Beat the Bookies

By: Casper, Malte, Philip, and Sebastian

What is the goal?

- Predict outcome of football matches using ML
- Combine with clever betting strategies
- Get positive return on investment (ROI)

The Data

Football data*
The 11 top european leagues
Seasons from 2008 to 2016

25979 matches

For each match the data contains
- Outcome (modified to be H or XA)
- Teams
- Players
- Date
- Odds from 10 bookies

*Data from Kaggle: kaggle.com/datasets/hugomathien/soccer

Wrangling the data (FIFA data*)

9 features for each team

- buildUpPlaySpeed, buildUpPlayDribbling, etc.

35 features for each player

- Overall_rating, heading_accuracy, etc.

= 788 features for each match

Some missing entries (NaN values)

*Data from Kaggle: kaggle.com/datasets/hugomathien/soccer

Imputation of FIFA data

Data: FIFA → loads of NaN

LightGBM-regression to impute data for all NaN-values

More suitable data for NN

‘Target Feature’ LGBM-Regression

Ft. 1	Ft. 2	Ft. n
3	50	NaN
NaN	75	7
8	NaN	10
5	43	8

→

‘Target Feature’ LGBM-Regression

Ft. 1	Ft. 2	Ft. n
3	50	NaN
6.73	75	7
8	NaN	10
5	43	8

→

Feature engineering

Using only the final score of each match

Creating data that shows the teams' league standing and stats during the season

Calculating parameters which try show performance in previous matches

Resulting in 57 features per match

Referred to as:
Historical League Standings and Stats

W	D	L	GF	GA	GD	pld	pld left	pts
0	0	0	0	0	0	0	34	0
0	0	1	1	2	-1	1	33	0
0	0	2	2	6	-4	2	32	0
0	0	3	3	10	-7	3	31	0
0	1	3	3	10	-7	4	30	1
0	1	4	3	11	-8	5	29	1
0	1	5	3	12	-9	6	28	1
1	1	5	5	13	-8	7	27	4
1	1	6	5	14	-9	8	26	4
1	1	7	6	18	-12	9	25	4
1	1	8	7	21	-14	10	24	4

Small cutout of one teams stats calculations

ROI (Return On Investment)

ROI = (Money earned - #bets placed) / #bets placed

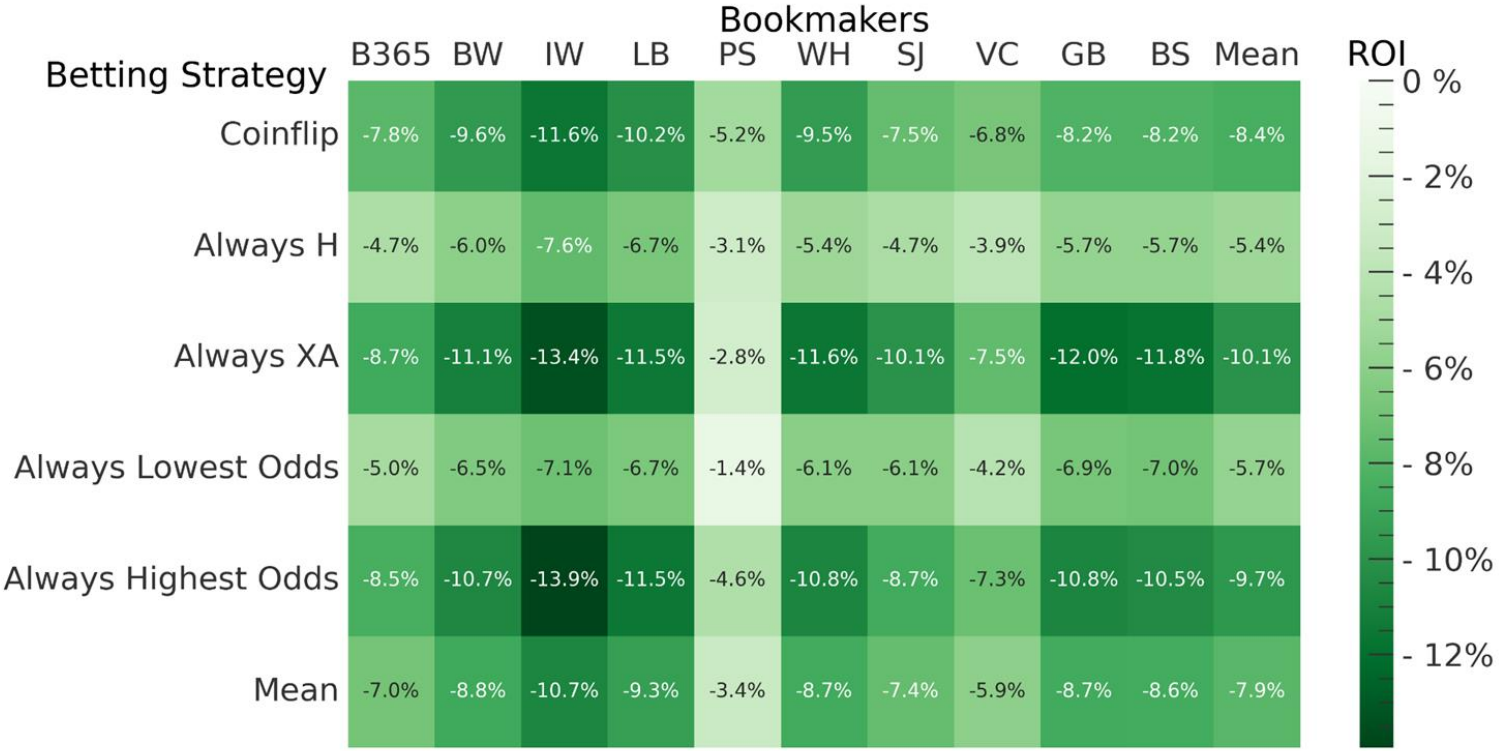
Assuming we always bet 1 kr

E.g. 1kr bet with odds 1.15.

Win: ROI = 15%

Lose: ROI = -100%

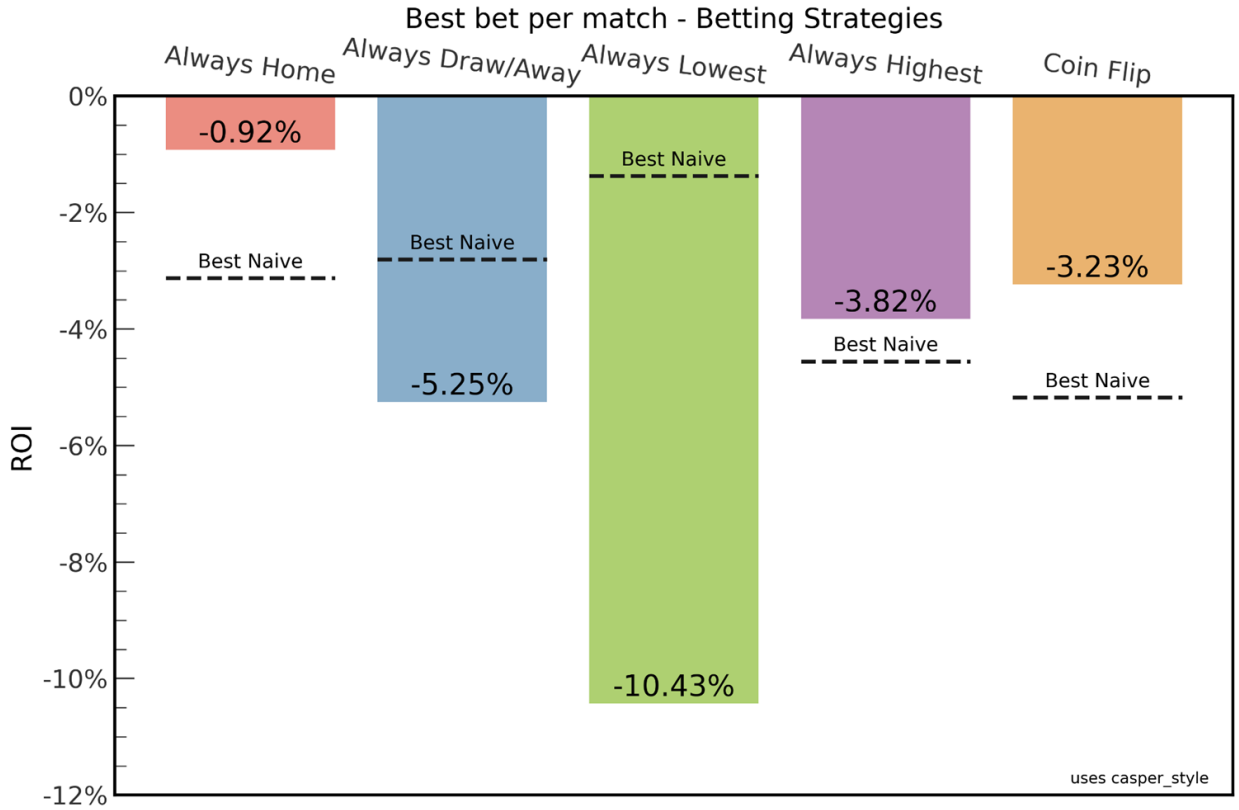
Naive betting strategies



Bookie Reference:

B365 - Bet365
BW - BetWay
IW - InterWetten
LB - LadBrokes
PS - Pinnacle Sports
WH - William Hill
SJ - Stan James
VC - Victor Chandler
GB - Gamebookers
BS - Blue Square

Can the betting strategies be improved?



Betting strategy & Custom Loss Function

Model probability prediction: $\{x \in R \mid 0<x<1\}$

Bookie probability prediction: 1/Odds

Confidence score: P_{model}/P_{bookie}

Select bet with highest confidence for each match, not necessarily the predicted winner!

Eg. we predict home win with probability 0.4, bookmaker predict 0.2: Conf-score = 2!

