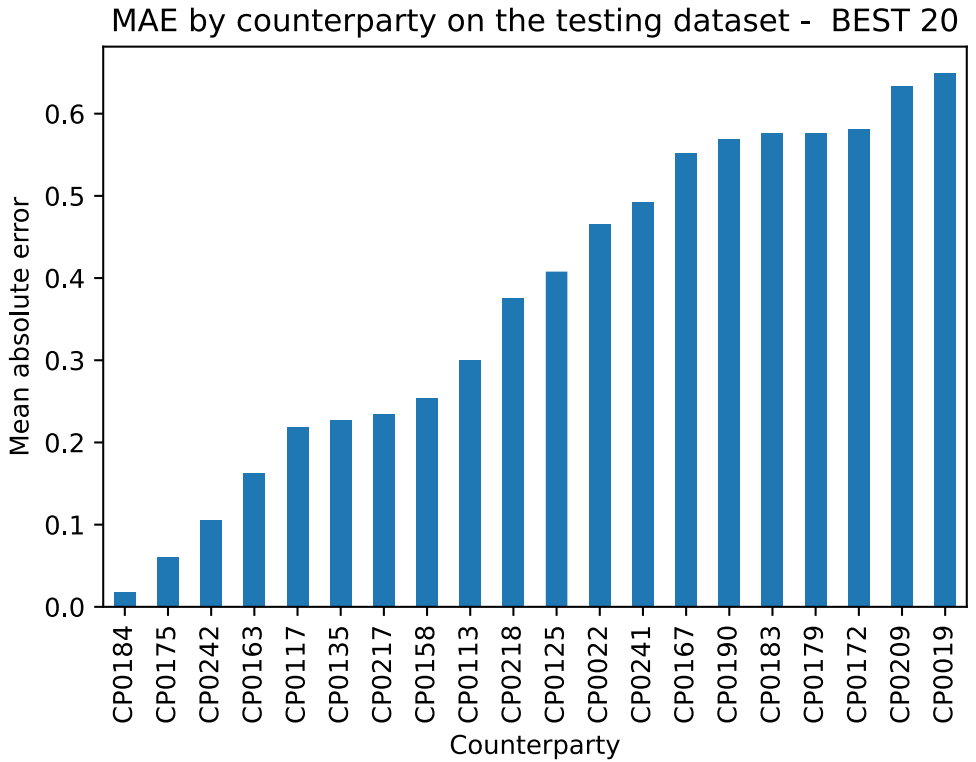




IV. IV. ERROR ANALYSIS (CONT.)

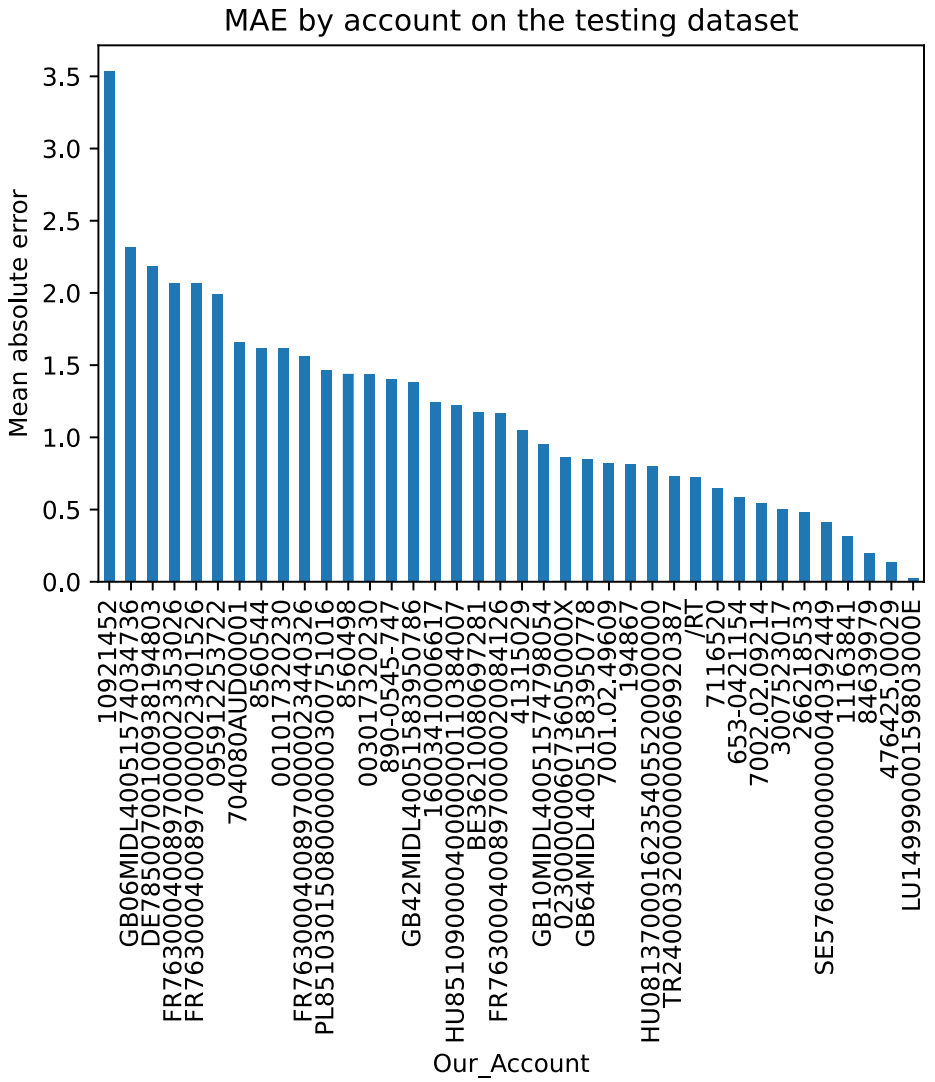
- Figures show the 20 counterparties associated with least/most predictable CF timings. CP0013, CP0051, and CP0071 seem to be the most unpredictable, while CP0184, CP0175, and CP0242 constitute most predictable ones.





