LEARNING TO PREDICTING RACING RESULTS

MSc Computer Science Dissertation
Oral Examination

Presented by
Ryan K.H. HO

Supervisor
Dr. Dirk Schnieders

OBJECTIVE

- To predict horse racing results for Hong Kong horse racing market using machine learning models.
- Search for important features
- Apply multiple models
- Develop a profitable and consistent betting system

PRIOR WORK

- Overview of Racing system and betting market
- Established the framework of developing the betting system
- Data collection from RaceMate and Initial analysis
- Introduced 2 Categories of Prediction models
  - Finishing Time Regressor
  - Discrete Choice Models
- Trained the some of the models and calculated prediction accuracy

NOTES

- Focus on Win/ Place betting due to data availability
- Pari-Mutuel Pools
  - Dividends will be shared by the winners after house-take of 17.5%
  - Final odds won't be available until the race starts!

PROJECT FRAMEWORK

Data Visualization → Variables Construction → Model Training → Back-Test Wagering

USEFUL FEATURES

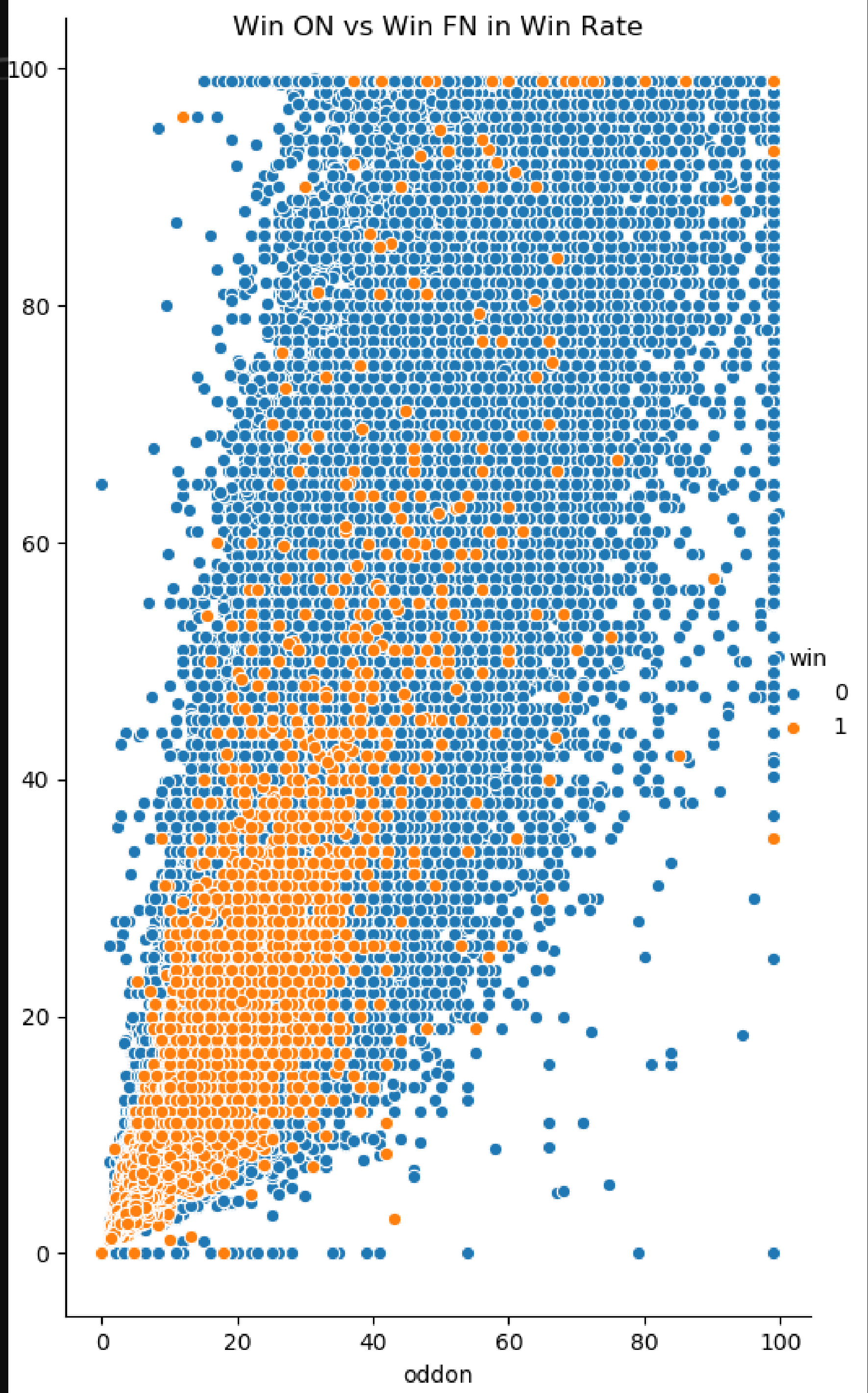
Directly sourced from the RaceMate Dataset. They are well understood and standard measures for the races

Fundamental:

- Horse Rating
- Age
- Draw
- Class Change
- Loading

Technical (market intelligence):