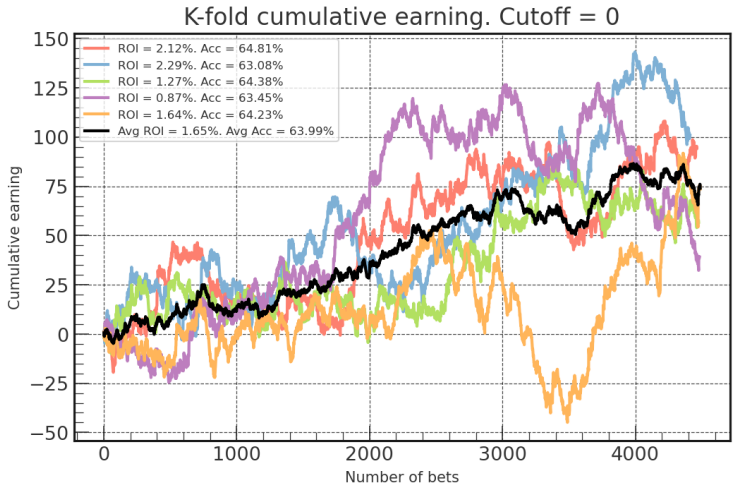
Custom loss function that maximizes the ROI during hyperparam optimization for LightGBM

