Regression Analysis
for
Professional Sports Teams Valuation

Sponsor : RCM-X

UIUC MSFE Practicum - Spring 2020

Yunxi Wu

Zihan Yu

Yuli Tang

Rakesh Reddy Mudhireddy

Background

- Sports play a major part in every society and attracts almost everyone.
- There is a long list of individuals who want to buy into major sports teams:
  - With the increasing TV contracts, advertising revenues and more attraction from audience, the value of teams shoots up continuously
  - It's a status symbol for lots of wealthy people to have a share in the sports teams
- We are trying to obtain a fair valuation of North America league teams for potential market investors
- **Potential features** for our regression analysis are listed as follows:

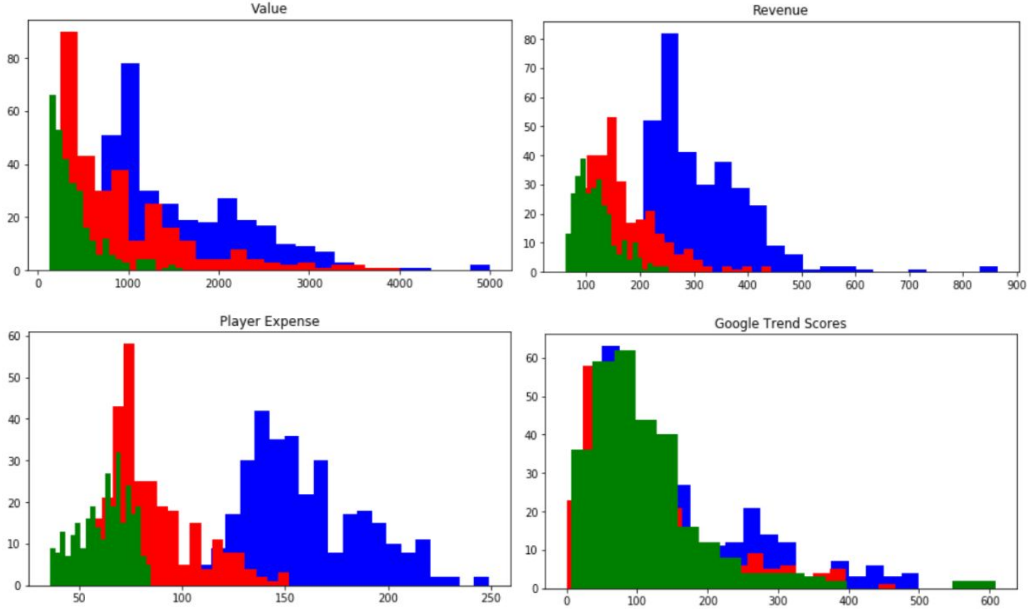
Financials	Team Management	Fans Analysis	Stadium
Valuation	Championships	Overall Favourite Rank	Stadium Year Open
Most Recent Purchase Price	Number of star players	Fan Equity	Stadium Capacity
Revenue	Coaches/Managers track record	Social Equity	Stadium Cost to Build
Average Revene per fan	Medical team	Road Equity (Sports Travel)	Executive club seats/luxury suites
Average ticket price	Index of parent company(Credit quality, Liquidity...)	Number of fans	Stadium facilities
Gate receipts	Sponsor's brand affinity	Attendance	Depreciation of stadiums/facilities
Operation Income	Youth academy	TV viewership	Arena ranking
Debt Approximation	Fines & Injuries	Mobile viewership(smartphone, tablet, computer)	Metro Area Population
Sponsorships		Facebook/Twitter/Instagram followers	
Telivision rights			
Player Expense			
Win-to-Player Cost Ratio			
EBITDA/Adjusted EBITDA			

Data Collection

- ***Dataset*:** 10-year data for 32 NFL teams, 30 NBA teams and 31 NHL teams; 919 data points, 14 features.
- ***Data Sources*:** Forbes, ESPN, Sports Market Analytics (SMA), nba.com, nfl.com, nhl.com.
- Since we merge the data across three leagues and ten years for more data points, we need to add control variables (Year, NFL, NBA) in our model to control difference between league and time effect.

