LYU1603
Predicting Horse Racing Results using TensorFlow

1155051298 CHENG Tsz Tung

1155051346 LAU Ming Hei

Supervised by Prof. LYU Ring Tsong Michael

Overviews

- Last Semester Summary
- This Semester Goal
- Standardization
- Data Extraction
- New Dataset (X and Y)
- New Modeling
  - K-nearest-neighbor regression
  - Linear Regression

Last Semester Summary

- Predicting whether a horse will win the races
- Classification Problem
- Two approach
  - Pattern Matching
  - Linear Classification
- Generate net profits is possible

This Semester Goal

- Improve accuracy of the model
- Evaluate in different bet types
  - Place Bet
  - Quinella Bet
  - Quinella Place Bet

Homework Problem

(CSCI 3320, Spring 2016-17, Homework 1, Q5)

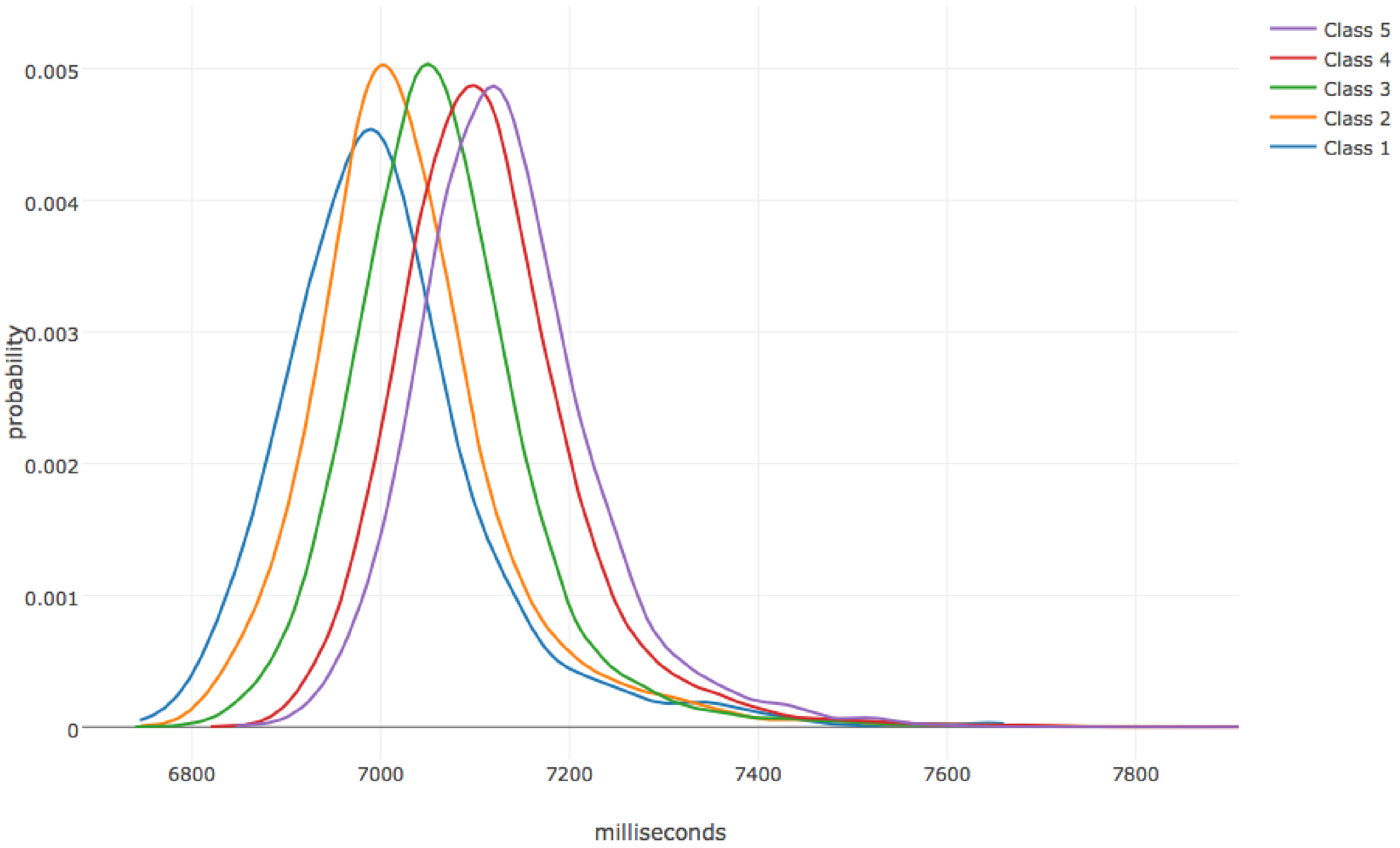
In estimating the price of a used car, it makes more sense to estimate the percent depreciation over the original price than to estimate the absolute price. Why?

Car	Original Price	Age	Price	Loss Percent
Lamborghini	$	5 Years	$	30%
Toyota	$	5 Years	$	30%

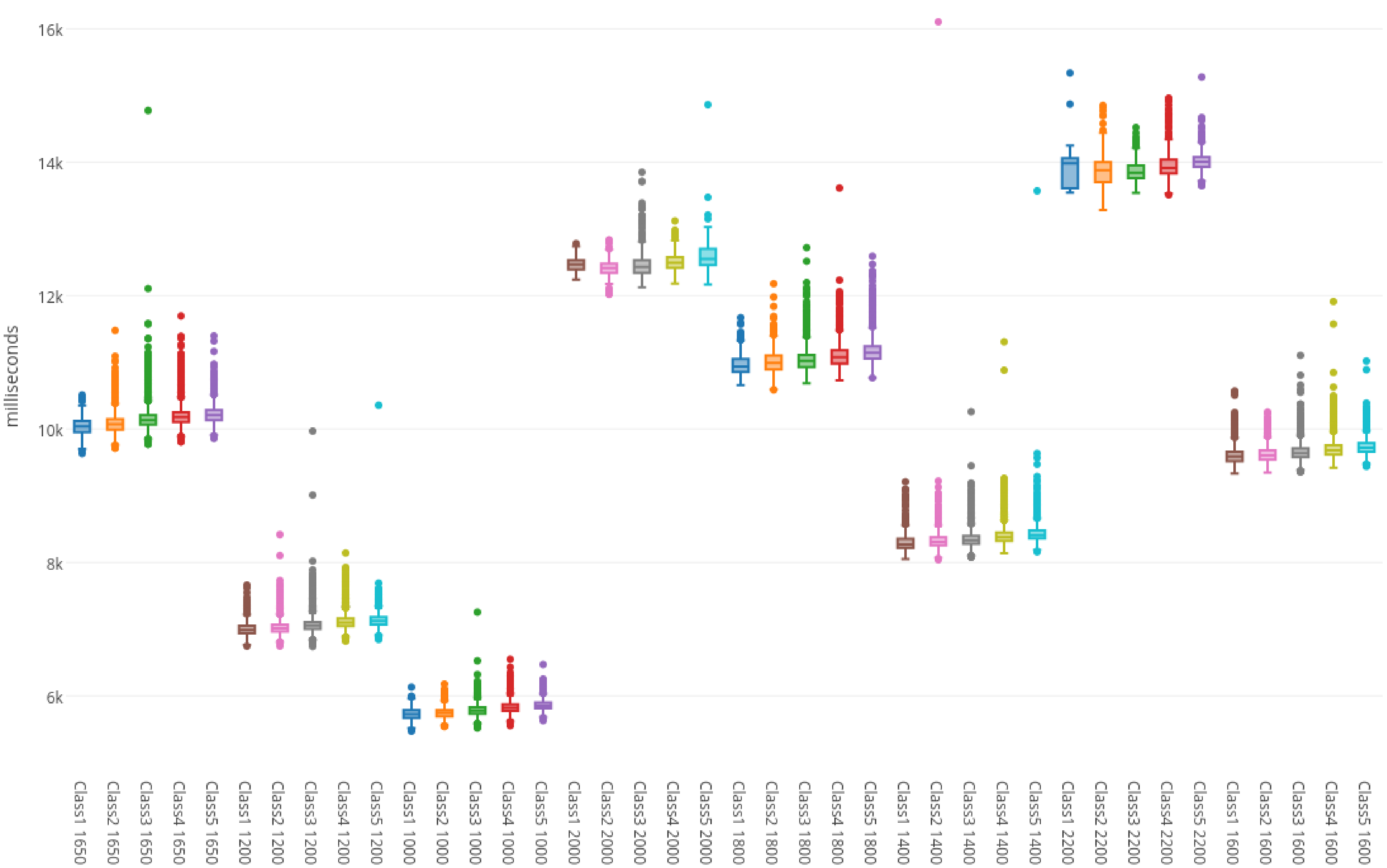
Normalization in Last Semester

- Standardized the data globally
    - Without consider it’s situation

1200-Meters Finishing Time Distribution



Finishing Time Distribution by (Class, Distance)



Standardization

- $L = \{locations\}$
- $T = \{cources\}$
- $C = \{classes\}$
- $D = \{distances\}$
- $z_{i_{l,t,c,d}}^{(j)} = \frac{x_{i_{l,t,c,d}}^{(j)} - \mu_{i_{l,t,c,d}}}{\sigma_{i_{l,t,c,d}}}$

ELO System in Last Semester

- Failed to capture recent performance

Horse	Last 3 Races	Last 2 Races	Last 1 Races	Final ELO
Horse A	WIN	LOSE	LOSE	1600
Horse B	LOSE	WIN	WIN	1600