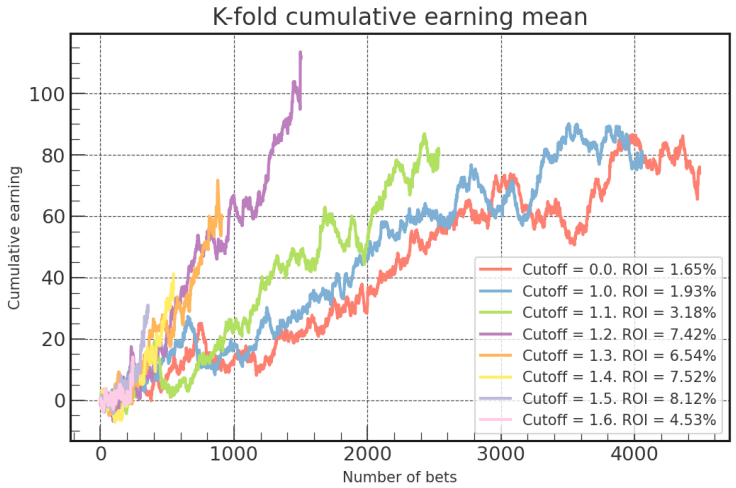
Results

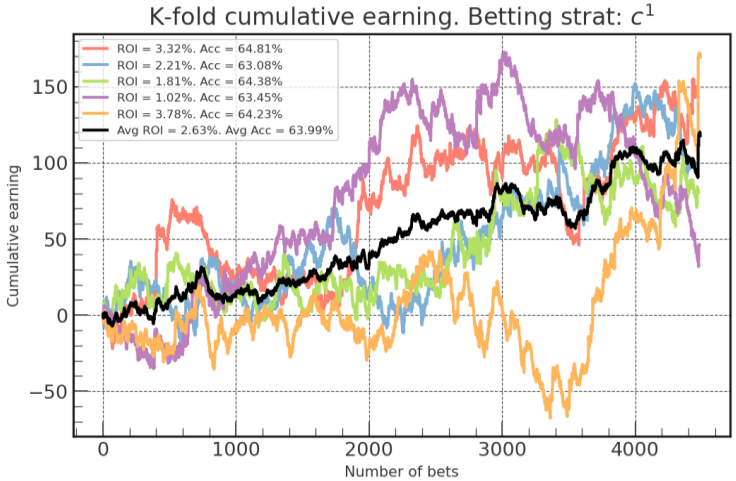
Performance

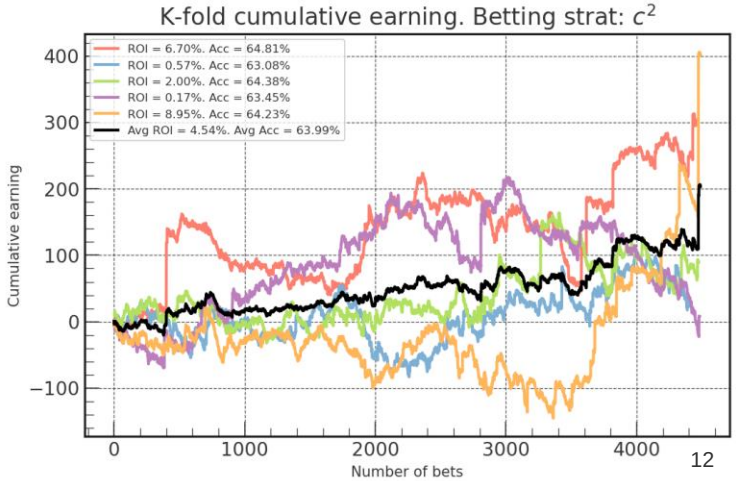
Model: LightGBM

Features: FIFA Ratings





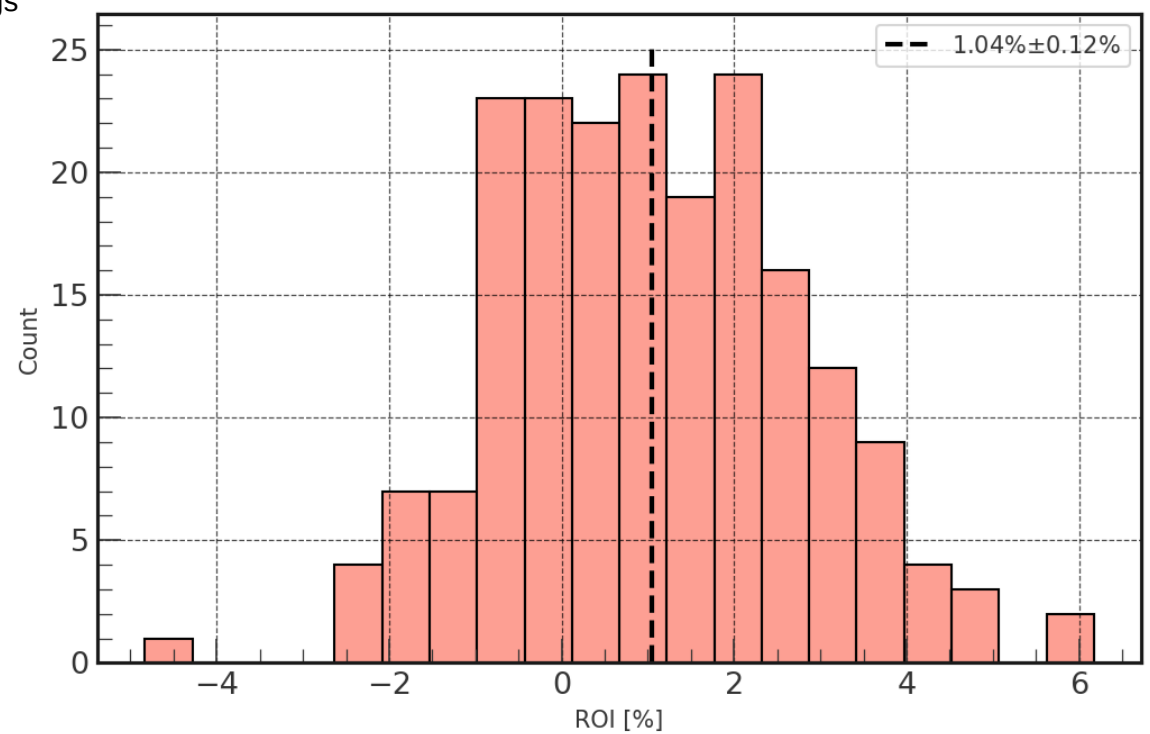




Performance

Model: LightGBM

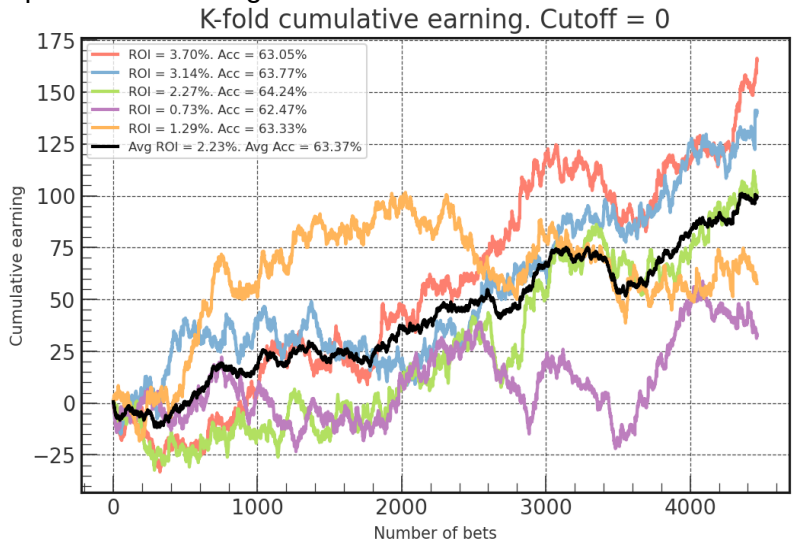
Features: FIFA Ratings

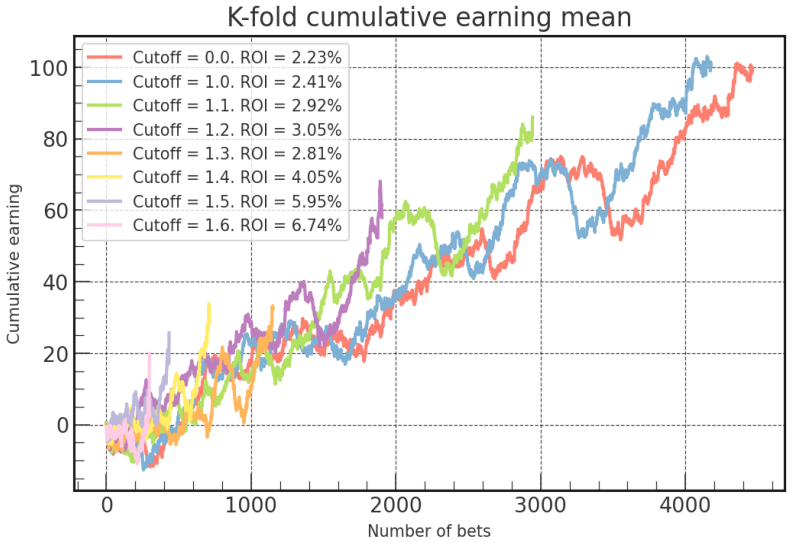


Performance

Model: LightGBM

Features: Imputed FIFA Ratings

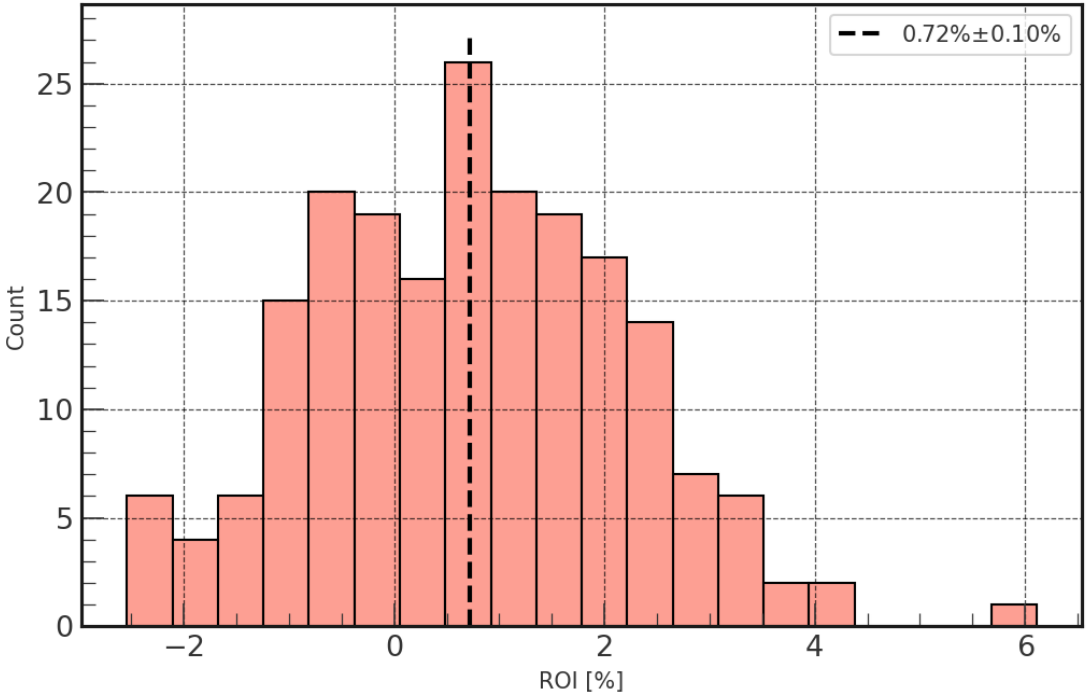




Performance

Model: LightGBM

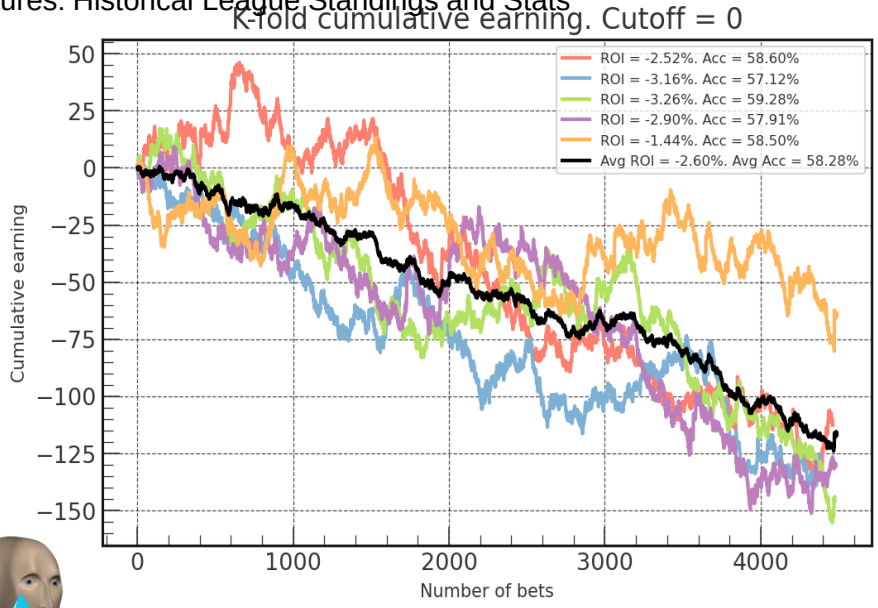
Features: Imputed FIFA Ratings

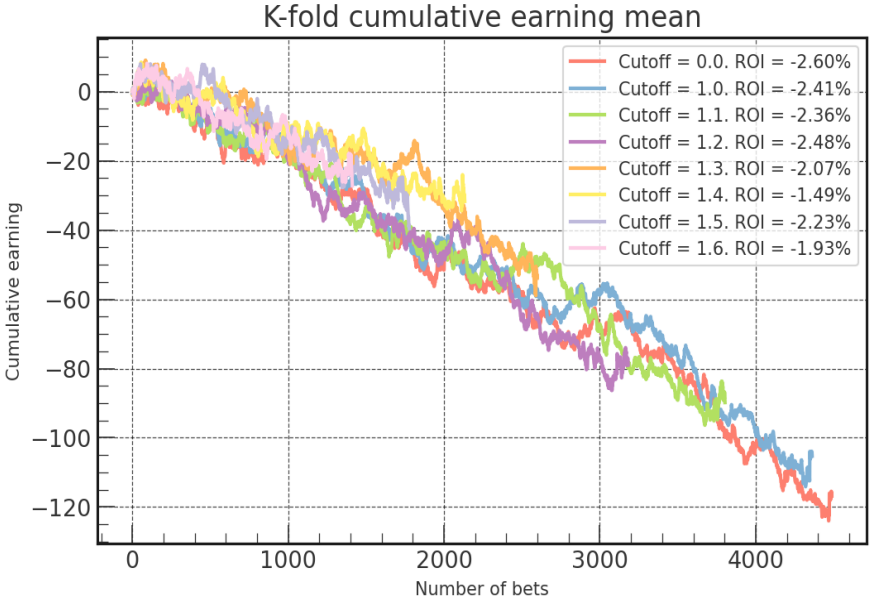


Performance

Model: LightGBM

Features: Historical League Standings and Stats

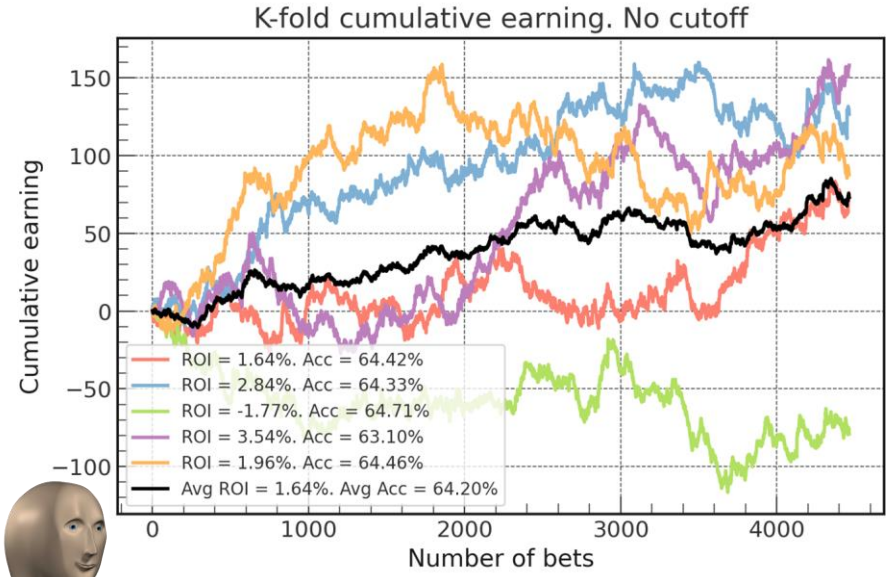


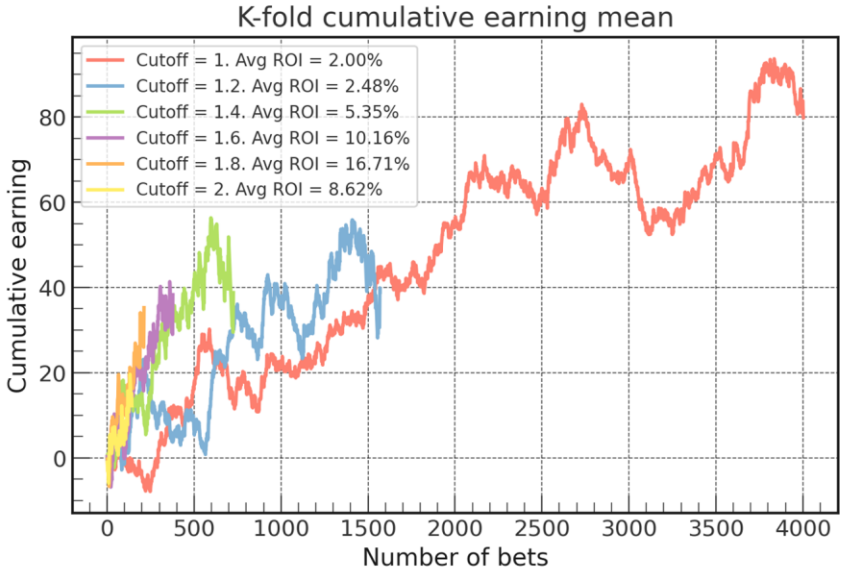


Performance

Model: TensorFlow Feed Forward-NN:

Features: Imputed FIFA ratings

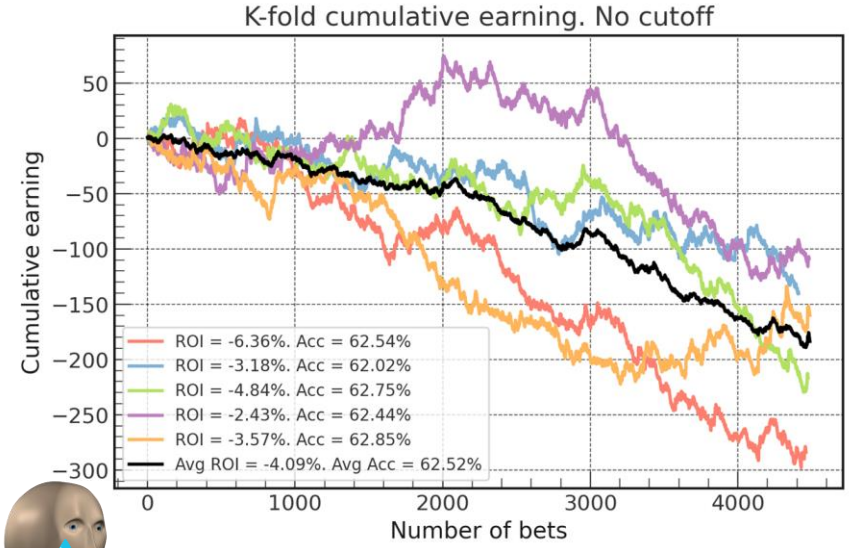


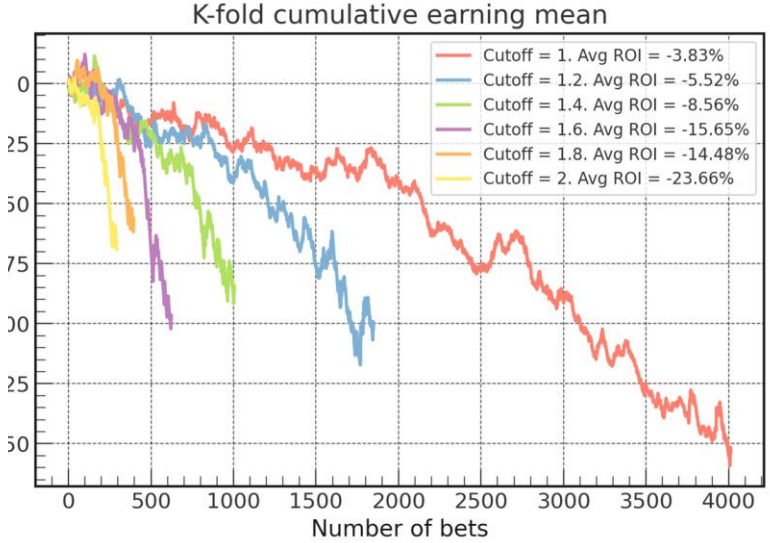


Performance

Model: TensorFlow Feed Forward-NN:

Features: Historical League Standings and Stats

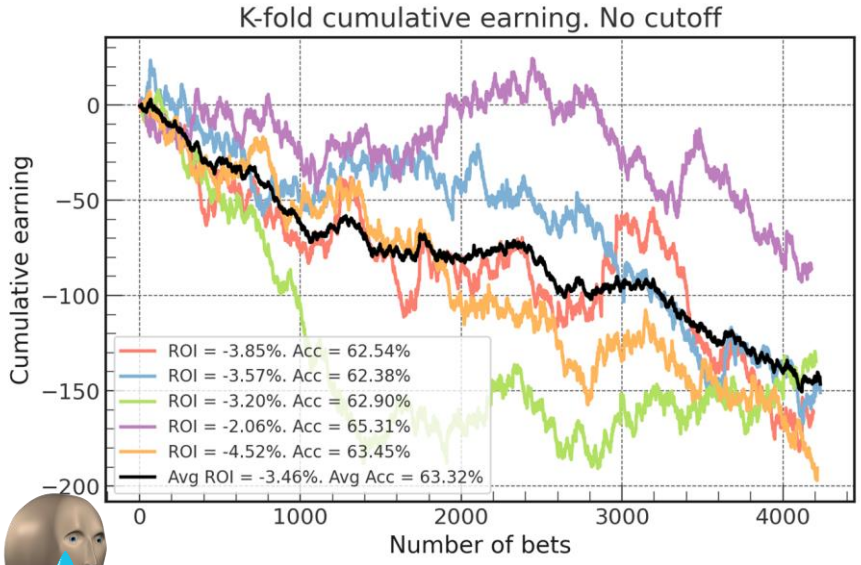


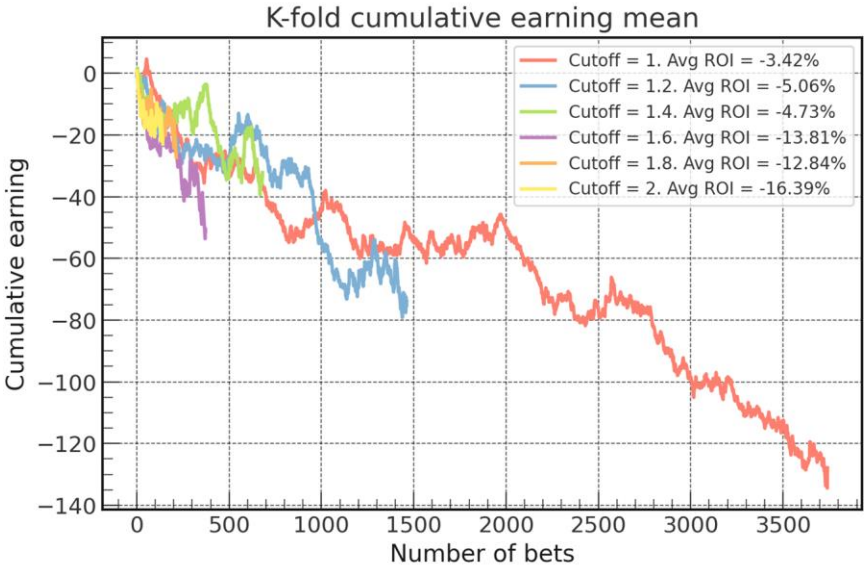


Performance

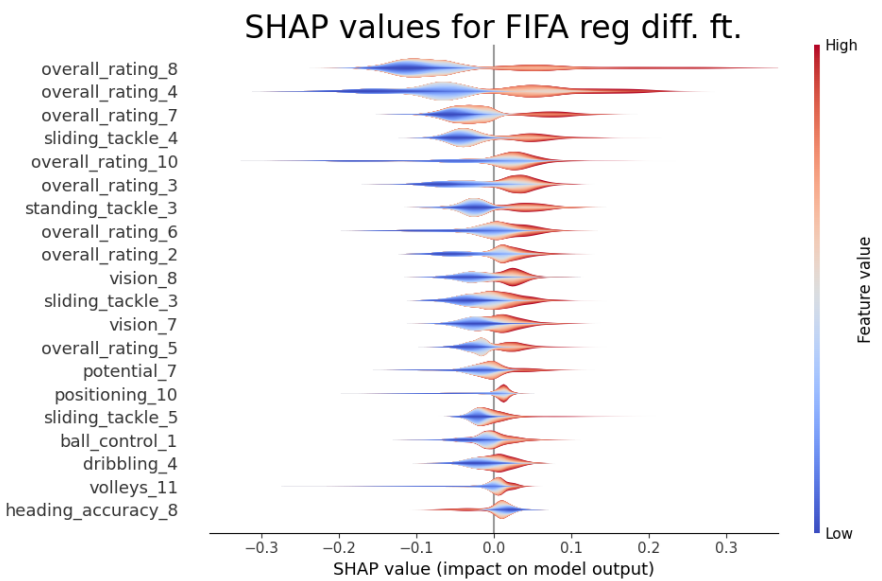
Model: TensorFlow Recurrent-NN with Long Short Term Memory layers:

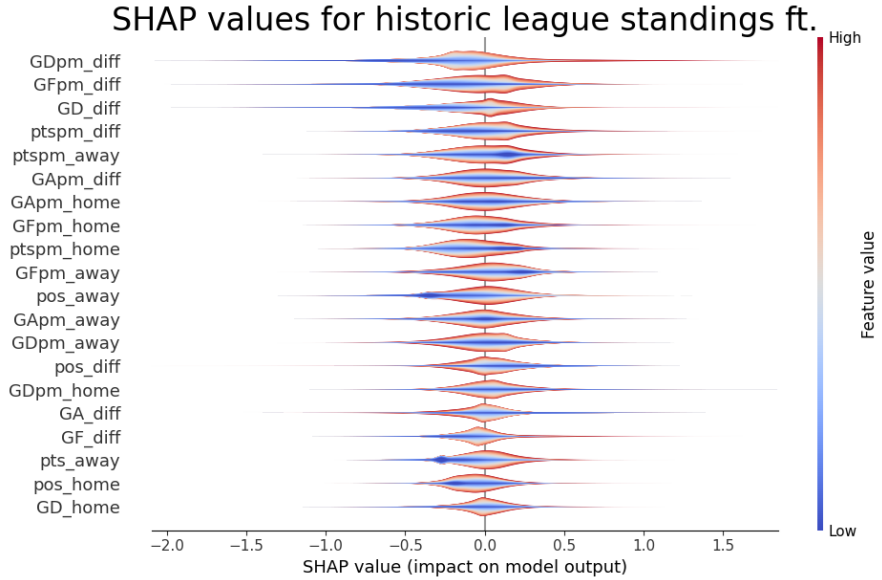
Features: Historical League Standings and Stats with Time-step = 10





SHAP Values

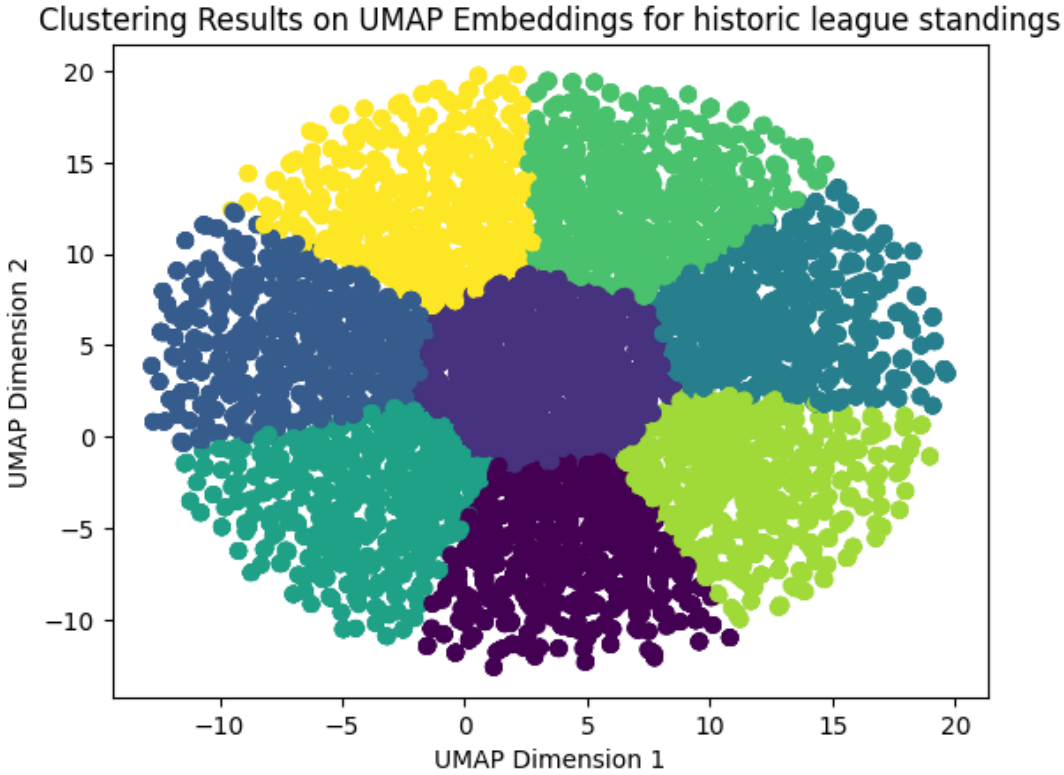

SHAP values for FIFA reg diff. ft.
overall_rating_8
overall_rating_4
overall_rating_7
sliding_tackle_4
overall_rating_10
overall_rating_3
standing_tackle_3
overall_rating_6
overall_rating_2
vision_8
sliding_tackle_3
vision_7
overall_rating_5
potential_7
positioning_10
sliding_tackle_5
ball_control_1
dribbling_4
volleys_11
heading_accuracy_8
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
SHAP value (impact on model output)
High
Feature value
Low


SHAP values for historic league standings ft.
GDpm_diff
GFpm_diff
GD_diff
ptspm_diff
ptspm_away
GApm_diff
GApm_home
GFpm_home
ptspm_home
GFpm_away
pos_away
GApm_away
GDpm_away
pos_diff
GDpm_home
GA_diff
GF_diff
pts_away
pos_home
GD_home
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
SHAP value (impact on model output)
High
Feature value
Low

Summary and further work

- Positive ROI for FIFA features with both LightGBM and NN:
  - Update features based on SHAP values and improve ROI
  - Do more rigorous statistical tests to veri
- Negative ROI for historical league standings features with both LightGBM, NN and RNN:
  - Calculate better performance parameters
  - Remove matches early in the season
  - Introduce more (or less) timesteps for RNN data
- Get more data!