- Overnight odds
- Before Race odds



CREATED FEATURES

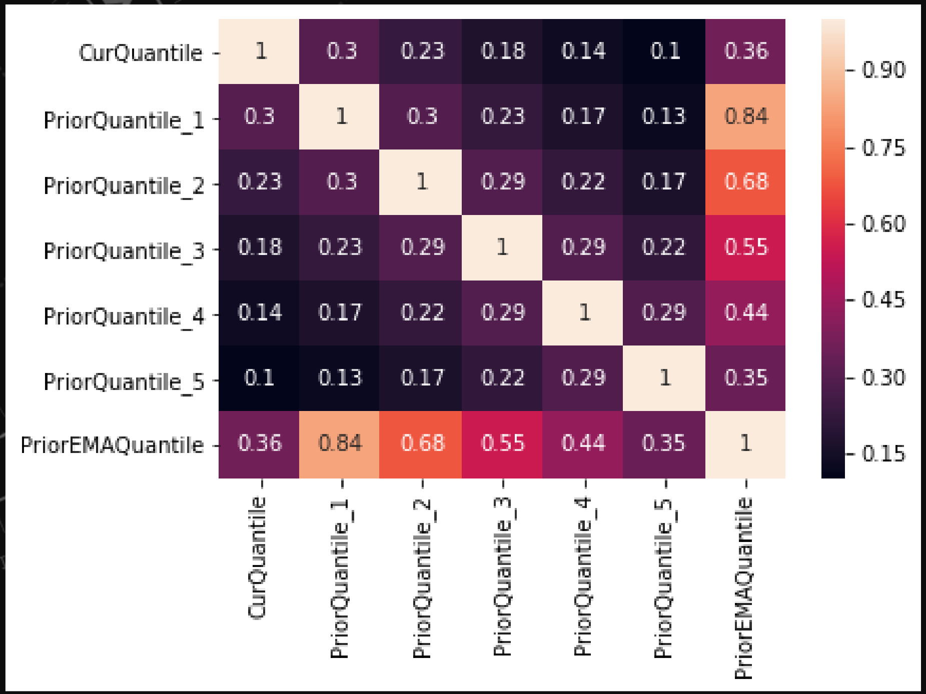
From the literature or supported by visualization, we have created the below features as inputs to the model

Fundamental:

- EMA_Quantile
- Jockey Win Rate
- New Distance Running
- Weight Difference

Technical (market intelligence):

