Predicting handicap result of Soccer using betting odds via Machine Learning

LYU 2005

Sun Ka Ho, 1155098418

Chan Cheong, 1155100189

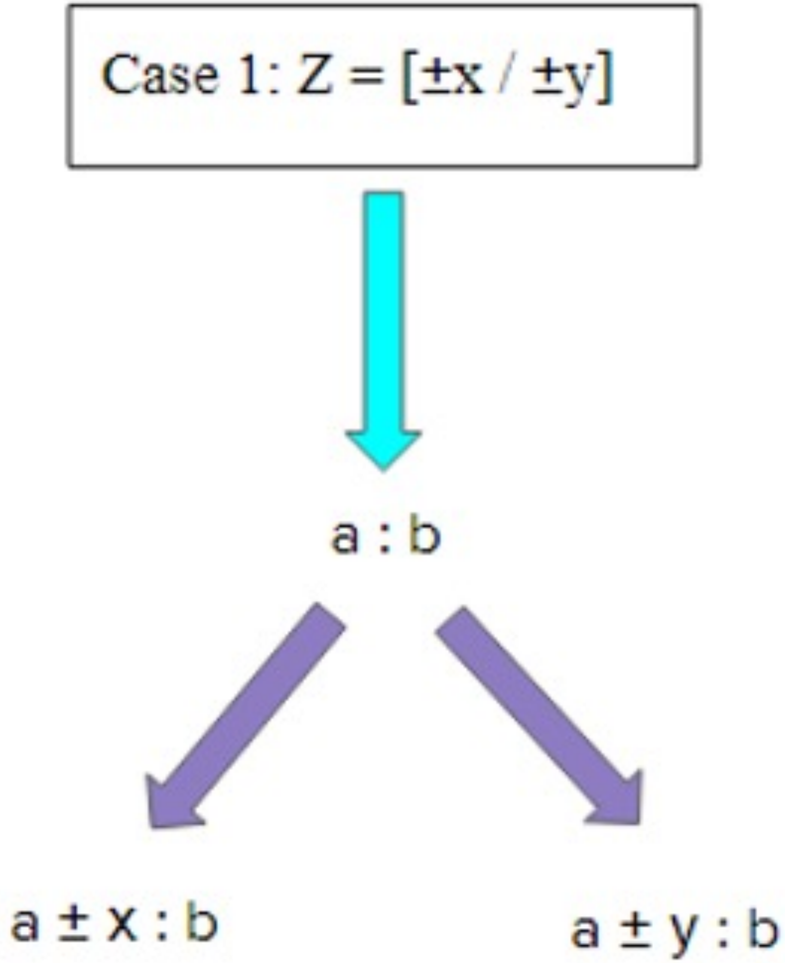
Agenda

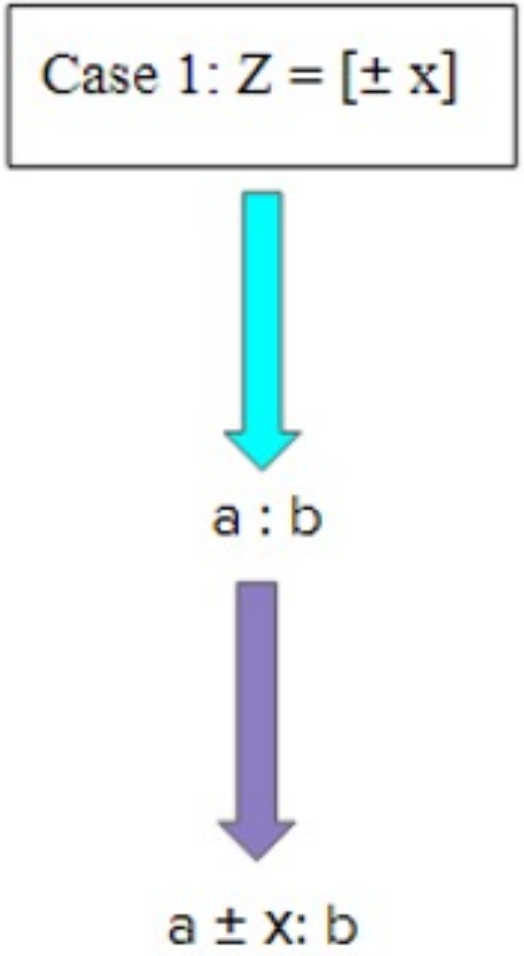
- Objective
- Data Introduction
- Data pre-processing & Feature engineering
- Project workflow – pipeline
- Model Introduction
- Evaluation
- Extra experiments
- Conclusion & Future work

Objective

- Semester 1
  - All useful data will be collected and cleaned.
  - Finalize the model pipeline and choose the baseline model.
  - Different models for comparison will be deployed.
- Overview
  - Profitable model

What is handicap?





What is handicap?

Date	Home Team	Away Team	Match score	Handicap Line (Home Team)
23/11/2020	A	B	3 : 3	0/-0.5
3/11/2020	C	D	1 : 4	+2

Match score	Handicap Line (Home)	Score under handicap
3 : 3	0/-0.5	3 : 3 ; 2.5 : 3
1 : 4	+2	3 : 4

Buy Home	Buy Away
loss half	win half
loss all	win all

We are buying this

Handicap Results

Win all	Win half	Draw	Loss half	Loss all
2	1	0	-1	-2

Dataset

1. Home field indicator
2. Handicap odds
    1. HKJC
    2. Non-HKJC
3. Past 10 records
    1. Home(H)/Away(A): A vs _ , _ vs A , H vs _ , _vs H
    2. Encounter: A vs H , H vs A
4. Statistics of teams
5. Player Score
    - [weight/height/age/value/hit/power/potential_power]
6. FIFA Team Score
    - ATT/MID/DEF/power/rating
7. Lineup
8. Weather
9. Temperature

Feature Engineering - HKJC time relative odds

handicap line

HKJC

#	時間	主勝	盤口	客勝
#7	10-28 02:47	2.02	[0]	1.81
#6	10-27 22:31	2.05	[0]	1.78
#5	10-27 22:28	2.09	[0]	1.75
#4	10-27 22:28	2.07	[0]	1.77
#3	10-27 21:38	2.03	[0]	1.80
#2	10-27 21:12	1.99	[0]	1.83
#1	10-27 11:41	1.96	[0]	1.86

Non-HKJC

Same

克魯	盤口	林肯城	變化時間
0.74	受讓平手/半球	0.96	10-27 21:40
0.90	平手	0.80	10-27 21:39
0.84	平手	0.86	10-27 17:02
0.77	平手	0.93	10-26 21:59

Same

Feature Engineering - Past 10 record

0 : 1
半/一 --> -0.75
-0.75 : 1
Score = 1.75

李斯特城 近 10 場 同客 英超 英聯盃 球會友誼

類型	日期	主場	比分(半場)	角球	客場	Bet365 初盤 主	盤口	客	平均歐賠 終盤 主	和	客	全場 勝負	讓球	大小
英超	18-12-29	李斯特城	0-1(0-0)	10-4	卡迪夫城	0.85	半/一	1.05	1.54	3.92	7.10	負	輸	小
英超	18-12-26	李斯特城	2-1(1-1)	3-7	曼城 1	0.90	*球半	1.00	9.21	5.36	1.33	勝	贏	走
英超	18-12-22	車路士	0-1(0-0)	9-5	李斯特城	1.00	球半	0.90	1.32	5.18	9.92	勝	贏	小
英聯盃	18-12-19	李斯特城	1-1(0-1)	4-7	曼城	1.09	*一球	0.81	7.47	4.63	1.41	平	贏	小
英超	18-12-15	水晶宮	1-0(1-0)	4-4	李斯特城	0.85	平手	1.05	2.69	3.10	2.83	負	輸	小
英超	18-12-09	李斯特城	0-2(0-1)	6-5	熱刺	0.99	*平/半	0.91	4.02	3.48	1.95	負	輸	小
英超	18-12-06	富咸	1-1(1-0)	10-8	李斯特城	1.10	平手	0.80	2.78	3.31	2.59	平	走	小
英超	18-12-01	李斯特城	2-0(2-0)	4-8	屈福特 1	0.97	半球	0.93	2.26	3.26	3.35	勝	贏	小
英聯盃	18-11-28	李斯特城	0-0(0-0)	6-6	修咸頓	1.05	半/一	0.85	1.98	3.41	3.75	平	輸	小
英超	18-11-24	白禮頓	1-1(1-0)	6-1	李斯特城1	0.85	平手	1.05	2.85	3.12	2.64	平	走	小

Data analysis

Is our project workable?