Thank you for listening



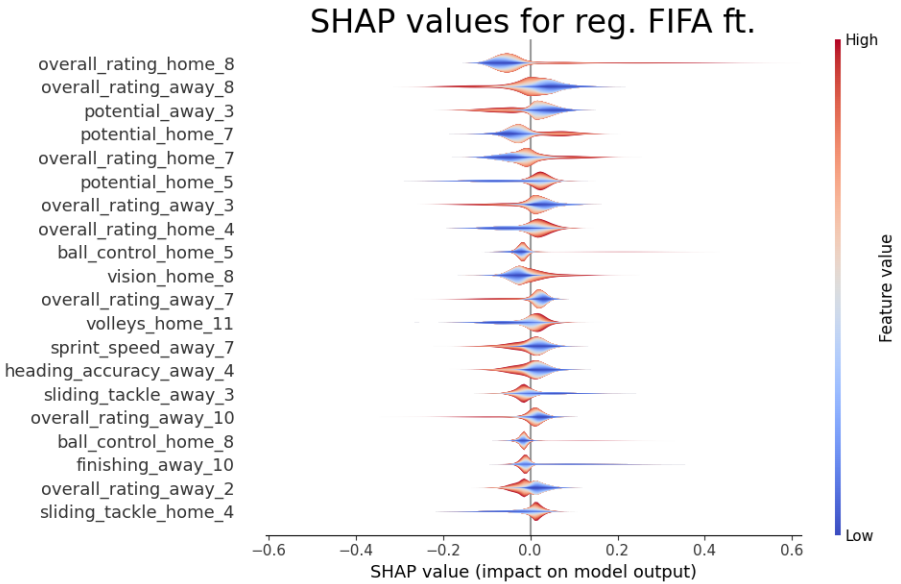
Appendix

All participants of the group have contributed equally in this project, and all the models were conjured and evaluated in collaboration with each other.

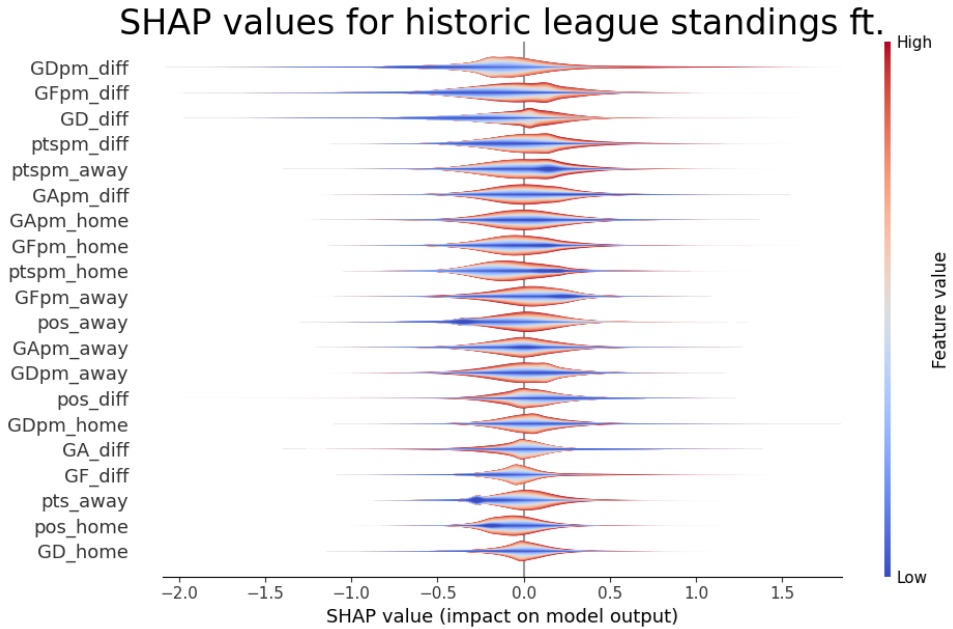
The group has consisted of the following members:

Sebastian Koza (wtj465)
Casper Wied (nqs117)
Philip Kofoed Djursner (tkv976)
Malte Wettergren Andreasen (srl902)

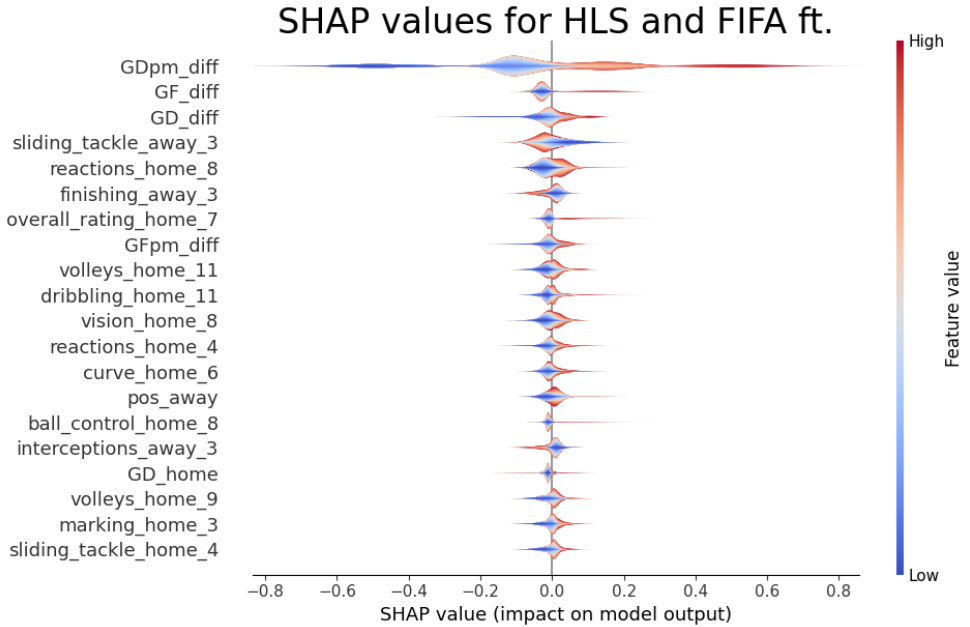
SHAP-values:


SHAP values for reg. FIFA ft.
overall_rating_home_8
overall_rating_away_8
potential_away_3
potential_home_7
overall_rating_home_7
potential_home_5
overall_rating_away_3
overall_rating_home_4
ball_control_home_5
vision_home_8
overall_rating_away_7
volleys_home_11
sprint_speed_away_7
heading_accuracy_away_4
sliding_tackle_away_3
overall_rating_away_10
ball_control_home_8
finishing_away_10
overall_rating_away_2
sliding_tackle_home_4
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
SHAP value (impact on model output)
High
Feature value
Low

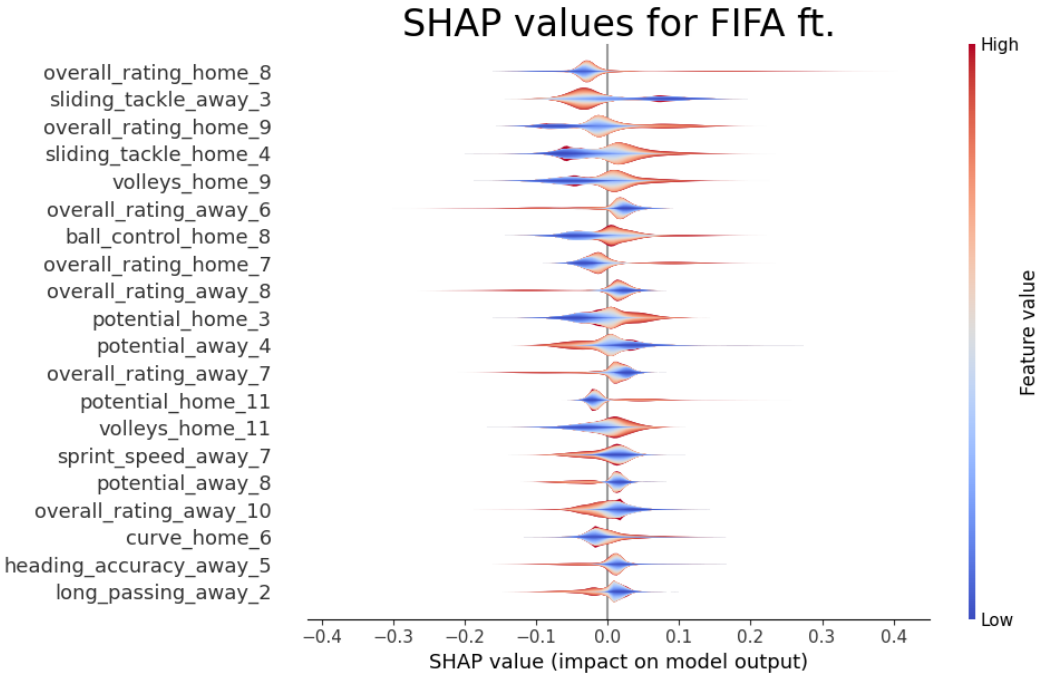
SHAP-values:


SHAP values for historic league standings ft.
GDpm_diff
GFpm_diff
GD_diff
ptspm_diff
ptspm_away
GApm_diff
GApm_home
GFpm_home
ptspm_home
GFpm_away
pos_away
GApm_away
GDpm_away
pos_diff
GDpm_home
GA_diff
GF_diff
pts_away
pos_home
GD_home
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
SHAP value (impact on model output)
High
Feature value
Low

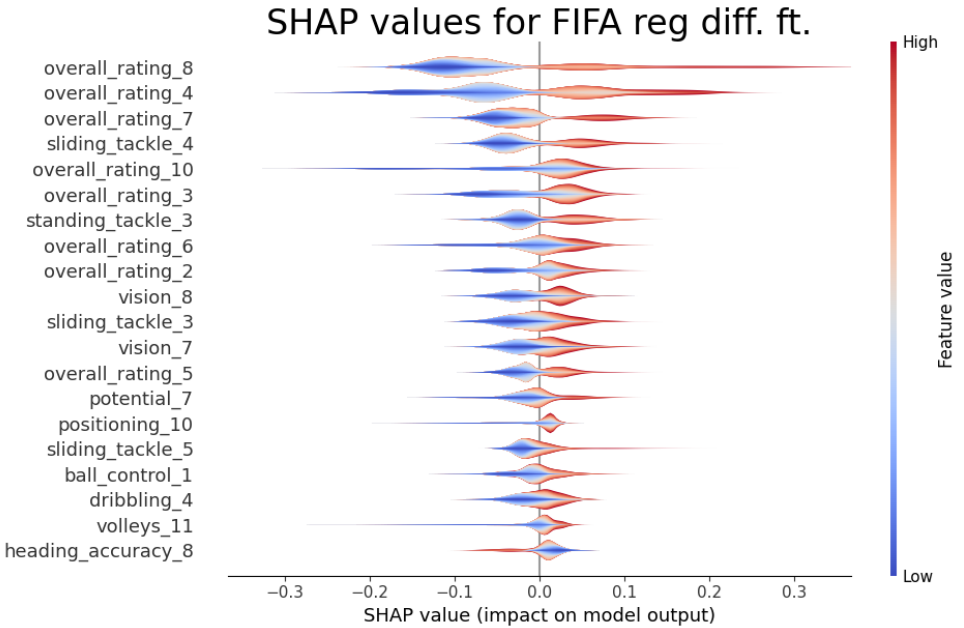
SHAP-values:


SHAP values for HLS and FIFA ft.
GDpm_diff
GF_diff
GD_diff
sliding_tackle_away_3
reactions_home_8
finishing_away_3
overall_rating_home_7
GFpm_diff
volleys_home_11
dribbling_home_11
vision_home_8
reactions_home_4
curve_home_6
pos_away
ball_control_home_8
interceptions_away_3
GD_home
volleys_home_9
marking_home_3
sliding_tackle_home_4
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
SHAP value (impact on model output)
High
Feature value
Low

SHAP-values:


SHAP values for FIFA ft.
overall_rating_home_8
sliding_tackle_away_3
overall_rating_home_9
sliding_tackle_home_4
volleys_home_9
overall_rating_away_6
ball_control_home_8
overall_rating_home_7
overall_rating_away_8
potential_home_3
potential_away_4
overall_rating_away_7
potential_home_11
volleys_home_11
sprint_speed_away_7
potential_away_8
overall_rating_away_10
curve_home_6
heading_accuracy_away_5
long_passing_away_2
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
SHAP value (impact on model output)
High
Feature value
Low

SHAP-values:


SHAP values for FIFA reg diff. ft.
overall_rating_8
overall_rating_4
overall_rating_7
sliding_tackle_4
overall_rating_10
overall_rating_3
standing_tackle_3
overall_rating_6
overall_rating_2
vision_8
sliding_tackle_3
vision_7
overall_rating_5
potential_7
positioning_10
sliding_tackle_5
ball_control_1
dribbling_4
volleys_11
heading_accuracy_8
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
SHAP value (impact on model output)
High
Feature value
Low

Custom loss function for optuna hyperparam optimization for LightGBM

Loss function for each K-fold:

- Calculate highest confidence for each match in val set.
- Check if the outcome was correct
- Add either -1 or Odds-1 to Money earned dependent on Loss/Win
- Calculate ROI

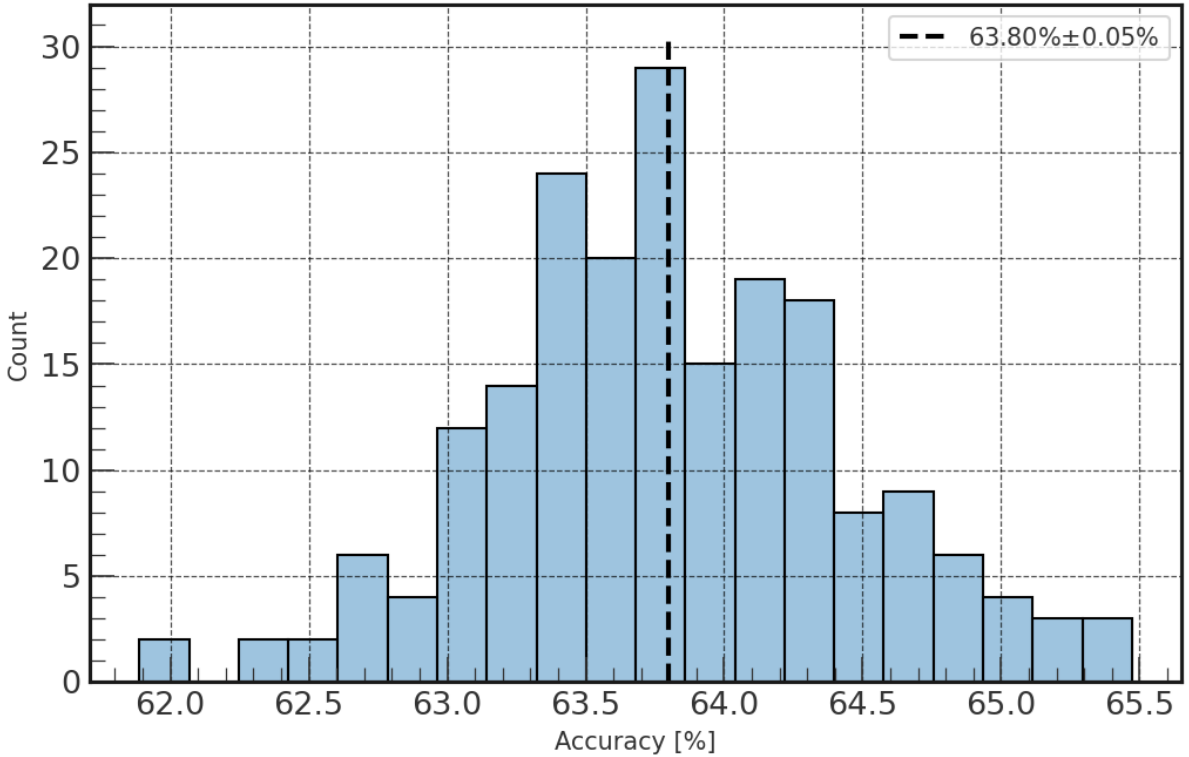
Average ROI over K-folds

Optuna maximises this loss function using Bayesian Optimization

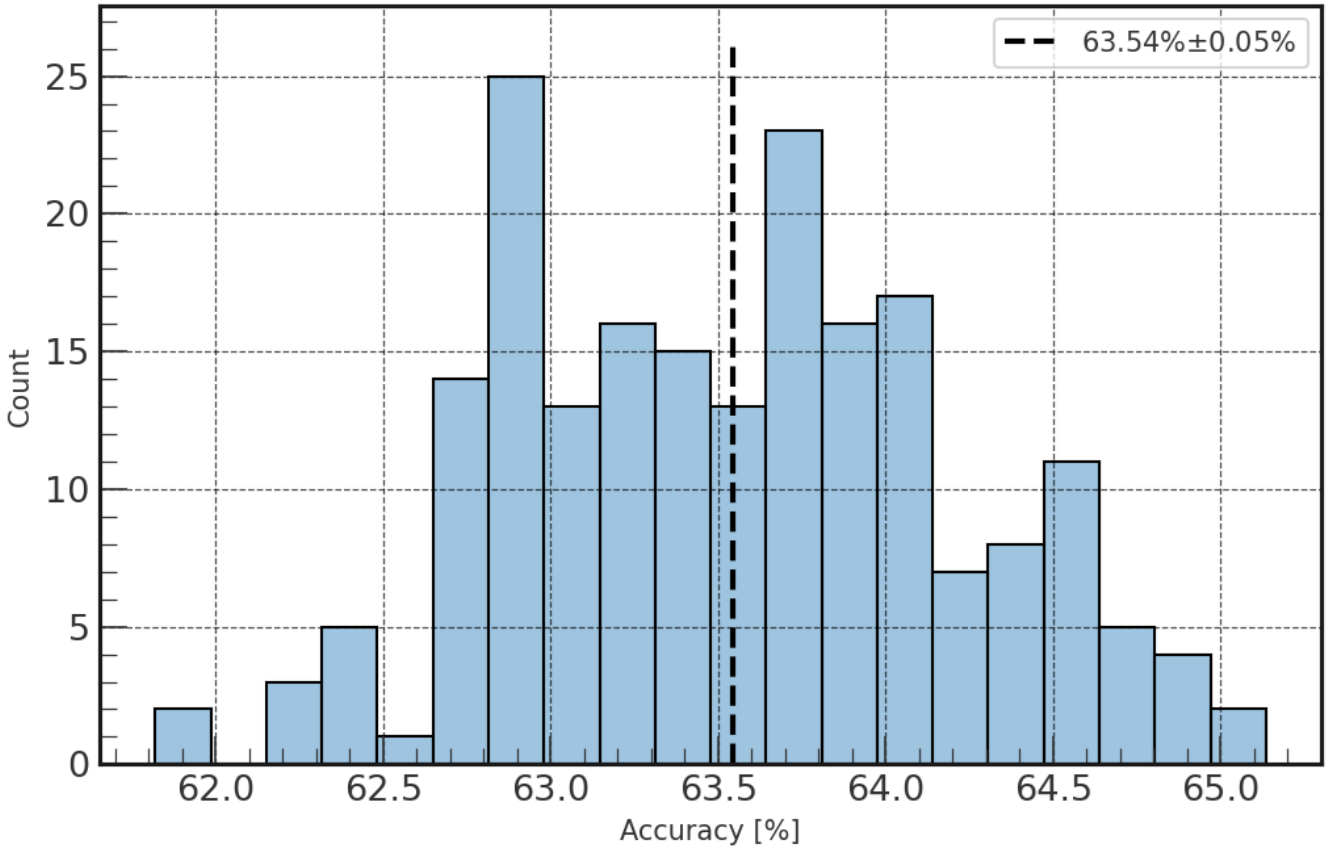
Dataset description

FIFA data	Team and player attributes taken from the FIFA game at the date closest to the match. 25629 matches with 788 features. Contains NaN values
Regressed FIFA data	The fifa data where regression have been used to guess the value in place of a NaN value and thereby give a complete dataset. 25629 matches with 788 features
Stats data	Standing and accumulative scores for teams before the match. Furthermore, contains information about performance in previous matches. 25969 matches with 57 features.
Long stats data	Adaptation of the stats dataset to include historical scores for use in RNN. For each match the standing before for home and away teams last 10 matches are included. 25969 matches with 10 timesteps each with 57 features.
Stats + Fifa	Combination of Fifa data and stats data. Excludes the matches not in fifa data. 25629 matches with 845 features

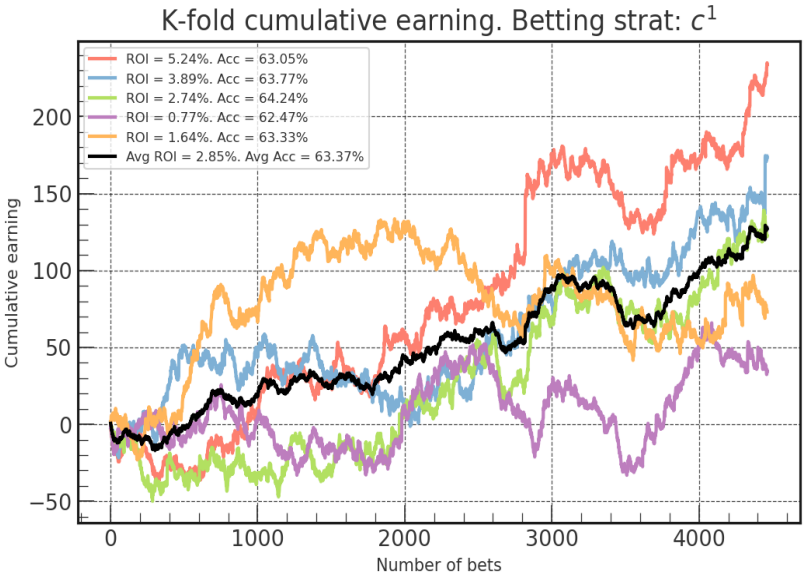
FIFA stats LightGBM accuracy distribution

