Pitch Prediction for Jacob deGrom

Jake Diamond, Jiatong Shi, Arunjyoti Sinha Roy

Introduction to Data Science Final Project

May 11th, 2021

Project Overview

- Jacob deGrom is a professional baseball player for the New York Mets
- He is one of the best pitchers in baseball
  - 3x All-star (2015, 2018, 2019)
  - 2x NL Cy Young Award (2018, 2019)
  - 2x NL strikeout leader (2019, 2020)
- **Can we predict what pitch he will throw?**

Data Cleaning

- Data set taken from MLB.com's Statcast database
- 18,648 observations with 92 features each
  - Data for every pitch deGrom has thrown in his entire career
- Clean data is reduced to 15,853 observations with 27 useful features

- Input features on game situation
  - Number of balls/strikes
  - Men on base
  - Score of game
  - Etc... 24 others
- Want to predict pitch class
  - 4-seam fastball
  - Slider
  - Changeup
  - 2-seam fastball
  - Curveball

Pitch prediction by classification

- The intuitive way for the problem is to train a classifier
- We adopt a wide range of classifier and preprocessing techniques to explore the best solution to the problem.
- **Classifiers**: KNN, DecisionTree, random forest, AdaBoost, Naïve Bayes, QDA and LDA
- **Preprocessing**: Standard Scaler, MinMax Scaler, Quantile Transform, Normalizer, Polynomial Feature expansion, and whitening
- **Dimension reduction**: PCA

Results for classification

Accuracy	AdaBoost	DecisionTree	GaussianNB	KNeighbors	LDA	QDA	RandomForest	Grand Total
MinMaxScaler	**14.10%**	**44.33%**	**37.90%**	**39.04%**	**44.03%**	**36.53%**	**40.69%**	**36.66%**
None	11.90%	44.57%	32.22%	39.68%	44.02%	31.30%	40.58%	34.90%
PCA	16.31%	44.09%	43.59%	38.40%	44.03%	41.77%	40.79%	38.43%
None	**15.82%**	**44.43%**	**39.11%**	**39.47%**	**44.13%**	**39.51%**	**40.09%**	**37.51%**
None	11.90%	44.57%	34.87%	39.48%	44.02%	38.02%	40.22%	36.16%
PCA	19.75%	44.29%	43.35%	39.46%	44.23%	41.00%	39.95%	38.86%
Normalizer	**35.12%**	**44.27%**	**40.52%**	**39.29%**	**43.96%**	**38.69%**	**40.90%**	**40.39%**
None	35.80%	44.42%	38.41%	39.34%	43.80%	36.47%	41.60%	39.98%
PCA	34.43%	44.12%	42.63%	39.24%	44.12%	40.92%	40.20%	40.81%
PCA	**22.98%**	**43.76%**	**42.91%**	**40.00%**	**44.05%**	**41.41%**	**40.93%**	**39.43%**
None	27.44%	43.88%	42.12%	40.13%	44.03%	40.09%	41.42%	39.87%
PCA	18.53%	43.63%	43.69%	39.86%	44.07%	42.73%	40.44%	38.99%
PolynomialFeatures	**21.74%**	**44.21%**	**30.55%**	**39.02%**	**43.98%**	**24.79%**	**39.87%**	**34.88%**
None	16.30%	44.53%	23.65%	39.36%	44.28%	16.04%	40.36%	32.07%
PCA	27.18%	43.89%	37.46%	38.68%	43.68%	33.55%	39.38%	37.69%
QuantileTransformer	**13.36%**	**44.30%**	**38.85%**	**39.40%**	**44.01%**	**41.77%**	**39.65%**	**37.33%**
None	11.90%	44.57%	33.50%	40.02%	43.93%	40.23%	39.83%	36.28%
PCA	14.82%	44.03%	44.21%	38.78%	44.09%	43.32%	39.47%	38.39%
StandardScaler	**18.18%**	**44.19%**	**37.53%**	**40.16%**	**43.82%**	**35.57%**	**40.36%**	**37.12%**
None	11.90%	44.57%	32.22%	40.35%	44.02%	29.74%	40.41%	34.74%
PCA	24.46%	43.80%	42.83%	39.98%	43.61%	41.40%	40.32%	39.49%
Grand Total	**22.05%**	**44.22%**	**38.49%**	**39.46%**	**43.99%**	**37.12%**	**40.42%**	**37.96%**