Wins Odd

- Capture public expectation
- Cannot use it before end of betting period

Recap

- Want to capture recent performance
- Use win odds as a feature

- Solution:
  - Add past records of a horse
  - Use past win odds

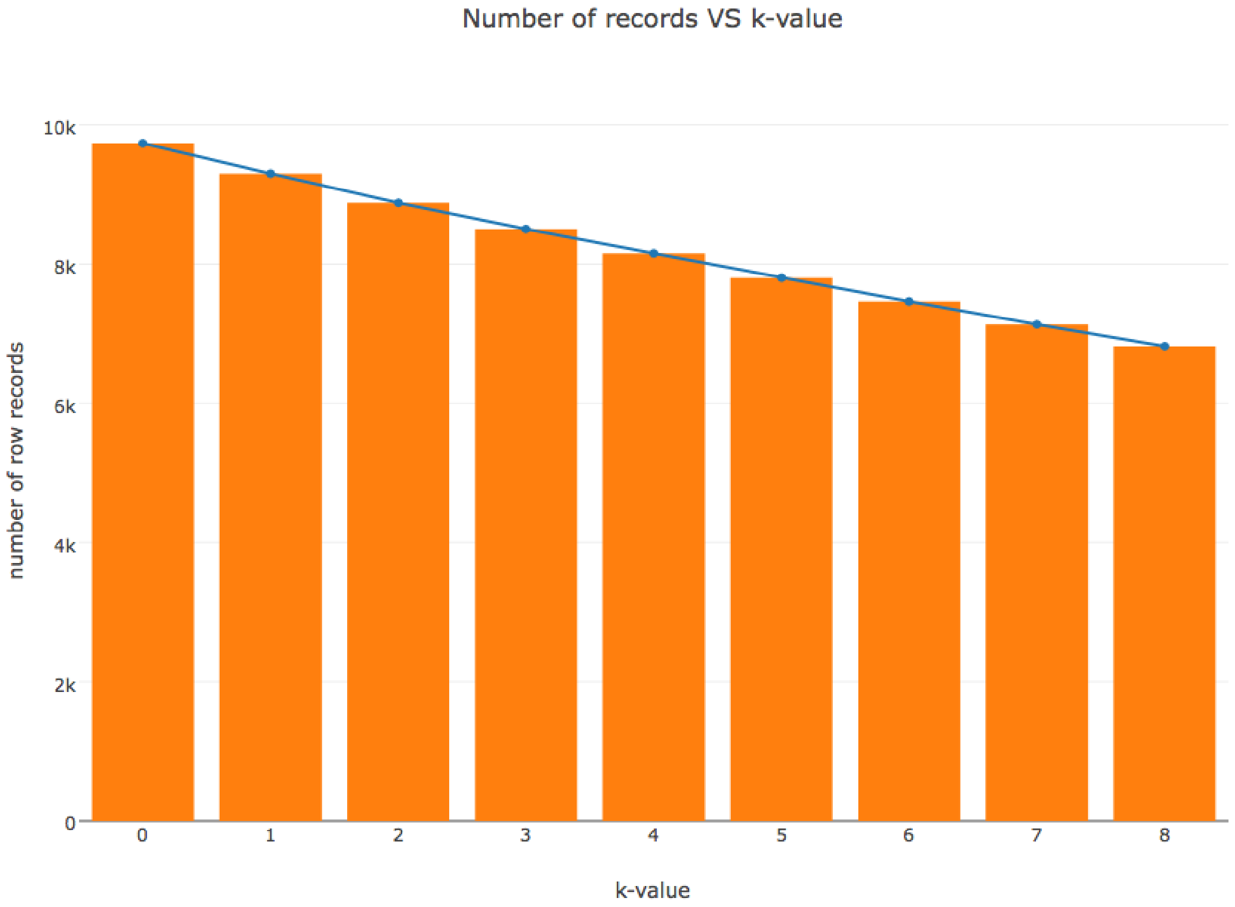
Data Extraction

- Extraction past k records of a horse

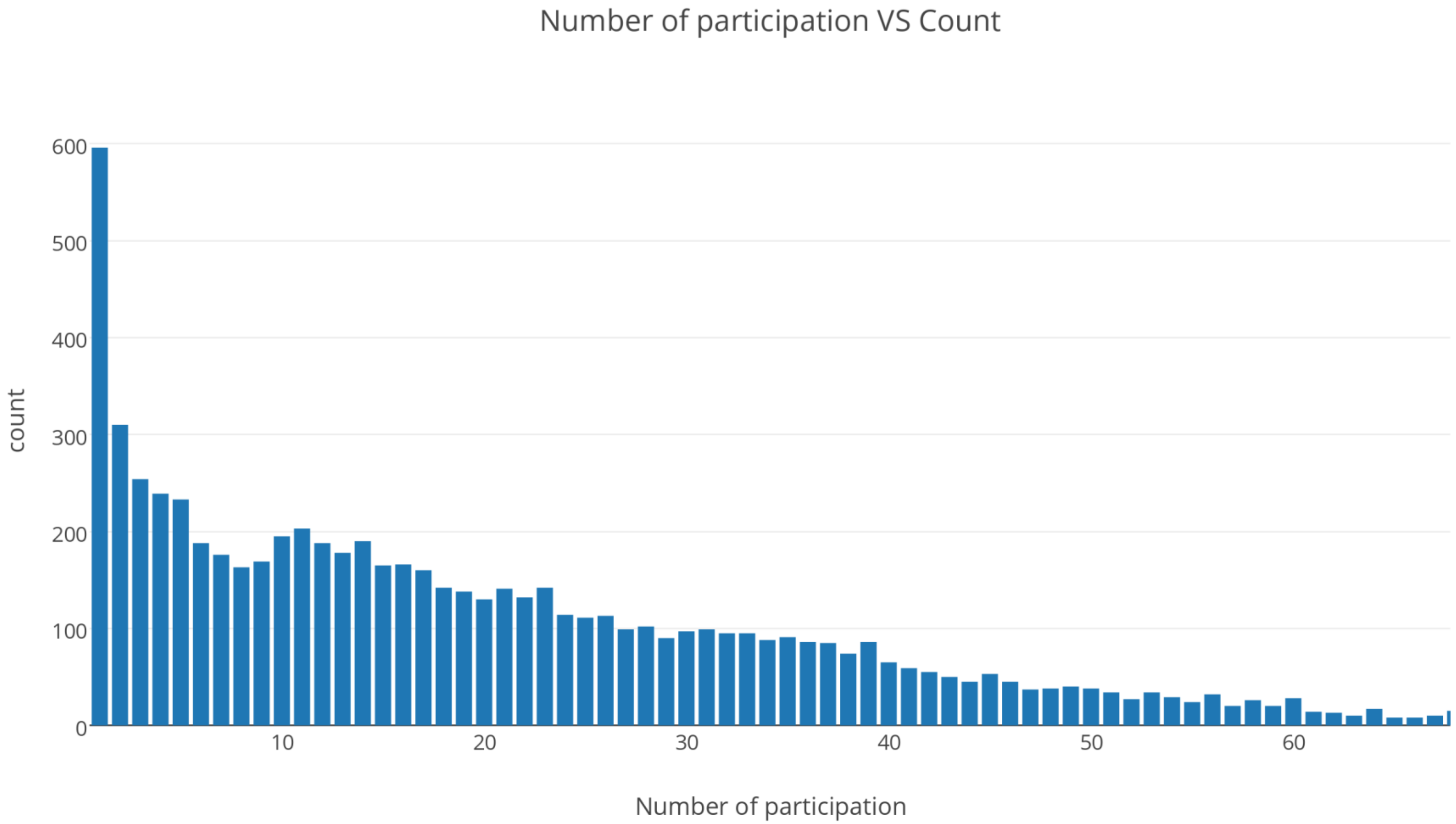
- $p = past\ k - records$
- $k_i = \begin{bmatrix} p_i^{(1)} & \cdots & p_i^{(k)} & x_i^{(j)} \end{bmatrix}$ … select particular feature
- $x^{(j)} \coloneqq k_1 \oplus k_2 \oplus \cdots \oplus k_n$ … append to those records

Potential Problem of Data Extraction

- K-value vs amount of data



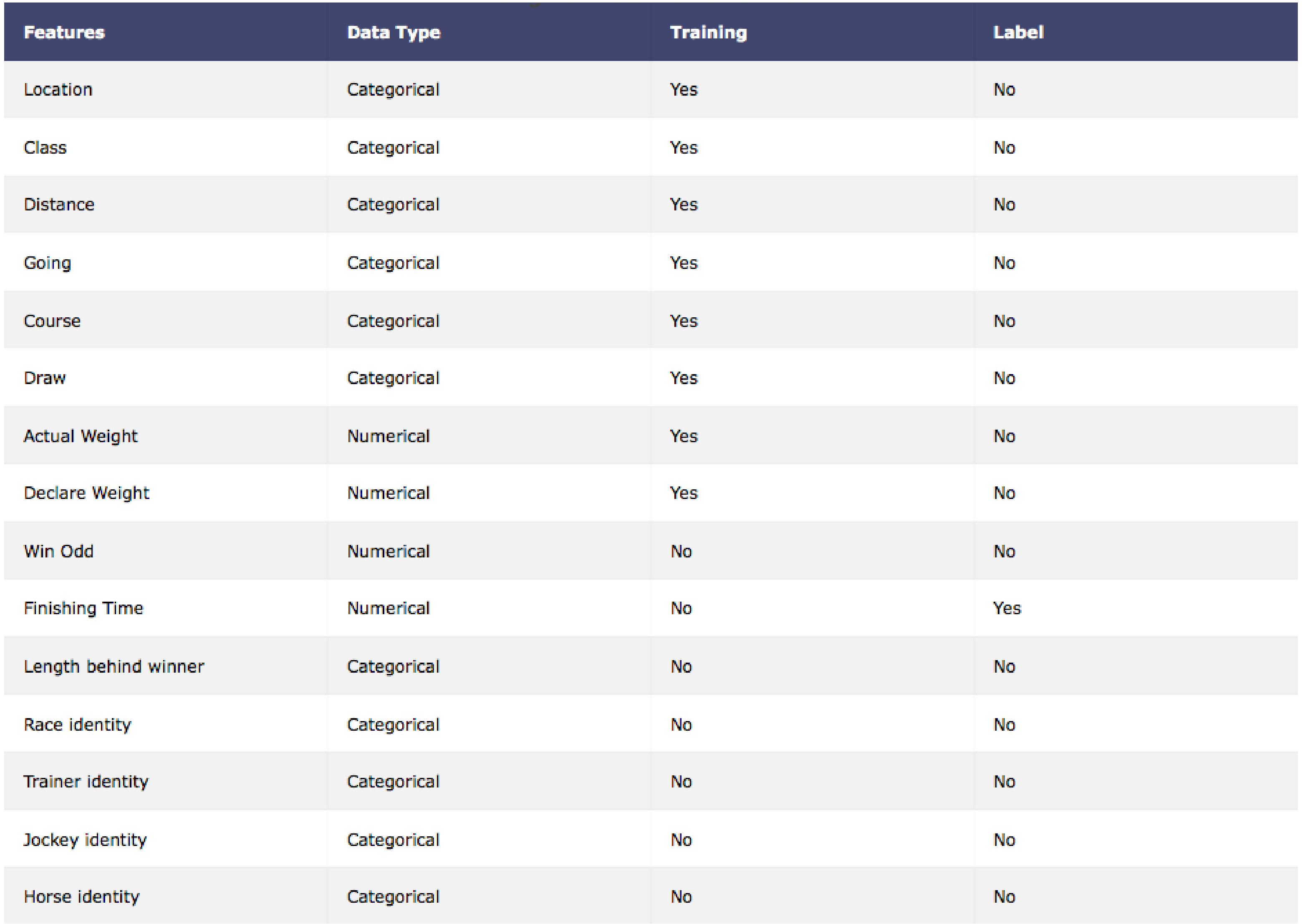
Potential Problem of Data Extraction



Potential Problem of Data Extraction (Cont.)

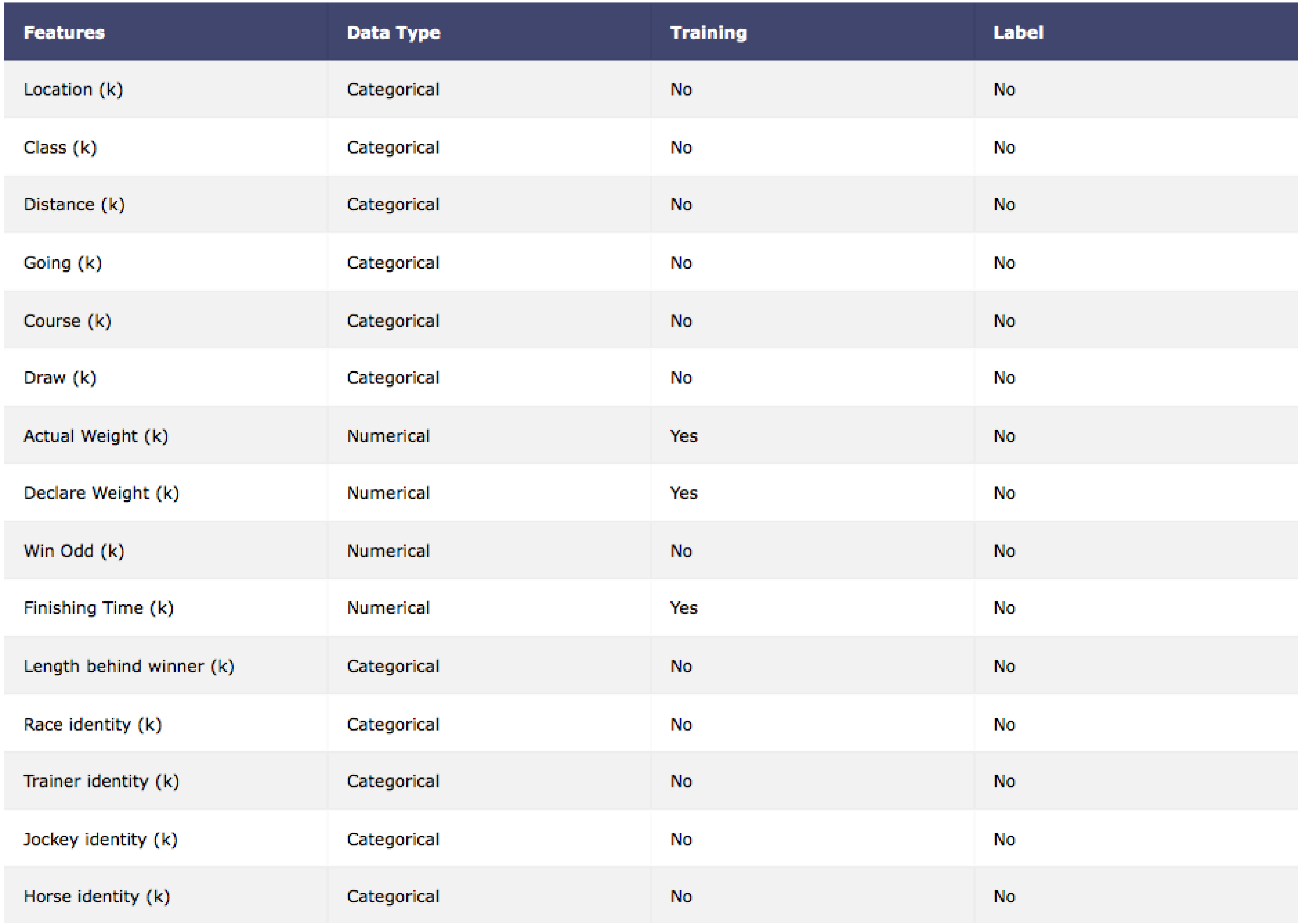
- K-value vs the coverage of recent performance

Feature Selection



Features	Data Type	Training	Label
Location	Categorical	Yes	No
Class	Categorical	Yes	No
Distance	Categorical	Yes	No
Going	Categorical	Yes	No
Course	Categorical	Yes	No
Draw	Categorical	Yes	No
Actual Weight	Numerical	Yes	No
Declare Weight	Numerical	Yes	No
Win Odd	Numerical	No	No
Finishing Time	Numerical	No	Yes
Length behind winner	Categorical	No	No
Race identity	Categorical	No	No
Trainer identity	Categorical	No	No
Jockey identity	Categorical	No	No
Horse identity	Categorical	No	No

Feature Selection(cont.)