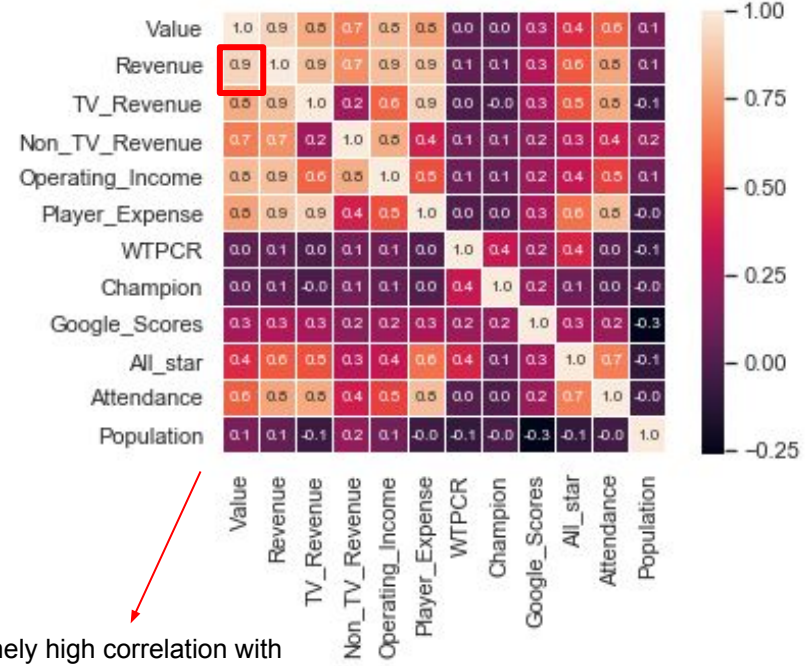
Team	Value	Revenue	National TV Revenue	Non-TV Revenue	Operating Income	Player Expenses	WTPCR	Champion	Google Trend Scores	All_star	Attendance	Population	NFL	NBA	Year
Arizona Cardinals	900	223	95	128	24	142	158	0	116	6	61323	1429285	1	0	2009
Atlanta Falcons	900	214	95	119	28	129	128	0	77	2	71601	420823	1	0	2009
Baltimore Ravens	1100	240	95	145	44	122	184	0	110	5	70627	621568	1	0	2009
Buffalo Bills	900	222	95	127	40	127	83	0	243	2	68839	261896	1	0	2009
Carolina Panthers	1000	238	95	143	23	138	130	0	78	4	72220	722228	1	0	2009
Chicago Bears	1100	241	95	146	42	137	98	0	230	1	61916	2686849	1	0	2009
Cincinnati Bengals	1000	222	95	127	35	131	52	0	74	0	47179	297654	1	0	2009
Cleveland Browns	1000	235	95	140	20	143	42	0	152	3	67431	399186	1	0	2009
Dallas Cowboys	1700	280	95	185	9	165	82	0	234	6	90929	1182424	1	0	2009
Denver Broncos	1100	240	95	145	40	123	97	0	166	3	75937	586195	1	0	2009

