Project 4 Team 6

MLB: Data Analysis

Cherno Jallow, Jason McHone, Kevin Ybarra, Omar Espinoza & Paul Brichta

Index

- ✔ Intro
- ✔ Data
- ✔ Classification
- ✔ Regression Analysis
- ✔ Site Demo
- ✔ Challenges - Next Steps
- ✔ Questions

Road Trip

Introduction to Our Project

Questions we were trying to answer:

- Can unsupervised learning classify teams using the data we have available?
- Could regression analysis be used to determine the anticipated number of wins in a season?
- What teams have performed best historically?

The Data

Source: Lahman's Baseball Database

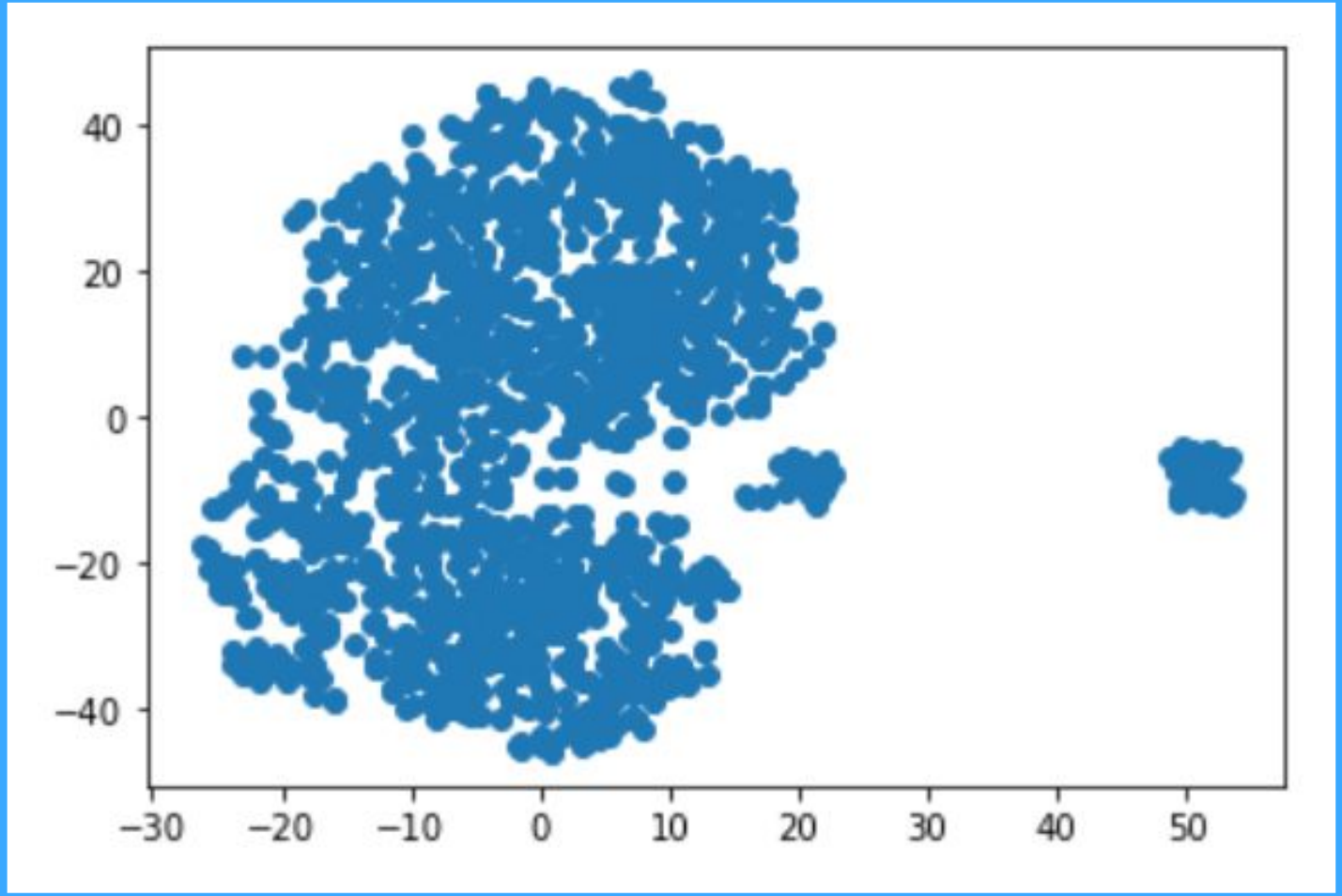
https://www.openintro.org/data/index.php?data=mlb_teams