Kruskal-Wallis H Test

	p-value	
	Initiated odd difference	Last odd difference
HKJC	3.27E-08	1.329E-02
Bet365	5.85E-05	1.096E-02
Crown	9.12E-04	3.302E-03
Macau	5.04E-06	1.617E-02

Data analysis

- Handicap result
  - 3 classes (+1, 0, -1)
  - 5 classes (+2, +1, 0, -1, +2)

Pairwise Wilcoxon Rank Sum Test

```
        Pairwise comparisons using Wilcoxon rank sum test

data:  hkjc_diff and football_data$hkjc_hdc_results

   -2      -1      0       1
-1 0.00028 -       -       -
0  0.01833 0.60497 -       -
1  0.60497 0.00139 0.02506 -
2  2.6e-06 0.62821 0.62821 0.00040

P value adjustment method: BH
```

Pipeline

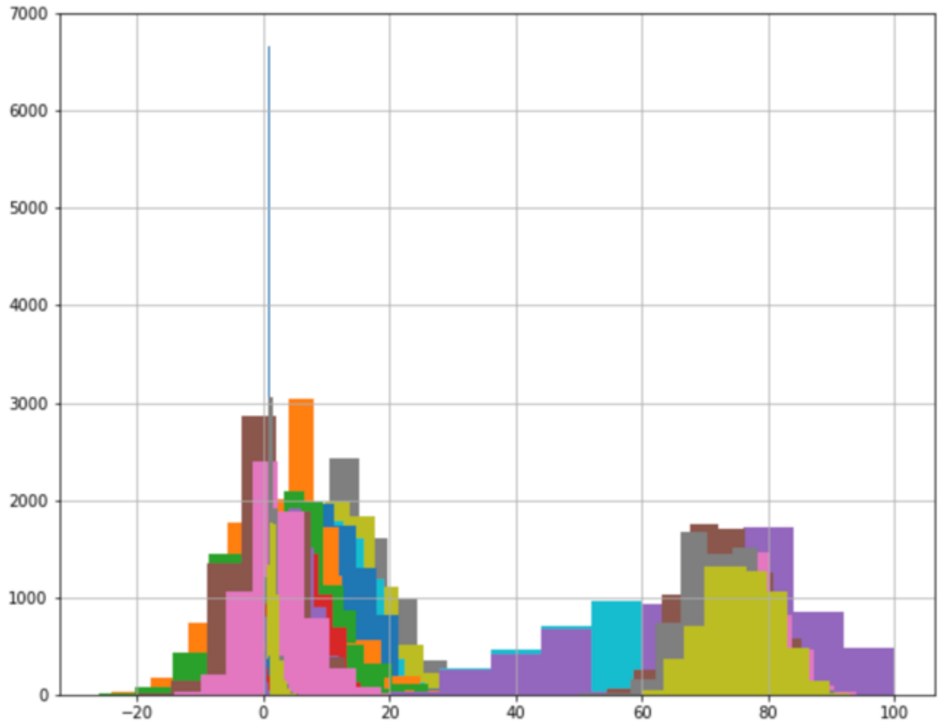

win007
hkjc
FIFA
json
csv
Cleaning
Pre-processing
Modeling
Evaluation

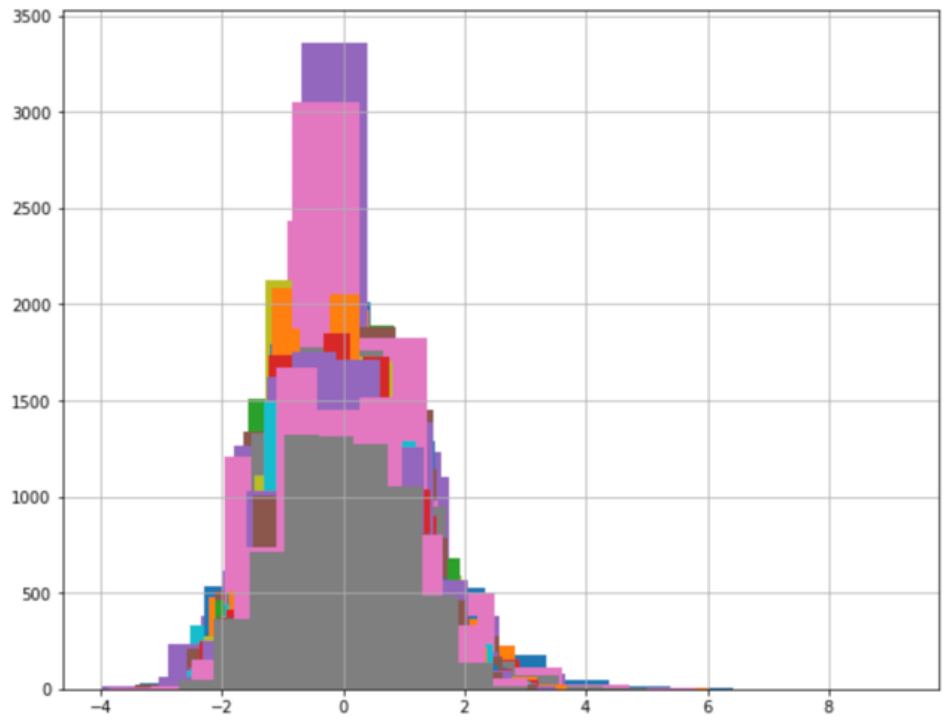
Pre-processing and Modeling

- Standardization
  - Zero mean and standard deviation equal to 1
- Imputation
  - KNN imputer
- Dimension reduction
  - PCA, autoencoders
- Models
  - Four statistical models, three neural network models
- Feature selection
  - AIC/BIC, RFE-SVM, FNN-FS

Standardization

- No standardization on One-hot values





Imputation

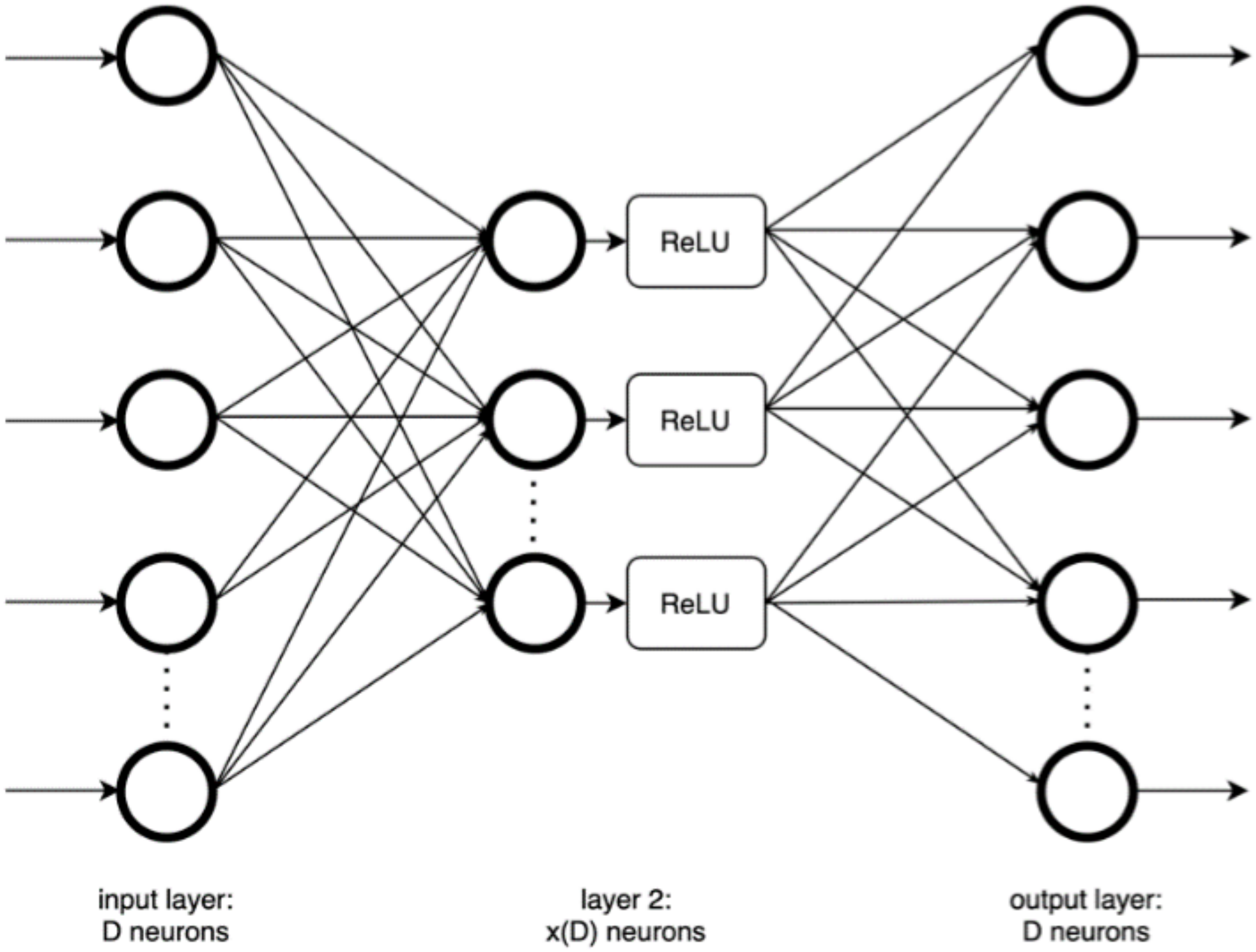
- Fill in the NAN values
- KNN imputer with k = 2
- Mainly on Player Scores and Team Scores

