**Predicting City Quality Score Based on Venue Frequency**

Sunjin David Kim

August 23, 2019

# Introduction

## Background

Every year, UN News analysis the 125 most populous cities of United States to determine the best cities to live. These cities are given an overall score from 0-10, and a ranking from 1-125. In addition to original surveys conducted by UN News, the following data sources were used to determine the city scores.

 [United States Census](https://www.census.gov/)

 [Gallup-Healthways Well-Being Index](http://www.well-beingindex.com/)

 [Federal Bureau of Investigation Uniform Crime Report](https://www.fbi.gov/about-us/cjis/ucr/ucr)

 [Bureau of Labor Statistics](http://www.bls.gov/)

The following Indexes were used to determine the overall score of each city:

* Job Market Index
* Value Index
* Quality of Life Index
* Desirability Index

For more information concerning the methodology of the research, please go to:

<https://realestate.usnews.com/places/methodology>

## Problem

In order to be competitive in the Realestate business, or to personally purchase home in a nice neighborhood before the prices of homes rise, it is important to use data to predict potential cities that will become popular in the future. Although the Rankings provided by USNEWS is very helpful in determining current hotspots, as potential home buyers, it is imperative to be ahead of the curve and purchase a home in a city before the popularity and prices rise. Therefore, using the data gathered from Foursquare, we will attempt to identify variables that correllate with city popularity represented by the City Scores, which could potentially help us find cities that will become popular in the future.

## Interest

This analysis may be useful to the following groups:

1. Real-estate Agents looking for potential cities to work in.

2. Individuals looking for a good place to live.

3. City developers and officials seeking to improve their cities can also use this information to determine what types of venues to increase or decrease.

# Data acquisition and cleaning

## Data sources

I will be conducting an analysis the 5 top ranked and the 5 lowest ranked cities to live in the US according to the 2019 Best Places to Live Research conducted by US NEWS.

The Following Cities will be analyzed:

Rank #1: Austin, TX

Rank #2: Denver, CO

Rank #3: Colorado Springs, CO

Rank #4: Fayetteville, AR

Rank #5: Des Moines, IA

Rank #125: San Juan, PR

Rank #124: Bakersfield, CA

Rank #123: Stockton, CA

Rank #122: Shreveport, LA

Rank #121: Mobile, AL

I will be obtaining the city scores from the research conducted by US News and the city coordinates from the web. It should be noted that the Lowest Ranked Cities, are not the worst cities to live in. Rather, they are the cities ranked 123-124th in the list of best places to live.

**Venue Data**

I will be using the Foursquare API to determine the frequency and ratio of venue types in each of the cities to determine whether there are any significant correlation between the venues found in a city and its overall score. If possible, we will use a machine learning algorithms to make predictive model that can can predict for us the city's overall score based on their venues

I will pick 4 random cities to test this model.

## Data cleaning

We began by creating a dataframe with the five top scoring cities and the 5 lowest scoring cities. We included their name, score, rank, latitude, and longitude. 

Next, using the Foursquare API, we gathered venue information of each city, setting the limit to 1000 and radius to 2000. Several cities had to be dropped from the data due to a lack of venue information. Cities with less than 10 venues were dropped, which includes Bakersfield, Colorado Springs, Des Moines, Fayetteville, and Shreveport.

The frequency of each venue category for each city was calculated using the information gathered from Foursquares. A new dataframe was created with the city name, score, rank, and the venue frequency. The rank of the city was later removed as it was redundant and a relative number based on comparison of other cities.



This process was repeated later with the test cities that were picked at random. (San Francisco, Phoenix, Memphis, and Scranton) 

## Feature selection

After data cleaning, data from 234 venues in 5 cities were left.

For the test, 255 venues from 4 cities were used.

For future research, these number should be adjusted to get a better result.

# Methodology

# 3.1 Exploratory Data Analysis

## 3.1.1 K-Cluster Analysis

To explore the data gained from Foursquare, we conducted a K-cluster analysis to see if the venue frequencies by themselves would create clusters that more or less aligned with their city scores. The initial analysis with the ten cities revealed that there were cities with not enough data that was skewing the clusters. The initial cluster visualized onto a map can be found below.



Upon eliminating the cities that did not have enough venues and reducing the k to 2, the cluster formed according to city scores, with Austin and Denver in once cluster and San Juan, Stockton, and Mobile in the other.

## 3.1.2 Correlation

## Using Pandas .corr() function, the correlation of the venue frequency and the city scores were calculated. Upon inspection, seven venues were found to be positive correlated with the city scores with a correlation score greater than 0.83.

# 3.2 Predictive Modeling

## 3.2.1 Multi Linear Regression Model

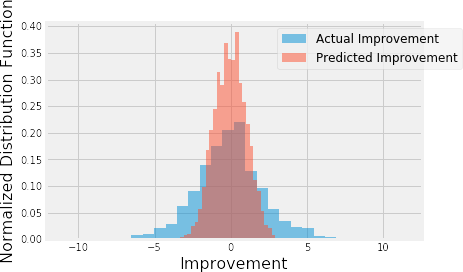


Figure 8. Distribution of actual and predicted improvement using linear regression with equal weights of samples.

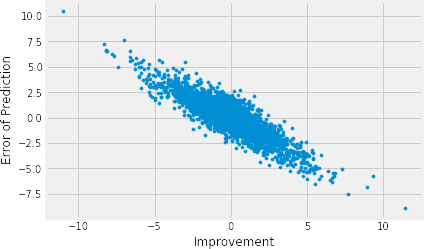


Figure 9. Scatterplot of prediction errors vs. actual target values using linear regression with equal weights of samples.

