Estimating Soccer Player Performance with Similarity Search Regression

Max Henry

April 28, 2016

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Motivation

- Money has flooded into European soccer
  - Most expensive player deal
    - 1992: £18 million
    - 2013: £85.3 million
- Lack of competitive balance like American sports
  - Unbounded budgets
  - No requirement to “balance the books”
  - Uneven distribution of commercial revenue
  - No limit to wages

Motivation (cont.)

- But there's hope!
- 40% success rate [7]
- Problem
  - Increase success rate
  - Player impact = goals scored
- Solution
  - *k*-Nearest neighbor regression model
    - Predicts goal output based on "similar" players' goal output

Hypothesis

A *k*-nearest neighbor regression model will output perform linear regression and ridge regression in the task of predicting single season goal production from a qualitative skills data set of professional players.

Data Sets

- Proprietary, *quantitative* data sets

- Free, *qualitative* data sets

EA Sports Data

Curve

Finishing

Balance

Positioning

“Tactical” Reactions

Data Sets (cont.)

- Qualitative skills data
  - EA Sports data set
  - Last names only
  - 36 attributes == high dimensionality?
- Goal production
  - 2014/2015 season
  - Covered all 39 leagues
  - Transfermarket.co.uk
- Full names
  - Needed to cross reference goal data
  - futwiz.com

Background: *k*-Nearest Neighbor

- Lazy learning

- “Similarity” searches

- Regression

Background: Distance Metrics

- Intuitions in high dimensions [12][5]