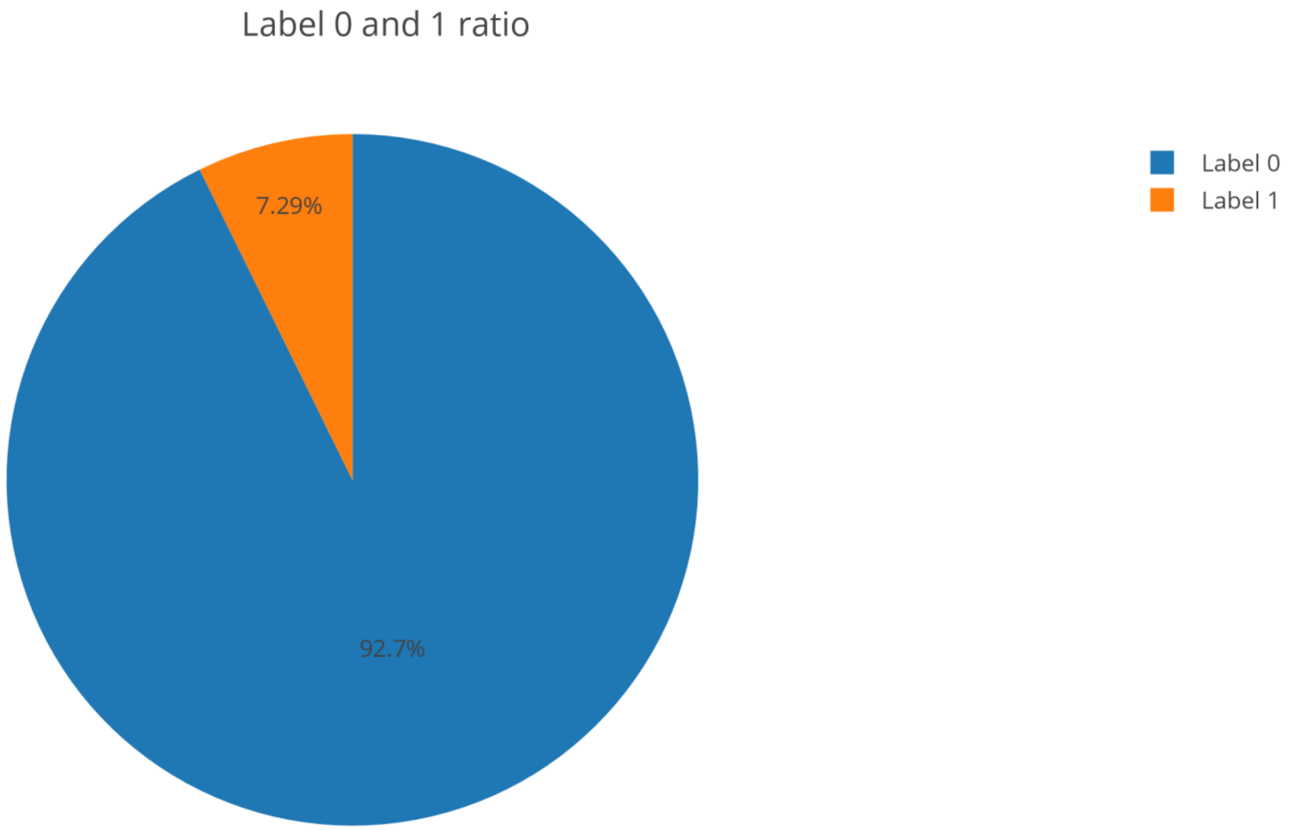


Features	Data Type	Training	Label
Location (k)	Categorical	No	No
Class (k)	Categorical	No	No
Distance (k)	Categorical	No	No
Going (k)	Categorical	No	No
Course (k)	Categorical	No	No
Draw (k)	Categorical	No	No
Actual Weight (k)	Numerical	Yes	No
Declare Weight (k)	Numerical	Yes	No
Win Odd (k)	Numerical	No	No
Finishing Time (k)	Numerical	Yes	No
Length behind winner (k)	Categorical	No	No
Race identity (k)	Categorical	No	No
Trainer identity (k)	Categorical	No	No
Jockey identity (k)	Categorical	No	No
Horse identity (k)	Categorical	No	No

Definition of Y

- At Last Semester
  - We use classification
    - 1 ⇔ the horse is the 1st Place
    - 0 ⇔ Otherwise

Problem (Unbalance Dataset)



Problem (Cannot rank horses)

	1^{st}	2^{nd}	3^{rd}
Horse A	46%	44%	10%
Horse B	10%	60%	30%
Horse C	44%	10%	46%

Seq. (consider individual place): Horse A -> Horse B -> Horse C

Seq. (only consider 1^{st} probability): Horse A -> Horse C -> Horse B

Redefine Y

- Classification Problem => Regression Problem
- Use Standardized Finishing Time instead of Place for Regression
- Benefits:
  - Unbalance dataset problem avoided
    - No hyper-parameters
  - Use Predicted Standardized finishimg time to rank horse

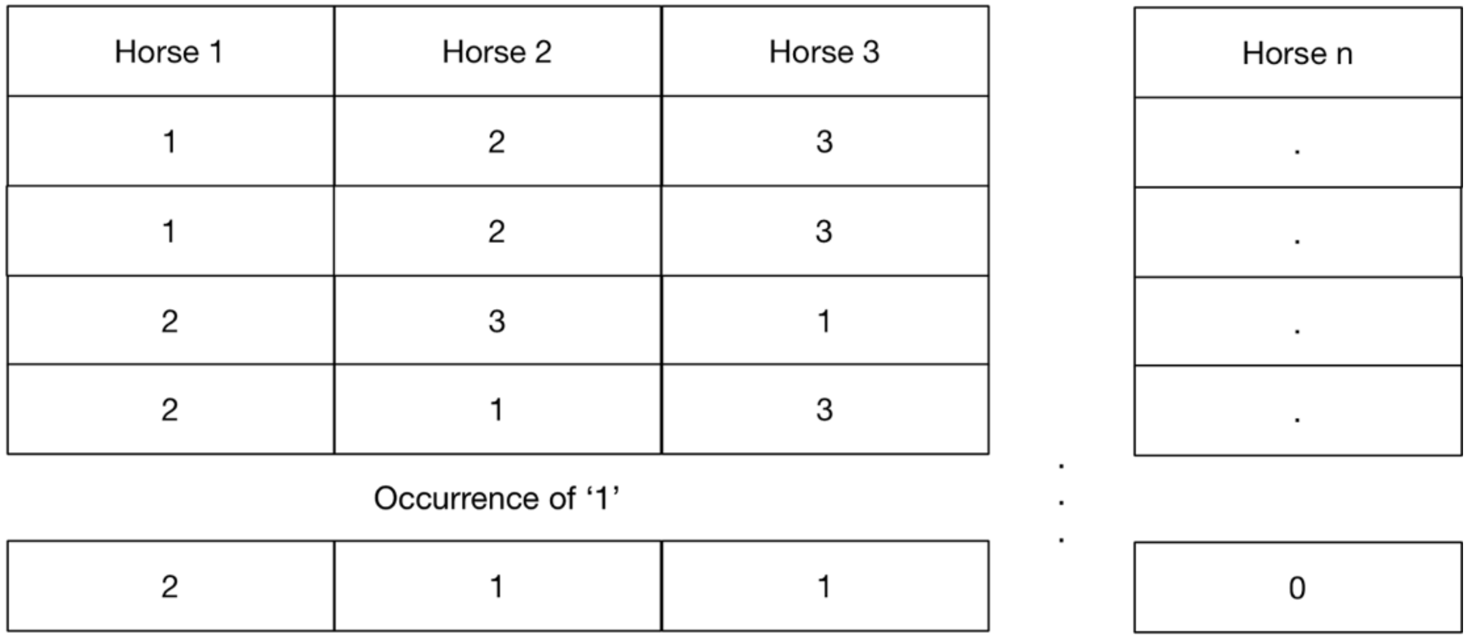
Two ways to modeling the Problem

- Pattern Matching
- Linear Regression

Pattern Matching (Last Semester)

- Build a races history index file
- Define $similarity(R_i, R_j) = \frac{R_i \cdot R_j}{\left| |R_i| |R_j| \right|}$

Example: Similar 4-races



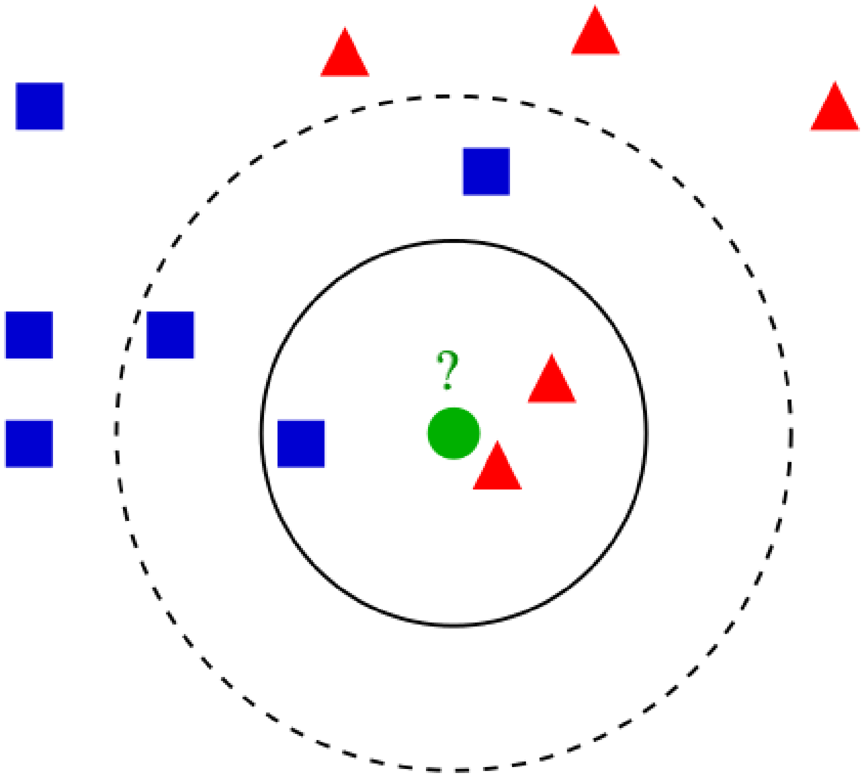
Horse 1	Horse 2	Horse 3	Horse n
1	2	3	.
1	2	3	.
2	3	1	.
2	1	3	.

Occurrence of ‘1’

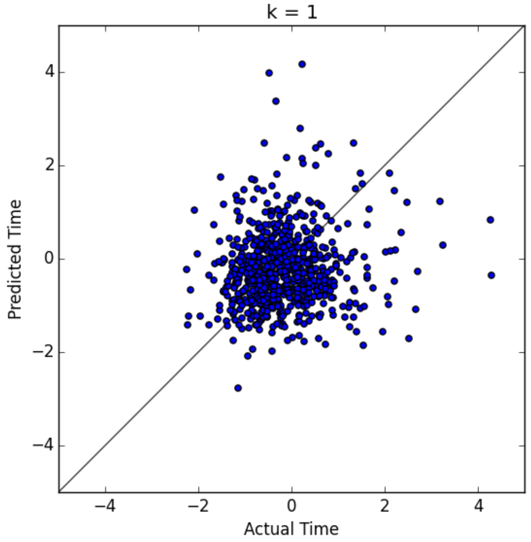
2	1	1	0

Pattern Matching (This Semester)

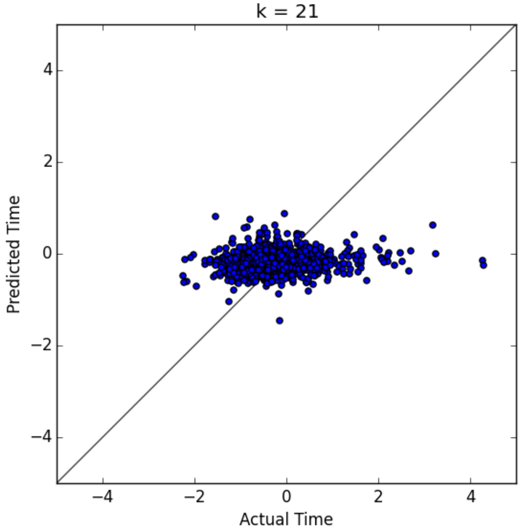
- Use k-nearest-neighbors algorithm
- To Find similar k-races
- Calculate finish time by apply distance weighting



