Trouble with the Curve: Predicting Future MLB Players Using Scouting Reports

Jacob Danovitch*
Department of Computer Science
Carleton University
Ottawa, Canada
jacob.danovitch@carleton.ca

Date: October 11, 2019

Abstract

In baseball, a scouting report profiles a player's characteristics and traits, usually intended for use in player valuation. This work presents a first-of-its-kind dataset of almost 10,000 scouting reports for minor league, international, and draft prospects. Compiled from articles posted to MLB.com and Fangraphs.com, each report consists of a written description of the player, numerical grades for several skills, and unique IDs to reference their profiles on popular resources like MLB.com, FanGraphs, and Baseball-Reference.

With this dataset, we employ several deep neural networks to predict if minor league players will make the MLB given their scouting report. We open-source this data to share with the community, and present a web application demonstrating language variations in the reports of successful and unsuccessful prospects.

1 Introduction

In the 2012 film *Trouble with the Curve*, Clint Eastwood portrays a scout for the Atlanta Braves whose traditional methods lie at odds with new statistical analyses. The film shares a look into the divide between the two major schools of thought in professional sports - the 'eye test', and 'advanced analytics.' The former method relies on closely studying players and making evaluations based on expert knowledge, whereas the latter advocates for data-driven decision making and the use of statistics to project future value. The work presented in this paper argues that this divide presents a false dichotomy - that both the eye test and advanced analytics have a place in the evaluation of baseball players.

To find a common ground between both sides, we use statistical and deep learning methods in combination with the expert knowledge contained in scouting reports to predict if an amateur baseball player will make it to

*Work primarily completed during internship at Microsoft.

the major leagues. Further, we demonstrate language variations in the reports of successful and unsuccessful prospects, and we open-source this data to share with the community**. This dataset contains almost 10,000 scouting reports for minor league, international, and draft prospects, consisting of written descriptions, numerical grades, and various pieces of metadata.

2 Related Work

Projecting the future performance of amateur athletes has long been of interest to analysts of all sports. However, most previous contributions have focused on the use of player statistics as features; very few have attempted a text mining approach. As such, we believe this dataset to be the first open-source collection of baseball scouting reports, and the largest open-source collection of scouting reports for any sport.

Statistical projections. Given the absence of scouting reports such as those presented here, most attempts to project the future performance of amateur baseball players have used box-score data and statistical measures as features. One such example is the KATOH system, developed to project the future performance of minor league baseball players. The system uses each player's statistics to model a probability distribution of future levels of performance, as well as estimating the probability that the player will make the major leagues. The author notes that the system fails to account for specific player traits like defense and speed, leading it to "underrate players who man premium defensive positions–like catcher, center field, and shortstop–whose offensive abilities may not be the most valuable part of their game" [6]. Similar work was presented in [1], which developed a method to translate minor league statistics to future performance. The method used Weighted Runs Created Plus (wRC+), a measure of offensive performance relative to league average, as well as the player's age and level of competition to project their future wRC+ in the major leagues. Lastly, in a three-part series [2] that appeared in the popular sports publication 'The Ringer', the authors were able to acquire 73,000 scouting reports from a former member of the Cincinnati Reds baseball organization. Unfortunately, the authors did not examine the written descriptions, opting instead to use numeric grades (also presented in our dataset). These grades demonstrated a moderate correlation to both player performance and career length, though use of the written descriptions could have created a stronger result.

Text mining approaches. Comparatively speaking, little work has been done using written descriptions to project future performance. [8] presents an excellent analysis of hockey scouting reports using text mining, performing keyword and topic extraction and modeling current player performance using written descriptions. Unfortunately, the dataset used in their work is not readily available. Their work was expanded upon in a later contribution very similar to that presented here. [3] constructs a dataset of scouting reports for amateur hockey

**https://github.com/jacobdanovitch/Trouble-With-The-Curve

players and uses the reports to predict the success of each prospect, defined as whether a player appeared in more games than the average for his cohort. This work shares the same intent as ours, and is entirely open source. The two distinctions of our work are (1) the size of the dataset (just under 1300 reports, versus nearly 10,000 for ours) and (2) the sport in question (hockey versus baseball). Hopefully, future work will replicate these studies for a wide variety of other sports.

3 The TWTC Dataset

3.1 Scouting Reports

The main contribution of this work is to open-source the *Trouble with the Curve* (TWTC) dataset for use by the community. The dataset contains almost 10,000 scouting reports for amateur baseball players of varying levels, such as minor leaguers, international prospects, and draft prospects. Each scouting report consists of several features.