Exploratory data analysis

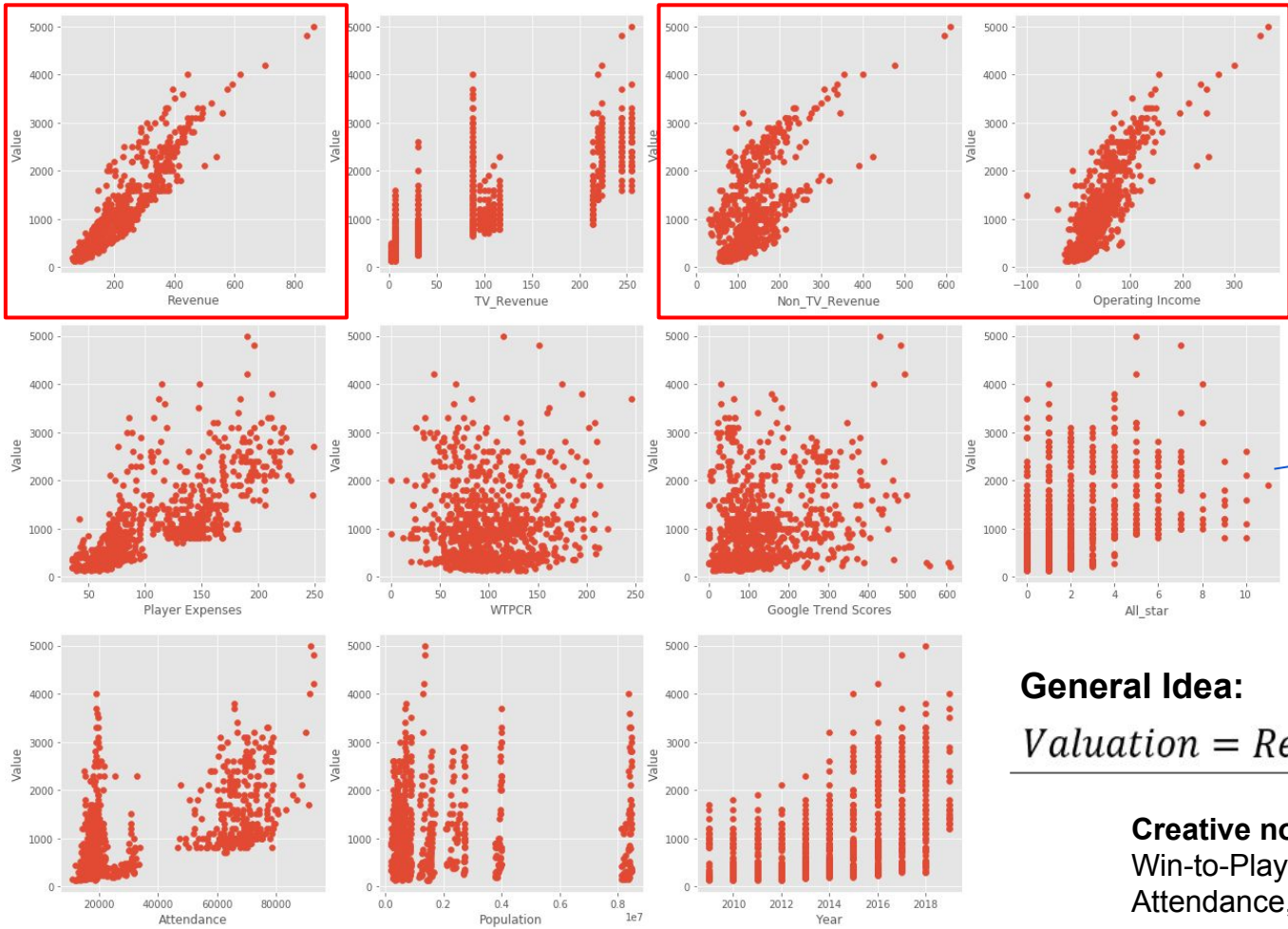


Histogram (blue=NFL, red=NBA, green=NHL)

Heatmap



Revenue has extremely high correlation with Valuation. The correlation between financial and non-financial variables are not strong.



Money Streams
(Linear relationship; explain large proportion of the variance)

Non-Financial Variables
(Nonlinear relationship)

General Idea:

$$Valuation = Revenue\ streams + X_{factor}$$

Creative non-financial variables:
Win-to-Player Cost Ratio, All Star Players, Attendance, Population, Google Trend Score, ……

Figure. Valuation vs Features

X-factor using PCA

- Principal Component Analysis (PCA)
  - Reduce the dimensionality of data while retaining most of the variation
  - In our case, compress the six non-financial factors into a single factor (i.e., the **X-factor**)
  - The following explains the rationales for how to generate the weights for non-financial factors, which will be used to determine the X-factor