Team Season TTL Stats by Year (1876 - 2020) 40 Columns of data per year per team (ex: Wins, Losses, Errors, Strikeouts By Pitcher, Saves etc) Initially limited to 1946-2020

Data Cleaned:
Nulls Removed & Years were narrowed from 1970 to 2019 (batters hit & sacrifice flies)

Un-Supervised Machine Learning - Classification -1

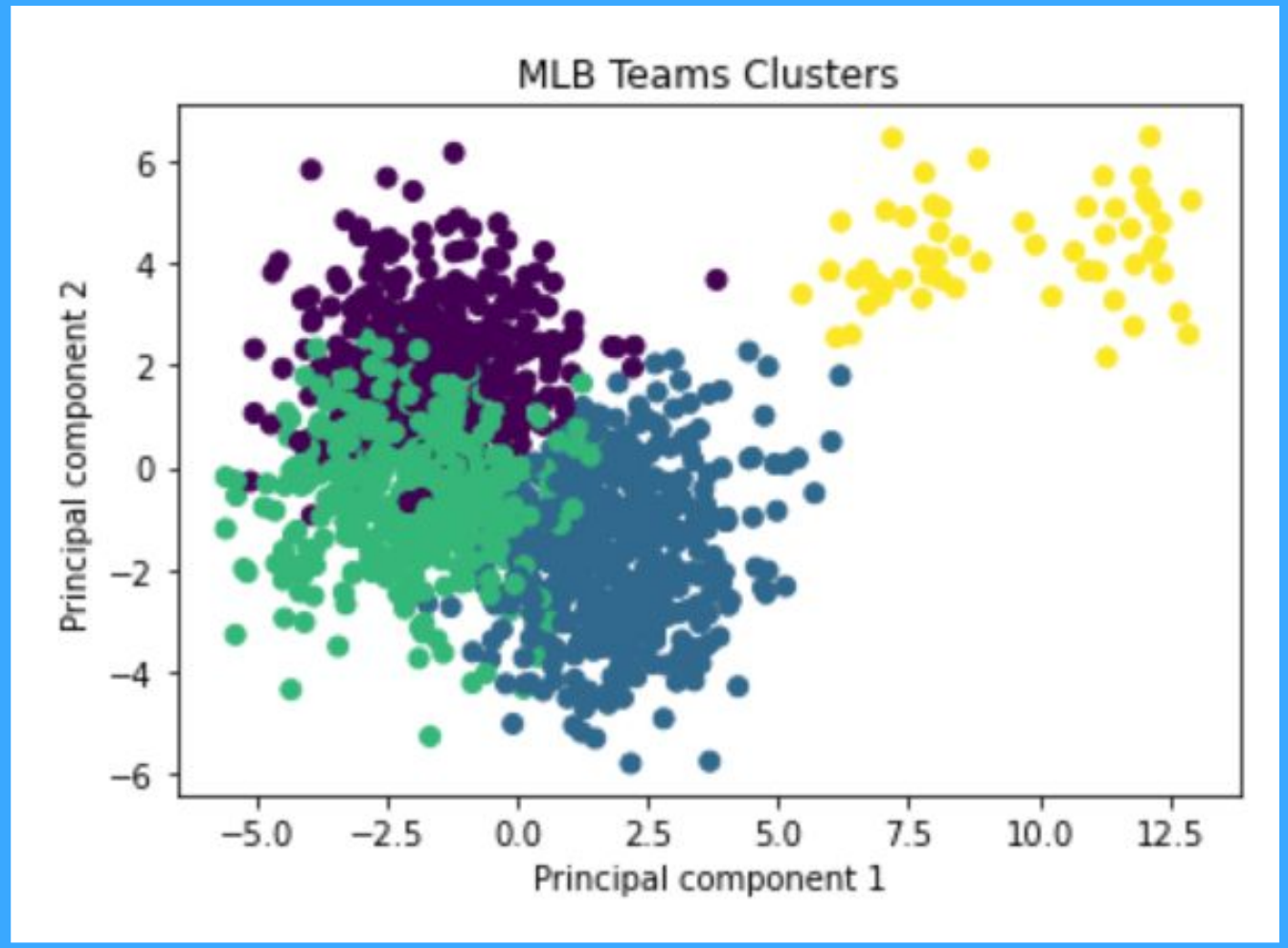
- Using Cleaned Data from 1970 - 2019 data was prepared for classification
- StandardScaler was used for data transformation
- PCA used to reduce dimensionality
- TSNE was used to reduce the dataset dimensions & produce initial cluster



