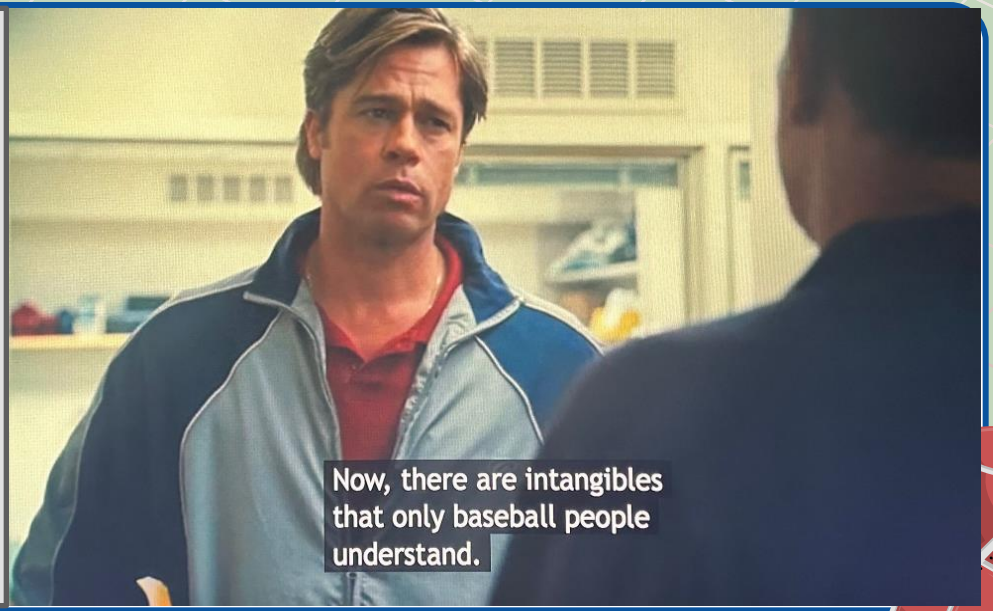
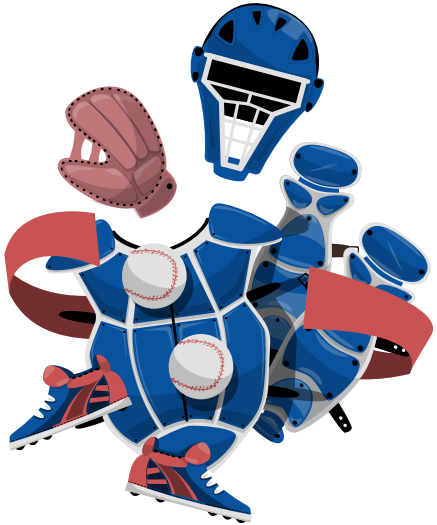
Baseball Analytics



Data Source

The Data is taken from www.retrosheet.org

Negro League Data



American League & National League

Overview

Problem Statement

Can we predict who will win the future scheduled baseball games ?

Proposed Vision

Based on the physical aspects like the ballpark location, game type, day night etc., and the players and team's dynamics using statistics, build a machine learning model to predict the win or lose for that scheduled day.

Impact

The Impact of this solution is how much hard the team need to prepare based on the opponent and other factors for that scheduled game ? Like How many wild pitches allowed ? How many balks allowed? How many triples allowed? Before you forfeit or lose …