Dimension reduction (PCA)

- Dimension reduction
- Mix odds features and other features
- Proportion of variation (POV) – 0.95 and 0.999
  - Selection features which can explain x% of variance

Dimension reduction (Autoencoder)

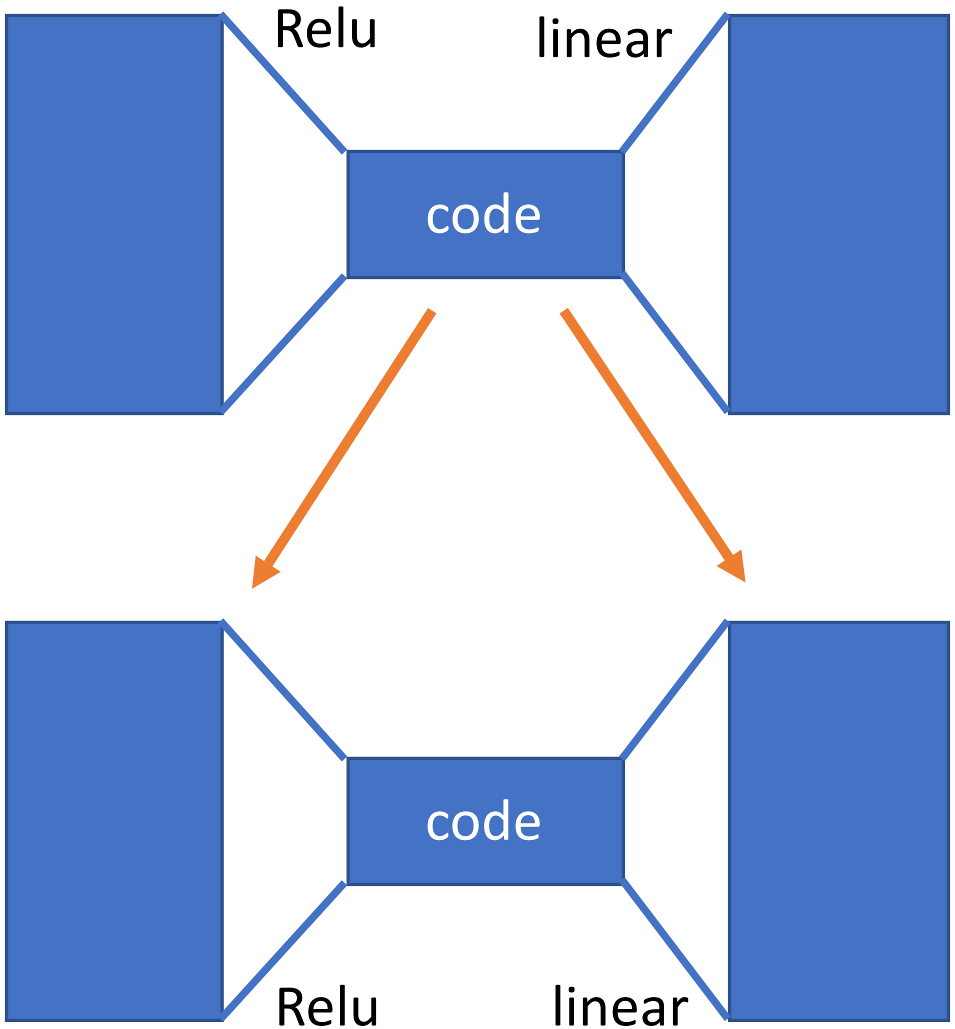
- One hidden layer
- Dynamic hidden layer uses percentage (%)



- Approximate x ≈ f(x)

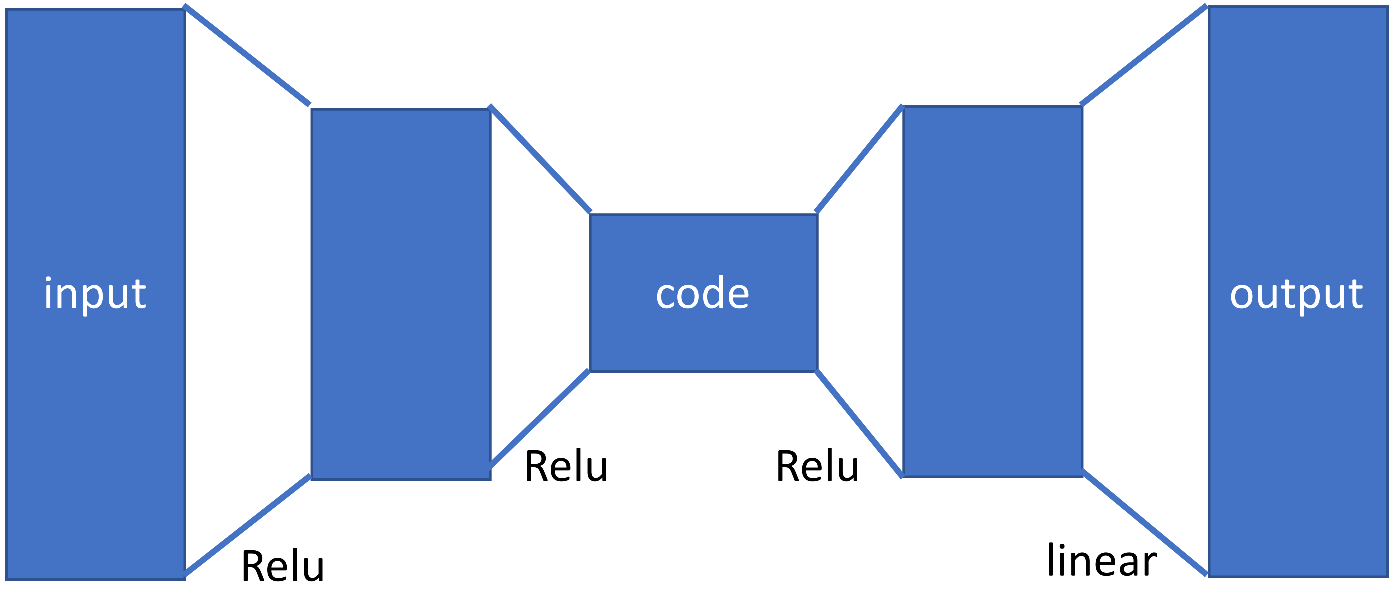
Dimension reduction (Stacked Autoencoder)

- Two hidden layers
- Dynamic hidden layer uses percentage (%)
- Greedy layer-wise pretrain



Dimension reduction (Deep Autoencoder)

- Two hidden layers
- Dynamic hidden layer uses percentage (%)
- Train as deep NN.