- Odds implied winning probability

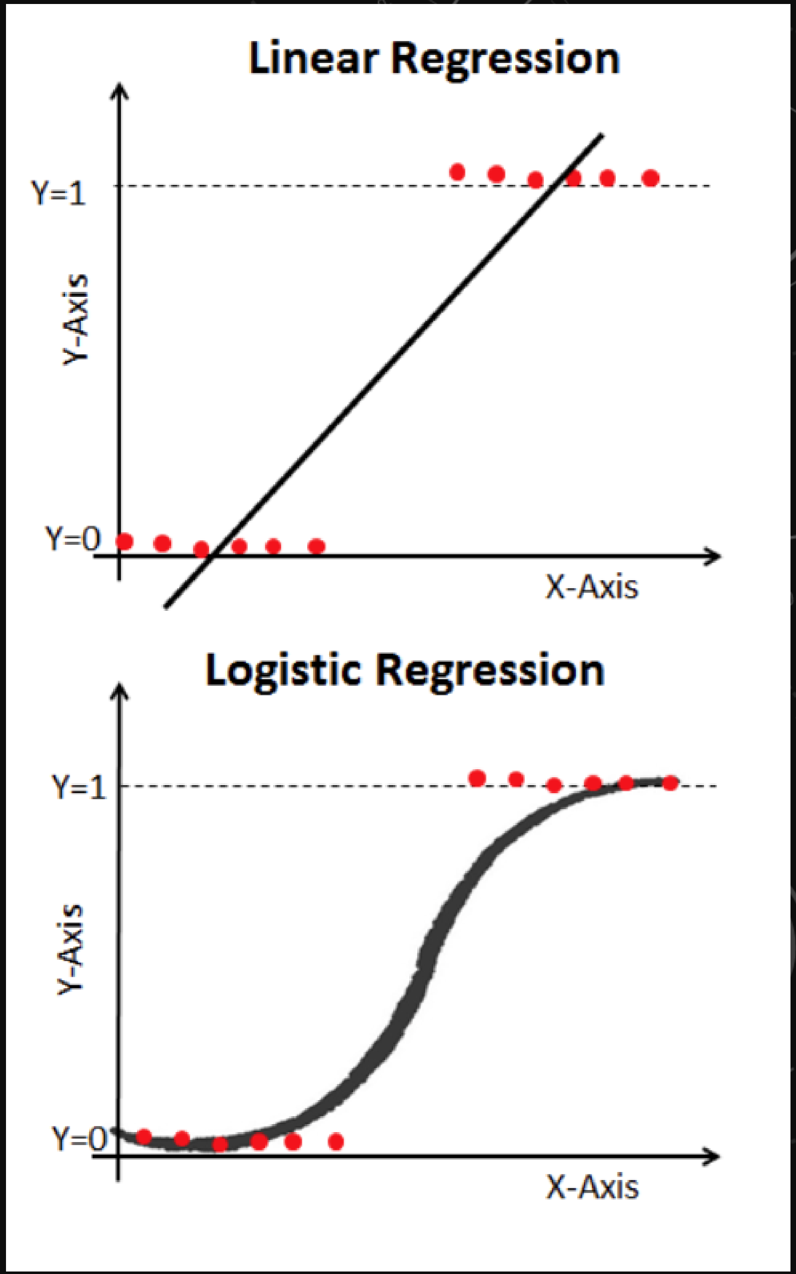


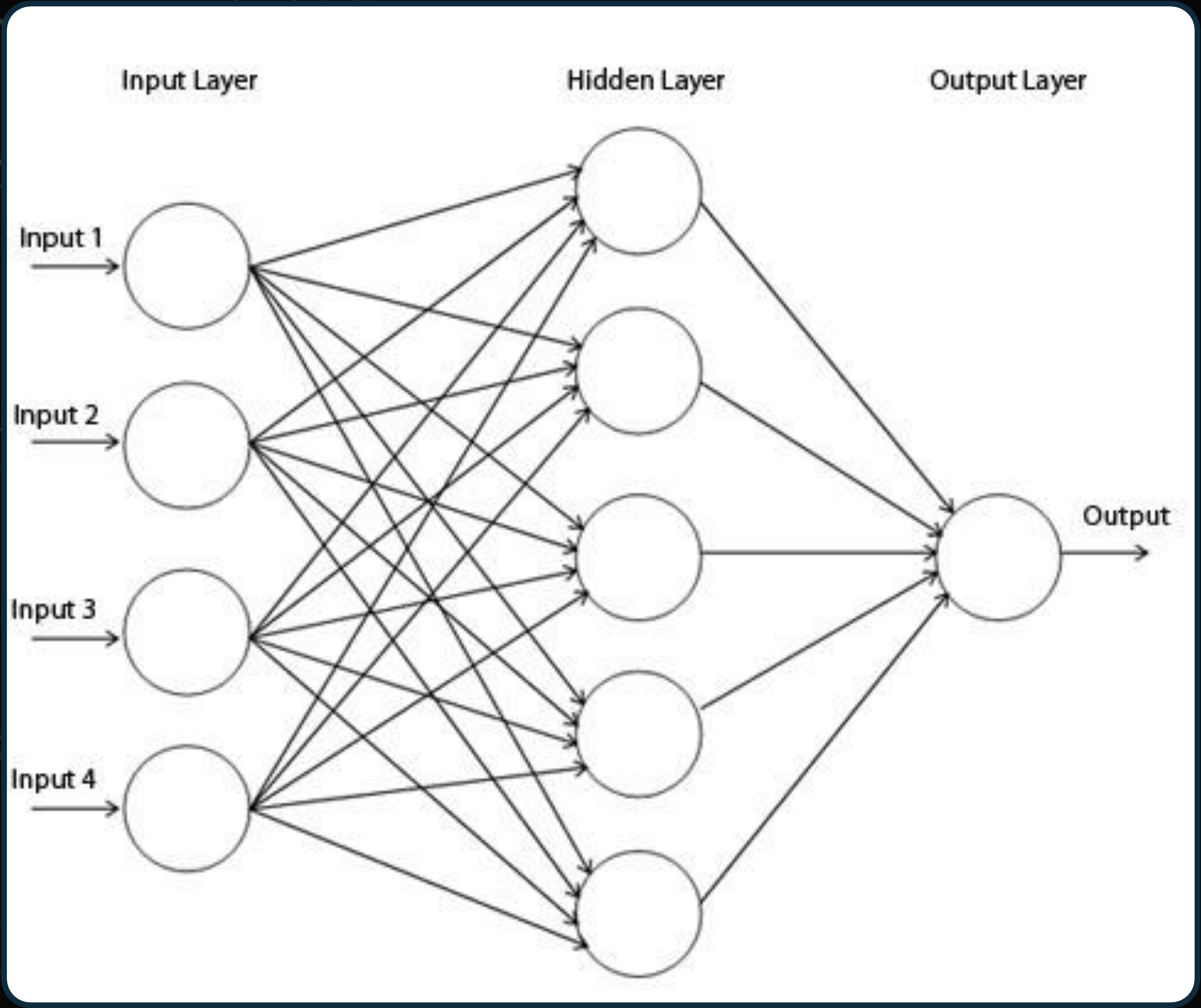
Jockey	race count	win count	win%	place count	plc%
Z Purton	408	86	21%	202	50%
J Moreira	221	44	20%	108	49%
S De Sousa	286	42	15%	111	39%
K Teetan	447	61	14%	151	34%
C Wong	254	30	12%	76	30%

PREDICTION MODELS

- Finishing Time Regression
  - Predict the finishing time of each runner
  - The runner with lowest predicted finishing time is the predicted winner!

- Discrete Choice Model
  - Estimate the conditional probability of winning
  - The runner with highest winning probability is the predicted winner!





FINISHING TIME REGRESSION

- A very intuitive way of prediction.
- Easy to implement
- More samples (each runner is a sample!)
- Finishing time regression can be simply achieved by various models.
- Did not consider the relative performance in the learning
- Models Applied:
  - Simple Linear Regression
  - Neural Network Regression
  - Random Forest Regression
  - K-Nearest Neighbor Regression

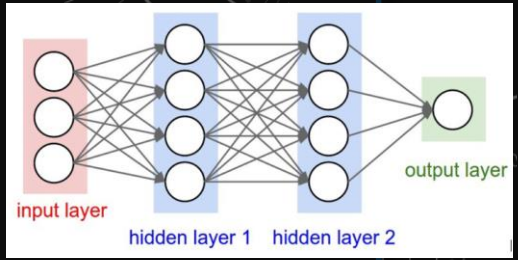
FINISHING TIME REGRESSION - DESCRIPTION

Simple Linear Regression

- Linear relationship between the Finishing time and input features
- $y_{R,H} = \beta_0 + \sum_{i=1}^{n} \beta_i \, x_{i,R,H}$

Neural Network Regression

- Multi-Layer Perceptron
- 4-hidden layers
- Minimizing the mean-squared error

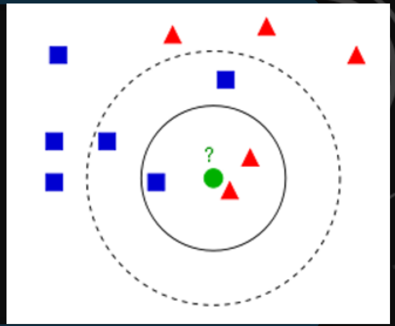


Random Forest Regression

- Predict by averaging the results of multiple Decision Trees
- Ensembled model to reduce the noise
- Depth = 10, avoid from overfitting