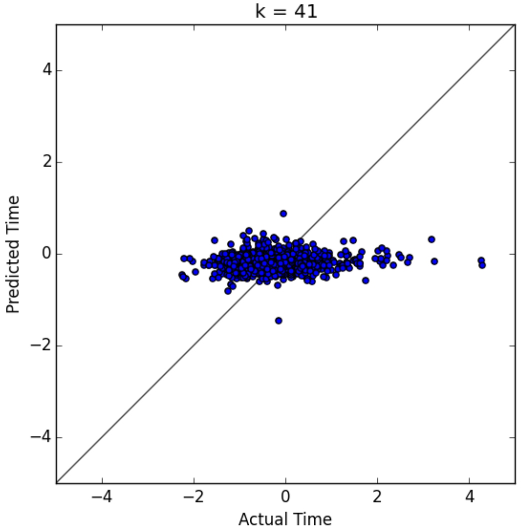
Results



Score: -0.605



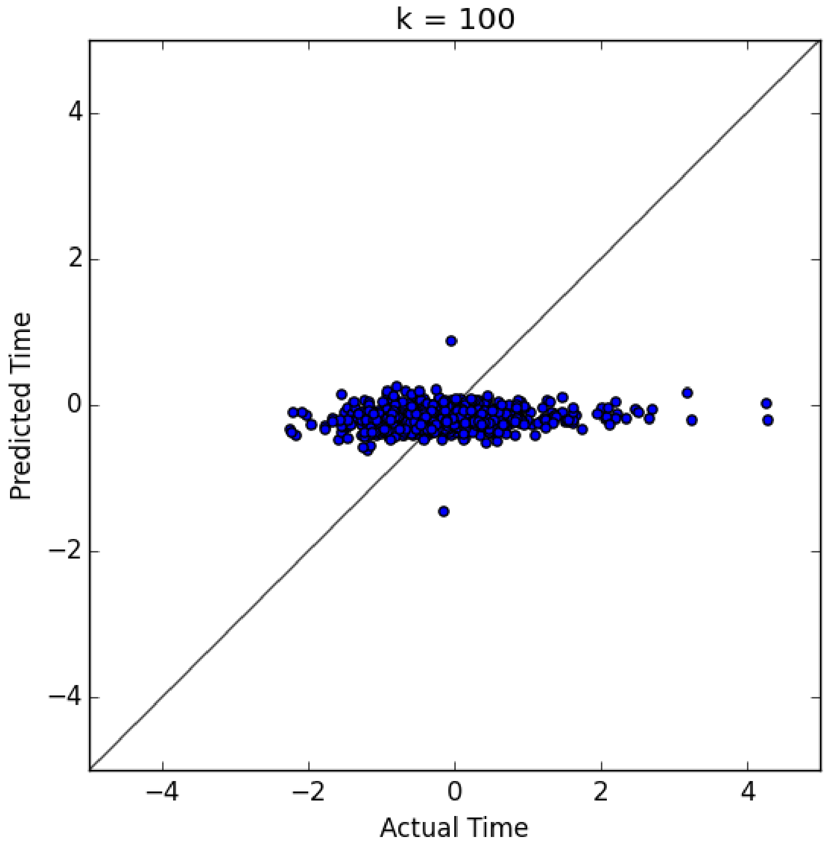
Score: -0.017



Score: 0.0052

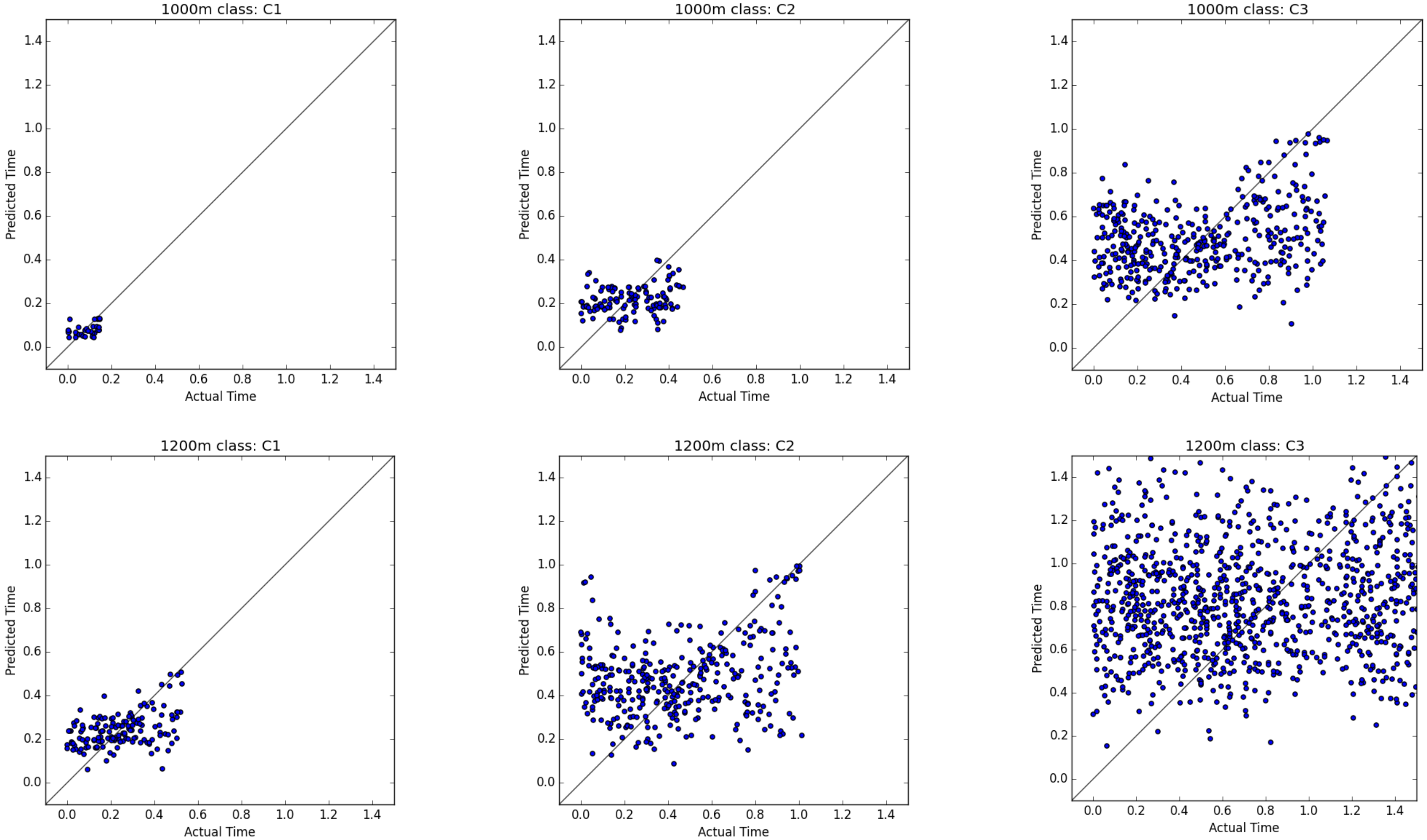
Value Of K?

- K = 100 reach the max score
- Score is too low



Score: 0.0097

Results of each Subset of data



Patterns

- Higher the class, the higher the score
- Longer the distance, the lower the score

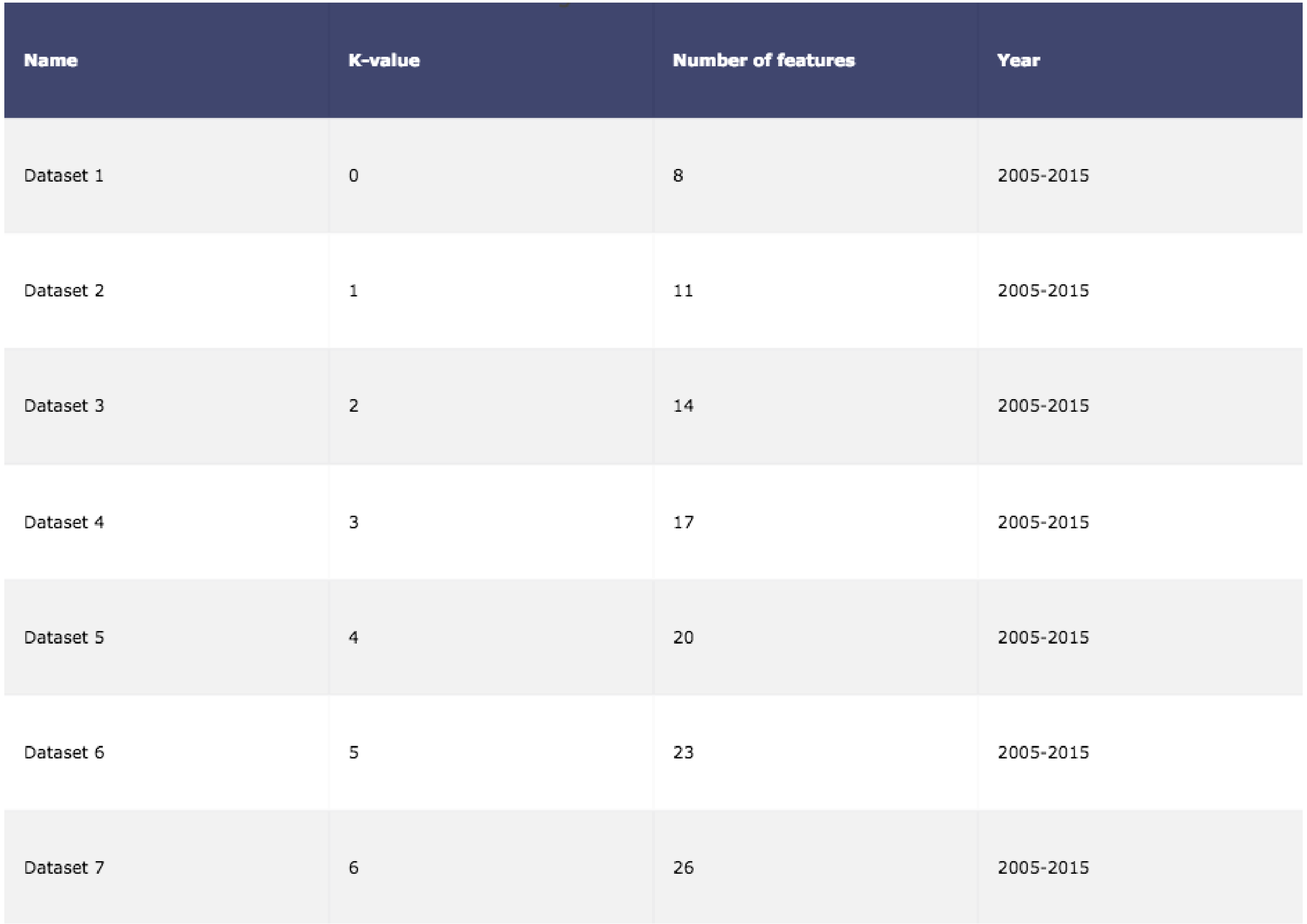
Linear Model (Last Semester)

- Classification
- $\theta^{\mathrm{T}}x = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n$
- $\hat{h}_\theta(x) = sigmoid\left(\theta^{\mathrm{T}}x\right) = \frac{1}{1+e^{-\theta^{\mathrm{T}}x}}$

Linear Model (This Semester)

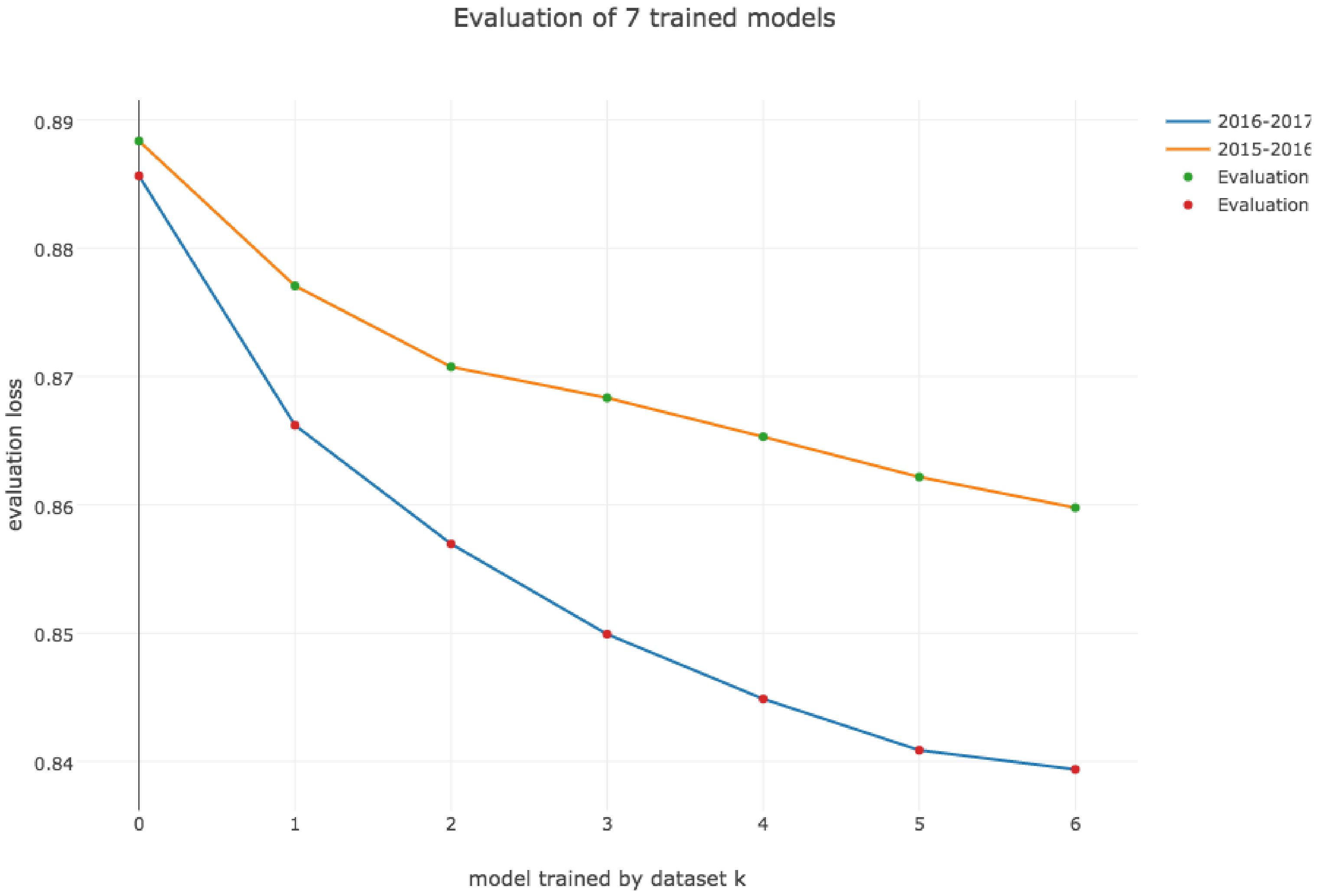
- Regression
- Predict standardized finishing time
- $h_\theta(x) = \theta^{\mathrm{T}} x$
- Only Different is the sigmoid function