Written descriptions. Each report features a parapgraph-length description of the player's strengths and weaknesses. To the best of our knowledge at the time of writing, this is the first public dataset containing these descriptions. The descriptions are written by MLB.com Prospect Pipeline and Fangraphs.com writers, well-respected staffs who provide extensive coverage of amateur baseball. Each description will generally summarize the recent performance of the player, describe strengths and weaknesses, and project future performance. The descriptions use highly domain-specific language, as shown in the example below (emphasis added):

> "Andujar received the highest bonus in the Yankees' 2011 international class, signing for $750,000 out of Venezuela. As a 19-year-old in the low Class A South Atlantic League, he struggled early in 2014 but rebounded to hit .319/.367/.456 in the second half. Andujar combines bat speed with an advanced approach for his age. He has shown an ability to catch up to quality fastballs and make adjustments against offspeed pitches. He has the potential to produce 20 or more homers per season, and if he can develop some more plate discipline, he should hit for average as well. Though he has committed 51 errors in 196 pro games at third base, Andujar has the tools to become a capable defender. He has *a cannon arm and good hands* but needs to learn to *not try to do too much at the hot corner*."

Numeric grades. Each report assigns numeric grades for a variety of skills to each player. Grades are assigned along the "20-80 scale", a fixture in baseball scouting. The player is assigned a grade between 20 and 80 for each skill, with 50 as the average score, and every 10 points representing one standard deviation from the mean. These skills can include a batter's ability to hit for power or run quickly, the speed and control

of a pitcher's fastball, and so on. A score of 20 would represent a skill that is severely inadequate for the major league level, whereas a score of 80 would represent a hall-of-fame level skill. These grades are then aggregated into an overall score for the player. Importantly, the grades are *projections*, rather than evaluations; the goal of scouting is to predict future performance. Table 1, the data for which was originally presented in [5], outlines the meaning of each score in detail, relating each score to an expected amount of Wins Above Replacement (WAR). The TWTC dataset presents overall grades, as well as grades for: batting (contact hitting, power); defense (speed, fielding, throwing strength); and pitching (a grade for each pitch type, as well as overall control).

Scouting Scale	**Hitter Role**	**Hitter WAR**	**Pitcher Role**	**Pitcher WAR**
20	Org guy	—	Org Guy	–
30	Up & Down	<-0.1	Up & Down	< -0.1
40	Bench Player	0.0 to 0.7	Backend starters, FIP typically close to 5.00	0.0 to 0.9
45	Low End Reg/Platoon	0.8 to 1.5	#4/5 starters, FIP approx 4.20	1.0 to 1.7
50	Avg Everyday Player	1.6 to 2.4	#4 starters. Approx 4.00 FIP, at times worse but then with lots of innings	1.8 to 2.5
55	Above Avg Reg	2.5 to 3.3	#3/4 starters. Approx 3.70 FIP along with about 160 IP	2.6 to 3.4
60	All Star	3.4 to 4.9	#3 starters, 3.30 FIP, volume approaching 200 innings	3.5 to 4.9
70	Top 10 overall	5.0 to 7.0	#2 starters, FIP under 3, about 200 IP	5.0 to 7.0
80	Top 5 overall	> 7.0	#1s. Top 1-3 arms in baseball. 'Ace' if they do it several years in a row.	>7.0

Table 1: WAR mapped to 20-80 scale. Data sourced from [5].

Identifiers. Lastly, each report contains several unique identifier keys to link each profile to the player's statistics on popular resources such as MLB.com, FanGraphs.com, and Baseball-Reference. This presents an exciting opportunity for further studies on the predictive value of these scouting reports, such as modeling future WAR instead of a binary classification.

3.2 Corpus Analysis

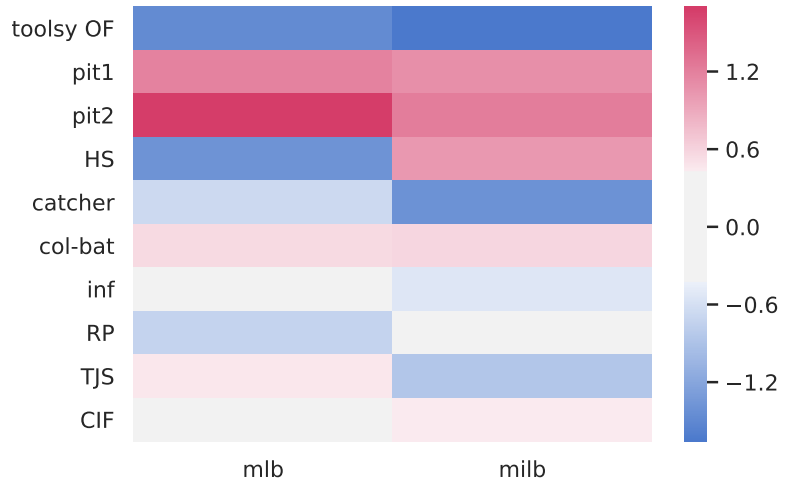

Figure 1: Latent topics identified by non-negative matrix factorization, by class label.