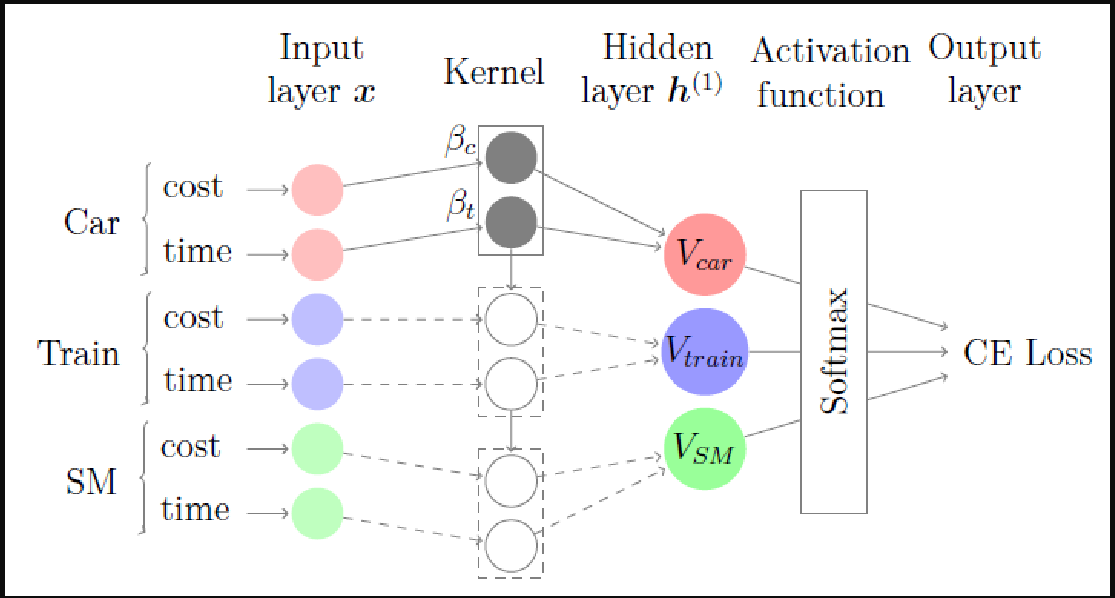
K-NN Regression

- N_neighbors chosen to be $\sqrt{n} = 400$
- Implemented using sklearn



DISCRETE CHOICE MODEL

- Conditional Logistic Regression is a very popular choice model, first proposed by McFadden (1974).
- Estimate the Conditional Winning Probability
- Incorporate the relative performance
- Less samples (each race is a sample!)
- Implemented using Keras, modelled as a Convolutional Neural Network (CNN)
- Models Applying
  - Conditional Logistic Regression (CL)
  - 2-Steps Conditional Logistic Regression (2-Steps CL)
  - Neural –Network Multinomial Logistic (NN-MNL)
  - Learning-Multinomial Logistic (L-MNL)



Basic structure of CL model (Sifringer, Lurkin, & Alahi, 2018)

DISCRETE CHOICE MODEL- DESCRIPTION

Conditional Logistic (CL)

- Linear estimation of Winningness
- $p_{R,H_j} = \frac{\exp(\sum_{i=1}^{n} \beta_i x_{i,R,H_j})}{\sum_m^M \exp(\sum_{i=1}^{n} \beta_i x_{i,R,H_m})}$

2-Steps Conditional Logistic (2-Steps CL)

- 1st Step: Estimate winning probability using Fundamental variables using CL
- 2nd Step: using prob in 1st Step and odds implied probability as input, to estimate the winning probability again using CL

Neural Network Multinomial Logit (NN-MNL)

- Neural network estimation of Winningness, to incorporate the non-linearity
- Tested with 4 and 8 hidden layers
- $p_{R,H} = \frac{\exp(NN(x_{R,H}))}{\sum_H \exp(NN(x_{R,H}))}$

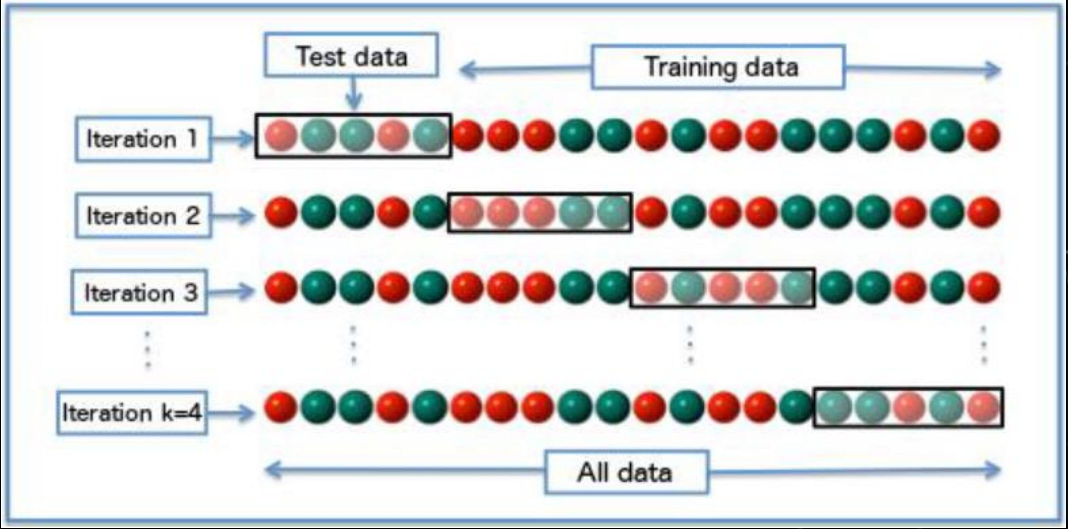
Learning – Multinomial Logit (L-MNL)

- A mixture of CL and NN-MNL
- To preserve the interpretability of the model, while incorporated with non-linearity

MODEL TRAINING

Cross Validation

- 10-Fold Cross Validation is applied to verify the consistency of models
- All models are found to be consistent



Data Scope

- Finishing Time Regression
  - All data is applied
- Discrete Choice Model
  - Races with 14 runners are chosen
  - No significant difference of accuracy when applied to other races

RESULTS – MODEL EVALUATION

- Prediction accuracy

- Wagering back-testing
  - Betting on the predicted winner
  - Betting on Positive expectation
  - Betting with Kelly Criterion

RESULTS – PREDICTION ACCURACY

- FTR Accuracy: 8 – 20%
- DCM Accuracy: about 28%
- DCM in general outperformed the FTR