Dataset preparation

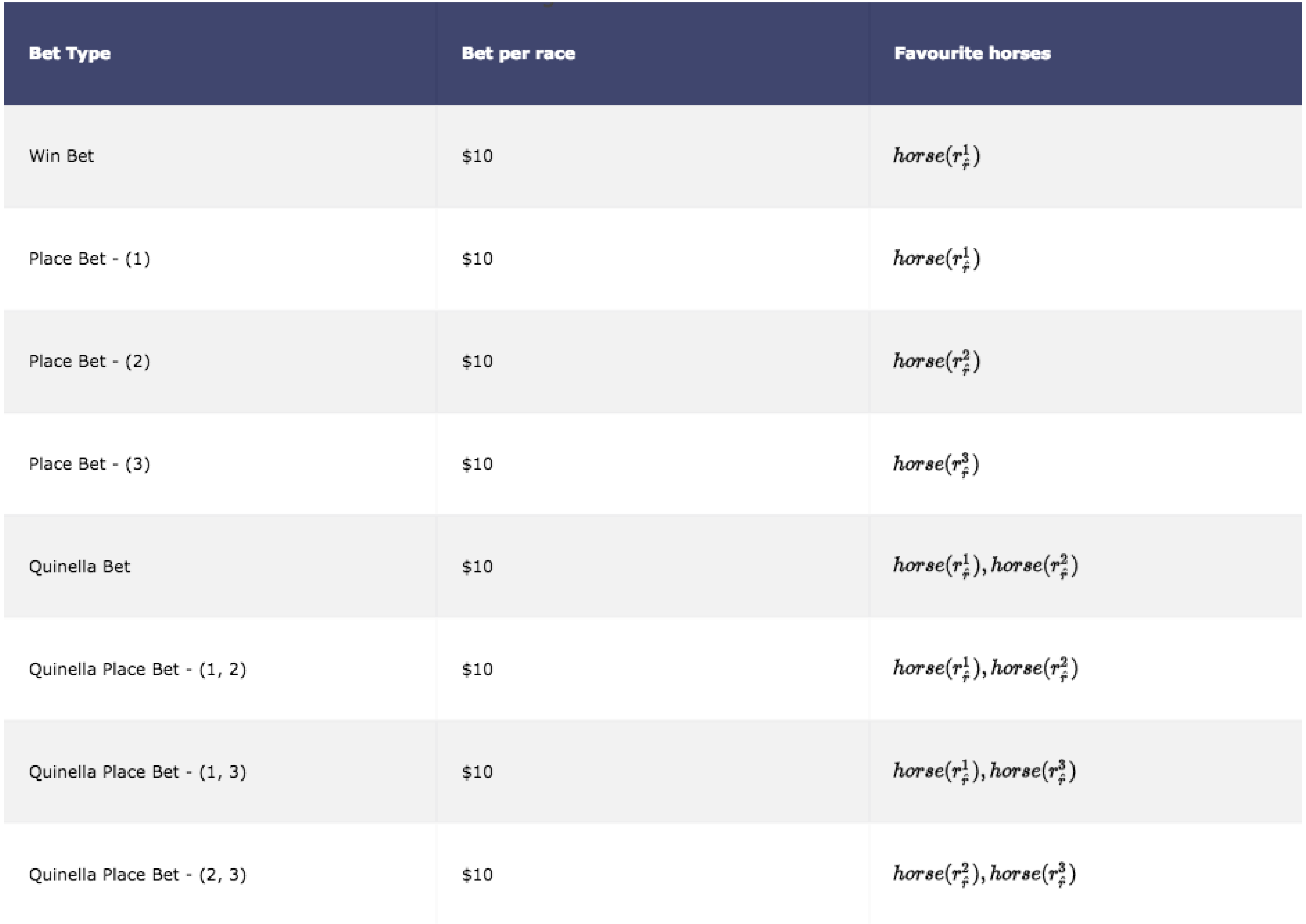


Name	K-value	Number of features	Year
Dataset 1	0	8	2005-2015
Dataset 2	1	11	2005-2015
Dataset 3	2	14	2005-2015
Dataset 4	3	17	2005-2015
Dataset 5	4	20	2005-2015
Dataset 6	5	23	2005-2015
Dataset 7	6	26	2005-2015

Evaluation Loss of 7 trained models

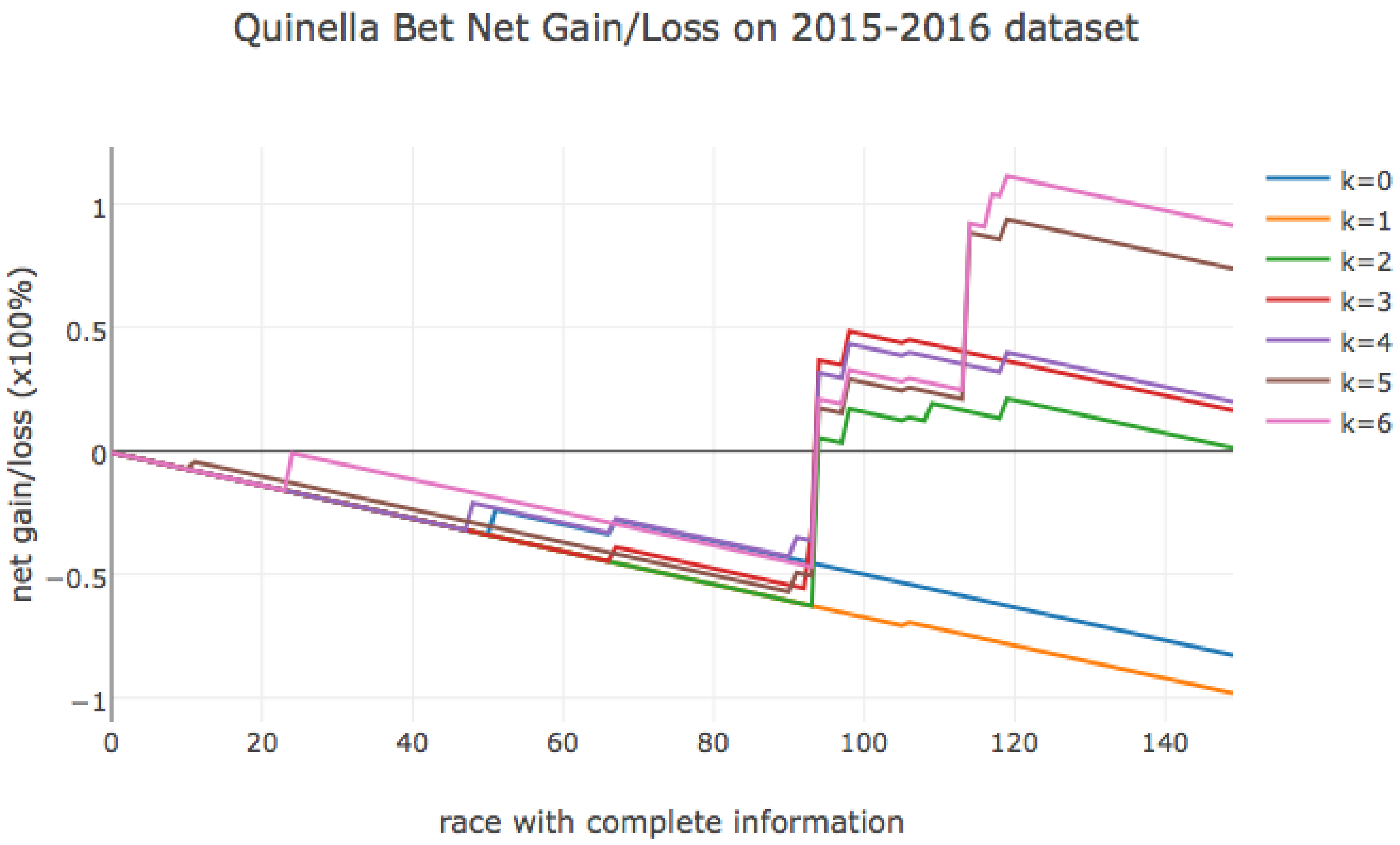


Evaluation by Race

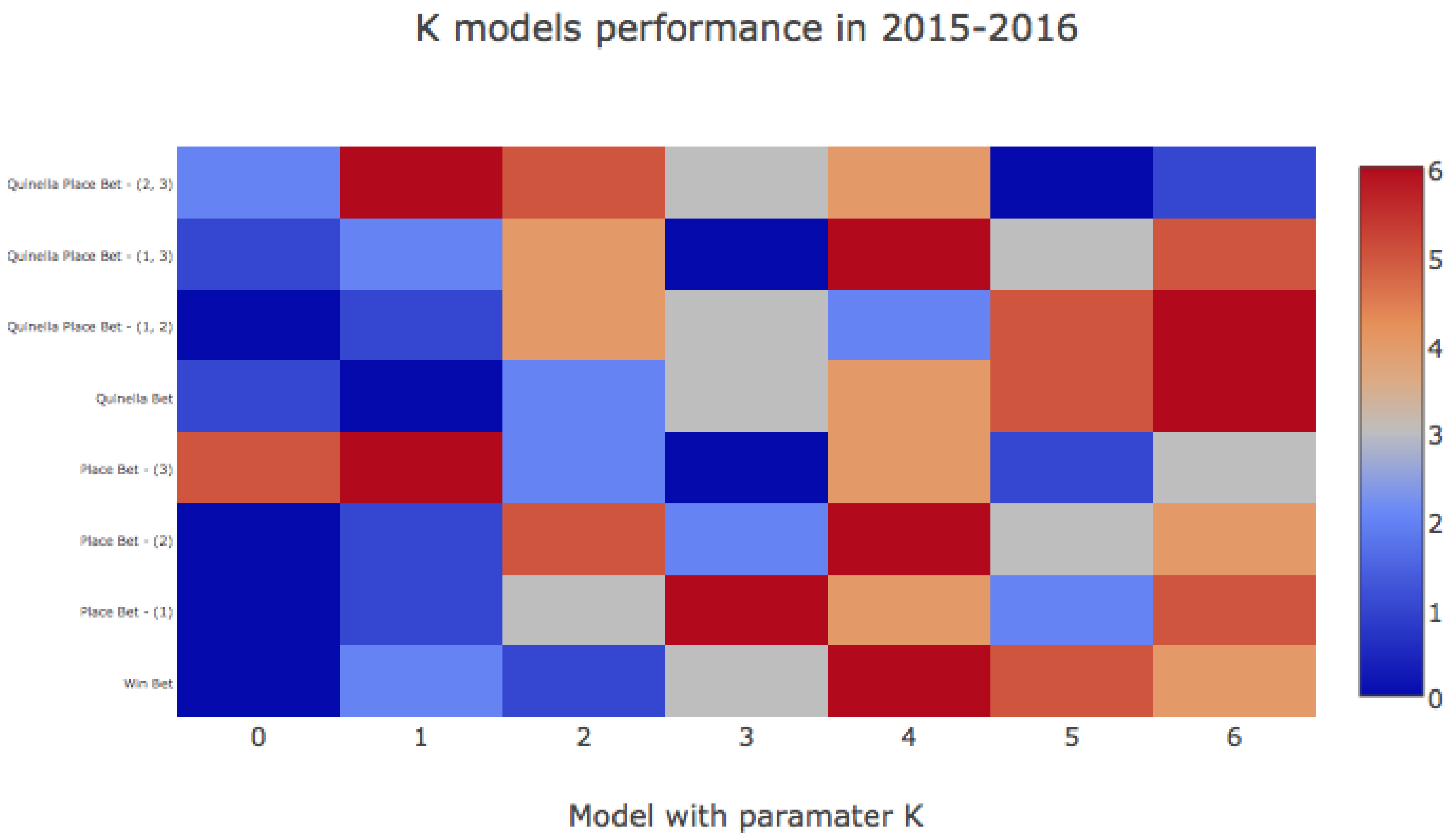


Bet Type	Bet per race	Favourite horses
Win Bet	$10	$horse(r_{\hat{r}}^{1})$
Place Bet - (1)	$10	$horse(r_{\hat{r}}^{1})$
Place Bet - (2)	$10	$horse(r_{\hat{r}}^{2})$
Place Bet - (3)	$10	$horse(r_{\hat{r}}^{3})$
Quinella Bet	$10	$horse(r_{\hat{r}}^{1}), horse(r_{\hat{r}}^{2})$
Quinella Place Bet - (1, 2)	$10	$horse(r_{\hat{r}}^{1}), horse(r_{\hat{r}}^{2})$
Quinella Place Bet - (1, 3)	$10	$horse(r_{\hat{r}}^{1}), horse(r_{\hat{r}}^{3})$
Quinella Place Bet - (2, 3)	$10	$horse(r_{\hat{r}}^{2}), horse(r_{\hat{r}}^{3})$