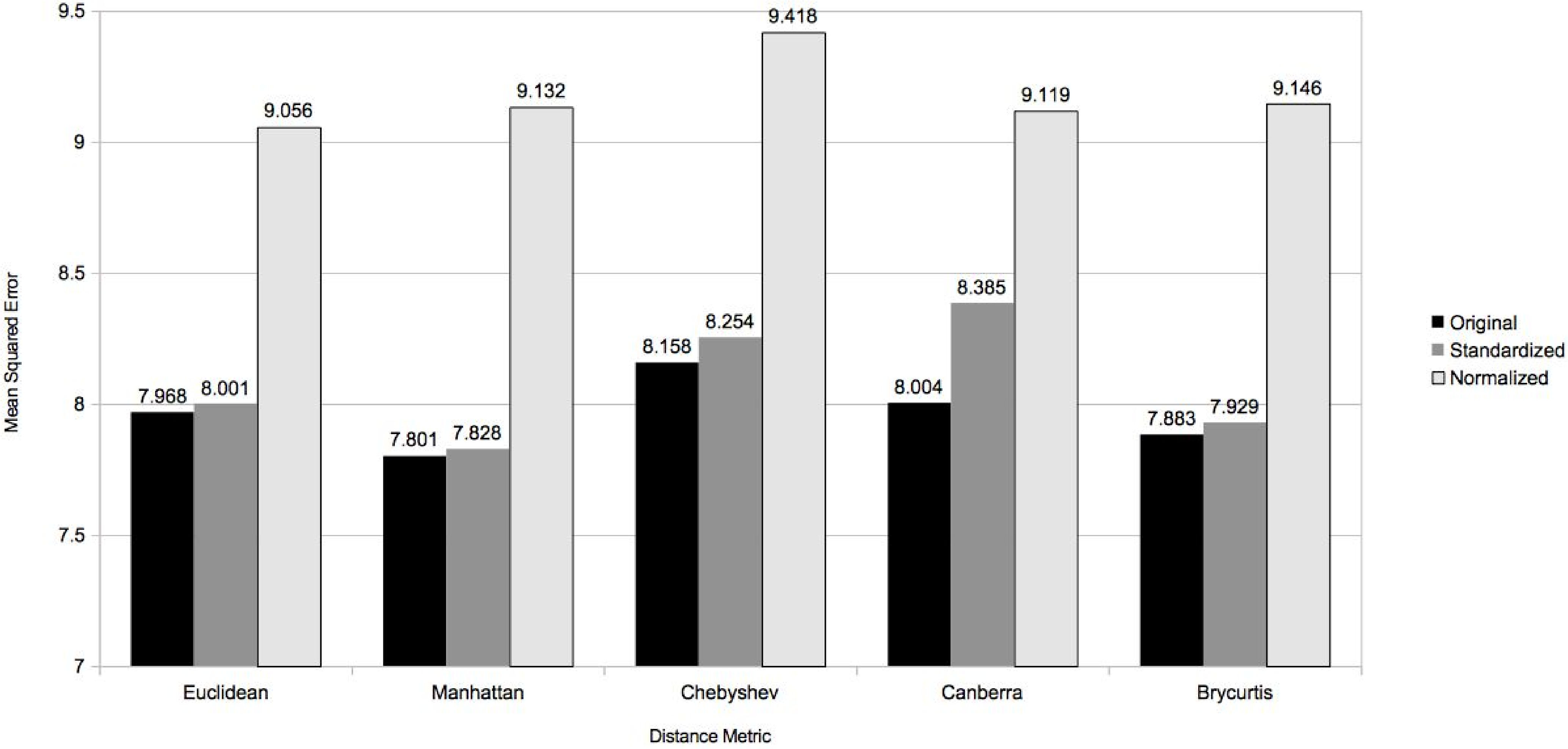
- Effect on space partitioning

- Adapt distance function or reduce data dimensionality [9]

Experimental Approach

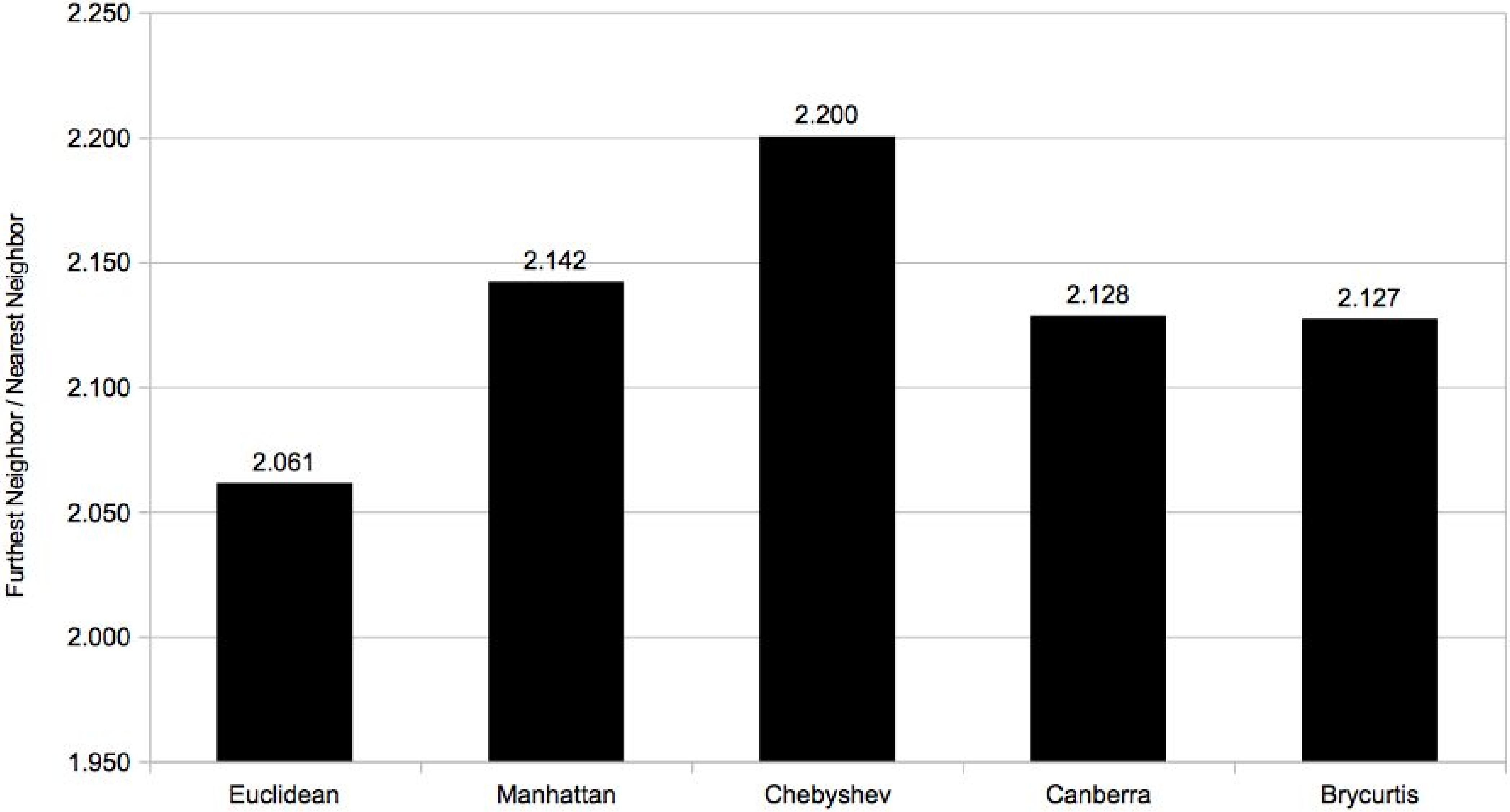
- Hyperparameter tuning
  - Distance functions
  - Feature selection
  - *k*
  - How much each neighbor contributes
- Algorithm Comparison
  - Linear regression
  - Ridge regression
  - Radius nearest neighbor
- Machine learning library ‘sklearn’