Hyperparameter tuning

- 5 folds cross validation
- Choose the best set of features on validation set
- Use it in testing set

20%	20%	20%	20%	20%
Validating	Training	Training	Training	Training
Training	Validating	Training	Training	Training
Training	Training	Validating	Training	Training
Training	Training	Training	Validating	Training
Training	Training	Training	Training	Validating

Benchmark Model

- Odds-based benchmark
  - Always bet on the team that has lower odds
- Expected value
  - $EV = \frac{(invest_{money} * home_{odd} + invest_{money} * away_{odd})}{2}$

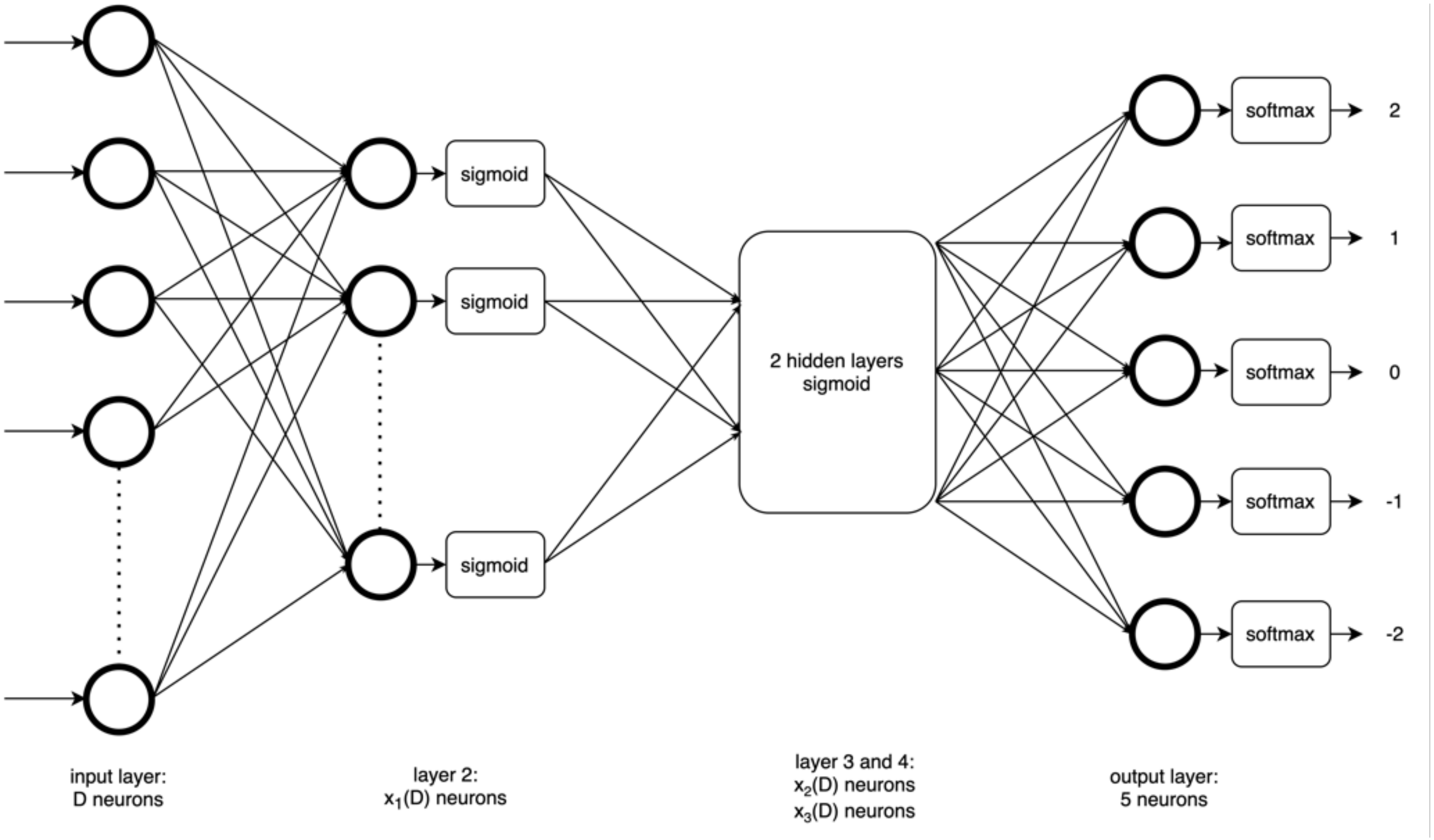
Statistical Model

- Linear Regression
- Logistic regression
- Random Forest
- K-nearest neighbour

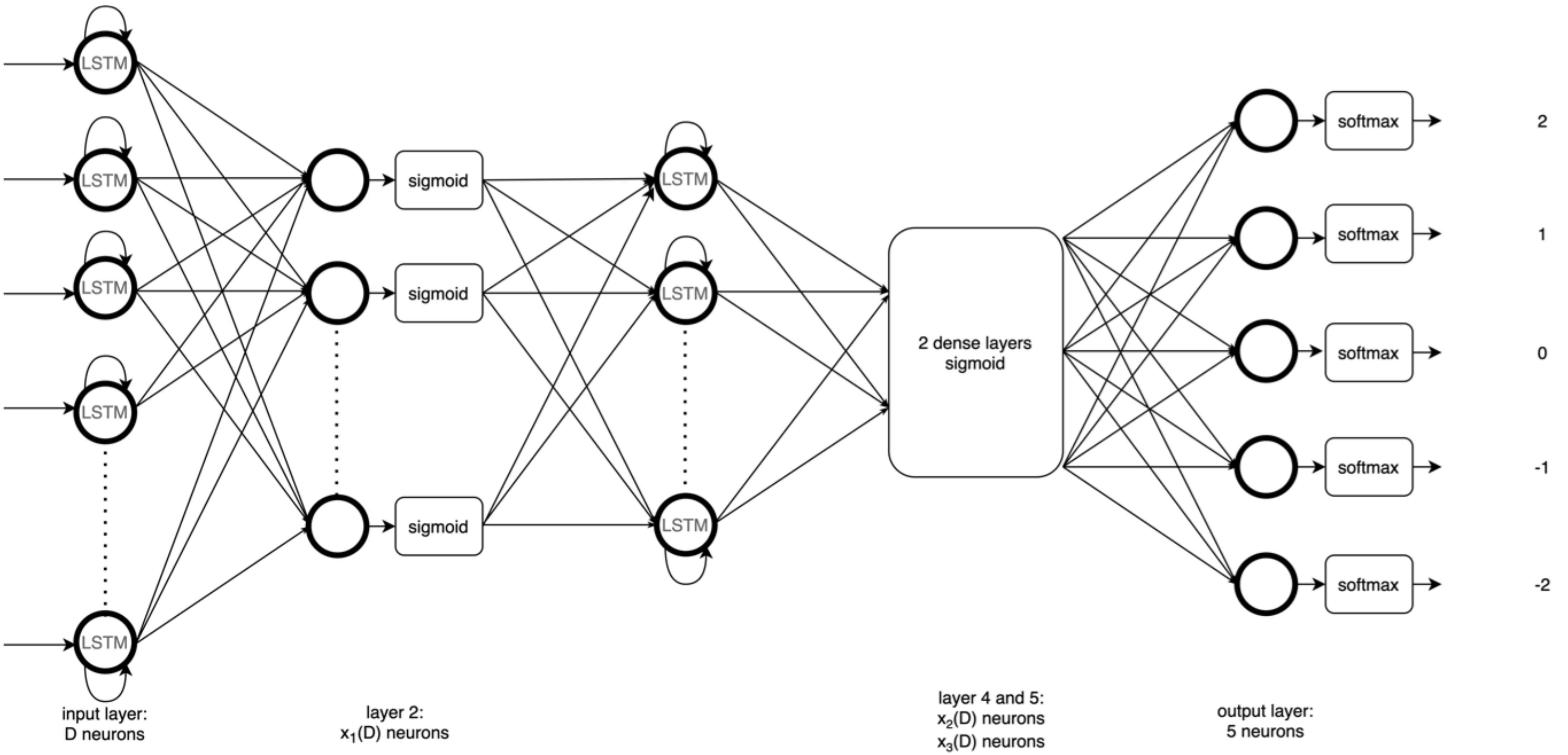
Neural Network Model

- Feedforward Neural Network (FNN)
- Long short-term memory (LSTM)
- Autoencoder (Supervised tuning) – (Sencoder)