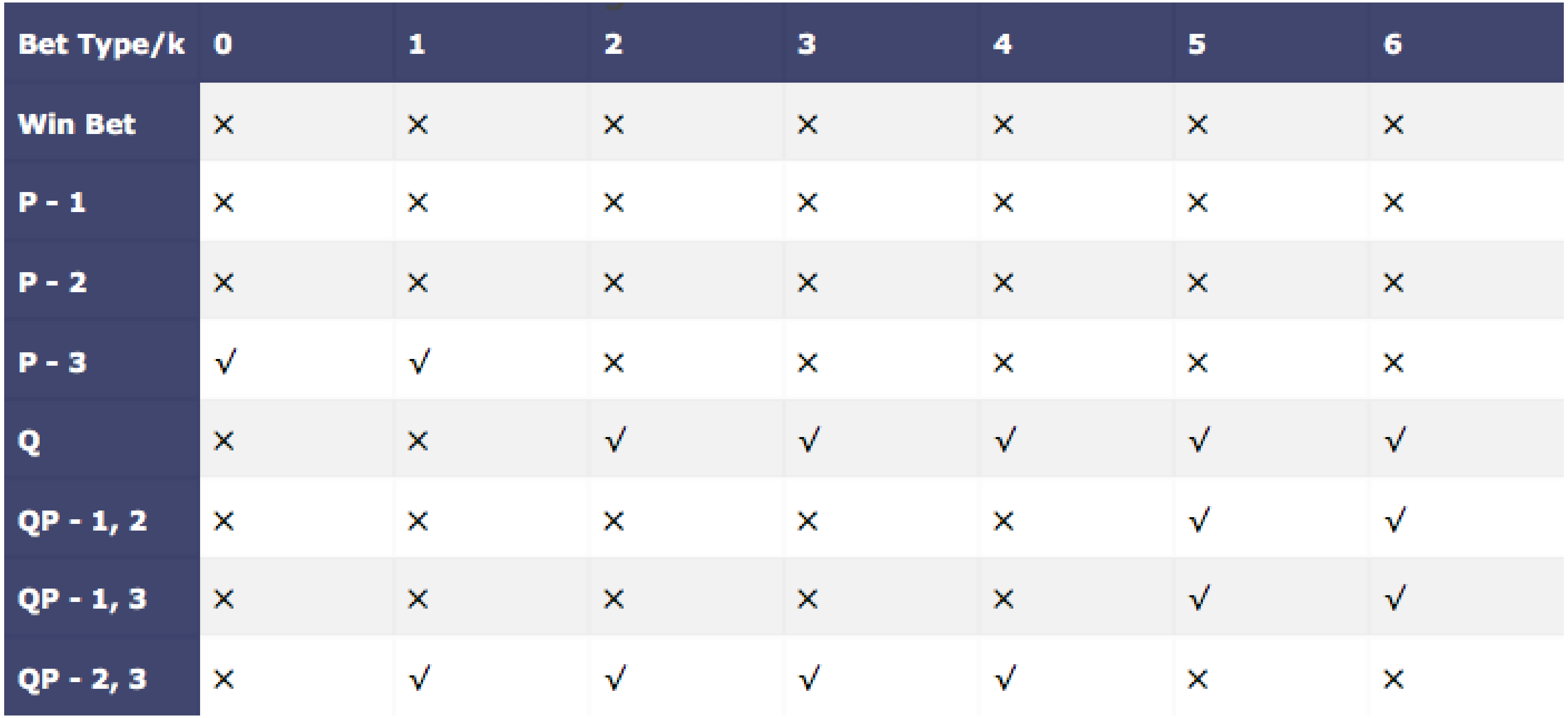
Evaluation by Race



Evaluation by Race



Overall results

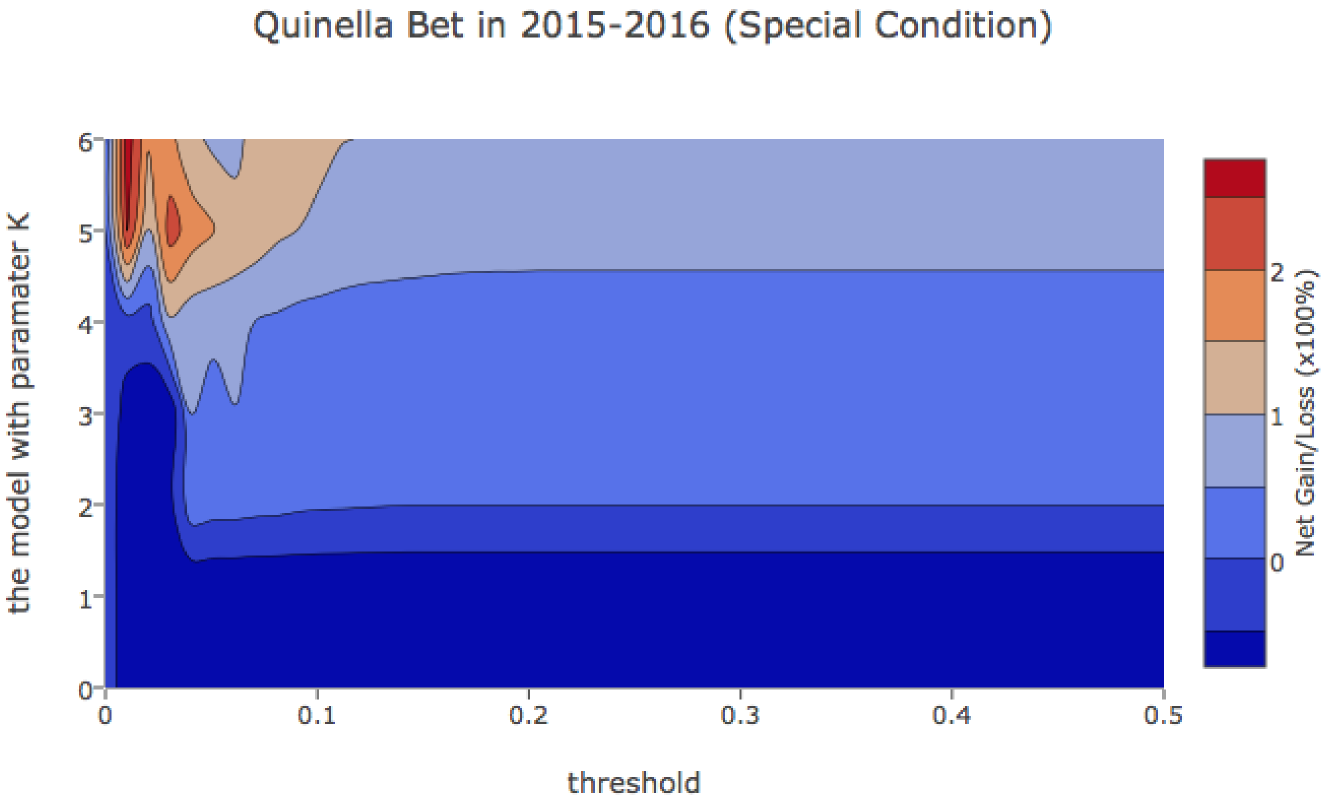


Bet Type/k	0	1	2	3	4	5	6
Win Bet	×	×	×	×	×	×	×
P - 1	×	×	×	×	×	×	×
P - 2	×	×	×	×	×	×	×
P - 3	√	√	×	×	×	×	×
Q	×	×	√	√	√	√	√
QP - 1, 2	×	×	×	×	×	√	√
QP - 1, 3	×	×	×	×	×	√	√
QP - 2, 3	×	√	√	√	√	×	×

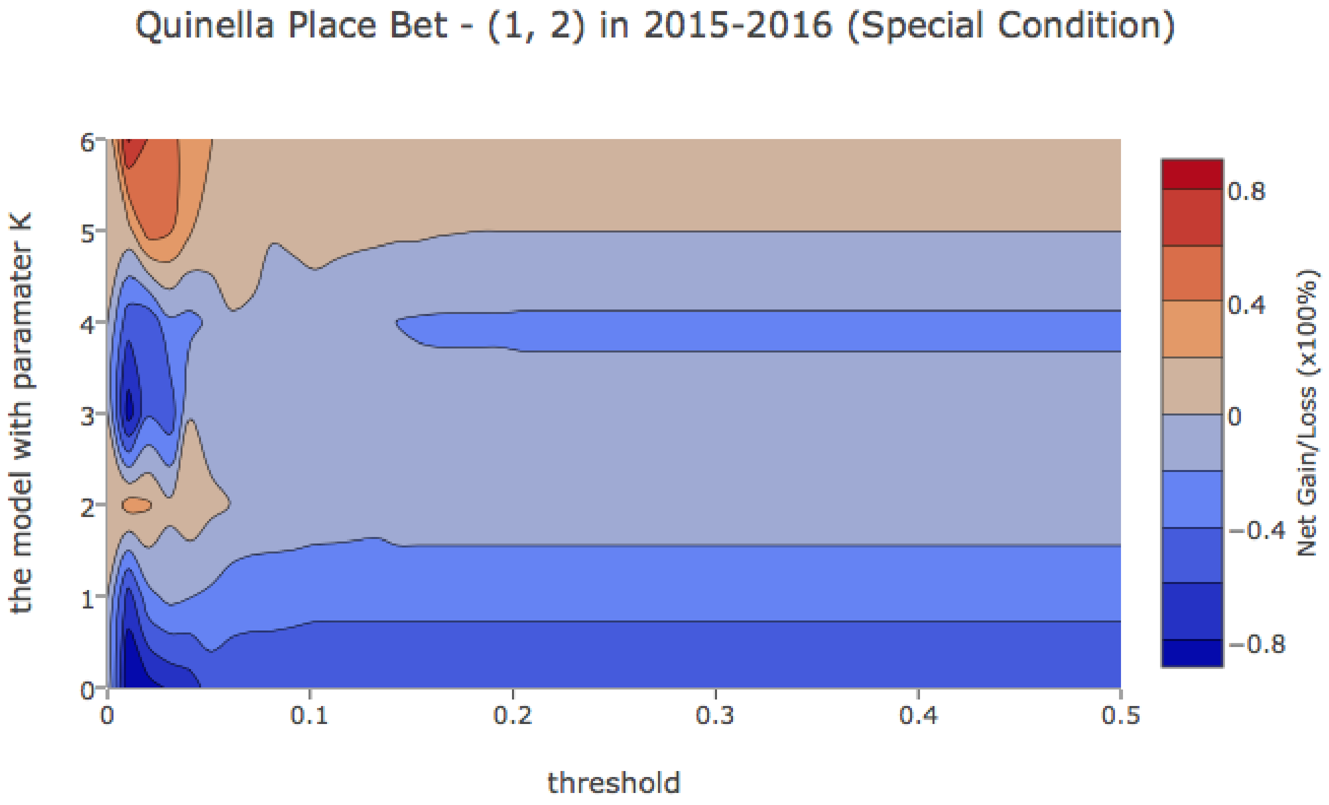
Special Condition

- $\alpha = abs(Predicted(r_{\hat{r}}^{(1)}) - Predicted(r_{\hat{r}}^{(2)}))$
- $\alpha < \varepsilon$
- $\varepsilon = threshold$