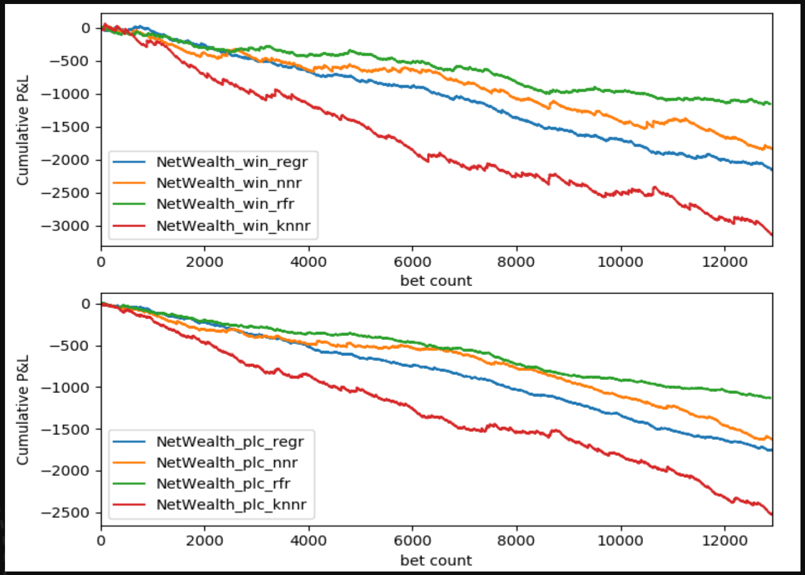
- Reason
  - DCM incorporated the relative performance
  - Hard to predict the time due to a lot of unobserved factors on the track

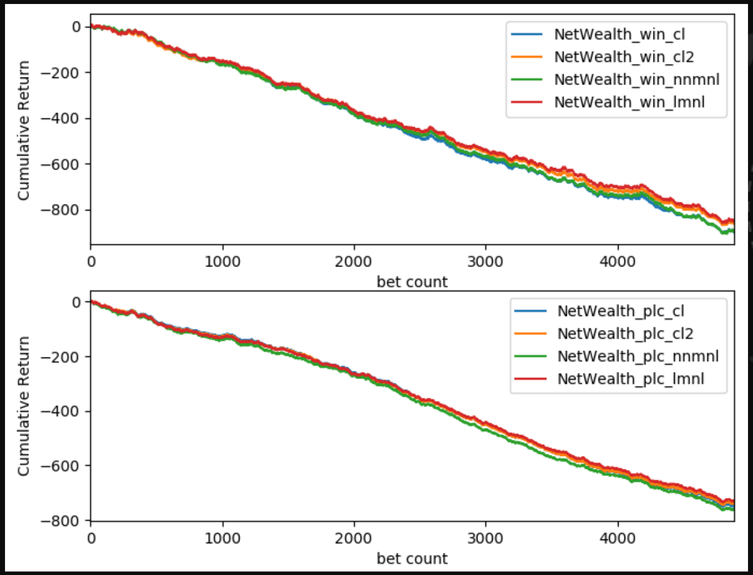
	Average Training Accuracy	**Average Testing Accuracy**
Linear Regression	20.52%	20.59%
Neural Network Regression (4-layers)	17.63%	15.68%
Random Forest Regression (Depth = 10)	22.77%	20.60%
K-NN Regression	8.47%	8.51%
Ensembled FTR	20.63%	20.53%

	Average Training Accuracy	**Average Testing Accuracy**
Conditional Logit	28.10%	27.83%
2-Steps Conditional Logit	28.07%	28.10%
Neural Net - Multinomial Logit	28.14%	28.19%
Learning - Multinomial Logit	28.03%	27.85%
Ensembled DCM	27.85%	30.02%

RESULTS – BETTING ON THE WINNER

- **Strategy:** Bet $1 to Win or Place on the predicted winner
- The strategy fails!
- Reason:
  - The prediction accuracy is not high enough
  - The models are optimizing the predicted time / winning probability, but not optimizing the expected return!





RESULTS – BETTING ON POSITIVE EXPECTATION

- Available only for DCM because they estimate the winning probability!
- **Strategy:** Bet $1 to Win or Place if the below is satisfied

$$EV = P(\widehat{winning}) \times Odds_{before\ race} > 1$$

	CL	2-Steps CL	NN-MNL	L-MNL
# of Bets	997	995	1044	959
# of Correct Win bet	166 (16.65%)	172 (17.29%)	152 (14.56%)	176 (18.35%)
# of Correct Place bet	386 (38.72%)	390 (39.20%)	375 (35.92%)	391 (40.77%)
P&L of Win bets	-74.70 (-7.49%)	-45.50 (-4.57%)	-54.90 (-5.26%)	-19.70 (-2.05%)
P&L of Place bets	-100.49 (-10.08%)	-110.34 (-11.09%)	-82.35 (-7.89%)	-90.19 (-9.40%)

RESULTS – BETTING ON POSITIVE EXPECTATION

- The cumulative P&L is mostly around -10%
- It is possible to achieve a positive return if rebate is applicable!

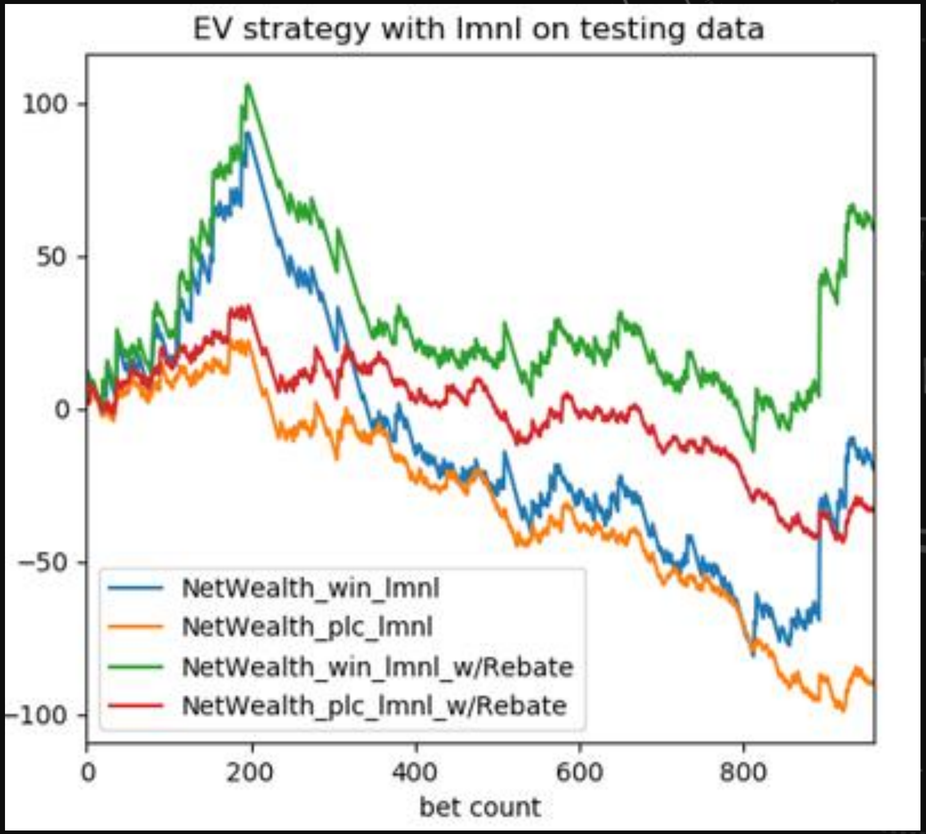
- **Rebate:** Any ticket with a total losing bet amount of HKD$10,000 or above will be eligible to receive a rebate of 10% of the total loss amount.