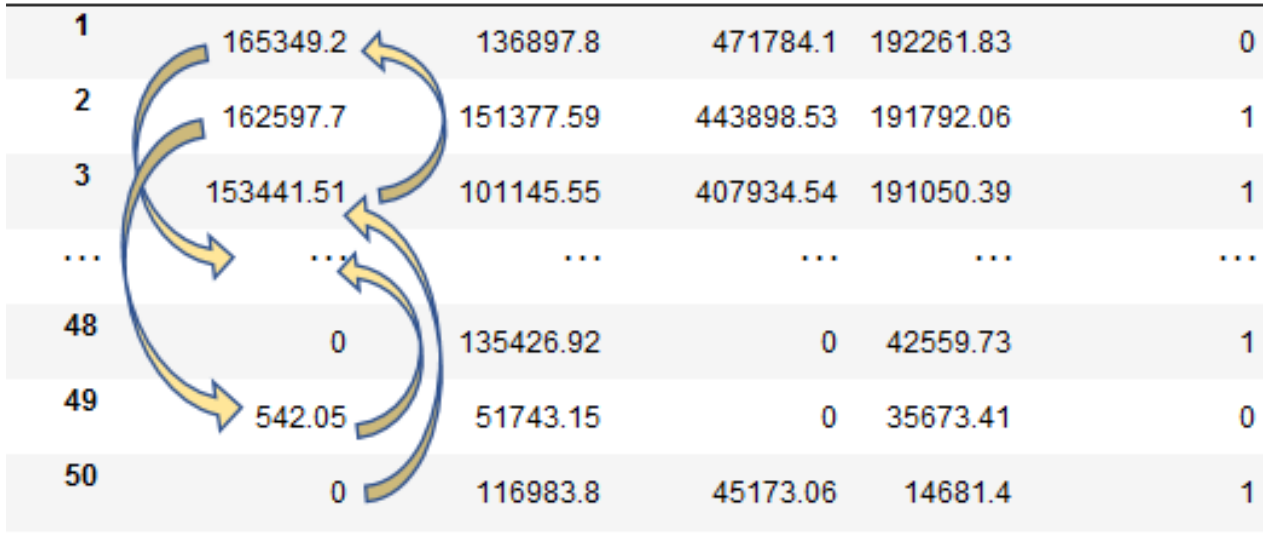
IV. IV. ERROR ANALYSIS (CONT.)

- Account “10921452" seems to be by far most problematic.
- Account LU1...0E => easiest to predict CFs
- Again, note significant differences in predictability across various accounts



