Analyzing NCAA Academic Scores

By: Jenny Alvauaje-Howard

Data

- NCAA Division I statistics (2004 - 2014)
- 57 attributes
- Source: https://www.kaggle.com/ncaa/academic-scores

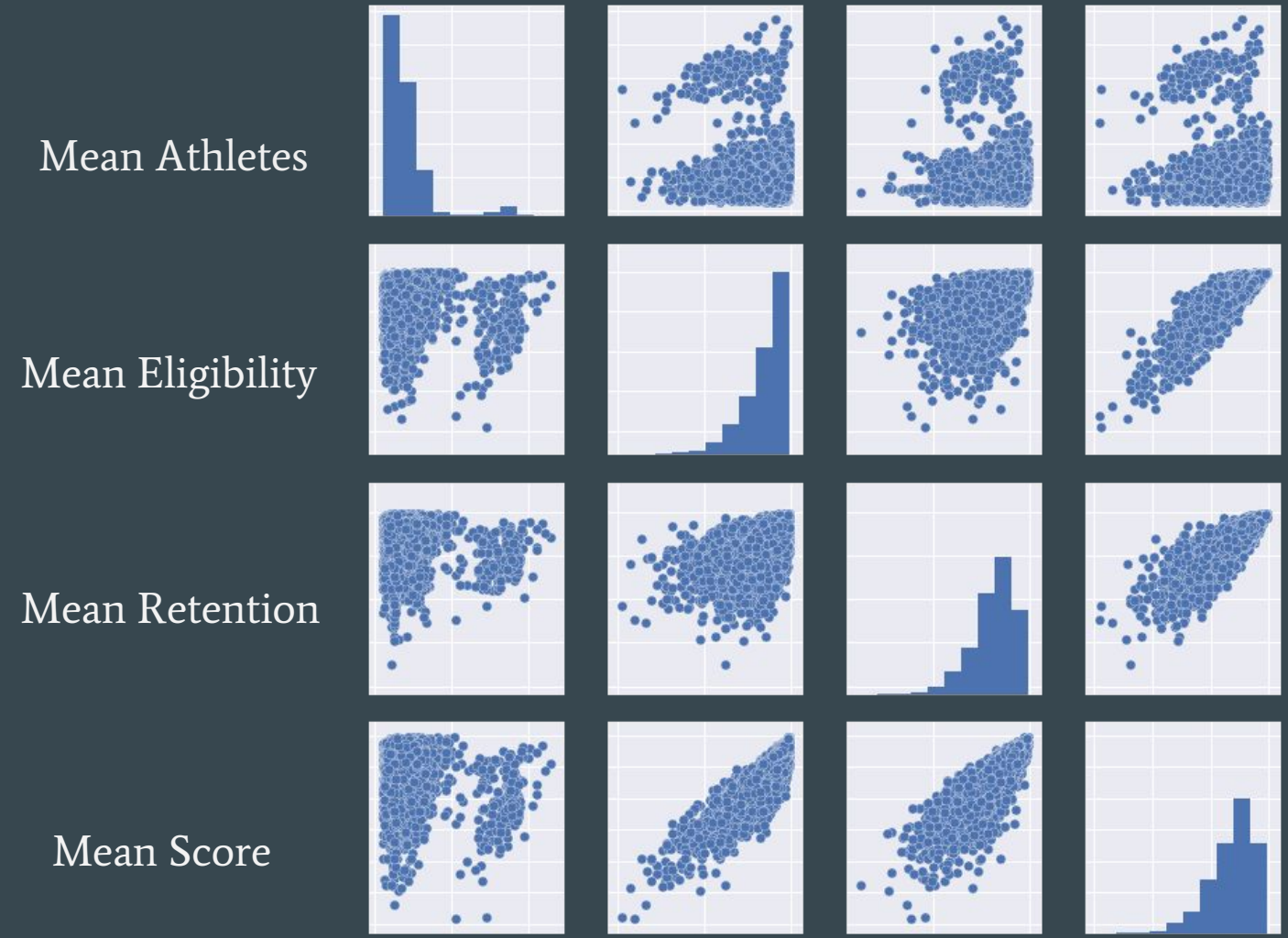
Relevance of our data

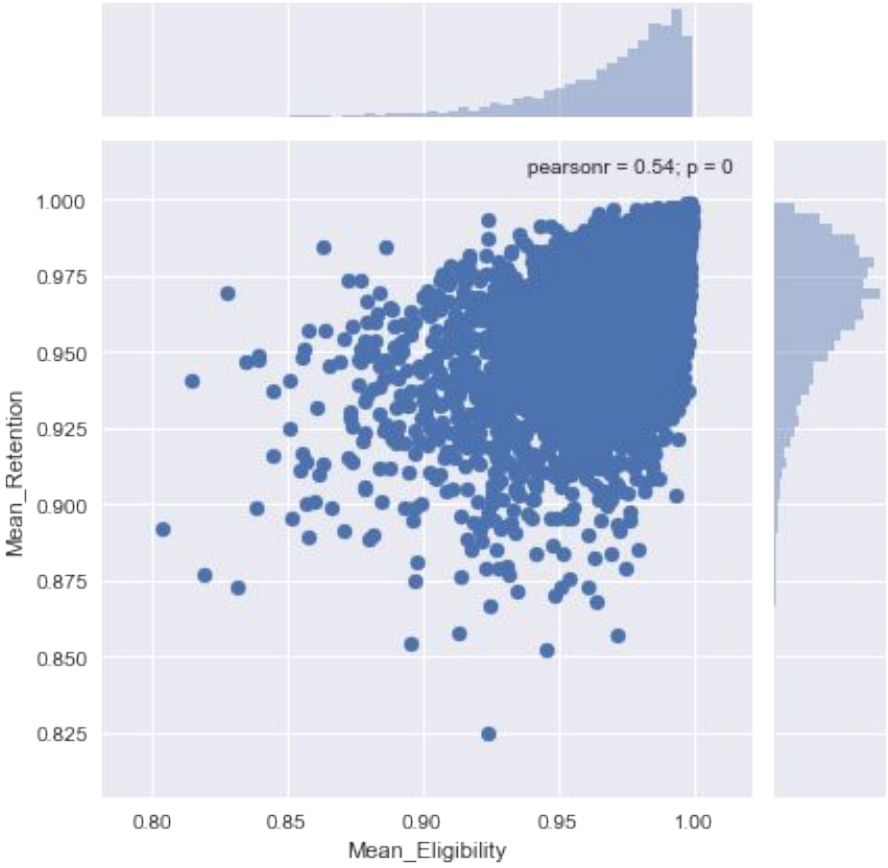
- The NCAA oversees 1000+ colleges and universities in the US and Canada
- Ensures that athletes are academically successful
- Measures APR, the Academic Progress Rate
- The APR rewards superior academic success and penalizes teams that don't achieve certain academic checkpoints

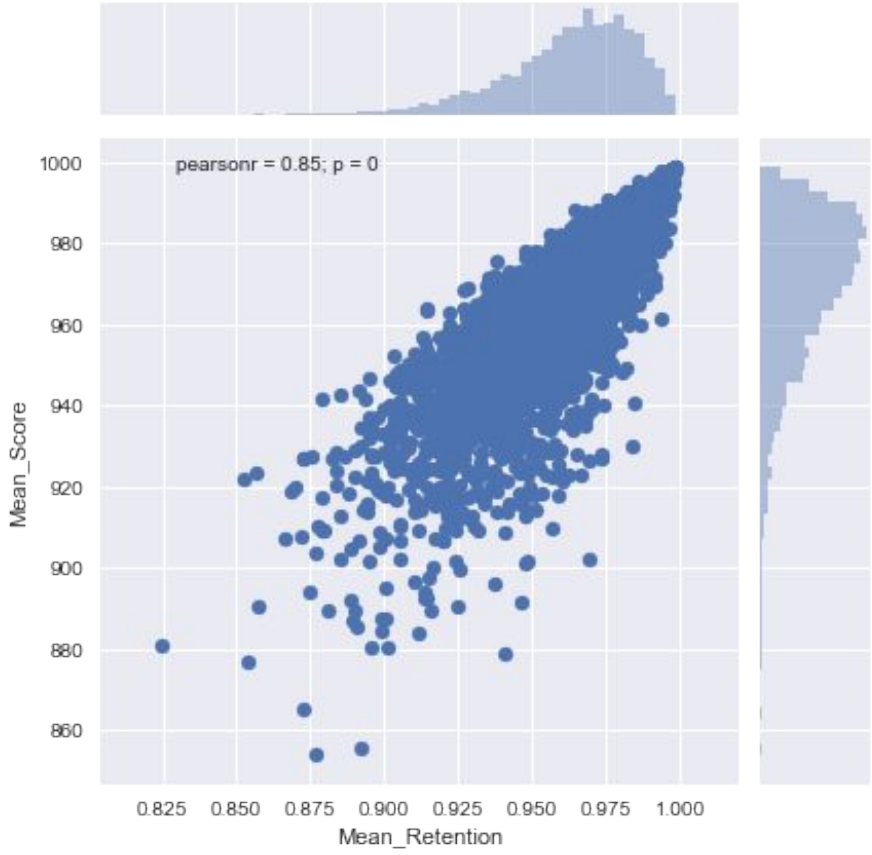
(NCAA.org)

How can we predict a team's retention rate?

- Using score and eligibility as predictors
- Feature engineering: adding squared correlations, adding mean retention, eligibility, score, and number of athletes for each school and sport.
- Since these are continuous variables, we'll use regression models


Mean Athletes
Mean Eligibility
Mean Retention
Mean Score


pearsonr = 0.54; p = 0
Mean_Retention
Mean_Eligibility
1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825
0.80
0.85
0.90
0.95
1.00


pearsonr = 0.85; p = 0
Mean_Score
Mean_Retention
1000
980
960
940
920
900
880
860
0.825
0.850
0.875
0.900
0.925
0.950
0.975
1.000

Potential models

- Linear Regression (StatsMod): **98.59%**
- Linear Regression (SKLearn): **98.75%**
- Ridge: **98.57%**
- Lasso: **98.57%**
- RFR: **91.83%**

Can we improve our worst accuracy?

- Random forest regressor
- Try adding more trees (30, 100, 150, 1000, 2500)
- Try feature engineering (taking means, squared means of predictors)
- Best accuracy score (without waiting for more than 5 minutes): **92.13%**

Classifying High Retention

- Creating high retention as a categorical variable
- Gradient boosting to predict whether or not a school will have high retention (>.95)
- Accuracy: 94.37%

Conclusions

- We can predict retention rates based off eligibility and score with incredible accuracy
- Teams (and we can assume, athletes) that score better and have higher NCAA eligibility are more likely to have high retention.