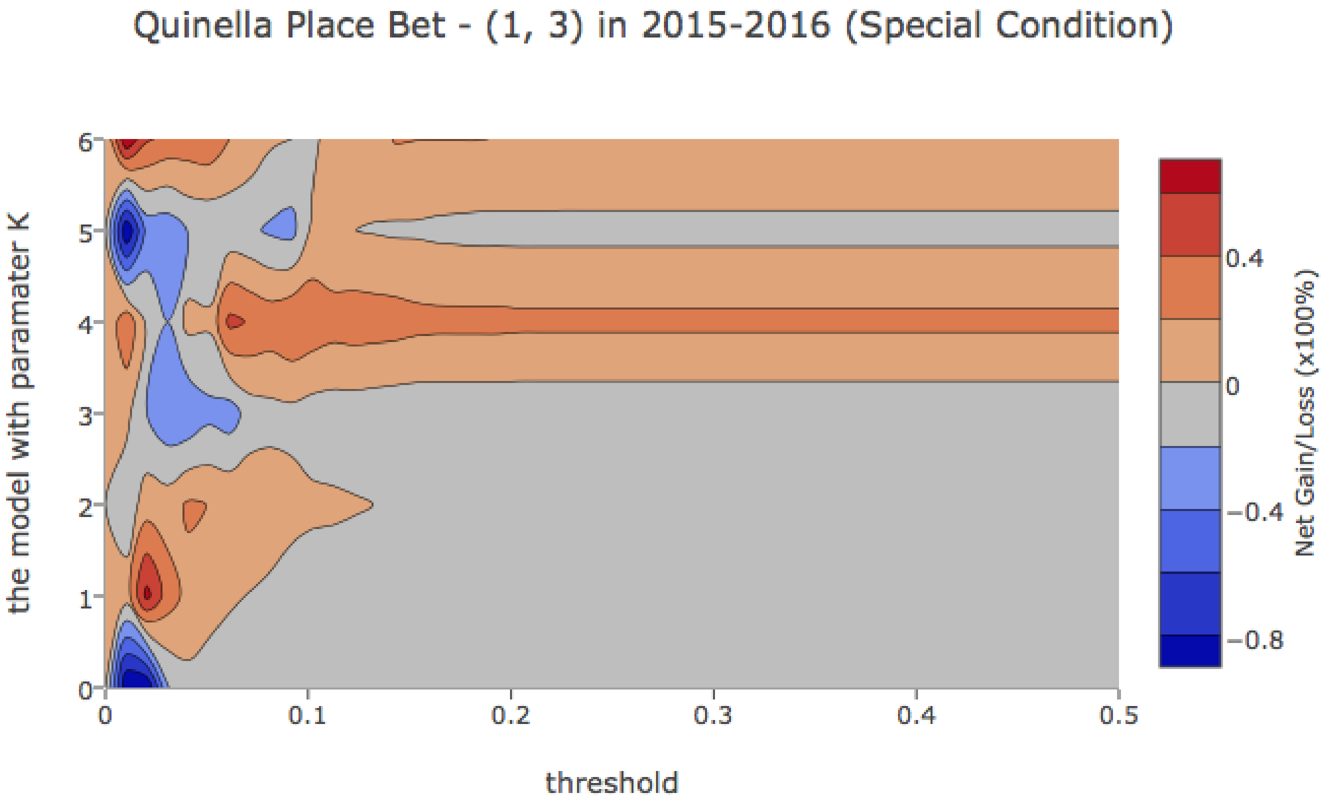
Special Condition in Bet 2015



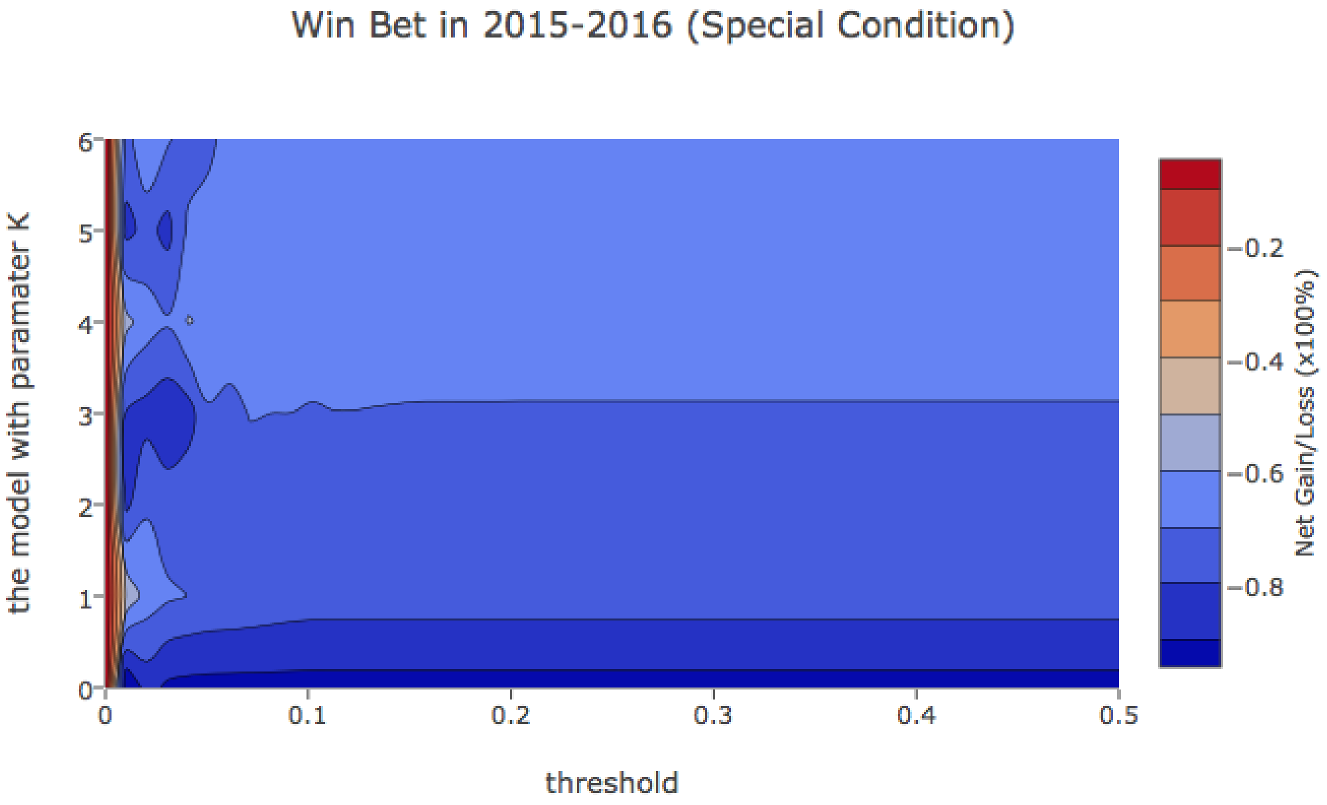
Special Condition in Bet 2015



Special Condition in Bet 2015

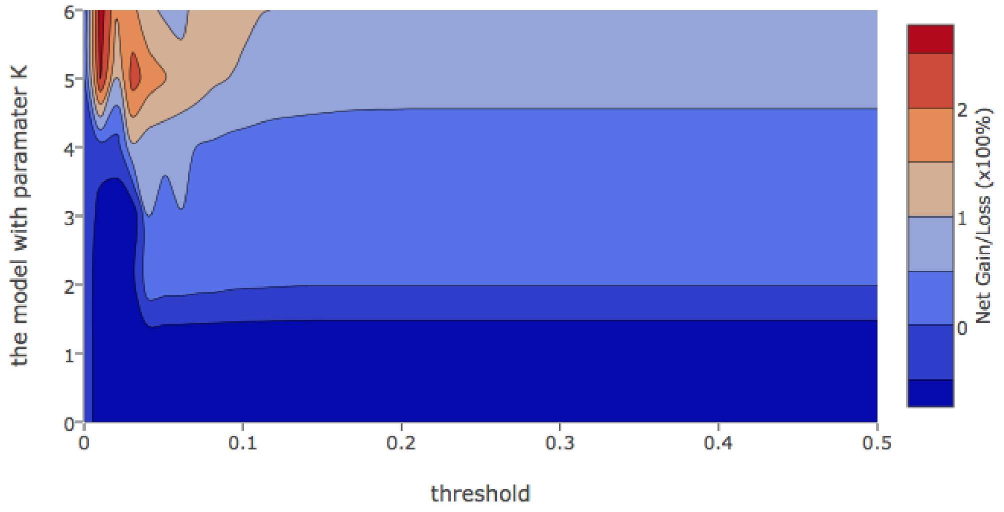


Special Condition in Bet 2015

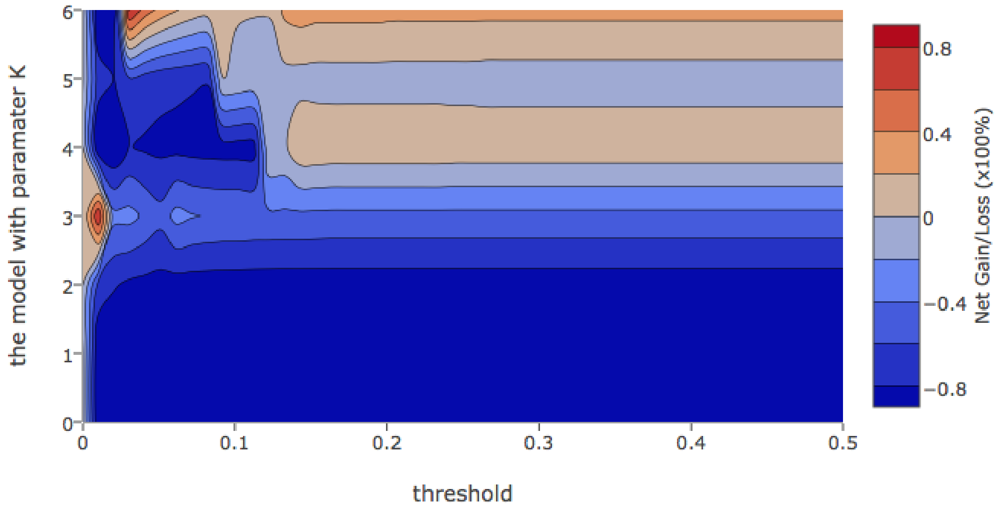


Special Condition

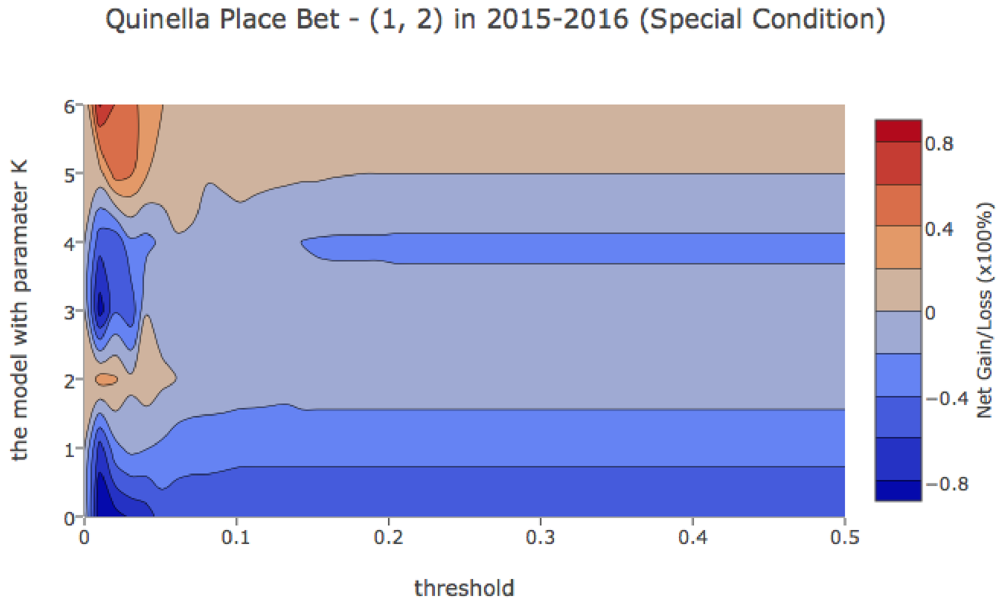
Quinella Bet in 2015-2016 (Special Condition)

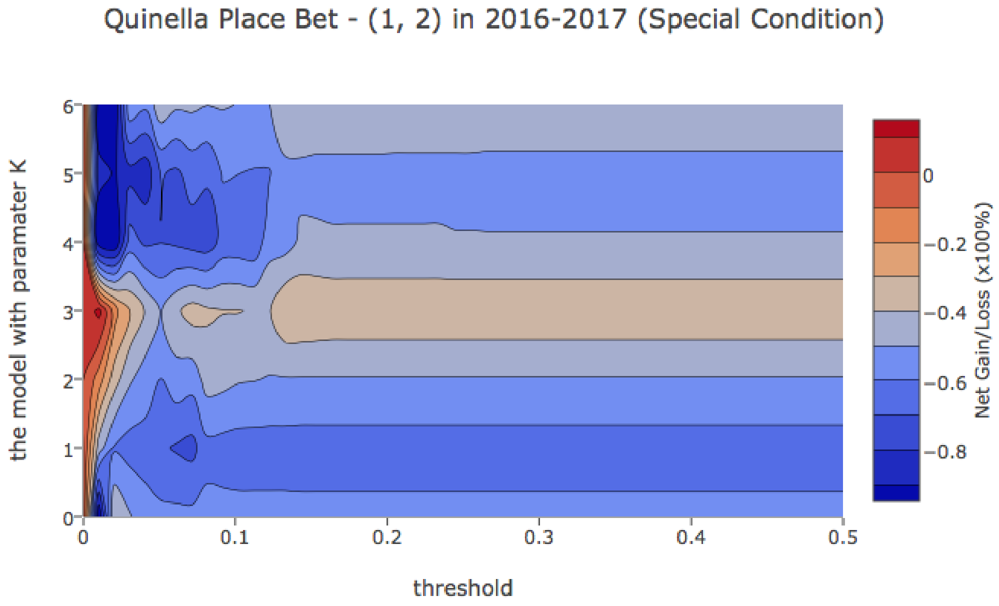


Quinella Bet in 2016-2017 (Special Condition)



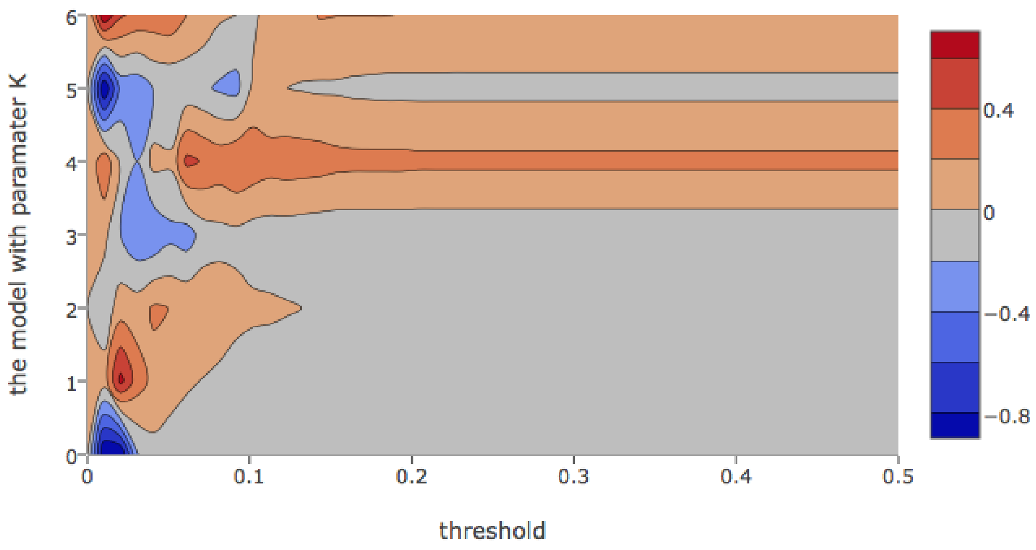
Special Condition



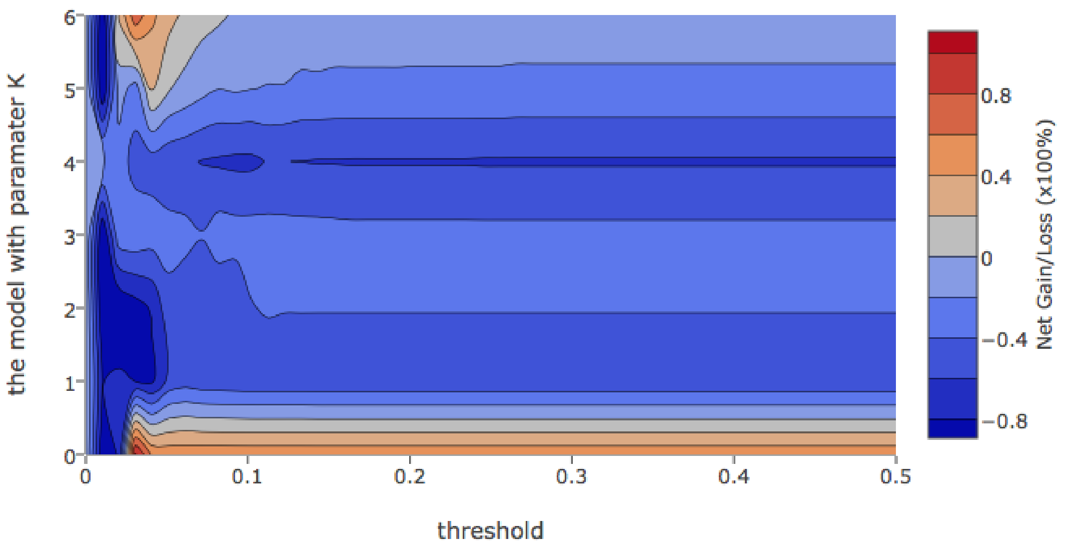


Special Condition

Quinella Place Bet - (1, 3) in 2015-2016 (Special Condition)