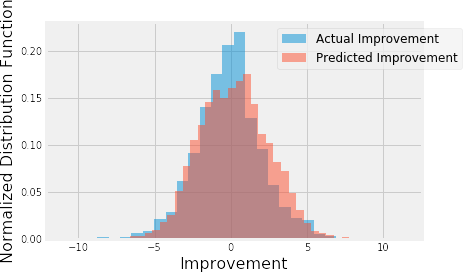


Figure 10. Distribution of actual and predicted improvement using linear regression with different weights of samples based on inverse of sample abundance.

## Performances of different models

Using the new approach of different sample weights, I built linear regression, SVM, random forest, and gradient boost models using weighted root mean squared error as the evaluation metric. For each model, hyperparameters were tuned using the same metric and cross validation. For comparison, I also built a simple linear regression model with just one independent variable (age) as the benchmark model. SVM had the best performance among all models, which had

~26% less error than the benchmark model (Table 2). The predicted improvements had linear relationship with the actual improvements (Figure 11).

Table 2. Performance of the regression models.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Benchmark (one feature) | Linear Regression | SVM | Random Forest | Gradient Boost |
| Weighted RMSE | 3.84 | 2.98 | 2.86 | 2.93 | 2.96 |

## 4.2 Classification models

The application of classification models was much more straightforward. I divided the samples into two classes (improvement>=0 or <0). The number of samples in each class were about the same. I chose logarithmic loss as the metric here because the results would probably be presented with probabilities and logarithmic loss puts more emphasis on the probabilities than other metrics. Logistic regression, SVM, random forest, gradient boost models and a voting model were tuned and built. Among the individual models, the SVM model performed the best (~67.5%

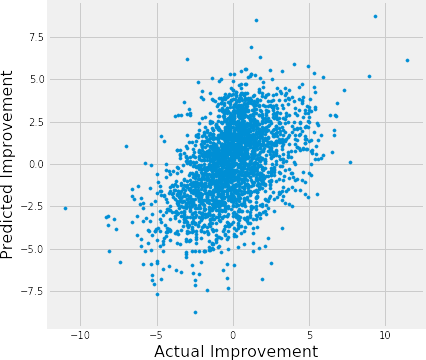
accuracy), and voting model performed similarly as the SVM model (Table 3), though the differences between models were small.

Figure 11. Scatter plot of predicted and actual player improvements of the SVM model.

Table 3. Performance of classification models. Best performance labeled in red.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Logistic Regression | SVM | Random Forest | Gradient Boost | Voting Model |
| Log Loss | 0.605 | 0.603 | 0.612 | 0.613 | 0.603 |
| Accuracy | 0.675 | 0.675 | 0.672 | 0.672 | 0.675 |
| No. of True Positives | 835 | 830 | 810 | 815 | 838 |
| No. of False Positives | 413 | 406 | 396 | 400 | 416 |
| No. of False Negatives | 438 | 443 | 463 | 458 | 435 |
| No. of True Negatives | 929 | 936 | 946 | 942 | 926 |

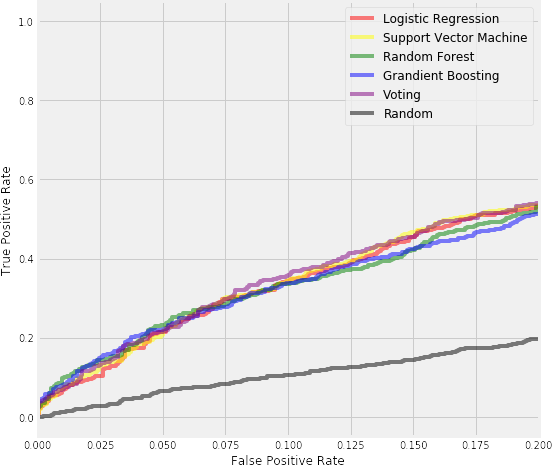


Figure 12. A section of ROC curves of different classification models.

I also evaluated the models using their ROC curves. In this particular problem, lower false positive rate is more important than higher true positive rate. In other words, it is more important to be sure that a player will improve as predicted, rather than predict all players who will improve, simply because a team can only have limited number of players. In the ROC curves with low false-positive rate, the voting model had slightly higher true positive rates than other models (Figure 12).

# Conclusions

In this study, I analyzed the relationship between NBA players’ improvement/decline and their performance and biographic data. I identified age, win share, minutes/games played, improvement last season among the most important features that affect a player’s improvement next season. I built both regression models and classification models to predict whether and how much a player would improve/decline. These models can be very useful in helping NBA team management in a number of ways. For example, it could help identify players to acquire, estimate the size of the contract to offer players, plan for performance changes of players already on the team, etc.

# Future directions

I was able to achieve ~26% improvement from the benchmark model in the regression problem, and ~68% accuracy in the classification problem. However, there was still significant variance that could not be predicted by the models in this study. I think the models could use more improvements on capturing players’ individual traits. For example, two players might have similar performance metrics, but one might be more physical and the other might be more finesse. The future performance of these two types of players might be different. Another example is that players whose contracts are expiring might play harder/better than players who just signed hefty contracts. More data, especially data of different types, would help improve model performances significantly.

Models in this study mainly focused on individual features. However, interactions with teammates, coaches, might also contribute to a player’s performance. For example, if a player had a new teammate who is a superstar at the same position, his performance is likely to suffer because of competition. These interactions data are obviously more difficult to extract and quantify, but if optimized, could bring significant improvements to the models.