Re-consider the problem with time-information

- The current pitch might have some correlations with recent pitches' information.
- Therefore, we adopted a sliding window with very recent pitches and its' features

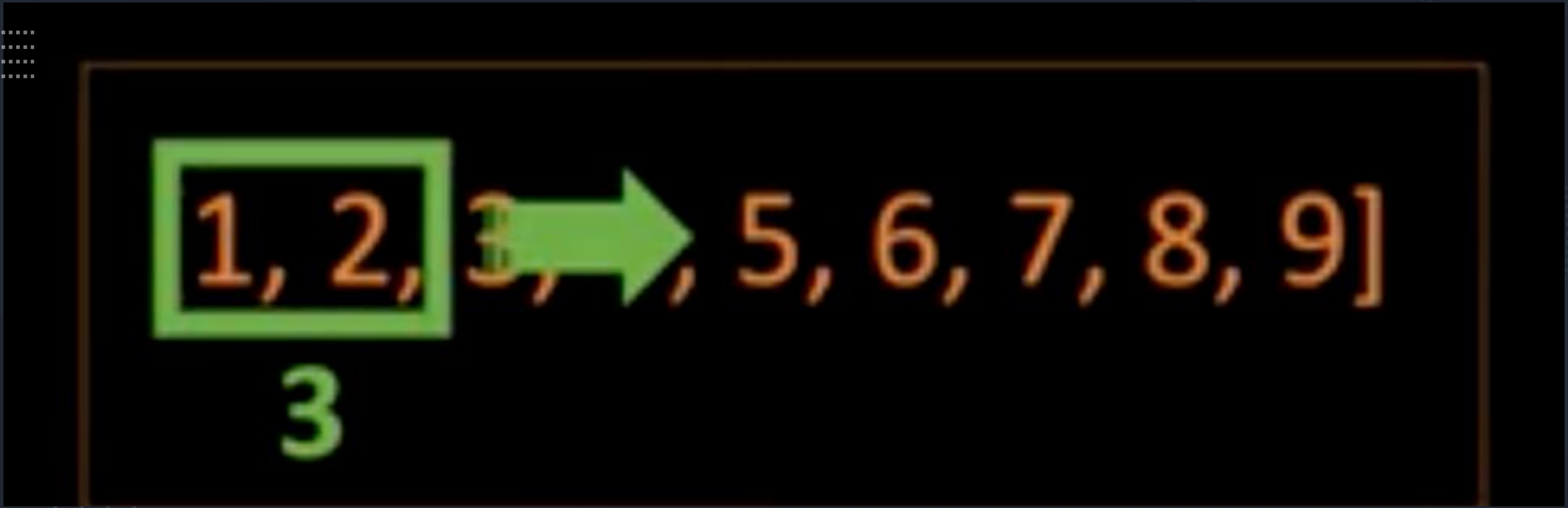


Figure from https://quanticdev.com/algorithms/dynamic-programming/sliding-window/

Results with Sliding Window

- The table presents each window length configuration with its best system

Window Length	Preprocessing	Dimension Reduction	Classifier	Accuracy
0	StandardScaler	None	Decision Tree	44.57%
3	StandardScaler	PCA	LDA	**44.80%**
5	Quantile Transformer	PCA	LDA	44.48%