Initial Cluster

Un-Supervised Machine Learning - Classification -2

- Cluster Analysis was performed using KMeans
- Elbow Curve Plotted
- Best Clusters Determined
- K-3 used to predict clusters & PCA for classification
- Classes assigned to team by year
- Teams plotted by class/cluster



Teams Classified

Supervised Machine Learning - Regression Analysis -1

After data processing, cleaning, and and classification, our dataset contained 1384 rows and 31 columns, corresponding to MLB stats from 1970 to 2019, including 36 teams, and the following metrics and variables:

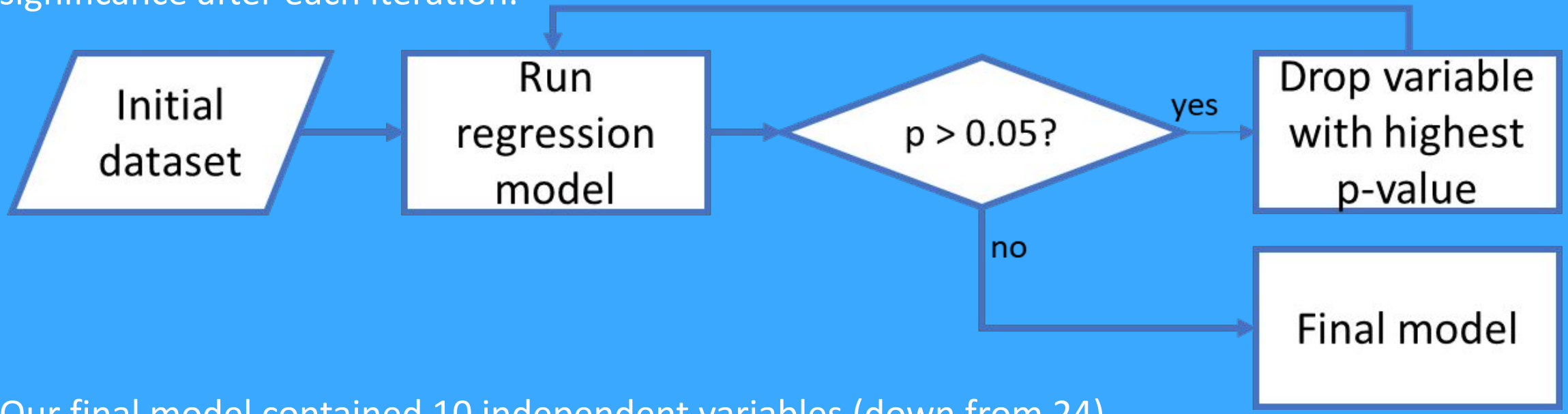
year	walks	saves
team_name	strikeouts_by_batters	outs_pitches
games_played	stolen_bases	hits_allowed
wins	caught_stealing	homeruns_allowed
losses	batters_hit_by_pitch	walks_allowed
runs_scored	sacrifice_flies	strikeouts_by_pitchers
hits	opponents_runs_scored	errors
doubles	earned_runs_allowed	double_plays
triples	complete_games	fielding_percentage
homeruns	shutouts	class

We wanted to understand how much of the variability in the response variable **y** (number of "**wins**") can be explained by changes in an **X** number of variables.

Additionally, we wanted to find whether the number of variables can be reduced without affecting the model score significantly.

Supervised Machine Learning - Regression Analysis -2

Approach: we performed a Stepwise Regression Analysis, which is the iterative construction of a model where independent variables are removed in succession and testing for statistical significance after each iteration.