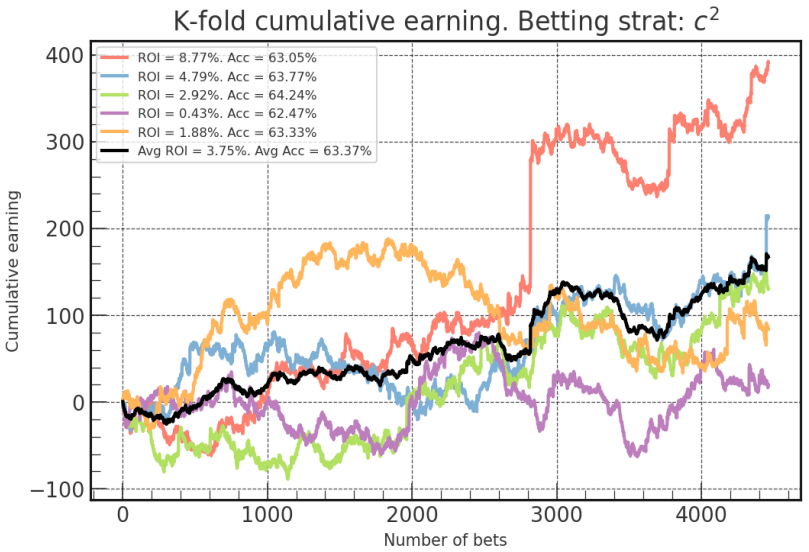


Imputed FIFA stats LightGBM accuracy distribution

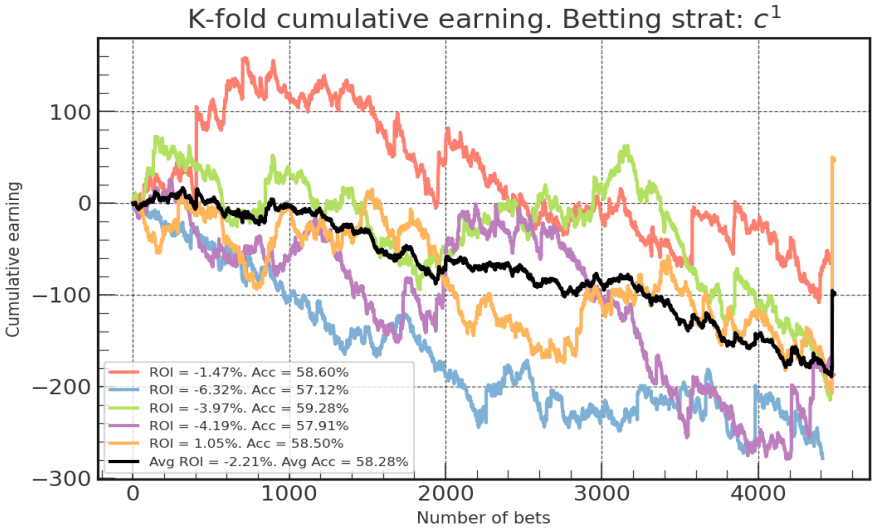


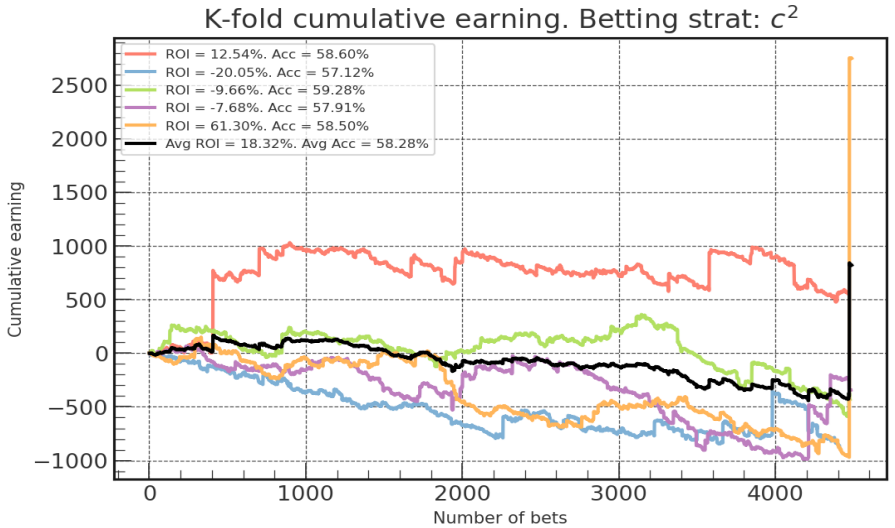
Imputed FIFA stats - Betting strategy





Historical league standings and stats - Betting strategy

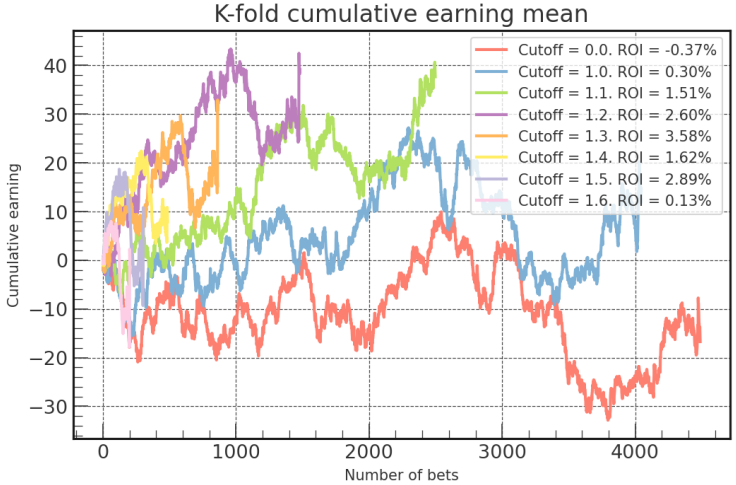


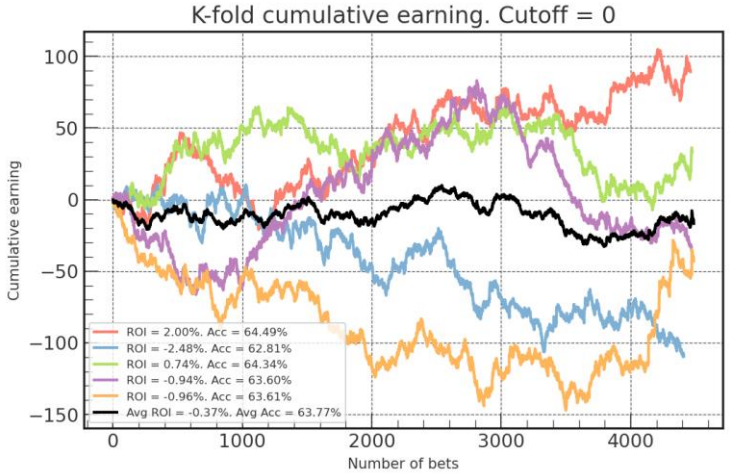


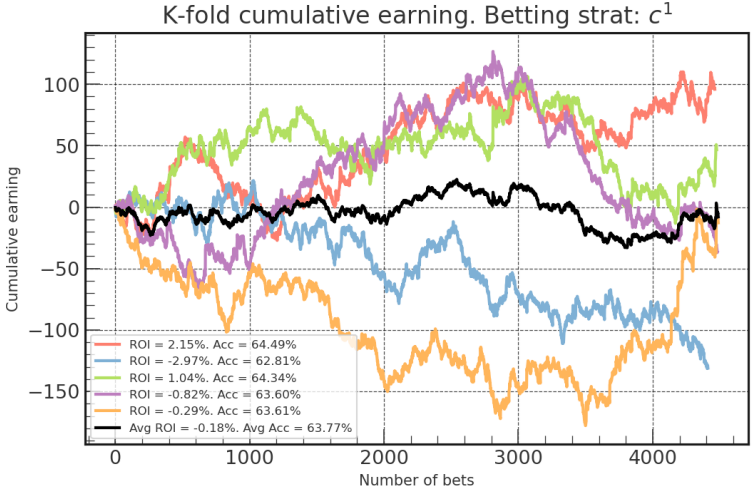
Performance

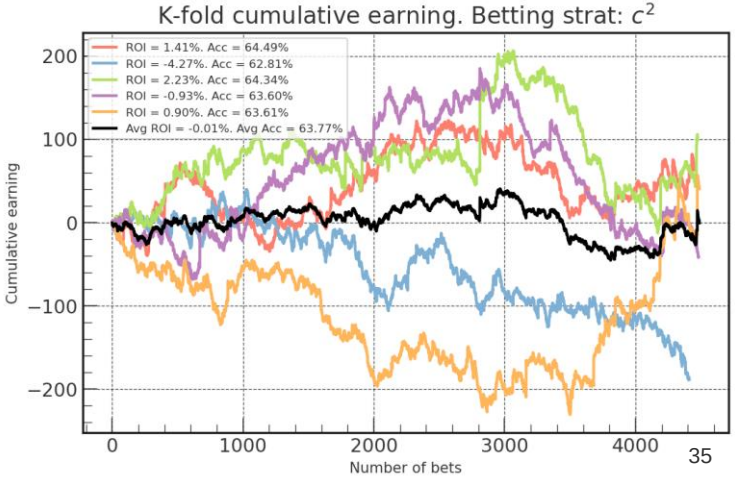
Model: LightGBM

Data: Fifa + Stats





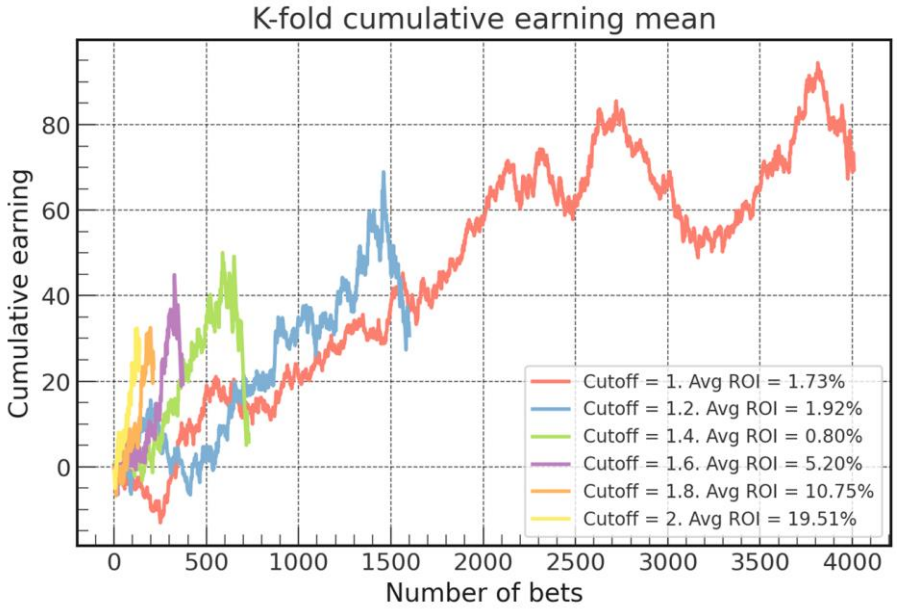


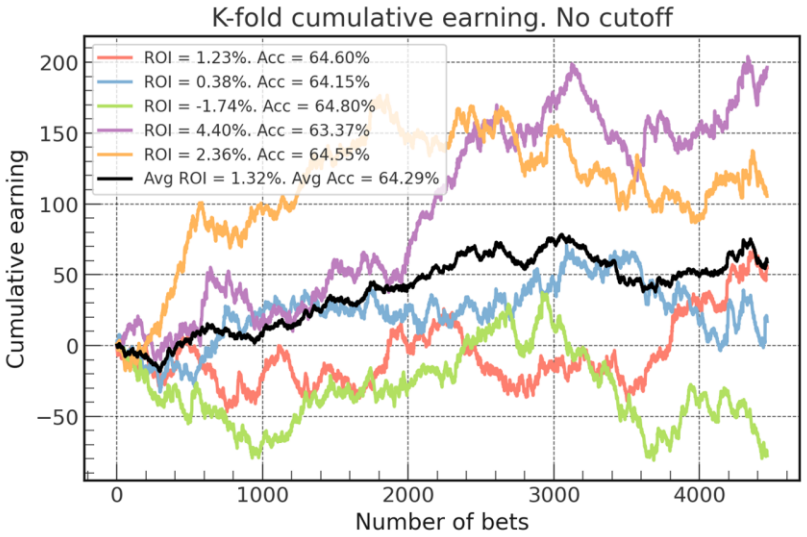


Performance

Model: TensorFlow Feed Forward-NN:

Data: Regressed FIFA

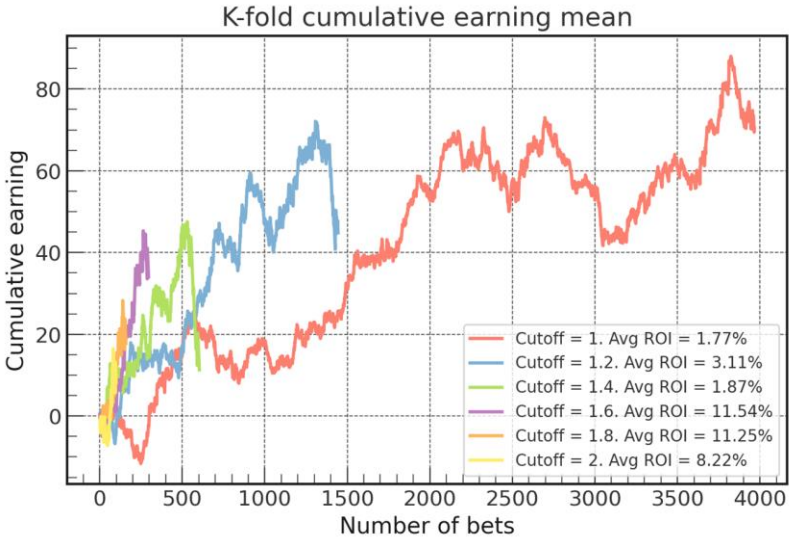


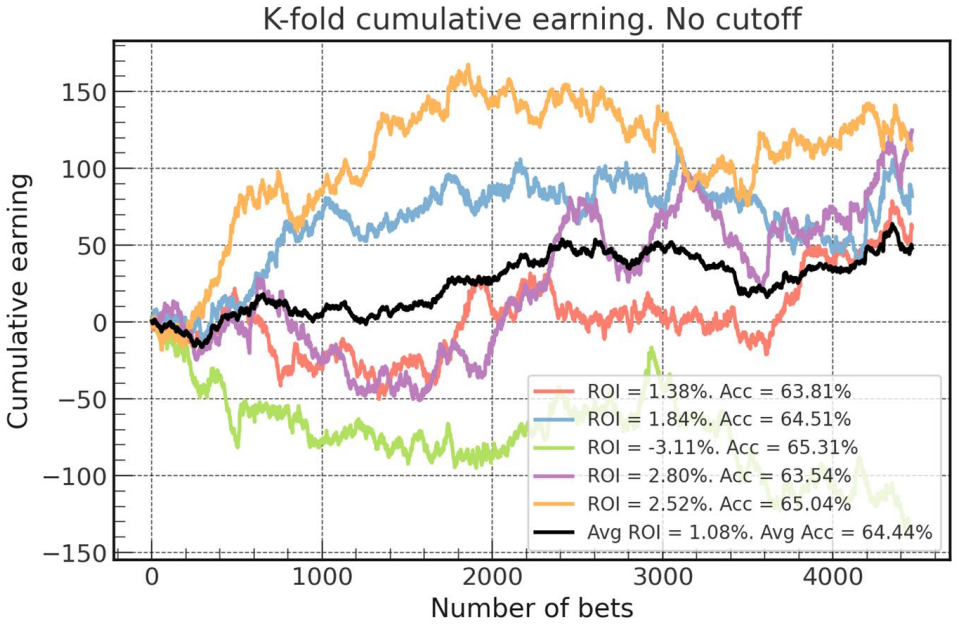


Performance

Model: TensorFlow Feed Forward-NN:

Data: Regressed FIFA

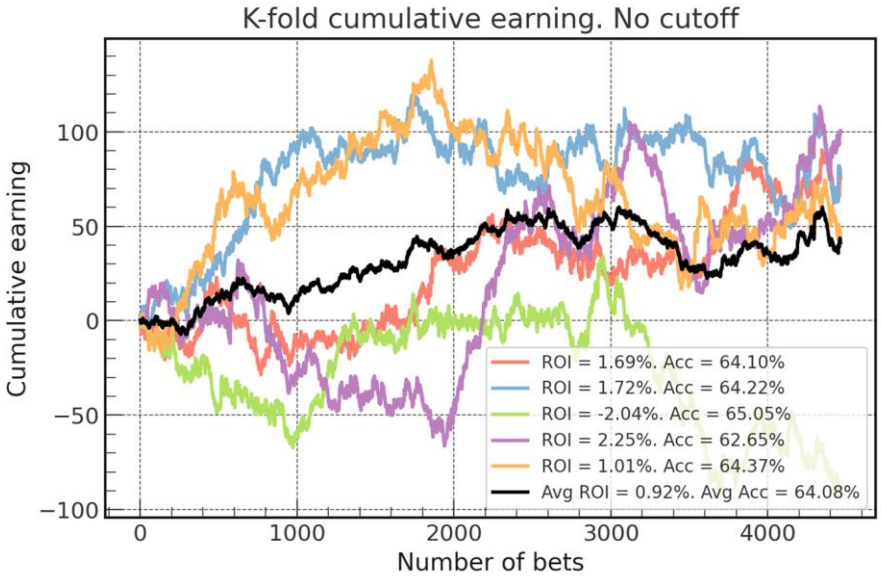


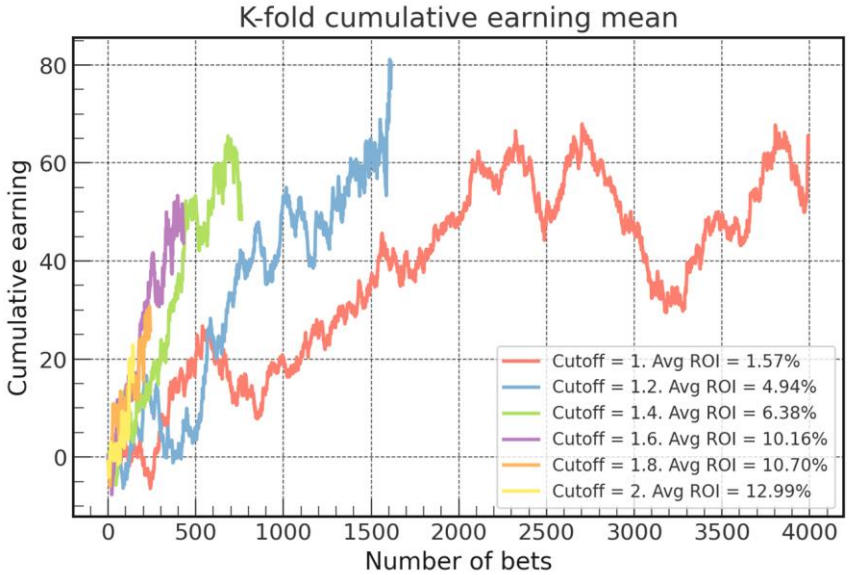


Performance

Model: TensorFlow Feed Forward-NN:

Data: Regressed FIFA





Hyperparams - LightGBM - optimized using Optuna with Bayesian Opt

LightGBM with FIFA ratings	{'bagging_fraction': 0.9579291293788237, 'bagging_freq': 2, 'feature_fraction': 0.6905869538273885, 'lambda_l1': 3.0207950743679935e-06, 'lambda_l2': 2.8471435761366433e-06, 'learning_rate': 0.08070953882212376, 'min_child_samples': 21, 'num_leaves': 60}
LightGBM with Imputed FIFA ratings	{'bagging_fraction': 0.9001065227814082, 'bagging_freq': 4, 'feature_fraction': 0.8571709403386009, 'lambda_l1': 1.3055455944442806e-07, 'lambda_l2': 0.9946641932692142, 'learning_rate': 0.08020792259086053, 'min_child_samples': 24, 'num_leaves': 231}
LightGBM with Historical League standings and stats	{'bagging_fraction': 0.8815773640303611, 'bagging_freq': 1, 'feature_fraction': 0.9331008825258347, 'lambda_l1': 0.00023013094300225074, 'lambda_l2': 9.29984411243807e-08, 'learning_rate': 0.4659936425277095, 'min_child_samples': 32, 'num_leaves': 117}
LightGBM with FIFA ratings + Historical	{'bagging_fraction': 0.7869734846747195, 'bagging_freq': 6, 'feature_fraction': 0.946364174886954, 'lambda_l1': 0.0010554319601229364, 'lambda_l2': 4.278160517318616e-05, 'learning_rate': 0.06269262282530857, 'min_child_samples': 23, 'num_leaves': 119}

Hyperparams - NN - Optimized using Bayesian Optimization

FF-NN with Historical League standings and stats	'batch_size': int(210.93147806917756), 'learning_rate': 0.0001031647889407131, 'num_layers': int(1.6527764683653896), 'num_nodes1': int(795.2791694689053), 'num_nodes2': int(981.1525219447333), 'num_nodes3': int(129.07866106404288)
FF-NN with Imputed FIFA ratings	'batch_size': int(264.05031761942996), 'learning_rate': 0.00090657875624180566, 'num_layers': int(1.5990497723377082), 'num_nodes1': int(1268.3567408393797), 'num_nodes2': int(2049.7155220818804), 'num_nodes3': int(115.84809550410591)
RNN - Historical League standings and stats	'num_layers': int(1.3354462066384936), 'num_nodes': int(81.26932393398478), 'learning_rate': 0.001999564797802341, 'batch_size': int(543.5715311507388), 'dropout_rate1': 0.37090214789269504, 'dropout_rate2': 0.2729329990267305