FNN

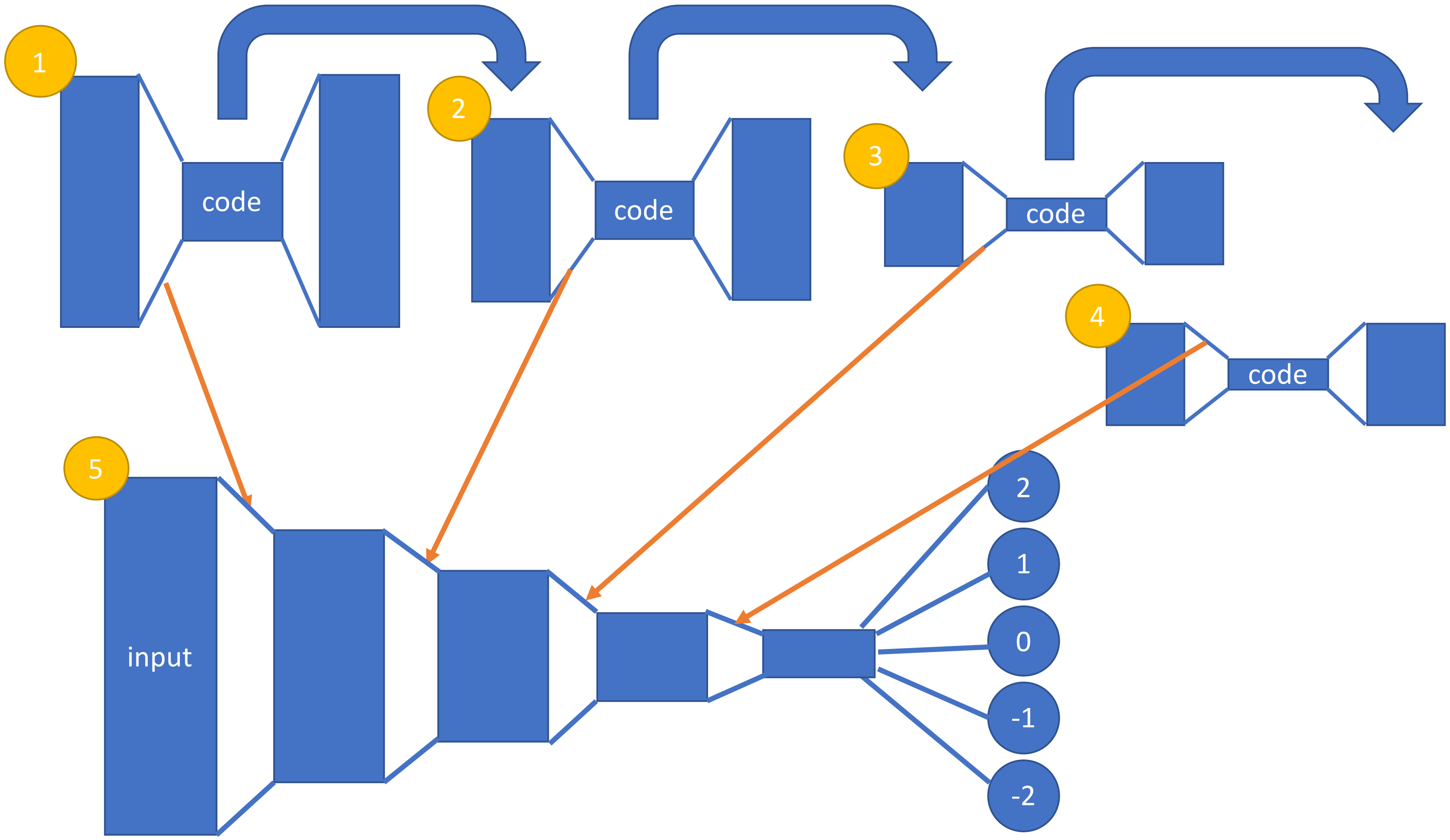

sigmoid
sigmoid
sigmoid
2 hidden layers
sigmoid
softmax
softmax
softmax
softmax
softmax
2
1
0
-1
-2
input layer:
D neurons
layer 2:
x_1(D) neurons
layer 3 and 4:
x_2(D) neurons
x_3(D) neurons
output layer:
5 neurons

LSTM



LSTM
LSTM
LSTM
LSTM
LSTM
sigmoid
sigmoid
sigmoid
LSTM
LSTM
LSTM
2 dense layers
sigmoid
softmax
softmax
softmax
softmax
softmax
2
1
0
-1
-2
input layer:
D neurons
layer 2:
x_1(D) neurons
layer 4 and 5:
x_2(D) neurons
x_3(D) neurons
output layer:
5 neurons

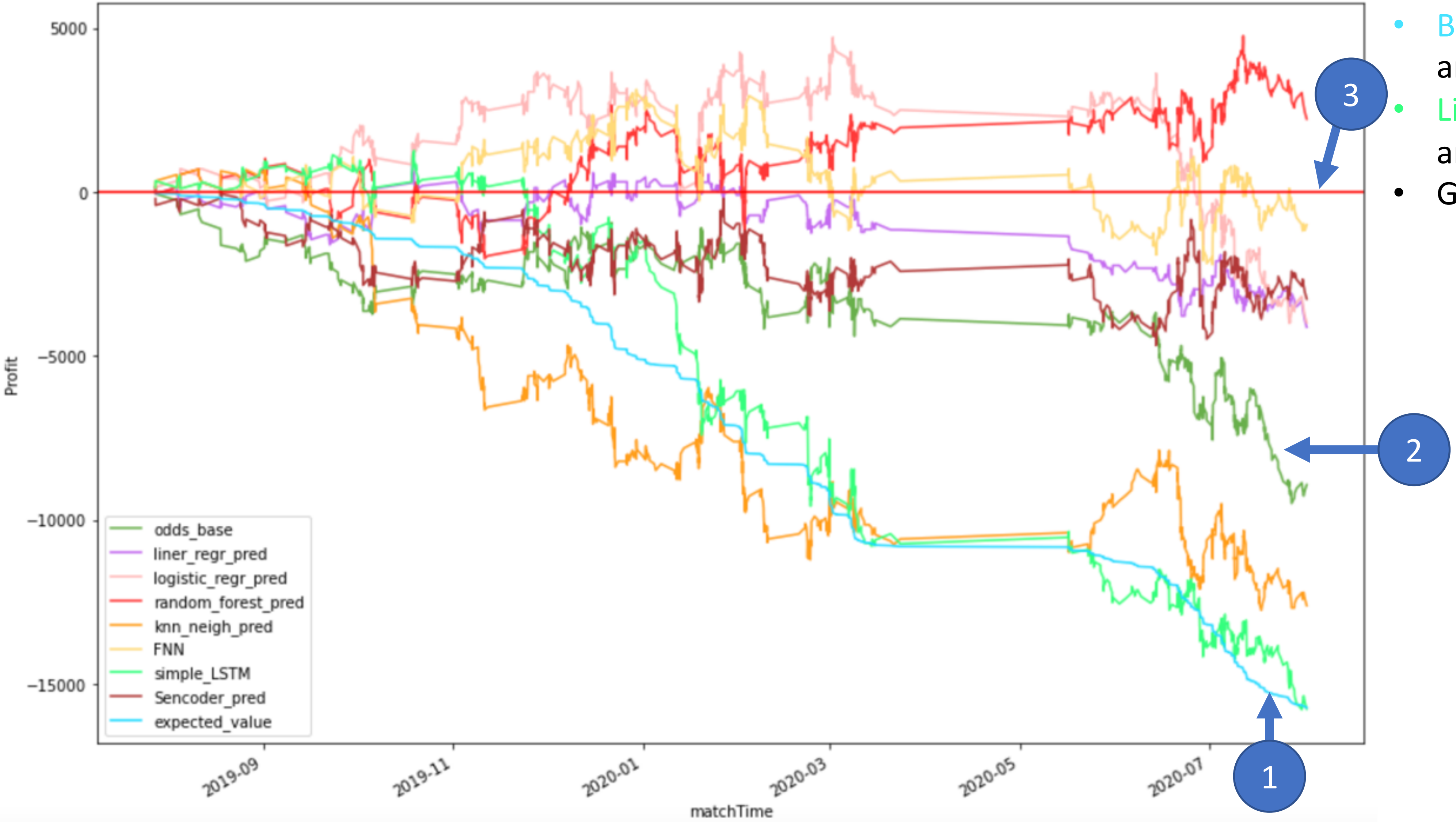
Autoencoder (Supervised tuning)