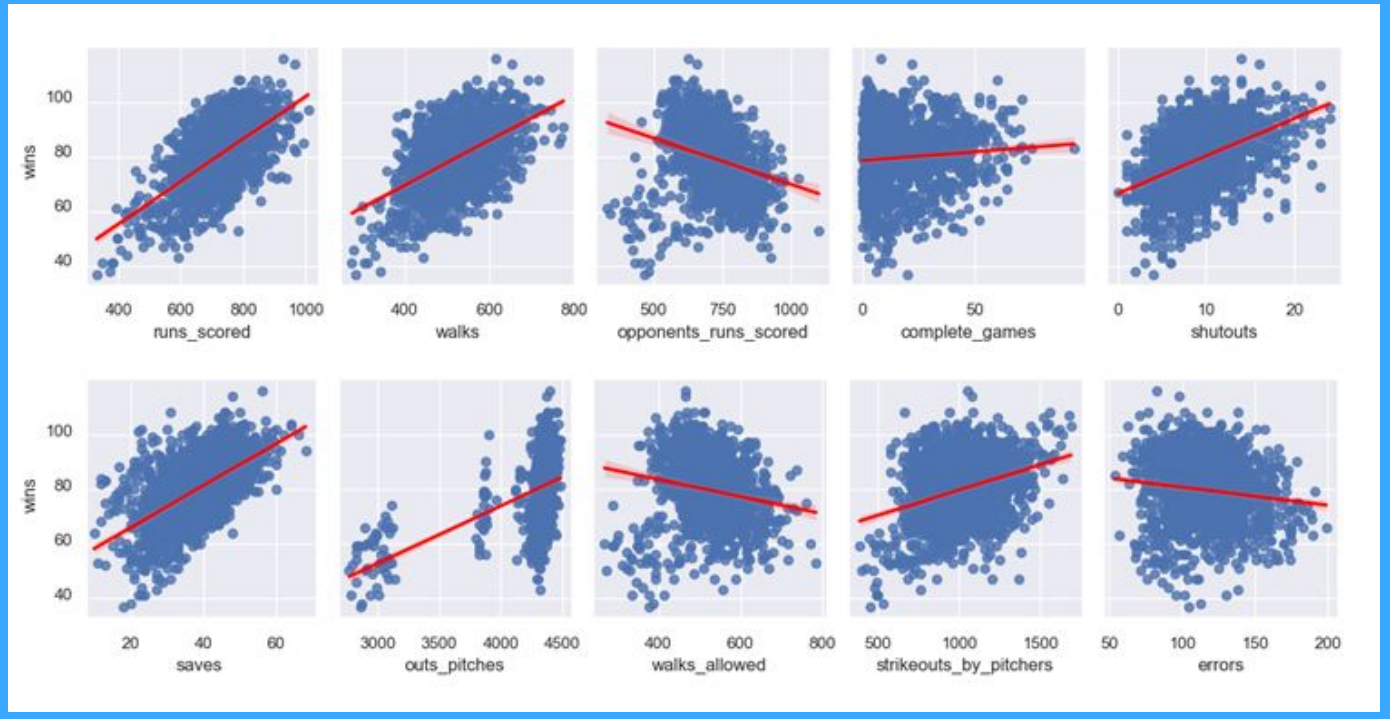
Our final model contained 10 independent variables (down from 24).

Furthermore, collinearity was evaluated and one additional variable was dropped.

Supervised Machine Learning - Regression Analysis -3

Final independent variables:

X	Coefficient	X	Coefficient
1 runs_scored	0.085938	6 saves	0.363825
2 walks	0.003817	7 outs_pitches	0.011554
3 opponents_runs_scored	-0.06977	8 walks_allowed	-0.00408
4 complete_games	0.15875	9 strikeouts_by_pitchers	0.003135
5 shutouts	0.19562	10 errors	-0.01533



Model performance:

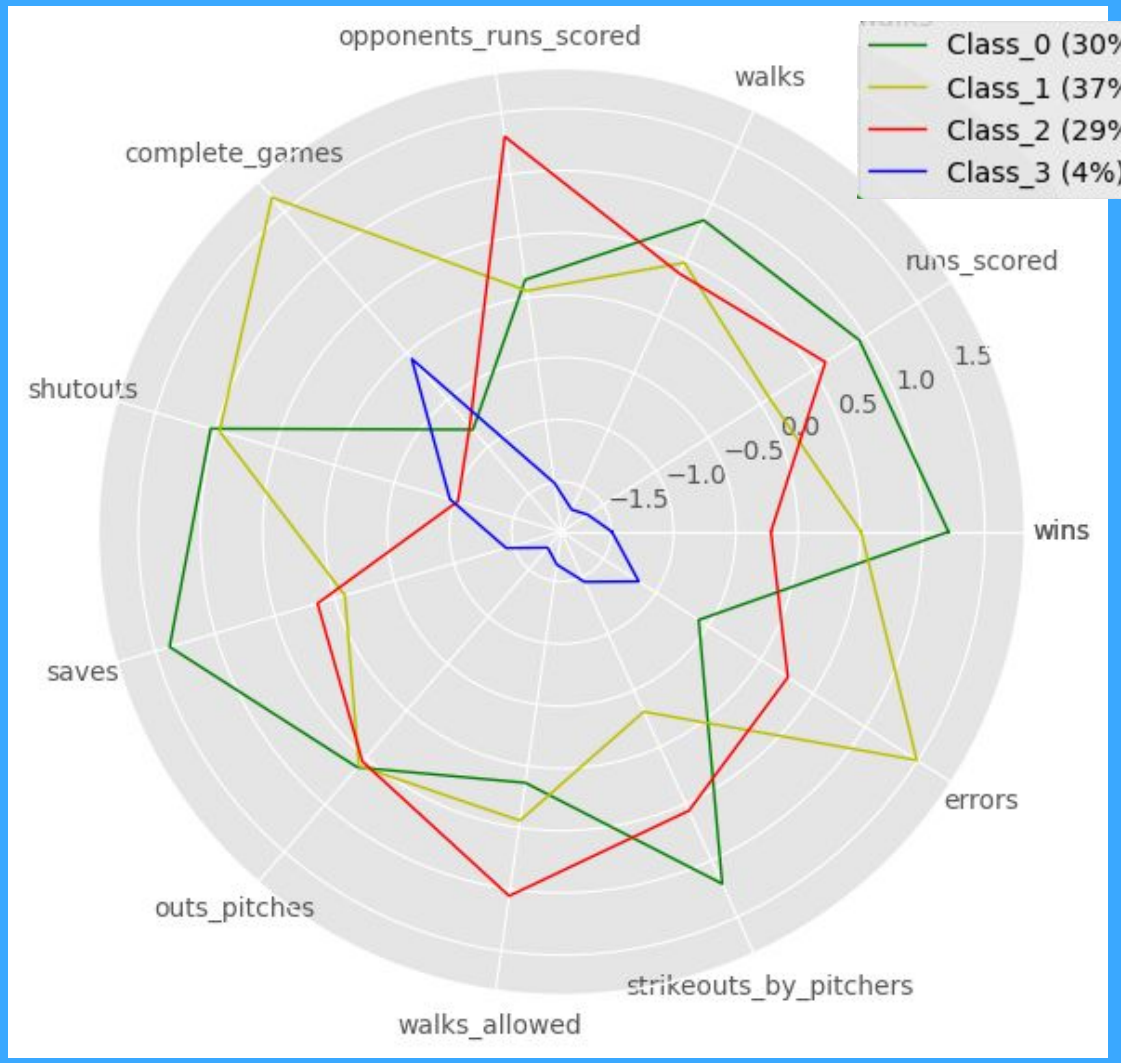
R2 Score: **0.9304714673783749**

Training Score: **0.9303466731994289**

Testing Score: **0.9297116942632093**

Combining Cluster and Regression Analysis

In this part of the project, we wanted to combined the classification work (unsupervised learning) and regression analysis (supervised learning) in one visualization of the different clusters and the 11 baseball statistics. We decided for using **matplotlib** to create a “radar chart,” which shows the relative values of all metrics and for the four clusters included. Data was normalized to the same scale using the **spicy** library, and its **zcore** tool. The resulting chart is shown next.



Site Demonstration

DEMO

Tableau Story

Tableau Story

Challenges & Next Steps

Challenges:

- Data was easy to locate & in fact was so prolific and wide reaching that narrowing the focus presented difficulties of its own
- Webpage building is always a challenge when we are working with new formats

Next Steps:

- Creating a predictor tool for the number of wins you can expect in a season based upon input team stats
- Addition of player data/coaching tenures to each team-year would be a way to tell a more complete story
- Creating a predictor tool for number of wins based on player data - if applicable

Extra Innings

Questions