Variational Autoencoders for Player Evaluation

Brooke Arnold, Geoff Converse, Siddarth Kannan, Zexing Song

University of Iowa

Abstract. In the sporting world, baseball has been quicker to embrace the use of data analytics than any other sport, as detailed baseball statistics have become readily available in large and diverse quantities to the general public. Professional baseball teams use this data to develop game plans and evaluate players. In this work, we explore the latter by using a Variational Autoencoder (VAE), a special class of artificial neural networks. Specifically, we wish to relate a player's season-long offensive statistics with the latent skills that a professional athlete needs to succeed in the MLB. This same model has also been used in educational data, using a VAE to predict latent traits of students based off their exam scores. While neural networks often suffer from a lack of interpretability, we are able to interpret the activation values of nodes in a hidden layer as distinct underlying skills because of a modification to the VAE architecture. In addition to assessing specific skills of athletes, we also rate a player's overall performance in a given season. In the growing field of sports analytics, we find this work incredibly important as it allows us to predict specific athletic skills based on years of recorded statistics.

Keywords: Sports Analytics, Neural Networks, Player Evaluation

1 Introduction

Baseball analytics has been steadily growing and expanding for decades now, as new methods are constantly being formed and the game is continuously changing. Specifically, professional teams have began using statistical analysis to evaluate player performance. Doing this allows them to identify the top performers to build their teams with. Many methods of player evaluation rely on perfecting and adapting certain evaluation statistics such as Wins Above Replacement (WAR) or Weighted Runs Created Plus (WRC+), which measure the overall value of a player. In this work, a variational autoencoder (VAE) model is proposed for player assessment. To do this, we relate a player's season-long batting statistics with four latent skills (Contact, Power, Baserunning, and Pitch Intuition) that a professional athlete needs to succeed as an offensive player in the MLB. To determine how well our model works, we compared these four latent abilities to similar offensive statistics. We also created a composite offensive trait using a VAE that only learns a single skill, and compared it to WRC+ to determine if our skill makes sense compared to a widely used offensive evaluation statistic. "Skill" is a very abstract concept, and representing it as a simple linear combination of statistics (as WAR and WRC+ attempt to do) would be rather difficult. A

VAE is not interpretable, but is useful to capture the complicated relationship between batting stats and offensive skills.

To begin, we describe previous methods of player evaluation, as well as previous work that uses VAE in educational cognitive assessment. We then describe our data collection and cleaning process through an MLB player statistics storage website, Fangraphs.com [1]. After describing the model, we give the results of our latent skill prediction and our composite talent prediction. Finally, we summarize the importance of our work, and how it can be improved and expanded upon in the future.

2 Background

2.1 Player Evaluation

Sports analytics is quickly growing in both popularity and in analysis methods. In 2011, the movie "Moneyball" really brought this idea to the surface as it tells the story of the 2002 Oakland Athletics who used player statistics to identify potential in underrated players [2]. Statistician Bill James developed a Multiple Linear Regression of a few independent variables that take into account (i) a player getting on base, (ii) that player advancing, and (iii) the player's opportunity to score. This method was adopted by the Oakland Athletics. Though simple, it was a very new and exciting method in the sporting world that allowed the team to greatly build their success and their monetary value as they purchase these cheap, underrated players showing potential to score runs.

Since then, sports analytics has taken off in the world of baseball. New techniques and approaches have been formed to enhance the game and improve player evaluation. Though many teams do not share their methods with the public, others have created their own models of analysis. One model created in 2015 by Benjamin Baumer, Shane Jensen, and Gregory Matthews measures player performance by creating a new version of the Wins Above Replacement (WAR) statistic called openWAR [3]. WAR is interpreted as how many more/less wins a team would have if the certain player was not playing. Their model is beneficial because it uses data that is reliable and reproducible, as "current versions of WAR depend upon proprietary data, ad hoc methodology, and opaque calculations" [3]. This model adjusts for offensive run values by calculating both baserunning values and hitting values, as well as adjusting for defensive run values by calculating both pitching and fielding values. Player k's WAR statistic is calculated as follows:

$$WAR_k = \frac{RAA_k - RAA_k^{repl}}{10}$$

where RAA_k is the Runs Above Average of all plate appearances involving a player k and RAA_k^{repl} is the Runs Above Average of the replacement player for player k. RAA is calculated by a combination of formulas that adjusts for the player's hitting, pitching, baserunning, and fielding, [3] and is interpreted as the number of more/less runs that player contributes to their team offensively compared to the average player. A replacement player is defined as “the typical player that is readily accessible in the absence of the player being evaluated.” By combining all of these facets of the game together, they found that the average openWAR value among players in the 2012 season was 0.91 with the best player being Mike Trout at 8.6 wins above replacement and the worst being Nick Blackburn at -2.6 wins above replacement. This model is extremely helpful as it allows us to rank players in certain facets of the game (i.e. as a pitcher, as a batter, etc.), and as overall baseball players.