Build a Model To Predict Win Or Lose

Data Cleaning

Remove Nulls, Nans and Impute

EDA

Statistics, Co-Relations, Characteristics

Train/Test Split

Split the data into 2 portions

Pre-Process

Scale the data, Check for multi-collinearity

Train & Test Model

Fit data into feature dimensions & Test

Improve

Go back and refine features of the data

Stolen Bases

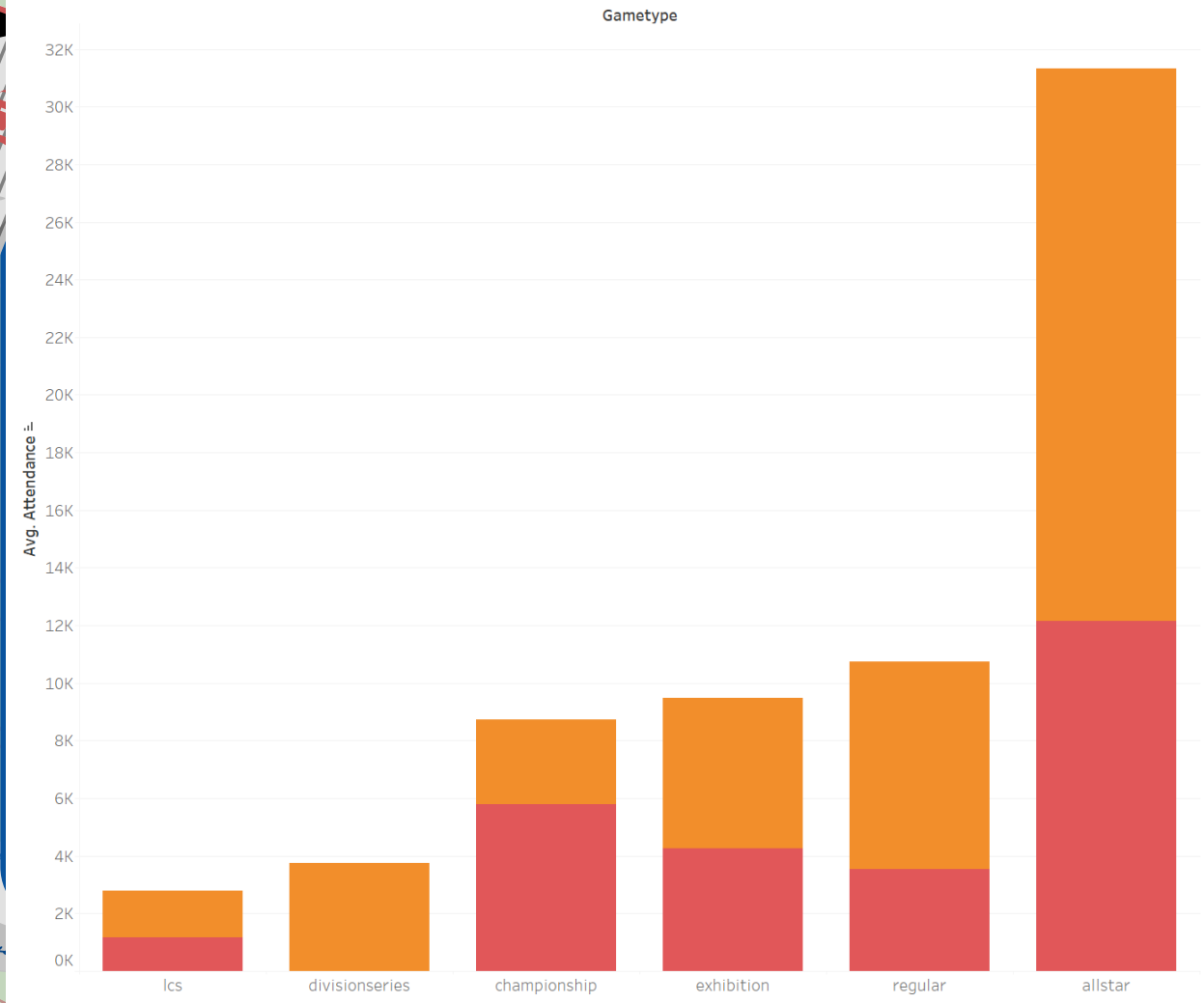
Double Play

Wild Pitch

Sample data – So Many Measures

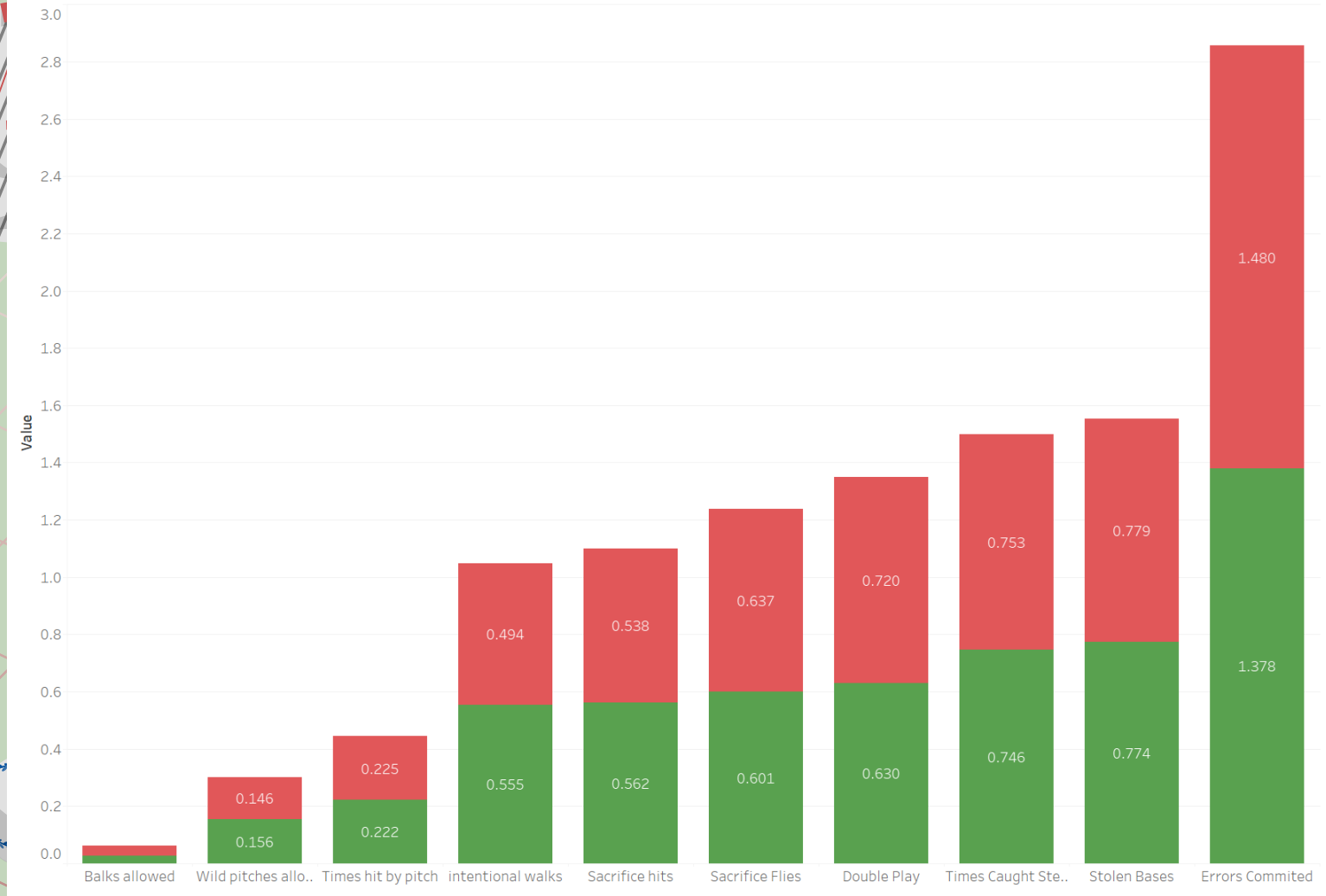
Home Team	Day Night	Game Type	Site	b_sh	b_sf
PHG	Day	Championship	CHI09	3	1
CUX	Night	Regular	NYC22	0	0
CAG	Day	Exhibition	IND11	2	0
NYL	Day	LCS	DET07	3	1