1. This is explained variance ratios. Drop the last entry to obtain 95% var.

2. This is the pca.components_ matrix. With rows correspond to **PCA features** 1, 2, 3…; columns correspond to **original features** 1, 2, 3…
Drop the last row in correspond to the drop of the last entry in ***step 1.***

```
[0.35  0.21  0.18  0.12  0.1   0.04]
[[0.42  0.28  0.57  0.45  0.42  0.19]
 [0.42  0.59  0.33  0.57  0.18  0.15]
 [0.27  0.29  0.16  0.06  0.38  0.82]
 [0.56  0.37  0.24  0.2   0.57  0.35]
 [0.33  0.59  0.02  0.28  0.56  0.39]
 [0.4   0.09  0.7   0.59  0.04  0.02]]
```

3. This row shows the 'importance' of each **PCA feature** to the **whole dataset**. Denote as $W_{original\ to\ PC}$

4. This matrix shows the 'importance' of each **original feature** to the corresponding **PCA feature.** Denote as $W_{PC\ to\ all}$

5. The following multiplication should hold if we consider 'importance' as a measure for weights :

$$W_{origin\ to\ PCA} * W_{PCA\ to\ all} = W_{origin\ to\ all}$$

Continued..

$$\begin{bmatrix}0.35 & 0.21 & 0.18 & 0.12 & 0.1\end{bmatrix} \begin{bmatrix}0.42 & 0.28 & 0.57 & 0.45 & 0.42 & 0.19\\ 0.42 & 0.59 & 0.33 & 0.57 & 0.18 & 0.15\\ 0.27 & 0.29 & 0.16 & 0.06 & 0.38 & 0.82\\ 0.56 & 0.37 & 0.24 & 0.2 & 0.57 & 0.35\\ 0.33 & 0.59 & 0.02 & 0.28 & 0.56 & 0.39\end{bmatrix} = \begin{bmatrix}0.38 & 0.38 & 0.33 & 0.34 & 0.38 & 0.33\end{bmatrix}$$

$W_{original\ to\ PC} \quad * \quad W_{PC\ to\ all} \quad = \quad W_{origin\ to\ all}$

$$X_{factor}(PC_{score}) = 0.38 * W + 0.38 * C + 0.33 * S + 0.34 * A + 0.38 * G + 0.33 * P$$

W: *Win to Player Cost Ratio*
C: *Champion*
S: *Star Players*
A: *Attendance*
G: *Google Trend Scores*
P: *Population*

Ex: In an arbitrary season, if **Arizona Cardinals** has a Win-to-Player-Cost ratio of 158 (***W=158***), did not win the championship (***C=0***), has 6 all-star players (***S=6***), average attendance of that season was ***61,323***, Google Trend Scores of the team was ***116***, and the population of the team's city was ***1,429,285***.

$$X_{Factor} = 0.38 * \mathbf{158} + 0.38 * \mathbf{0} + 0.33 * \mathbf{6} + 0.34 * \left(\frac{\mathbf{61,323}}{\mathbf{10,000}}\right) + 0.38 * \mathbf{116} + 0.33 * \left(\frac{\mathbf{1,429,285}}{\mathbf{10,000}}\right) = 155.21$$

Linear Regression

Model 1 - Without Revenue Division

$$ln(value) = -133.558 + 1.538 \times ln(revenue) - 0.569 \times ln(player\_expenses) + 0.0843 \times ln(X\_factor) + 0.066577 \times year + 0.461 \times NFL + 0.356 \times NBA$$

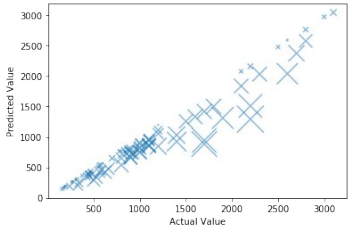
$$\downarrow$$

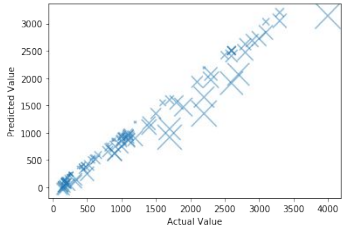
$$value = (revenue^{1.538} \times X\_factor^{0.0843} \times exp(-133.558 + 0.066577 \times year + 0.461 \times NFL + 0.356 \times NBA)) \div player\_expenses^{0.569}$$