Pitch prediction by clustering

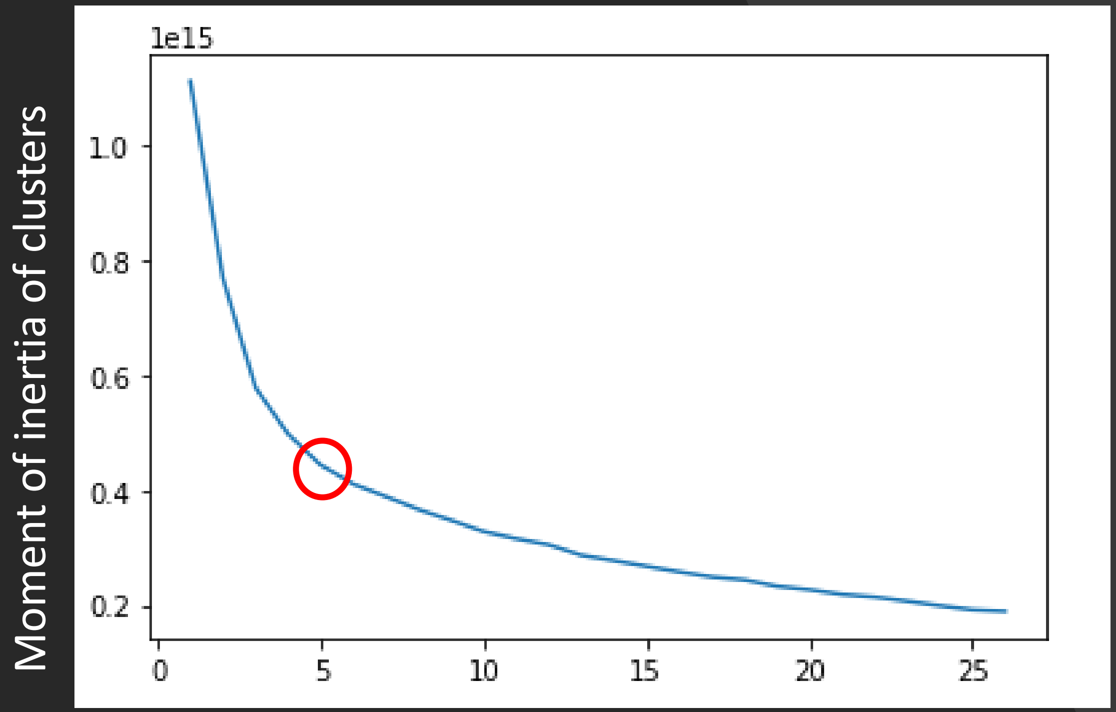
Best cross-validation score, amongst all the classifiers, was achieved with Linear Discriminant Analysis (~44.5%)

Attempt to improve the prediction accuracy using cluster-then-predict model

Computed the optimal number of clusters using the elbow region of the moment of inertia plot of the clusters

For all the clustering algorithms, the number of clusters was assumed to be 5.

Determination of appropriate number of clusters

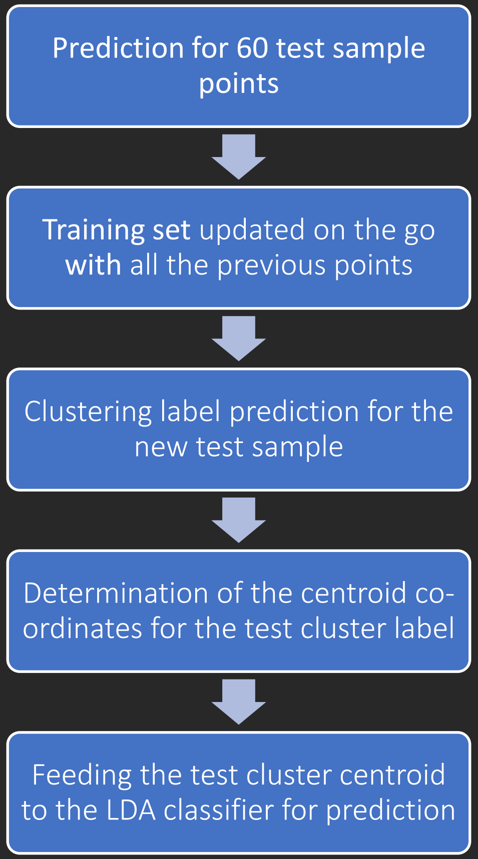


Number of clusters

Pitch prediction by clustering

- Types of clustering used:
  1. K-Means
  2. K-Medians
  3. Gaussian Mixture model
  4. Spectral Clustering
- Classifier used for prediction:
  - ➢ Linear Discriminant Analysis

