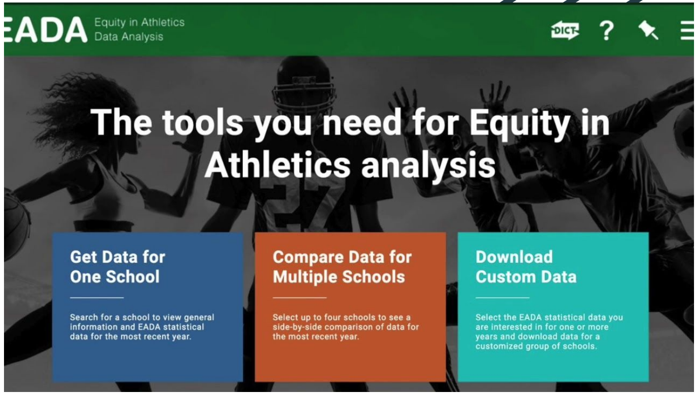



# Predicting collegiate sports' gender revnue

Wooyoung Jeong,
Equity in Athletics Data Analysis

# Design

Objective: Explore relationship area, population, section ID for each sports and schools to with gender's revenue to help companies invest decision making for collegiate sports or gender equality study.

Goal: Based on the feature, figure out which school and sports have higher revenue compare to other gender.

# Tools Used

- Numpy & Pandas
- Scikit-learn & Statsmodels
- Matplotlib

# Data

- All collegiate sports data from 2016 to 2019
- Each row represents particular sports in specific college
- Baseline probability = 0.586
- Features
  - Year, state, classification, sports, populations of each gender and others.

# Data Cleaning

- Delete all rows that have empty data
- There are total 132327 rows in the dataset
- Removed any irrelevant features
- Made another column that can show which gender has higher revenue
- Delete categorical features that have too many different values
    - ex) city and zip code
    - Made dummy variables for other categorical features

```
RangeIndex: 132327 entries, 0 to 132326
Data columns (total 28 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   year                  132327 non-null  int64
 1   unitid                132327 non-null  int64
 2   institution_name      132327 non-null  object
 3   city_txt              132282 non-null  object
 4   state_cd              132282 non-null  object
 5   zip_text              132282 non-null  float64
 6   classification_code   132327 non-null  int64
 7   classification_name   132327 non-null  object
 8   classification_other  1685 non-null    object
 9   ef_male_count         132327 non-null  int64
 10  ef_female_count       132327 non-null  int64
 11  ef_total_count        132327 non-null  int64
 12  sector_cd             132327 non-null  int64
 13  sector_name           132282 non-null  object
 14  sportscode            132327 non-null  int64
 15  partic_men            61865 non-null   float64
 16  partic_women          68885 non-null   float64
 17  partic_coed_men       767 non-null     float64
 18  partic_coed_women     767 non-null     float64
 19  sum_partic_men        132327 non-null  int64
 20  sum_partic_women      132327 non-null  int64
 21  rev_men               61865 non-null   float64
 22  rev_women             68883 non-null   float64
 23  total_rev_menwomen    87134 non-null   float64
 24  exp_men               61865 non-null   float64
 25  exp_women             68885 non-null   float64
 26  total_exp_menwomen    87136 non-null   float64
 27  sports                132327 non-null  object
```

# Logistic Regression

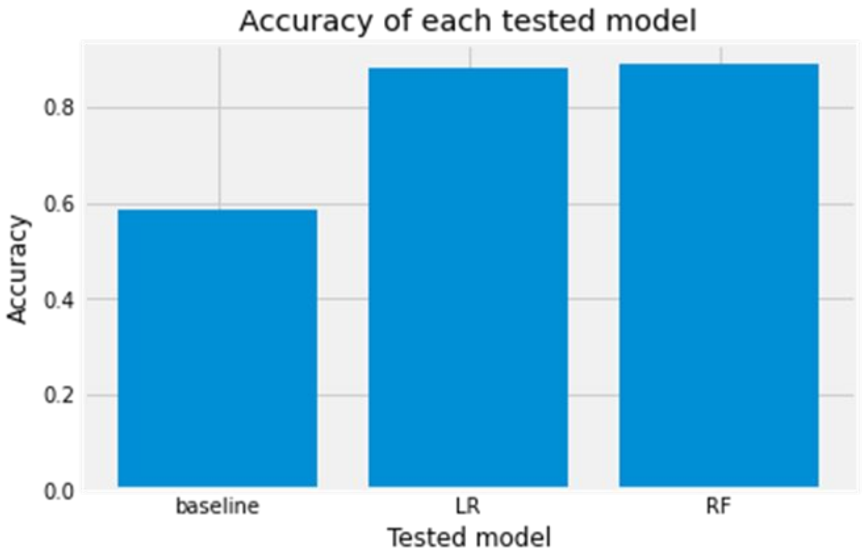
- Used StandardScaler to scale each features
- Accuracy score = 0.883
- Cross validation score = 0.882

# Random Forest

- Did not need to be scaled
- Accuracy score = 0.894
- Cross validation score = 0.891

# Results

- Each models have much higher accuracy compare to baseline
- Both cross validation cv = 10
- Random Forest has slightly higher score than Logistic Regression



# Future Work

- Hyperparameter tuning
  - For both logistic regression and random forest
- Compare many different other models ex) kNN
- Collect more data
  - There are only 4 years of data in the dataset
- Add more visualization
  - ex) confusion matrix