Distance Metric Comparison

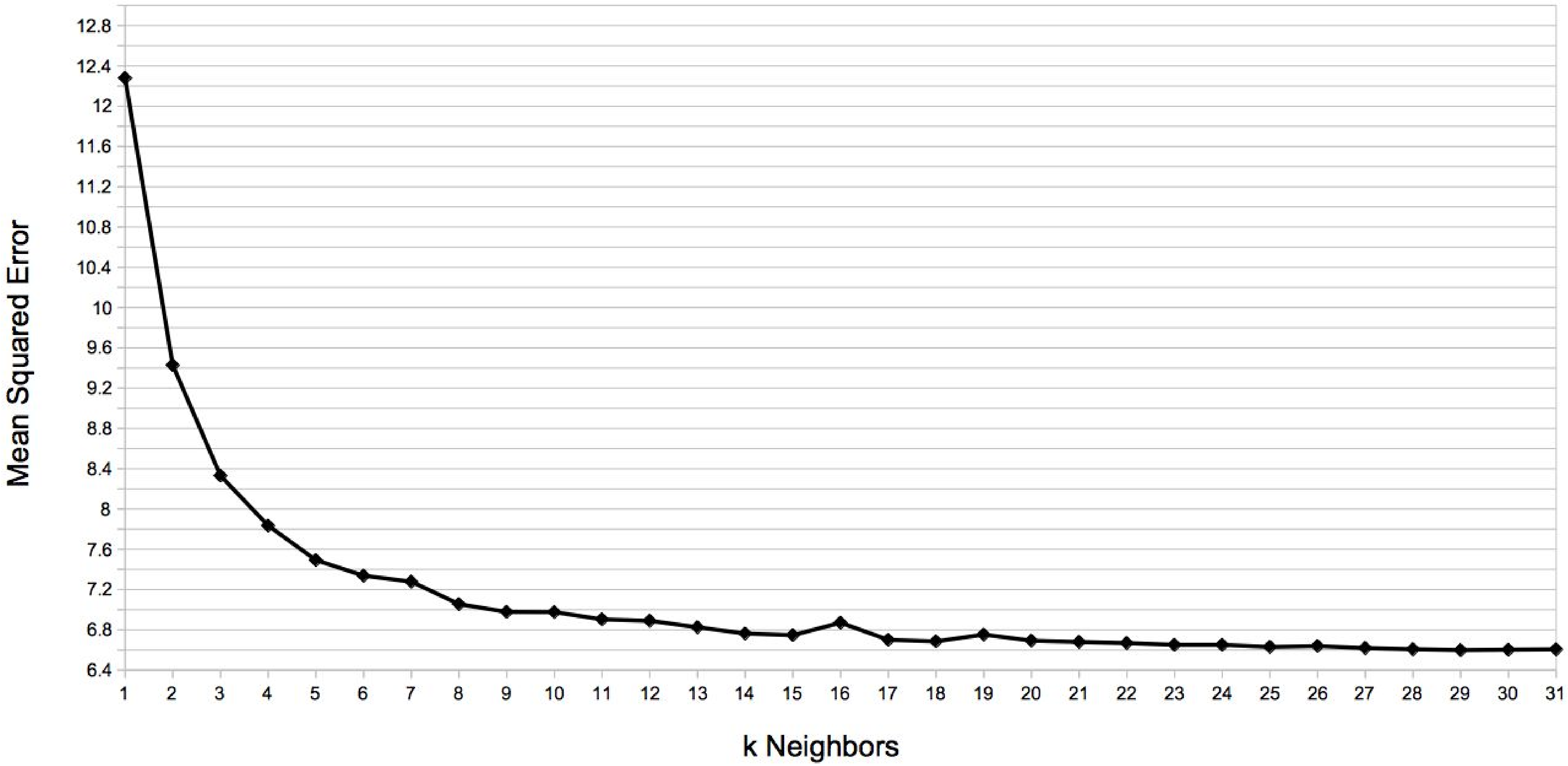


*Fractional distances performed poorly and aren’t reported

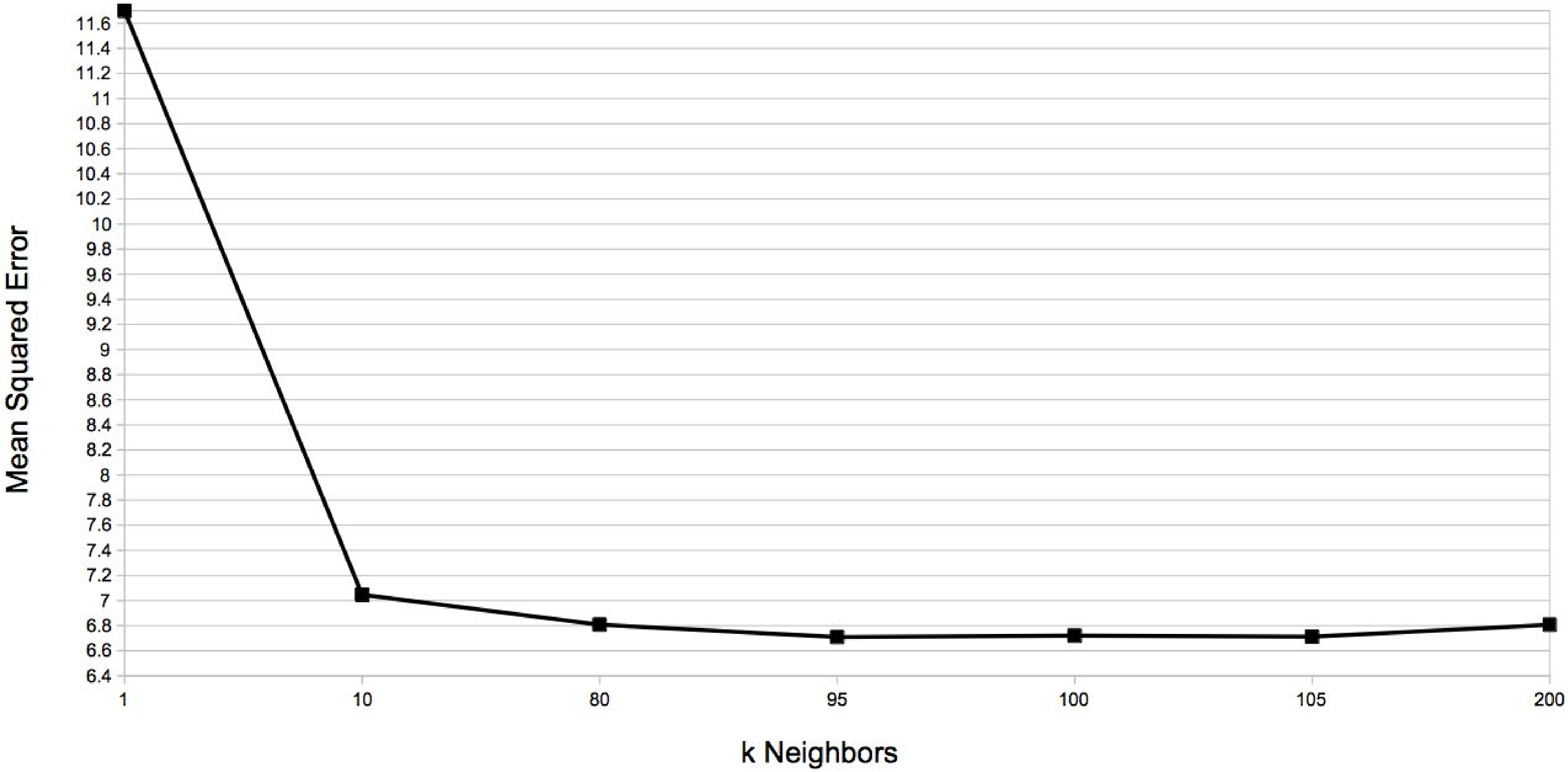
Contrast Ratio



Stepwise Forward Selection



Stepwise Backward Elimination

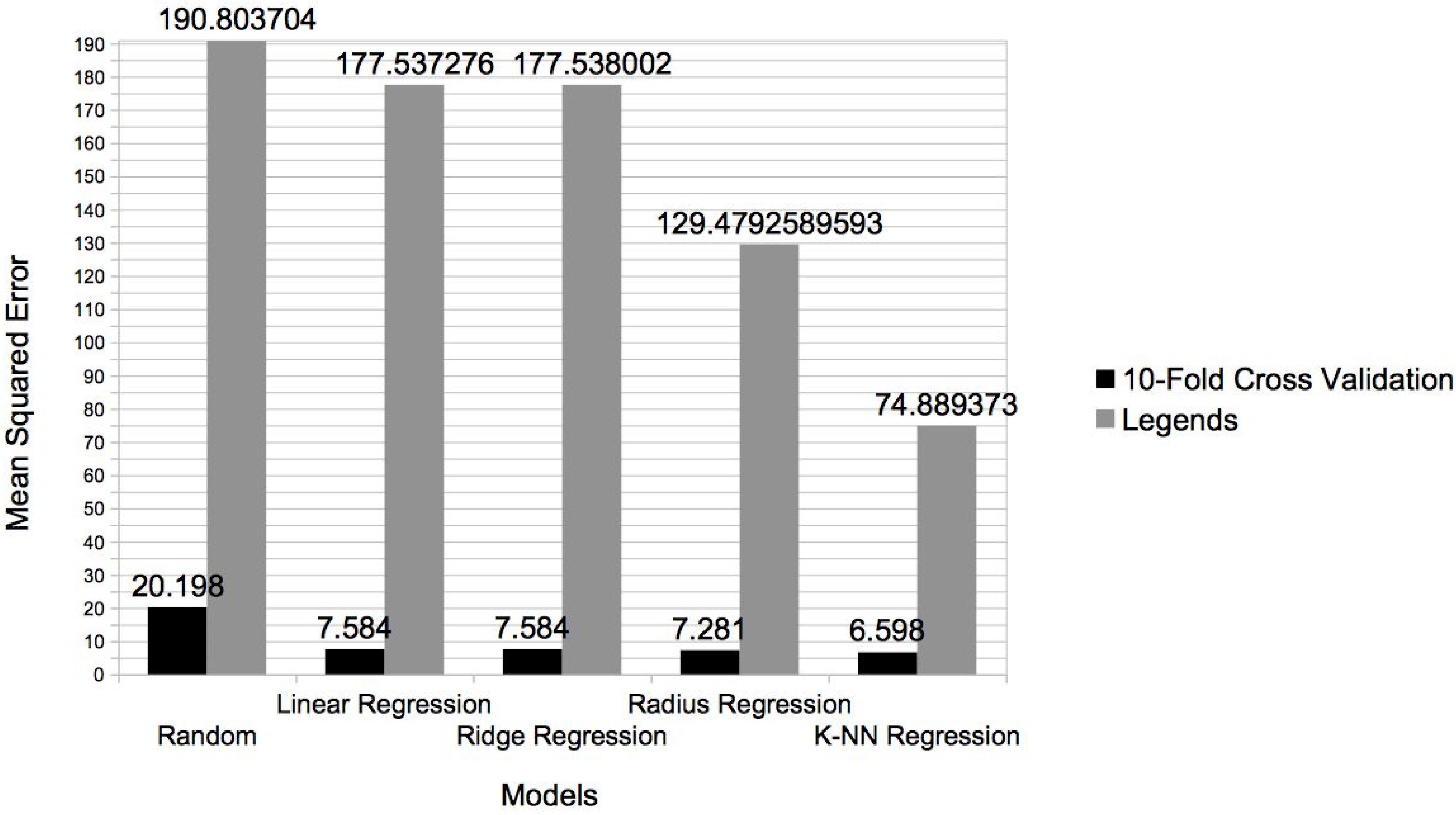


Other Hyperparameter Results

- Forward selection features + “expert”-chosen features

- Uniform weighting vs Inverse distance weighting

Linear vs. Non-linear Models



Analysis

- Qualitative features selected
- Increase in contrast
- Legendary data set
- Cross validation error

Future Work

- More feature selection
- Quantify selected features
- Weighting neighbor contributions
- MARS/LOESS

References

[1] Mitchell, Tom M. Machine Learning. New York: McGraw-Hill, 1997. Print.

[2] John Hopkins Unversity Engineering for Professionals, EN.605.746.81.SP16 Machine Learning, Dr. Sheppard, ”Instance-based Classification”, p. 93