	CL	2-Steps CL	NN-MNL	L-MNL
# of Bets	997	995	1044	959
# of Correct Win bet	166 (16.65%)	172 (17.29%)	152 (14.56%)	176 (18.35%)
# of Correct Place bet	386 (38.72%)	390 (39.20%)	375 (35.92%)	391 (40.77%)
P&L of Win bets w/Rebate	8.40 (0.84%)	36.80 (3.70%)	34.30 (3.29%)	58.60 (6.11%)
P&L of Place bets w/Rebate	-39.39 (-3.95%)	-49.84 (-5.01%)	-15.45 (-1.48%)	-33.39 (-3.48%)

RESULTS – BETTING ON POSITIVE EXPECTATION

- Betting on Win with Rebate returned positive P&L, while betting on Place with Rebate is still losing

- Reason could be the number of losing bets on Win pool (84%) is more than in Place pool (61%), hence a better "return" from rebates!



RESULTS – BETTING WITH KELLY CRITERION

- Kelly Criterion (1956) is a strategy of asset allocation. It determines the fraction to bet in the game such that the funds grow exponentially

$$f = \frac{BP - Q}{B}$$

$$where\ B = decimal\ odds - 1, P = P(Win), Q = P(Lose)$$

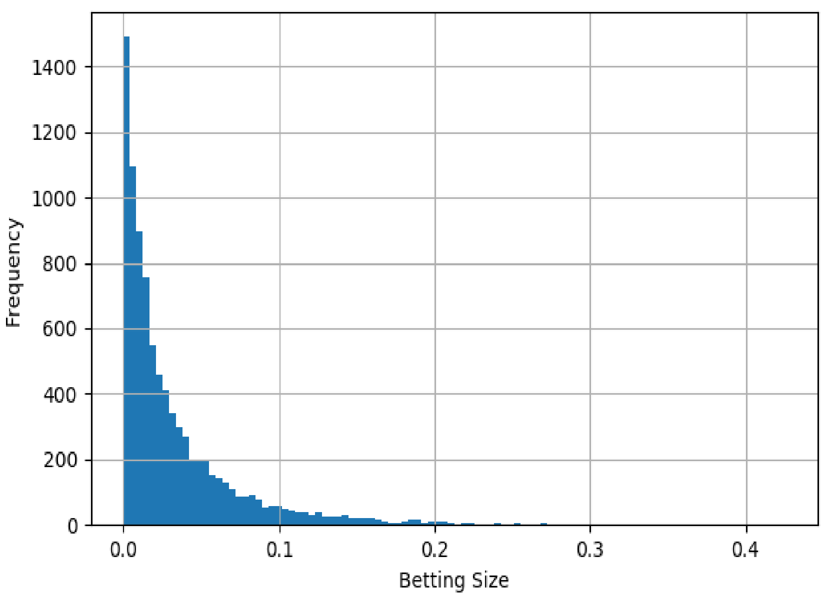
- **Strategy:** Bet $f * (Total Capital) to Win or Place if f is greater than zero

- Compare to positive Expectation Strategy:
  - The entry criteria is the same; entry only if positive expectation
  - Kelly assigns a higher bet amount if the expectation is very positive!