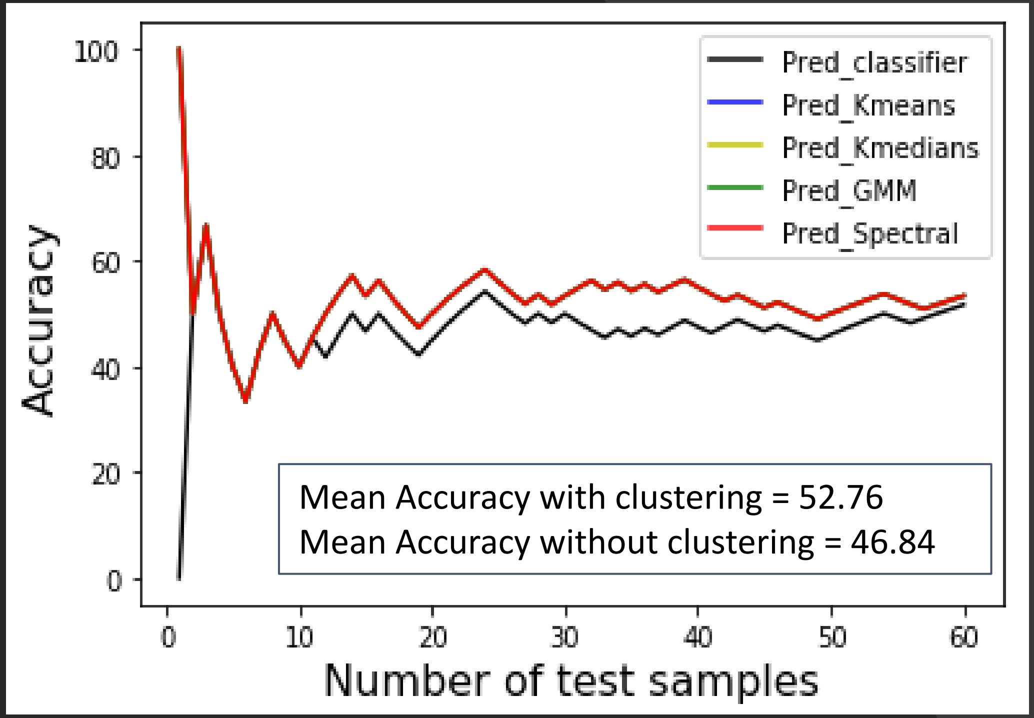
Cluster-then-predict approach



Pitch prediction by clustering

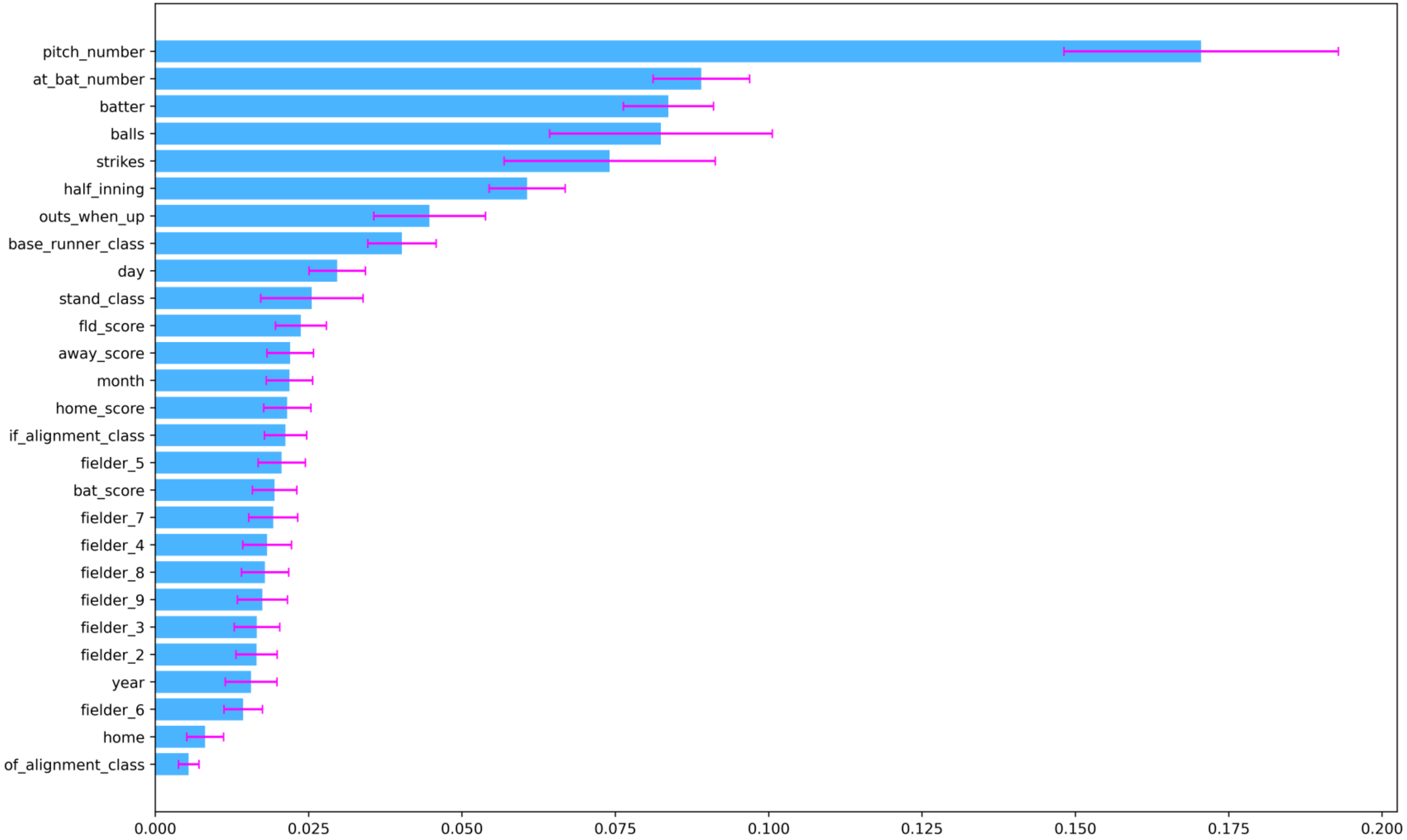
- Observations
  - ➢ Different clustering techniques, followed by the LDA prediction led to the same accuracy levels.
  - ➢ For 12 or more test samples, cluster-then-predict method yields better accuracy.
- Possible reason
  - ➢ For this problem, since LDA performs best prediction, the classes can be best linearly demarcated (i.e., clusters do not overlap too much).
  - ➢ Clustering before prediction adds the attribute of the cluster as an additional feature for better accuracy (i.e., outlier effect of some features which might wrongly influence the prediction is reduced by clustering).

Prediction Accuracy comparison



Feature Importance

- Feature importance determined by random forest classifier
- Surprising result: pitch number is by far the most important feature



Breakdown by Pitch Number

Conclusions

- The best classifier reaches 44.80% accuracy using standard-scaler, PCA, and LDA as a pipeline. The features are extracted using a sliding window of size 3
- For our problem, it is better to perform clustering before prediction using the classifier to improve the prediction accuracy
- Pitch number is the most important feature for classifying pitches
  - Not a great predictor on its own