Statistic	Training set (80%)	Test set (20%)
R-squared	94.08%	93.69%

Model 2 - With Revenue Division

$$value = -46246.278 + 9.732 \times tv\_revenue + 7.255 \times non\_tv\_revenue - 2.126 \times player\_expenses + 0.523 \times X\_factor + 22.766 \times year - 421.859 \times NFL + 74.489 \times NBA$$

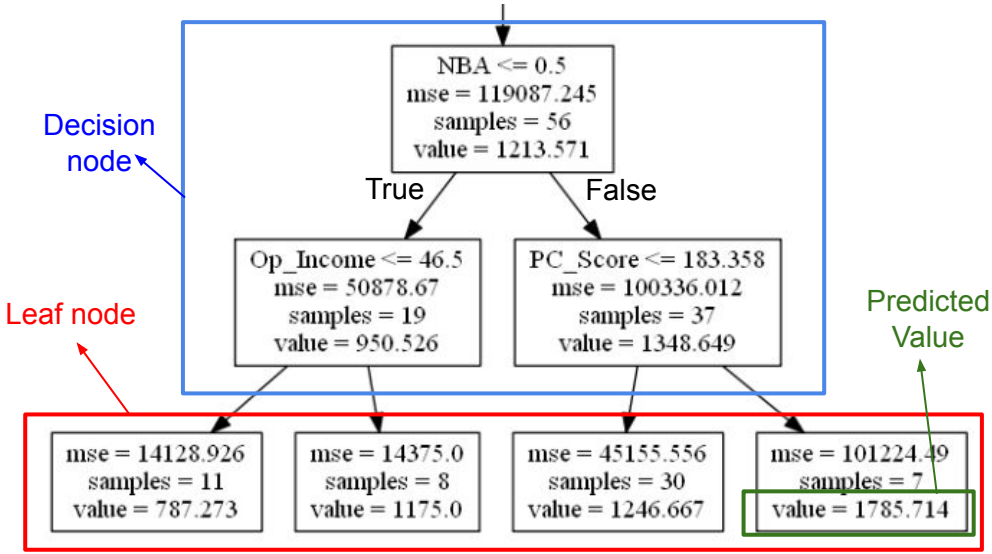




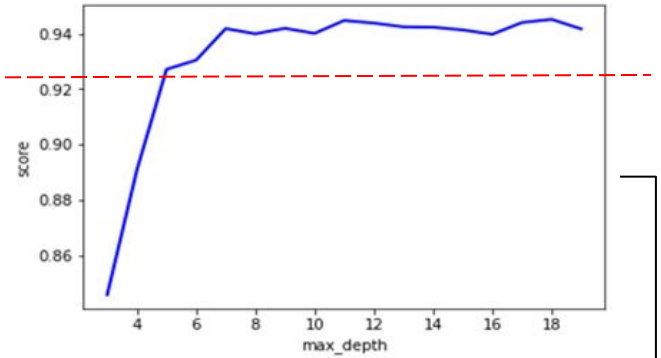
Statistic	Training set (80%)	Test set (20%)
R-squared	91.05%	93.54%

Decision Tree Regression

Regression Tree example:



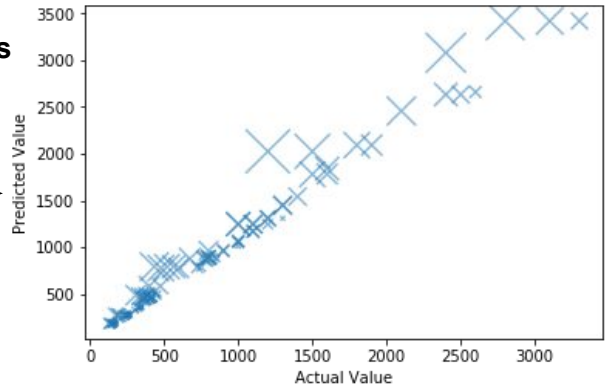
Statistic	Training set (80%)	Test set (20%)
R-squared	95.86%	94.55%