Evaluation

- Dataset contains 2017 to 2020-08, 5277 records
- Train on first 80% of data, test on the rest
- Invest $200 on each match
- Plot the cumulative profit over time

Evaluation (Example)



- Blue and Dark-Green are benchmark models
- Light-Green and Yellow are NN models
- Goals:
    1. Bit Blue curve
    2. Bit Dark-Green
    3. Positive return

Three approaches on training

- By-league
- All-league
- Cluster-then-Predict
- For By-league and Cluster-then-Predict:
  - Train on each subset, evaluate on them, plot the graph in one plot

By-league

- Odds in each league is different
- Separating them helps modeling
- Only select league with > 200 records

```
{'挪超': 215,
 '日職聯': 232,
 '德乙': 472,
 '美職業': 253,
 '法甲': 383,
 '英冠': 463,
 '歐冠盃': 214,
 '西甲': 691,
 '德甲': 520,
 '英超': 738,
 '荷甲': 250,
 '意甲': 594,
 '澳洲甲': 252}
```

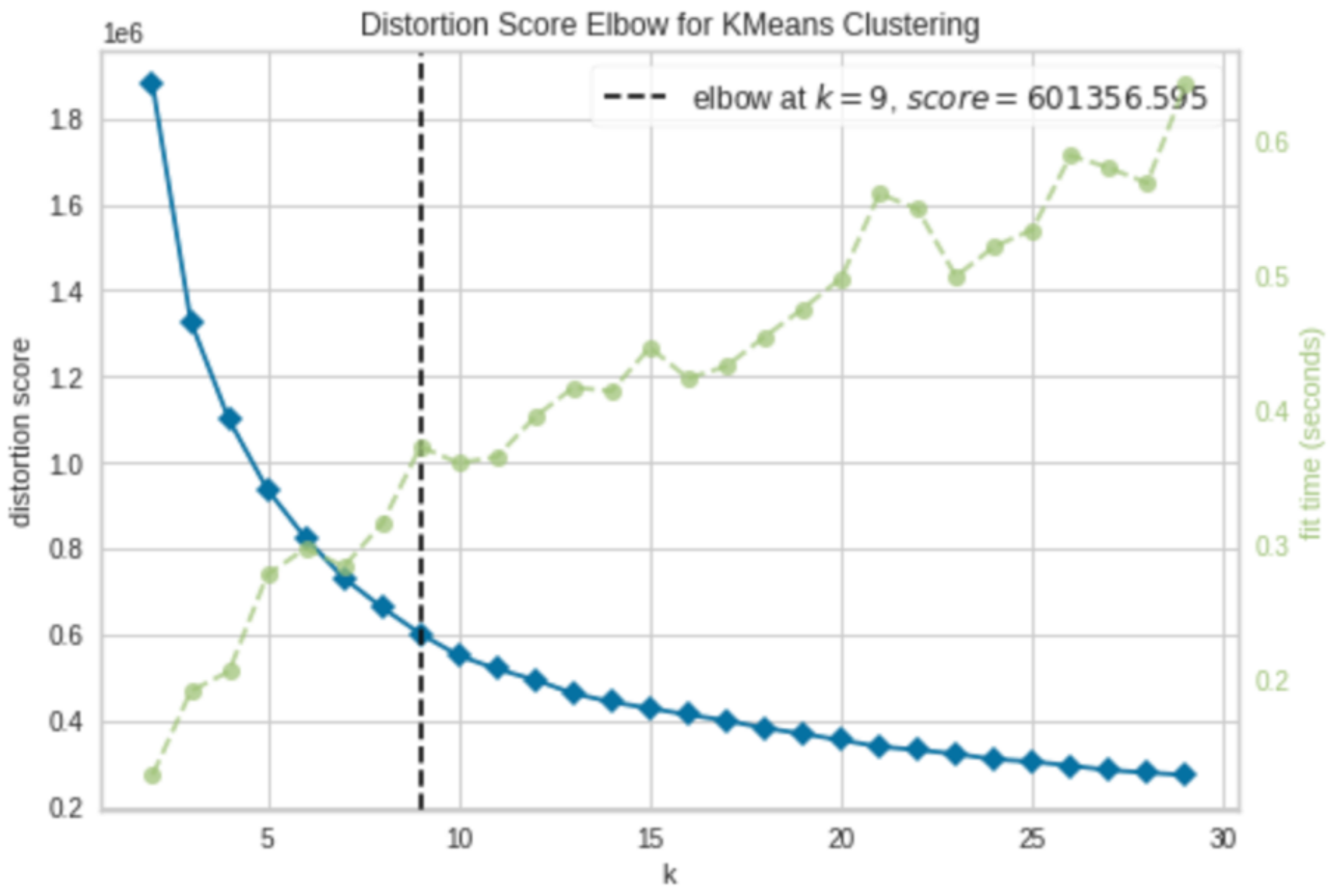
All-league

- Train all the data at once
- Only select league with > 200 records

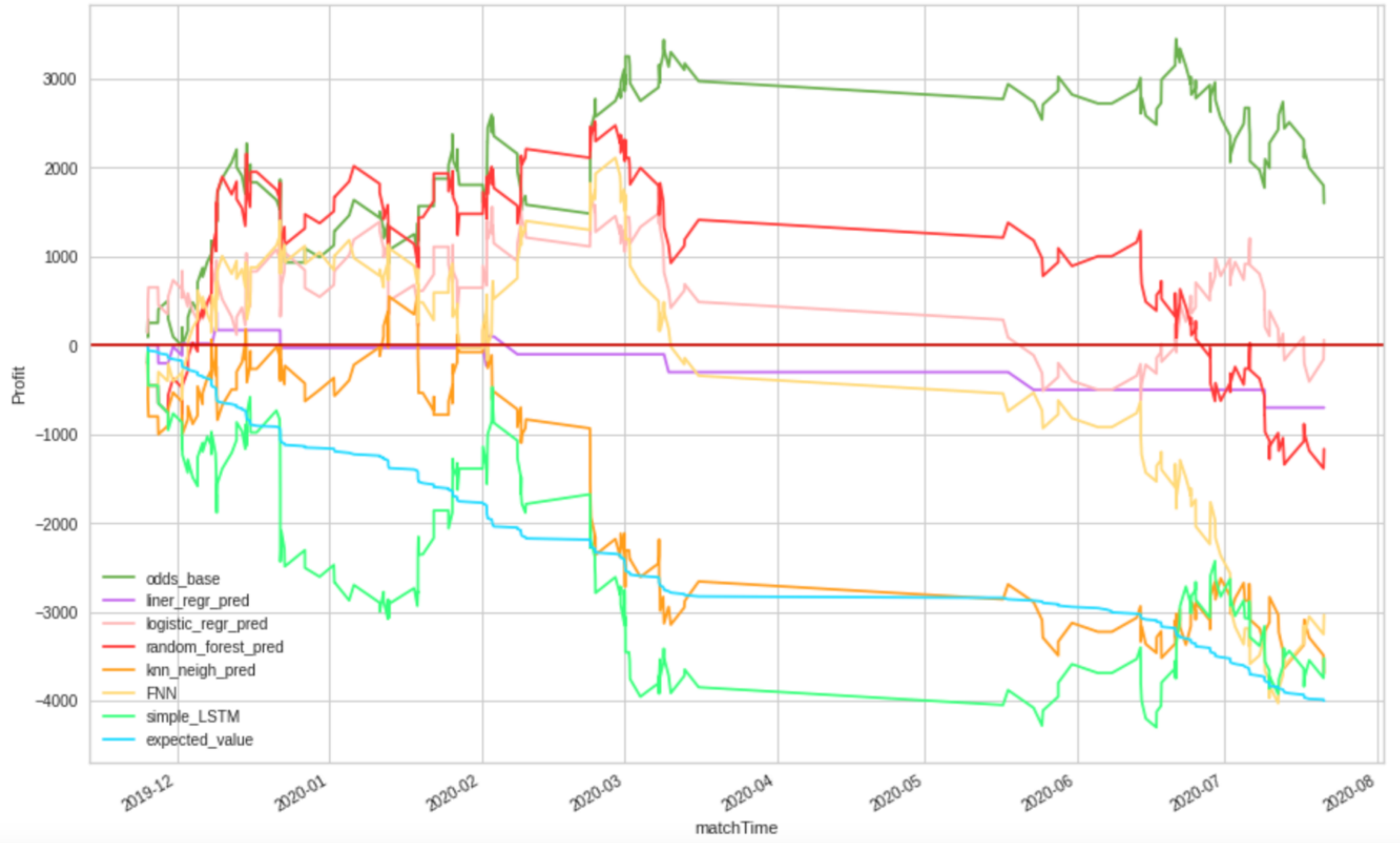
Cluster-then-Predict

- Kmeans clustering
- Modeling on each cluster
- Elbow Method to select k

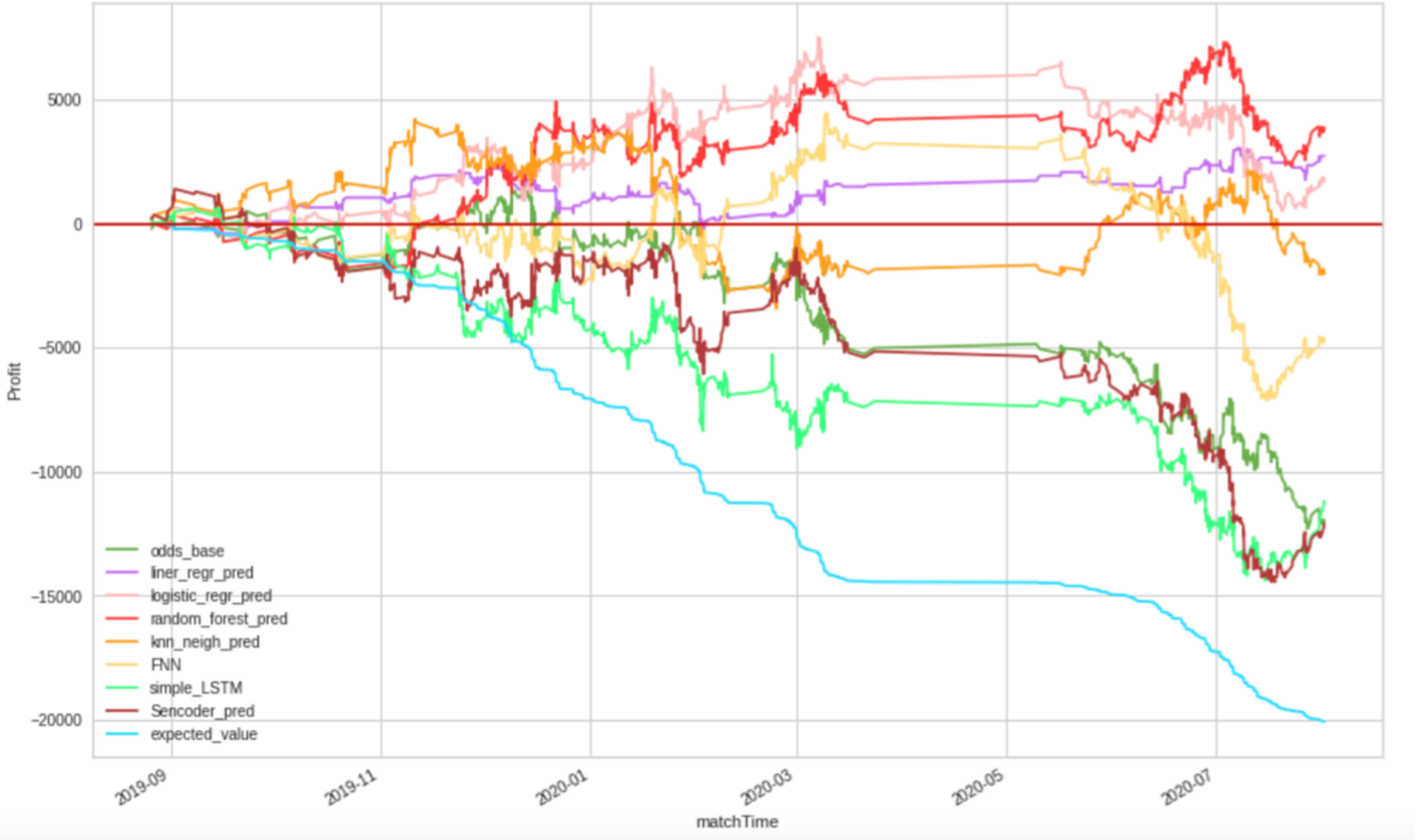
Cluster	Number of Records
1	1373
2	314
3	952
4	464
5	830
6	594
7	729
8	872
9	607



Cluster-then-Predict, the first cluster



Cluster-then-Predict, overall



Best Model

- All-league
- Reasons
  - Quite good performance of statistical models
  - Best average performance of FNN and LSTM

	Expected Value	Odds based	Linear Regr	Logistic Regr	Random forest	KNN	FNN	LSTM	Sencoder
Cluster-then-Predict	-20000	-12000	2500	1500	3750	-2000	-4950	-11000	-12000
by-league	-16000	-9000	-4250	-4250	2250	-12500	-750	-16000	-3000
all-league	-16000	-6250	-50	1250	100	-4500	-2000	-12500	-14000

Dimension Reduction

- 0.999POV better than 0.95 POV
- Deep autoencoder better than stacked autoencoder