Number of games played day vs night

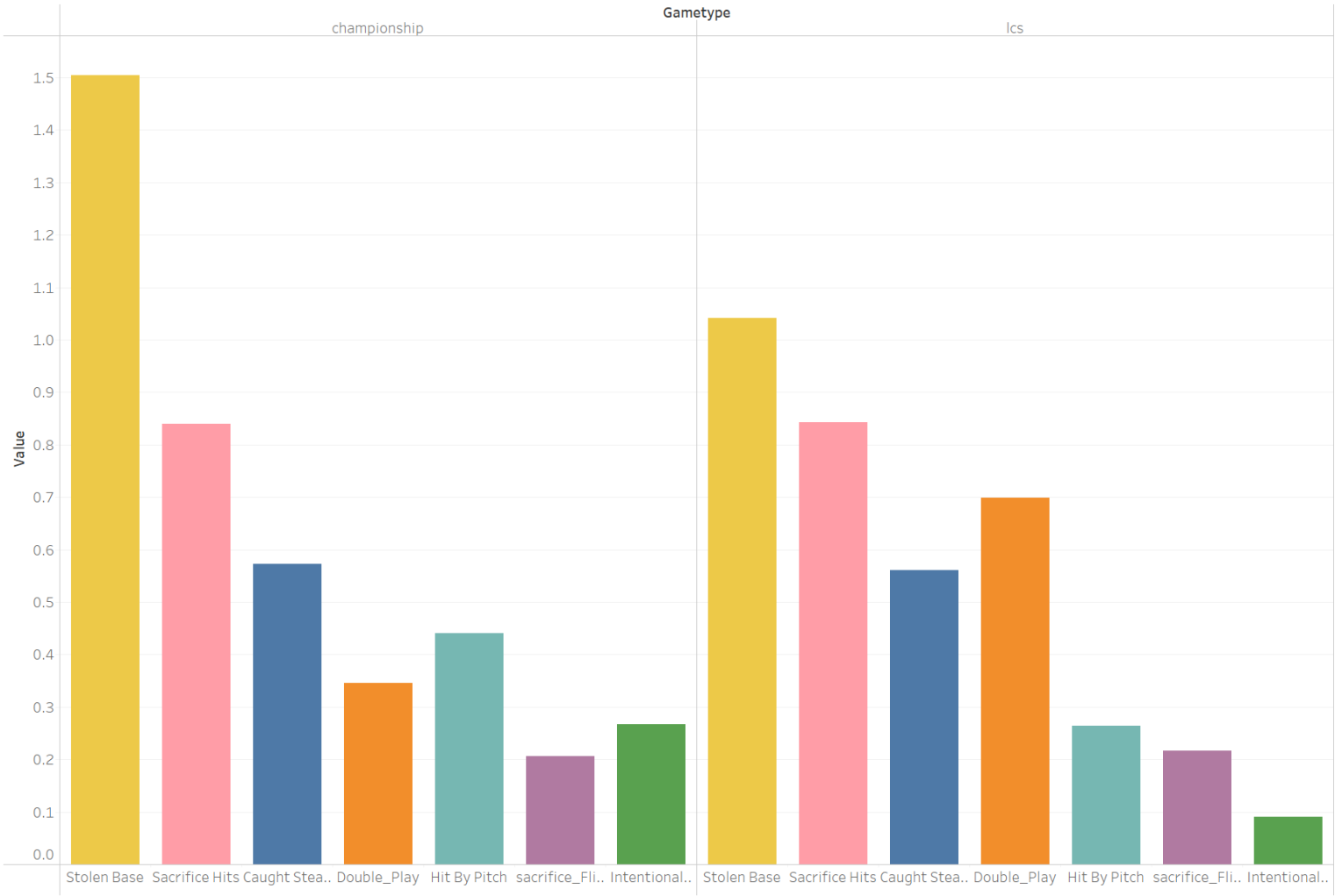


Daynight
day
night

Features Across Winning Games and Losing Games



Features Across Gametypes



Measure Names

- Stolen Base
- Sacrifice Hits
- Caught Stealing
- Double_Play
- Hit By Pitch
- sacrifice_Flies
- Intentional_Walks

Next Steps

Data Process

MLB Data Load to Database , EDA and Preprocess

Feature Engineering

Time Series, PCA

Base Model

Logistic Regression with Ridge and Random Forest