RESULTS – BETTING WITH KELLY CRITERION

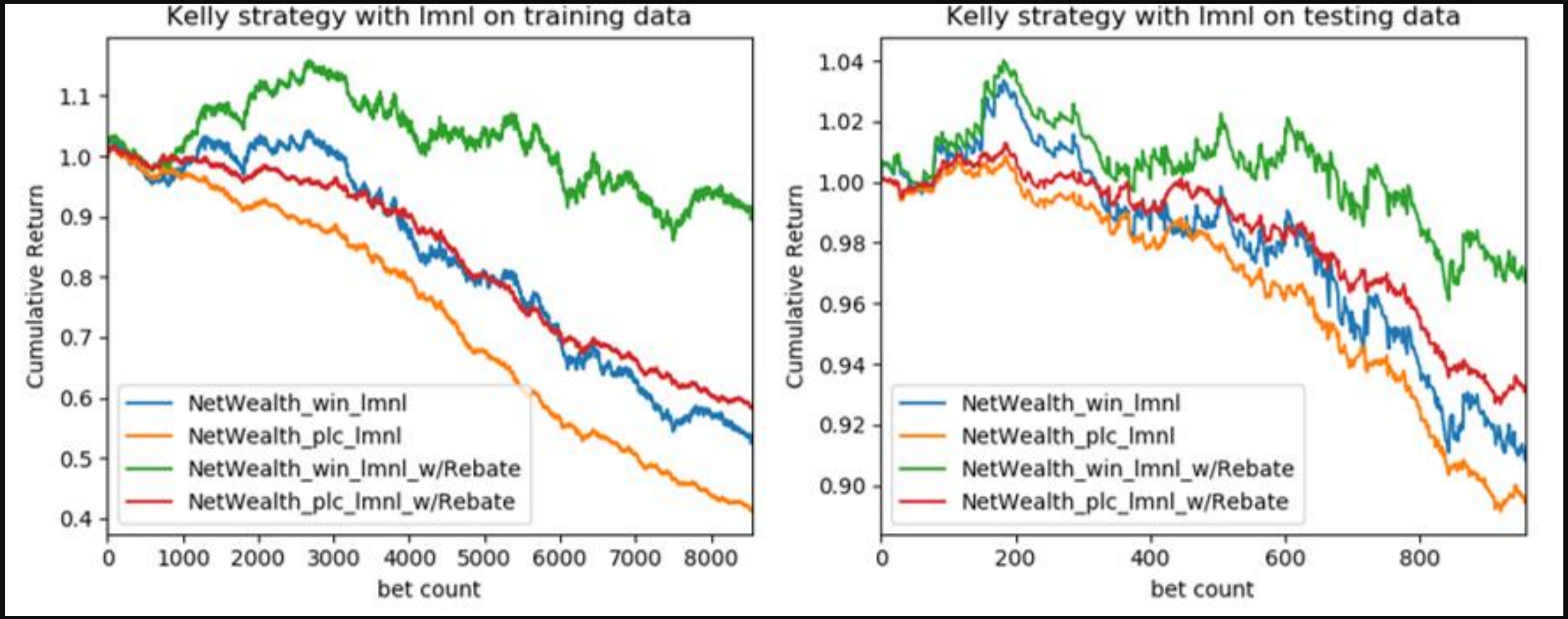
- Distribution of Kelly's fraction is highly skewed to the right, with maximum at 42.5%
- Kelly Criterion is often criticized due to
  - Aggressiveness
  - unable to cater the probability estimate uncertainty
- Fractional Kelly is applied
  - Betting fraction = $h \times f$
  - h is chosen such that none of the bets would NOT deploy more than 1% of total capital

Statistic	Value
Mean	3.5%
Standard Deviation	4.4%
Minimum	0.0%
25% Quantile	0.69%
50% Quantile	1.83%
75% Quantile	4.4%
Maximum	42.5%



RESULTS – BETTING WITH KELLY CRITERION

- No significant advantage brought by Kelly Criteria
- Implying the probability estimate is not sufficiently accurate given the current input features and model architecture!



Not a direct comparison with previous – Kelly based strategy is compounding with starting asset $1, while the previous strategies is arithmetic with starting asset $0

RESULTS – BETTING WITH ENSEMBLED MODEL

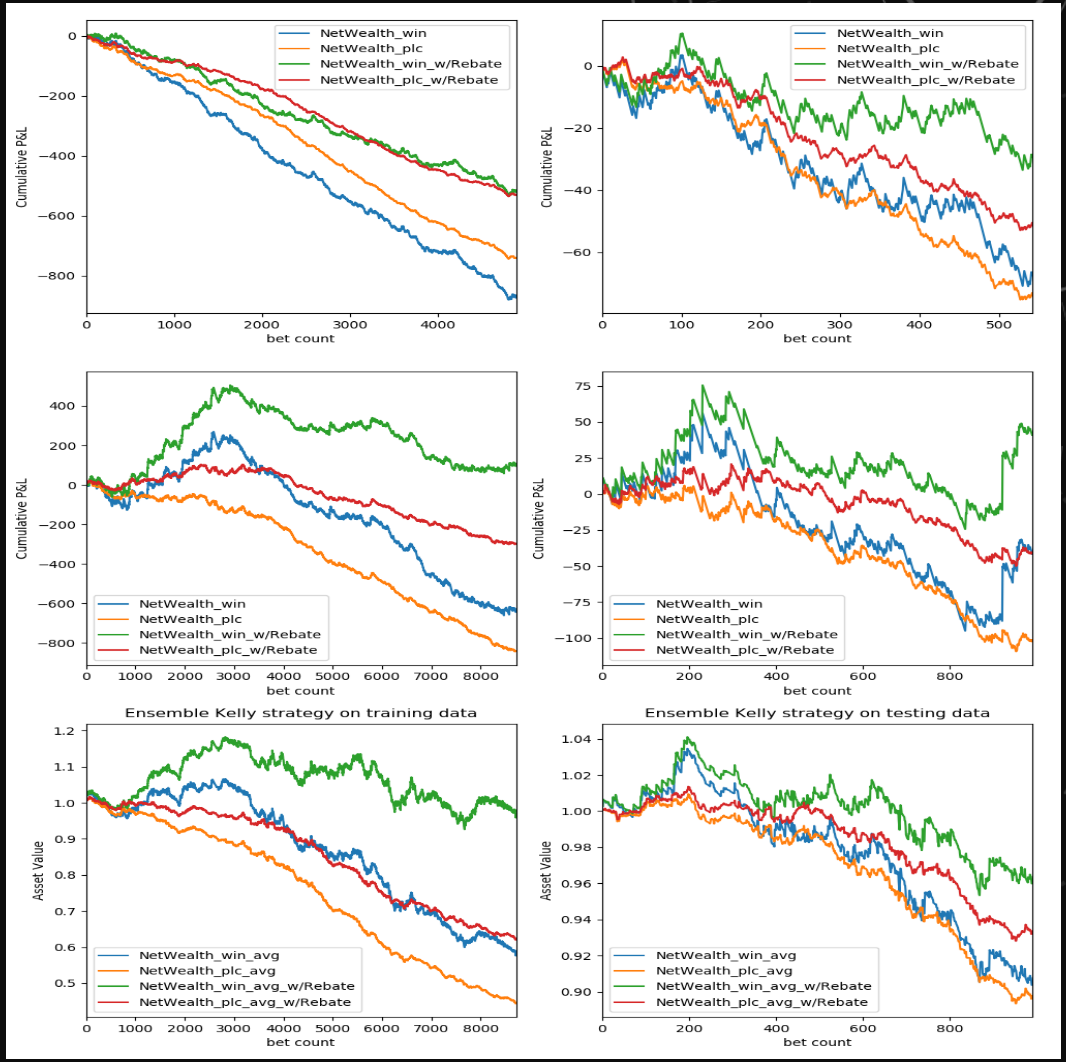
- Ensembled Modelling
  - Combine the results of related but different algorithms, such that the resultant predictions are less noisy and more accurate.