(all-league)	Expected Value	Odds based	Linear Regr	Logistic Regr	Random forest	KNN	FNN	LSTM	AVG
Base	-16000	-6250	-50	1250	100	-4500	-2000	-12500	-2950
PCA, POV=0.95	-16000	-6250	0	1250	-1000	-10000	-5000	-3500	-3042
PCA, POV=0.999	-16000	-6250	-50	2000	-5750	-4500	-1300	100	-1583
Stacked Autoencoder	-16000	-6250	-50	-1000	2000	-10000	-100	-6000	-2525
Deep Autoencoer	-16000	-6250	-50	-1800	2000	-2000	-14000	4000	-1975
Autoencoder	-16000	-6250	0	-100	-3000	-6000	1200	-5100	-2167

Feature Selection

```
Index(['home_away_label', 'hkjc_hdc_home_first', 'hkjc_hdc_away_first',
       'hkjc_hdc_home_last', 'hkjc_hdc_away_last', 'bet365_hdc_hkjcFirst_home',
       'bet365_hdc_hkjcFirst_away', 'crown_hdc_hkjcFirst_home',
       'crown_hdc_hkjcFirst_away', 'macau_hdc_hkjcFirst_home',
       'macau_hdc_hkjcFirst_away', 'bet365_hdc_hkjcLast_home',
       'bet365_hdc_hkjcLast_away', 'crown_hdc_hkjcLast_home',
       'crown_hdc_hkjcLast_away', 'macau_hdc_hkjcLast_home',
       'macau_hdc_hkjcLast_away', '總進球_away', '總進球_home', '總失球_away',
       '總失球_home', '淨勝球_away', '淨勝球_home', '場均進球_away', '場均進球_home', '勝率_away',
       '勝率_home', '平率_away', '平率_home', '负率_away', '负率_home', '同主客進_away',
       '同主客進_home', '同主客失_away', '同主客失_home', '同主客淨勝_away', '同主客淨勝_home',
       '同主客均進_away', '同主客均進_home', '同主客勝_away', '同主客勝_home', '同主客平_away',
       '同主客平_home', '同主客负_away', '同主客负_home', 'fifa_team_home_score_ATT',
       'fifa_team_home_score_MID', 'fifa_team_home_score_DEF',
       'fifa_team_home_score_能力', 'fifa_team_home_score_球隊評分',
       'fifa_team_away_score_ATT', 'fifa_team_away_score_MID',
       'fifa_team_away_score_DEF', 'fifa_team_away_score_能力',
       'fifa_team_away_score_球隊評分', 'home_player_power_mean',
       'home_player_hidden_power_mean', 'away_player_power_mean',
       'away_player_hidden_power_mean'],
      dtype='object')

Number of features: 59
```