[3] Yingying, L. I., Silvia Chiusano, and Vincenzo DElia. ”Modeling athlete performance using clustering techniques.” The Third International Symposium on Electronic Commerce and Security Workshops (ISECS 2010). 2010. 13 Henry

[4] ”Hilltop Analytics.” Hilltop Analytics. Web. 13 Mar. 2016. ¡http://www.hilltopanalytics.com/football/find-me-a-player-like-andres-iniesta/

[5] White, D. (n.d.). Liverpool owners looking to use baseball principles of statistical analysis. Retrieved March 13, 2016

[6] Gibson, O. (2015). Sky and BT retain Premier League TV rights for record 5.14bn. Retrieved March 13, 2016

[7] Tomkins’ Law: Only 40% of Transfers Succeed. (2014). Retrieved March 13, 2016, from https://tomkinstimes.com/2014/06/tomkins-law-only-40-of-transfers-succeed/

References

[8] Berkhin, Pavel. ”A survey of clustering data mining techniques.” Grouping multidimensional data. Springer Berlin Heidelberg, 2006. 25-71.

[9] Aggarwal, Charu C., Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional space. Springer Berlin Heidelberg, 2001.

[10] Hinneburg, Alexander, Charu C. Aggarwal, and Daniel A. Keim. ”What is the nearest neighbor in high dimensional spaces?.” (2000).

[11] Aggarwal, Charu C. ”Re-designing distance functions and distance-based applications for high dimensional data.” ACM Sigmod Record 30.1 (2001): 13-18.

[12] Domingos, Pedro. ”A few useful things to know about machine learning.” Communications of the ACM 55.10 (2012): 78-87.

[13] Bao, Yongguang, Naohiro Ishii, and Xiaoyong Du. ”Combining multiple k-nearest neighbor classifiers using different distance functions.” Intelligent Data Engineering and Automated LearningIDEAL 2004. Springer Berlin Heidelberg, 2004. 634-641.