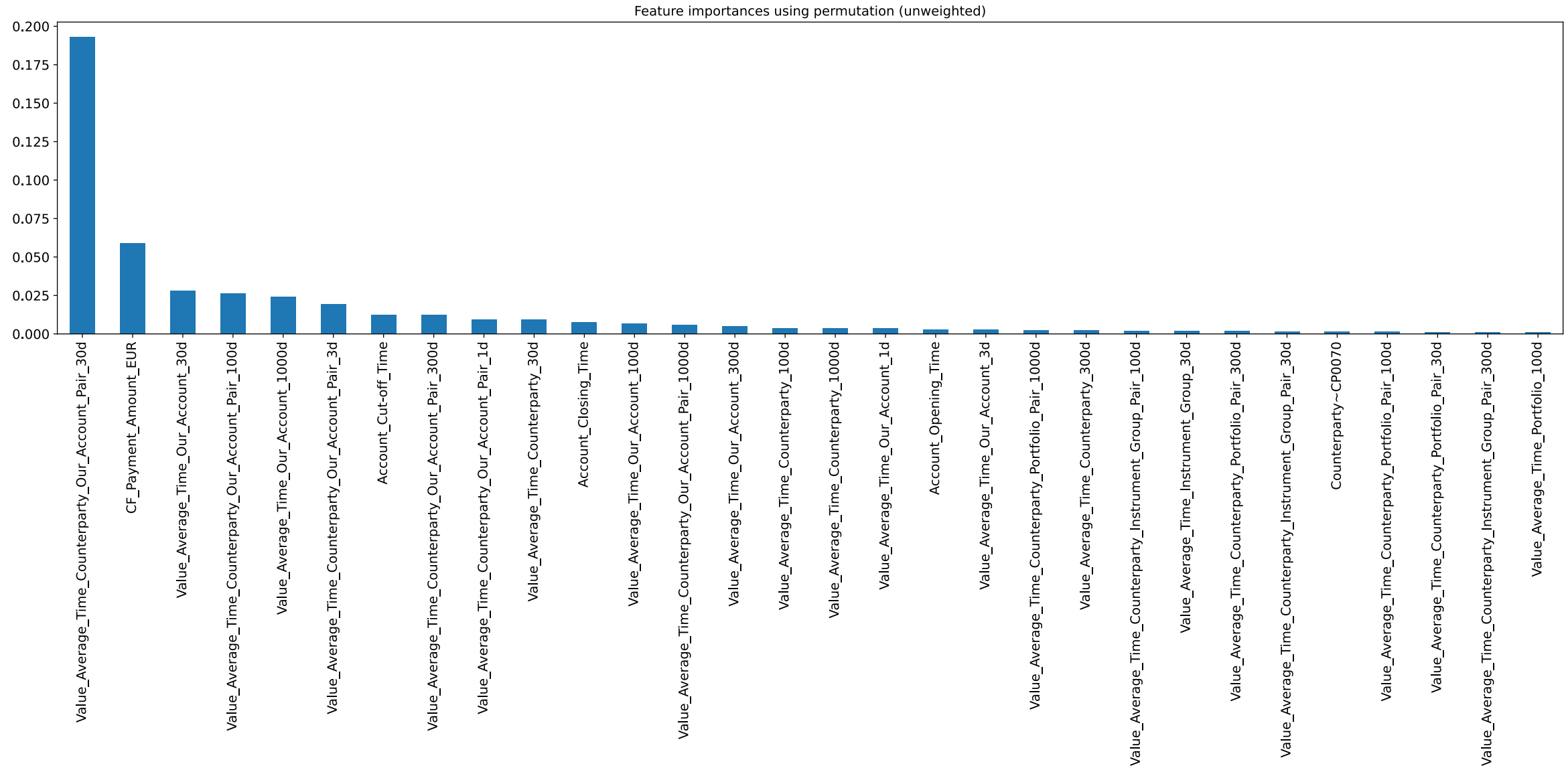
IV. V. FEATURE IMPORTANCE ANALYSIS

- We use feature importance based on feature permutation
- Permutation feature importance is the decrease in a model score when a certain (single) feature value is randomly shuffled
- Features are shuffled M times and the score is recomputed on corrupted (shuffled) testing data
- Permutation feature importance does not require retraining the model
- Adding correlated features can decrease the importance of the associated feature



1	165349.2	136897.8	471784.1	192261.83	0
2	162597.7	151377.59	443898.53	191792.06	1
3	153441.51	101145.55	407934.54	191050.39	1
...	...	...	...	...	...
48	0	135426.92	0	42559.73	1
49	542.05	51743.15	0	35673.41	0
50	0	116983.8	45173.06	14681.4	1

IV. V. FEATURE IMPORTANCE ANALYSIS (CONT.)



Note the disparity in importance.

V. Conclusion

KEY TAKEAWAYS

- ML based methods, in particular random forests, lend themselves particularly well for the problem of cashflow timing prediction
- Temporal features (handcrafted mean cashflow timings and opening/closing/cutoff time features) represent features with most predictive power
- Significant disparities in predictability between different counterparties, portfolios, instrument groups, payment amounts
- Prediction intervals provide a probabilistic perspective to the problem and enable quantifying the reliability of predictions
  - Humans using the system can decide whether the interval is too wide to trust the prediction
  - Intervals can be provided for different levels (90%, 95%, 99%)

KEY TAKEAWAYS (CONT.)

- The project shows the ability of the model to demonstrate predictive power
- More testing and live usage is needed to check usefulness under realistic conditions
- Combination of the prediction intervals with the EIB's information on the due financial obligations should be considered
- Modeling approaches (time series formulation), additional components and some questions left to future research
- The resulting ML framework hopefully provides a useful addition to the EIB's liquidity management arsenal

THANK YOU FOR YOUR ATTENTION.