- Prediction Accuracy
  - FTR: around 20% - improved in general
  - DCM: around 28% - No improvements

RESULTS – BETTING WITH ENSEMBLED MODEL

- No significant improvements for all 3 strategies:
  - Betting on the predicted winner (top)
  - Betting on Positive expectation (middle)
  - Betting with Kelly Criterion(bottom)

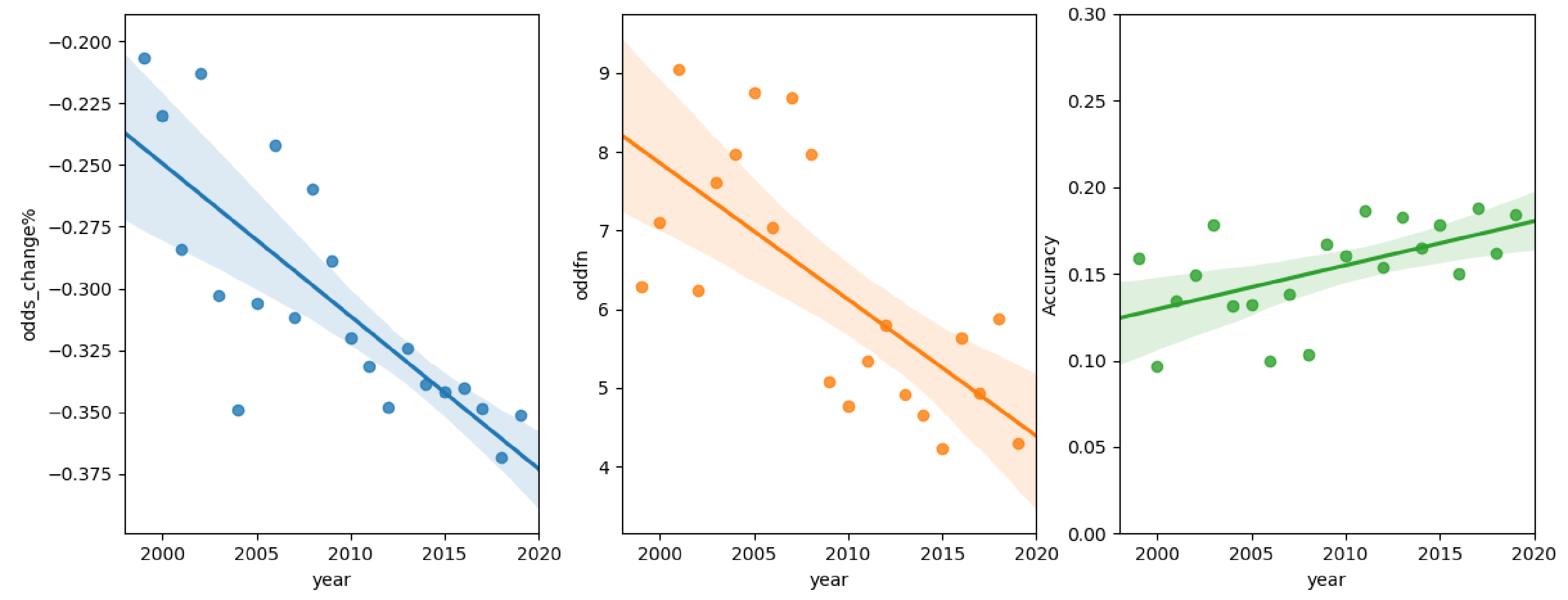
- Likely due to the predictions for all 4 DCM models are quite similar.



RESULTS – P&L EXPLANATION

Market is getting smarter!

- Final odds dropped further right before race
- Average final odds of our bets are going down

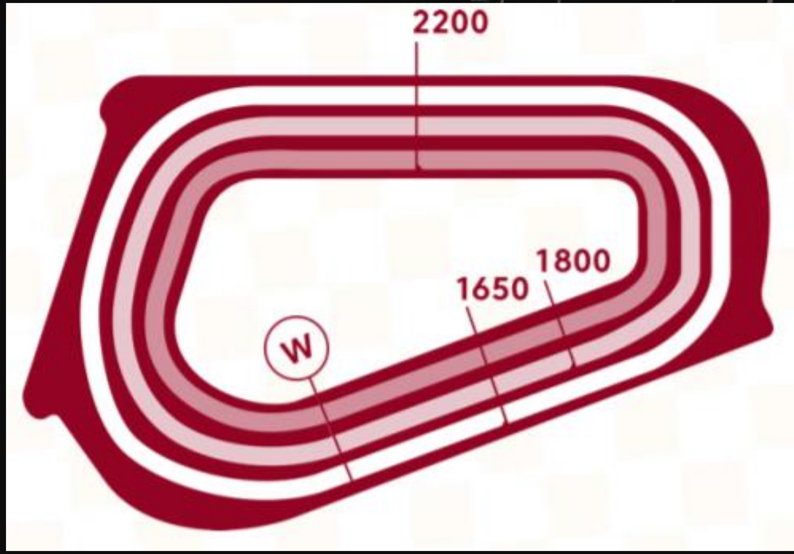


CONCLUSION

- Horse Racing is difficult to predict by nature due to the results can be affected by a lot of factors
- Created & Visualized useful features for prediction
- DCM Models (~28%) outperformed FTR Models (~20%) in terms of accuracy
- Betting on positive expectation with DCM could result a positive return if rebate is applicable
- Horseracing odds market is getting smarter over the years

FUTURE WORK

- Additional features
  - Morning exercise data
  - Barrier trail results
  - Emotion of the horse
  - …. And more!



- Further Analysis
  - Racecourse topology (angles, lengths of straight paths)
  - Racing style of horse and jockey
  - Difficulty: requires a lot of intra-race data and hard to quantify

Q&A

Dissertation Webpage: https://hokai999.wixsite.com/website