Feature Selection – AIC

- Forward selection
- La Liga League

Linear regression

```
Step:  AIC=161.3
fd$hkjc_hdc_results ~ fd$home_full_勝率_總 + fd$away_full_失_近6 +
    fd$fifa_team_home_score_ATT + fd$homeVsAwayPassTenRecord_converted +
    fd$crown_hdc_home_first + fd$同主客淨勝_home + fd$bet365_hdc_initialFirst_away +
    fd$bet365_hdc_initialFirst_home

                              Df Sum of Sq    RSS    AIC
<none>                                     390.52 161.30
```

Logistic regression

```
Step:  AIC=412.21
fd$hkjc_hdc_results ~ fd$away_full_勝率_客 + fd$home_full_負_總 +
    fd$home_full_勝率_總 + fd$away_full_平_近6 + fd$away_full_得分_近6 +
    fd$同主客均進_home
```

Feature Selection – AIC

- Backward selection
- La Liga

Linear regression

Logistic regression

```
fd$hkjc_hdc_results ~ fd$fifa_team_home_score_ATT + fd$fifa_team_home_score_MID +
    fd$fifa_team_home_score_DEF + fd$fifa_team_home_score_能力 +
    fd$fifa_team_home_score_球隊評分 + fd$fifa_team_away_score_ATT +
    fd$fifa_team_away_score_MID + fd$fifa_team_away_score_能力 +
    fd$hkjc_hdc_home_first + fd$hkjc_hdc_away_first + fd$hkjc_hdc_home_last +
    fd$hkjc_hdc_away_last + fd$bet365_hdc_home_first + fd$bet365_hdc_away_first +
    fd$bet365_hdc_away_last + fd$bet365_hdc_initialFirst_away +
    fd$bet365_hdc_initialLast_home + fd$bet365_hdc_initialLast_away +
    fd$crown_hdc_home_first + fd$crown_hdc_away_first + fd$crown_hdc_home_last +
    fd$crown_hdc_away_last + fd$crown_hdc_initialFirst_away +
    fd$crown_hdc_initialLast_home + fd$crown_hdc_initialLast_away +
    fd$macau_hdc_initialFirst_home + fd$bet365_hdc_hkjcFirst_home +
    fd$bet365_hdc_hkjcFirst_away + fd$crown_hdc_hkjcFirst_home +
    fd$crown_hdc_hkjcFirst_away + fd$bet365_hdc_hkjcLast_away +
    fd$crown_hdc_hkjcLast_home + fd$crown_hdc_hkjcLast_away +
    fd$macau_hdc_hkjcLast_home + fd$homePassTenRecord_converted +
    fd$總進球_away + fd$總進球_home + fd$總失球_home + fd$場均進球_away +
    fd$場均進球_home + fd$勝率_away + fd$勝率_home + fd$平率_away +
    fd$同主客進_away + fd$同主客進_home + fd$同主客失_away +
    fd$同主客失_home + fd$同主客均進_away + fd$同主客均進_home +
    fd$同主客勝_away + fd$同主客勝_home + fd$同主客平_away +
    fd$同主客負_away + fd$home_full_賽_主
    fd$home_player_weight_mean + fd$away_player_weight_mean +
    fd$home_player_value_mean + fd$home_player_age_mean + fd$away_player_age_mean +
    fd$home_player_hit_mean + fd$away_player_hit_mean + fd$home_player_power_mean +
    fd$away_player_power_mean + fd$away_player_hidden_power_mean
```

Feature Selection – RFE

- Backward selection using SVM
- Eliminate the feature that has lowest importance each time
- Grid search on “number of feature to retain”
- Only odds-related features retained

```
Index(['home_away_label', 'hkjc_hdc_home_first', 'hkjc_hdc_away_first',
       'hkjc_hdc_home_last', 'hkjc_hdc_away_last', 'bet365_hdc_hkjcFirst_home',
       'macau_hdc_hkjcFirst_home', 'bet365_hdc_hkjcLast_home',
       'bet365_hdc_hkjcLast_away', 'crown_hdc_hkjcLast_home',
       'macau_hdc_hkjcLast_home', 'macau_hdc_hkjcLast_away'],
      dtype='object')

Number of features: 12
```

Feature Selection – FNNFS

- Forward selection using our FNN
- Select up till 40 features

```
array(['home_away_label', 'hkjc_hdc_home_first', '平率_home',
       'bet365_hdc_hkjcLast_away', '場均進球_away',
       'away_player_hidden_power_mean', '同主客平_away', '负率_away',
       'home_player_hidden_power_mean', '同主客進_away', '同主客失_home',
       '總進球_home', 'fifa_team_home_score_MID', 'crown_hdc_hkjcLast_away',
       'hkjc_hdc_away_last', '平率_away', 'bet365_hdc_hkjcFirst_home',
       'fifa_team_away_score_球隊評分', 'fifa_team_home_score_能力',
       '場均進球_home', 'hkjc_hdc_home_last', '勝率_home', '總失球_away',
       'bet365_hdc_hkjcLast_home', '同主客均進_away',
       'crown_hdc_hkjcFirst_away', '同主客均進_home', '總失球_home', '勝率_away',
       '同主客失_away', '同主客勝_home', 'bet365_hdc_hkjcFirst_away',
       'away_player_power_mean', '同主客负_away', 'hkjc_hdc_away_first',
       'macau_hdc_hkjcLast_away', '同主客淨勝_home',
       'fifa_team_away_score_ATT', '同主客平_home', '淨勝球_home'], dtype='<U29')
```

Feature Selection

- Deep autoencoder better than stacked autoencoder

(all-league)	Expected Value	Odds based	Linear Regr	Logistic Regr	Random forest	KNN	FNN	LSTM	AVG
Base	-16000	-6250	-50	1250	100	-4500	-2000	-12500	-2950
Stacked RFE	-16000	-6250	0	-4500	500	2000	-500	-4000	-1083
Deep RFE	-16000	-6250	0	-100	2000	-100	-200	-10	265
Stacked FNNFS	-16000	-6250	0	-3000	-1500	-6300	-10	-13000	-3968
Deep FNNFS	-16000	-6250	0	750	3000	-800	-8000	2500	-425

Odds-only features

	Expected Value	Odds based	Linear Regr	Logistic Regr	Random forest	KNN	FNN	LSTM	Sencoder	AVG
by-league	-16000	-9000	-4250	-4250	2250	-12500	-750	-16000	-3000	-5500
by-league (odds only)	-16000	-9000	2000	-2200	-9000	-2700	2500	6300	-3000	-871
all-league	-16000	-6250	-50	1250	100	-4500	-2000	-12500	-14000	-4529
all-league (odds only)	-16000	-6250	-100	1250	-4900	-5100	2000	3000	-14500	-2621

Conclusion & Future Work

- Workable
  - Proved and 5 classes for handicap result
- Best features
  - Odds-related features
- Best model
  - NN models
- Future Work
  - Focus on deep learning
  - Follow up on Cluster-then-Predict
  - Betting strategy
  - Real-time prediction program

Q&A

Appendix

- These Pages are only for Q&A

Imputation

- KNN imputer with k = 2
- Mainly on Player Scores and Team Scores

Group by

H/A Team	League
Liverpool	EPL
Liverpool	EPL

Home player scores
int
int

Away player scores
int
int

Home team scores
int
int

Away team scores
int
int

Player scores example

H player power	H player pot. power
nan	1
4	10
6	1
nan	3

KNN imputer →

H player power	H player pot. power
6	1
4	10
6	1
5.56	3

KNN Imputation, n = 2

A	B
nan	1
4	10
6	1
nan	3

$$\sum_i \sum_j Euclidean\_distance(row_i, row_j)$$

R1	R2	R3	R4
0	11.29	0	2.51
11.29	0	11.29	8.8
0	11.29	0	2.51
2.51	8.8	2.51	0

Example
Fill in Row4, ColA that nan
Choose the closest two point(without nan)
4* (1 - 8.8/(8.8+2.51))+
6* (1 - 2.51/(8.8+2.51))
Output = 5.56

A	B
6	1
4	10
6	1
5.56	3