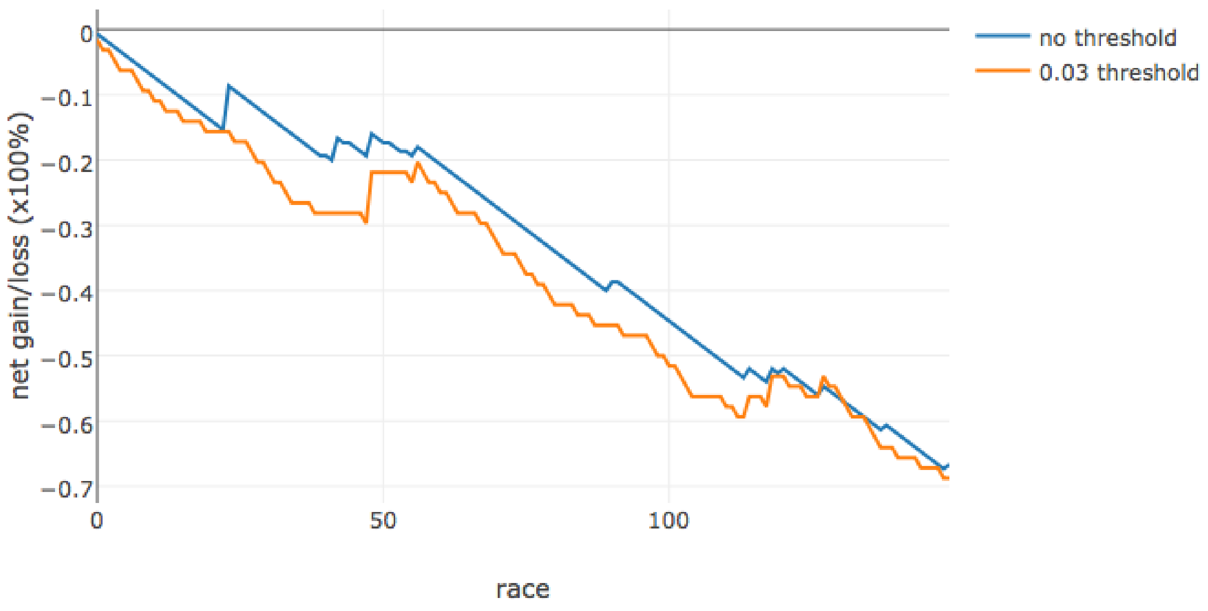


Quinella Place Bet - (1, 3) in 2016-2017 (Special Condition)

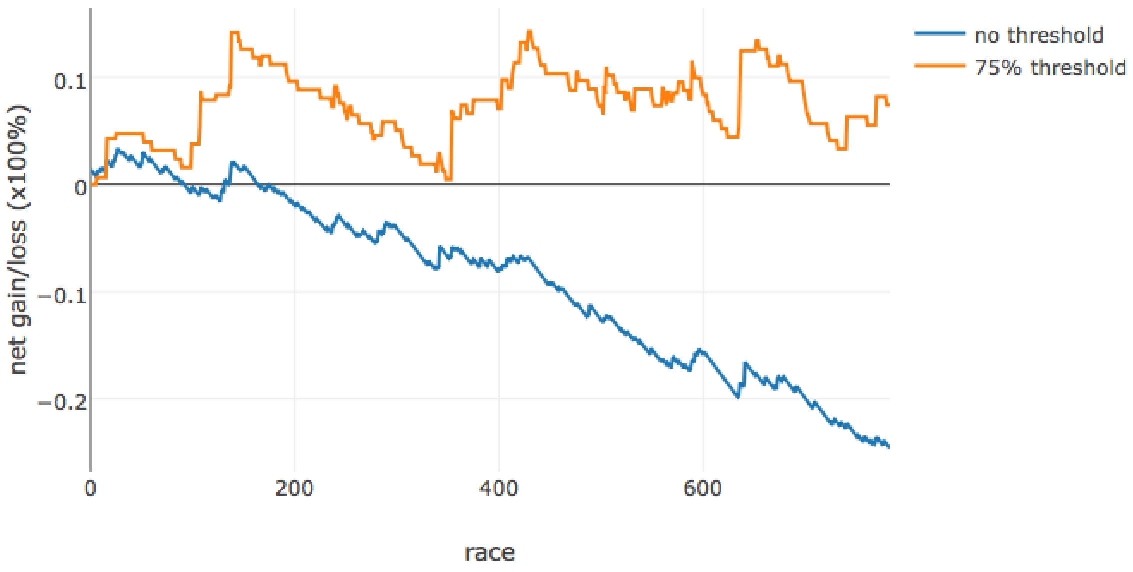


Compare with old model

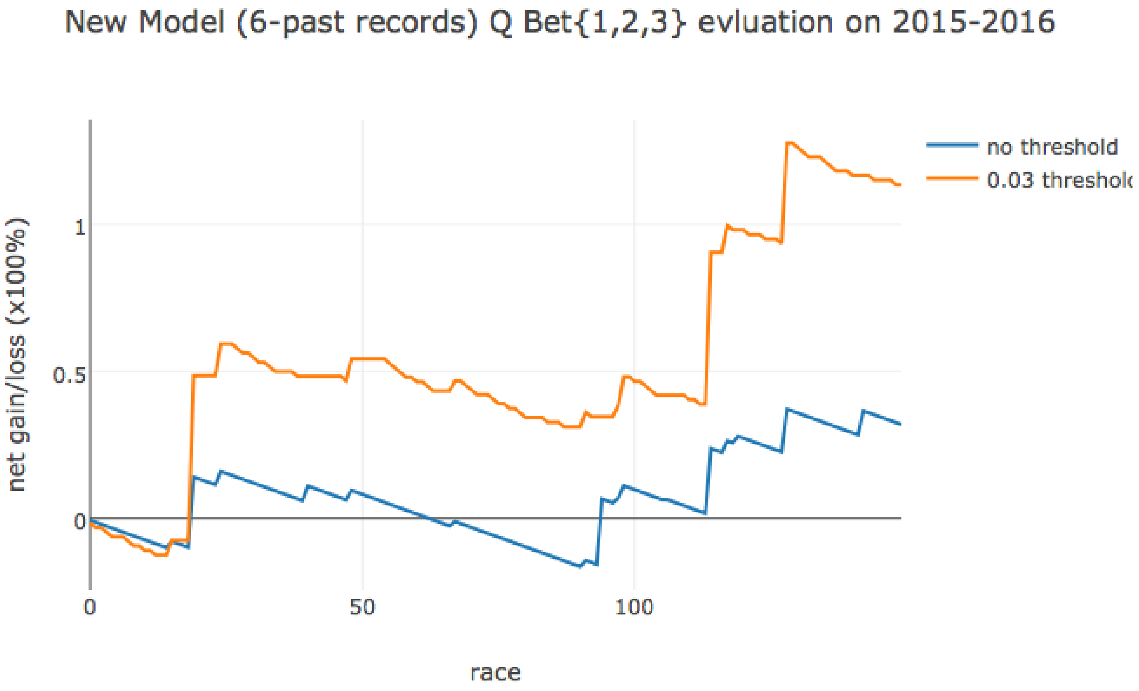
New Model (6-past records) win bet evluation on 2015-2016

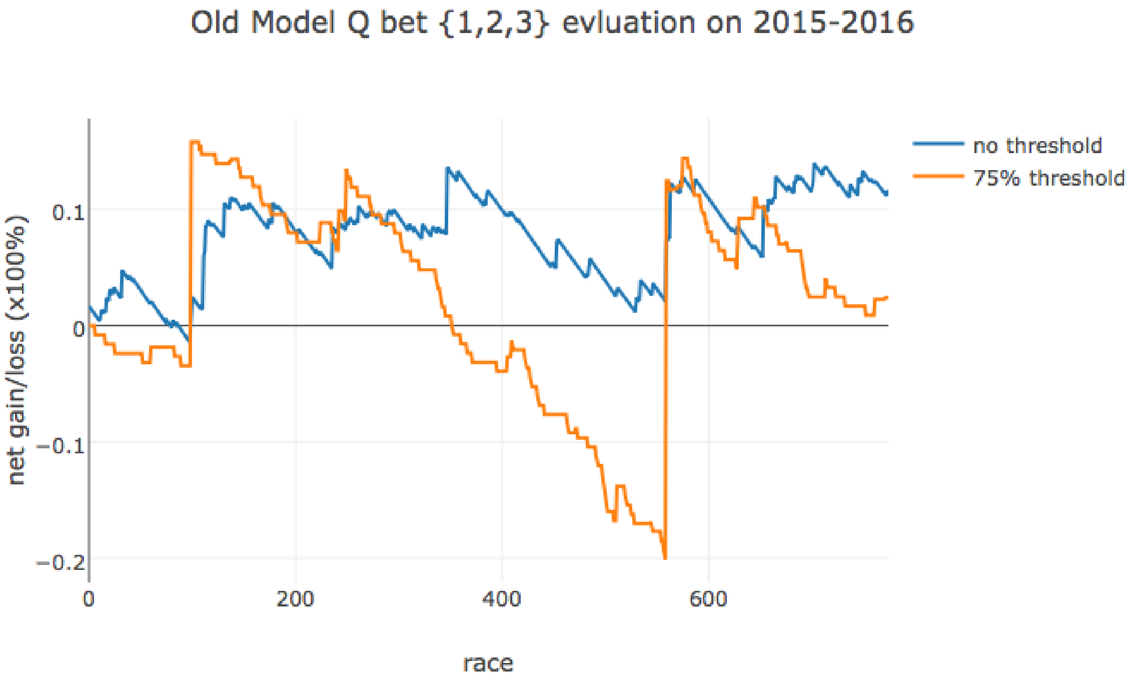


Old Model win bet evluation on 2015-2016

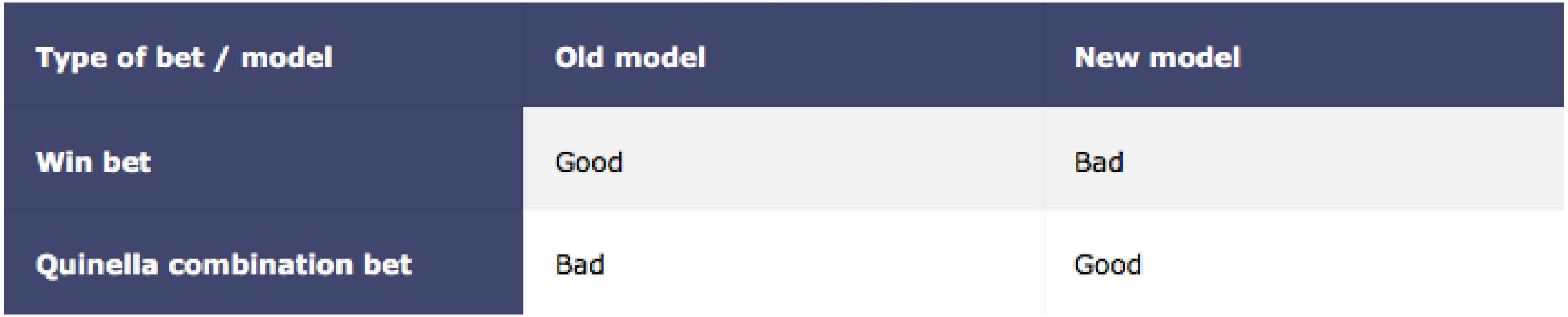


Compare with old model





Compare with old model



Type of bet / model	Old model	New model
Win bet	Good	Bad
Quinella combination bet	Bad	Good

Conclusion

- New Stardardization
- Extract past-k-records
- K-nearest-neighbor lacks of data
- Linear regression perform well on some betting methods