Before attempting classification, we can examine the linguistic differences in the scouting reports of prospects who were successful versus those who weren't. Non-negative matrix factorization was used to identify 10 latent topics, which were assigned a label based on the 10 most representative words of each topic. Further details are shared in a Jupyter notebook on the project's GitHub repository. The results of the model shows that reports for unsuccessful prospects mention catchers and high school prospects, two of the harder classes to predict. Interestingly, a report that discusses an injury - primarily elbow injuries associated with Tommy John Surgery - is more likely to belong to a successful prospect than an unsuccessful one. One potential reason for this could be selection bias, as players who fail to return from injuries are less likely to be written about (and thus appear in the dataset), as opposed to the notable stories of those who do.

4 Classification Task

To demonstrate the utility of our dataset, we seek to answer the research question: "Using a player's scouting report, can we predict if they will make the major leagues?" All metrics are evaluated on the same out-of-sample testing set. We use models implemented in the `Scikit-learn` [7] and `Pytorch Transformers`* python packages to perform simple text classification. The reports are encoded into fixed-size vectors using Term Frequency-Inverse Document Frequency (`Tf·Idf`) and the RoBERTa [4] language model, which are then passed to a neural network to perform binary classification.

*https://github.com/huggingface/pytorch-transformers

To prevent data leakage, it is crucial to mask certain tokens in the scouting reports such as player names, numeric quantities, and to a lesser extent, references to names of organizations. To illustrate the importance of this step, consider players who appear in the dataset more than once (for example, in multiple years). If the player has a unique name (for example, Byron Buxton), and he appears in both the train and test sets, the model can exploit mentions to his name in the report to make the correct classification. Numeric quantities can also have an important effect - higher numeric values (throwing harder, a higher signing bonus) will also likely bias the model. We elect to mask this data as well, though the necessity of doing so is less than that of masking the players' names.

Finally, we exclude players who are too young to yet be considered successful or unsuccessful in their careers. We make this distinction by fitting a simple regression to predict when a player should be expected to make the major leagues based on their age, and eliminate all prospects who currently have not yet made the major leagues and would not yet be expected to have done so. The process is demonstrated in a notebook on the project's GitHub.

Model	Accuracy	Balanced Accuracy	F-1 Score
`Tf·Idf`	0.8249	0.6558	0.4660
RoBERTa	0.8086	0.6861	0.5100

Table 2: Model performance for classifying success using scouting reports.

Table 2 demonstrates the performance of each classifier. While both show strong accuracy, it is important to note that the labels are heavily imbalanced, with 80% of the labels being negative. For this reason, we also consider the balanced accuracy (normalized with respect to the label distribution) and binary F-1 score. Considering the heavy imbalance as well as the aforementioned masking steps, the classifiers still perform fairly well. This result provides evidence for the possibility that scouts are able to author high-quality descriptions of players that can be used to predict the future performance of amateur baseball players.

5 Discussion

In this work, we present the *Trouble with the Curve* (TWTC) dataset of scouting reports for amateur baseball players. Each report contains written descriptions, numeric grades, and player identifiers. We then use this dataset to predict if an amateur baseball player will make the major leagues. We show that written descriptions provided by baseball scouts hold some predictive capacity for this task, providing an endorsement for the role of the scout in major league baseball. The dataset used for this task is shared with the community.

As seen in previous works, statistical approaches have also shown success in similar tasks. Future contributions could look to integrate player statistics with the written descriptions seen here to create a rich set of features, which we believe would greatly improve upon the results seen here. Other potential directions could

include analyses of the scouting reports themselves, analyzing the language use in the written descriptions and evaluating the predictive capabilities of the numeric grades.

References

[1] **Carruthers, Chris**, "Projecting Hitting Prospects' MLB Primes," *Breaking Blue*, 2014.

[2] **Lindbergh, Ben and Rob Arthur**, "Our Deep Dive Into 73,000 Never-Before-Seen MLB Scouting Reports," *The Ringer*, 2019.

[3] **Liu, Matthew**, "NHL Prospect Classifier," https://github.com/mattjliu/NHL-Prospect-Classifier 2019.

[4] **Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov**, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," 2019.

[5] **Longenhagen, Eric and Kiley McDaniel**, "The New FanGraphs Scouting Primer," *FanGraphs*, 2018.

[6] **Mitchell, Chris**, "KATOH: Forecasting Major League Hitting with Minor League Stats," *The Hardball Times*, 2014.

[7] **Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay**, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, 2011, *12*, 2825–2830.

[8] **Seppa, Timo, Michael E Schuckers, and Mike Rovito**, "Text Mining of Scouting Reports as a Novel Data Source for Improving NHL Draft Analytics."