Pruning parameters - Avoid Overfitting

- Max_depth=6 (maximum step the tree takes)
- Min_samples_leaf=5 (minimum number of samples for every leaf node), number of total nodes decrease by 17.39%

Error Analysis



Future Scope

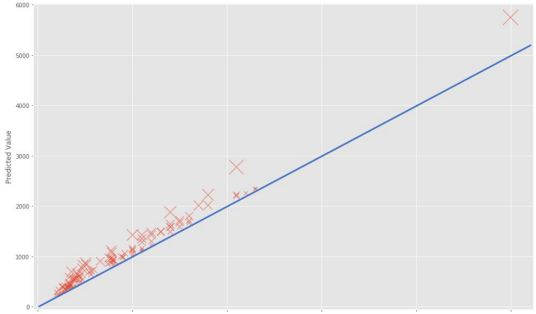
- Ridge, Lasso, and Elastic Net regression
  - These are **supervised learning** models
  - Ridge regression performs the best
    - Errors close to the diagonal line
    - High accuracy on the test set
    - Steady performance between training & test MSE

Statistic	Ridge	Lasso	Elastic Net
Training R^2	91.70%	91.75%	92.41%
Test R^2	92.66%	92.44%	89.94%
Training RMSE	229.89	225.63	230.26
Test RMSE	231.95	244.02	233.13

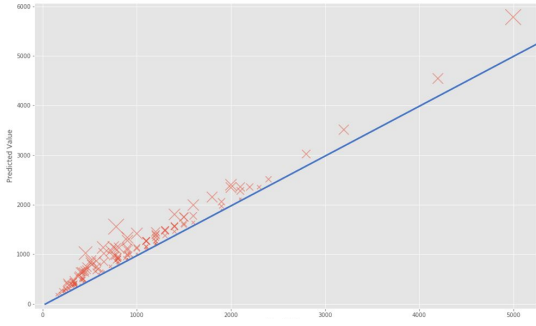
Regression Performance

Error Analysis

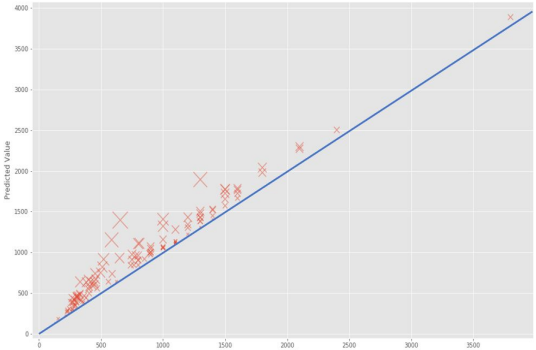
Ridge



Lasso



Elastic Net



Continued..

- Random Forest, XGBoost
  - **Ensemble learning** based on tree structure
  - For larger dataset, apply these models to benefit from their algorithms
  - Significant improvement over previous three models
- Neural Network
  - **Deep learning** algorithm
  - Performance not significant due to small dataset

Statistic	Random Forest	XGBoost	Neural Network
Training R^2	99.24%	98.46%	--
Test R^2	93.85%	95.62%	--
Training RMSE	69.69	99.29	264.27
Test RMSE	209.32	176.66	276.34

Regression Performance

Continued..

- Impact of Coronavirus
  - CNBC released findings on how the virus would affect ad spending in the sports industry
  - The analysis found that, due to coronavirus, NBA, MLB, and NHL together will lose **$1 billion** in broadcasting revenue from March to May
  - If the NFL season is delayed or cancelled later this year, it may lead to losses of up to **$6 billion** for around 3000 advertisers.
  - Not only on broadcasts and advertising, teams also lose a huge amount of money in gate receipts. (**e.g.,** NBA's playoff season would have started by this time, and NBA will lose an average of **$2 million** in gate receipts, **per playoff season game**, because of the coronavirus.)
- References
  - https://sports.yahoo.com/nba-season-cancelled-much-ticket-030034660.html
  - https://www.cnbc.com/2020/04/03/coronavirus-could-cause-1-billion-loss-for-nba-nhl-and-mlb-broadcasters.html