Another model by Nikolas Furnald is also centered around the WAR statistic [4]. Furnald’s work focuses on using a player’s age to predict his WAR stat by creating the following regression model:

$$\begin{aligned} WAR \approx y = \beta_0 &+ \beta_1 * age + \beta_2 * age^2 + \beta_3 * steroid + \beta_4 * steroid * age \\ &+ \beta_5 * steroid * age^2 + \beta_6 * AL + \beta_7 * HighMound + \beta_8 * DH \\ &+ \beta_9 * DH * AL \end{aligned}$$

where all β_is are constants. This model also uses dummy and interaction terms to take into account the impact of the steroid era, as well as the fact that the pitching mound was lowered in 1969, as that may have had an impact on both pitchers and hitters. Lastly, this model adjusts for the fact that only teams in the American League are able to utilize a designated hitter. This impacts batters as older players often contribute more offensively than defensively, and this rule allows them to solely bat in the game. Furnald found that “the steroid era allowed players to peak at a higher level as well as an older age.” [4] Overall, this work is specifically beneficial to player evaluation during the steroid era, but provides a good method of player evaluation.

2.2 Variational Autoencoders

An autoencoder is a neural network where the input and output layers are the same size. Here, the goal is to encode the data into a low-dimensional representation, and then reconstruct (decode) the original input. Similar to principal component analysis, the lower dimensional data can be used in some kind of regression model, or to store the data more efficiently.

A variational autoencoder (VAE) has a similar structure. But it also aims to map the low-dimension representation θ to some probability distribution, typically $\mathcal{N}(0, I)$ [5]. So if we want the encoded data to be k-dimensional, we actually need two sets of size k of densely connected nodes - one of these represents the mean, and the other the variance of the latent space. In training, we feed-forward data to the encoded space, then sample from this distribution, passing the sample through the decoder to be reconstructed. The idea is that given a latent representation θ, there exists some function $p_\beta(x|\theta)$ that maps the low-dimensional data to our original inputs. Similarly, given our input data x, there exists some approximating function $q_\alpha(\theta|x)$ that will fit θ to a normal distribution. A visualization of this architecture is shown in Figure 1.

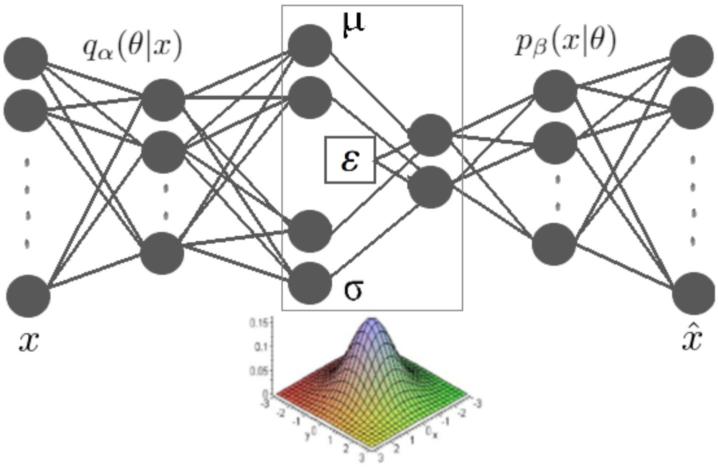

Fig. 1: A visualization of a VAE structure.

In order to fit this distribution, we need to have another term in our network's loss function in addition to the reconstruction error. The Kullback-Liebler Divergence measures the expectation of the log-difference between a two distributions [6]. In our case, we wish to measure the KL divergence between $p(\theta) = N(0, I)$ and our approximating function $q_\alpha(\theta|x)$, the encoder. So we have

$$KL(p||q_\alpha) = E[\log(p(\theta)) - \log(q_\alpha(\theta)) = \sum_{i=1}^{n} p(\theta_i) \log \left(\frac{p(\theta_i)}{q(\theta_i}) \right)$$

In our neural network, q_α is represented as the encoder, with parameters α being the weights and biases.

Though KL Divergence is difficult to compute, more convenient representations can be found [7]. Then when training our neural network, the aim is to minimize both reconstruction error and KL Divergence, thus ensuring that our low-dimensional representation is both accurate and normally distributed.

2.3 VAE in Item Response Theory

Formal examinations are given to students at every level of education in order to assess knowledge of course material. But it is difficult to quantify and accurately assess an abstract notion of "knowledge" - how can one guarantee that a subject's performance on certain items of an exam correlate with that subject's ability? Item response theory attempts to model this problem. One popular method is the Multidimensional Logistic 2-Parameter (ML2P) model[8], which gives the probability of a student answering a given question correctly, given their skill values. Let $X_{ik} = 1$ if student k answers item i correctly, and 0 otherwise. Define

$\Theta_k = (\theta_{1k}, ..., \theta_{Jk})^T \in \mathbb{R}^J$ be the set of student k's J latent skills. Then

$$Pr(X_{ik} = 1|\Theta_k) = \frac{1}{1 + \exp\left(-\sum_i a_{ij}\theta_{jk} - b_i\right)}$$

where the discrimination parameter a_{ij} represents how necessary skill j is to answering item i correctly, and the b_i represents the difficulty of item i. Typically, $a_{ij} \geq 0$, since it is assumed that more knowledge will always increase performance. Note that the ML2P model predicts student performance given their skills, rather than predict skills given examination results.

Notice that if $a_{ij} = 0$, then item i does not require skill j at all. In order to keep track of which skills relate to each item, we can define a binary Q-matrix [8]:

$$Q_{ij} = \begin{cases} 1 & a_{ij} > 0 \\ 0 & a_{ij} = 0 \end{cases}$$

Recently, Curi et. al. [9] proposed using a VAE to predict latent skills of students, given their exam results. The input to the network is a binary vector representing correct (1) or incorrect (0) answers to an exam. The low-dimensional distribution represents the skills of students. This work used a modified neural network architecture, with no hidden layers in the decoder. Instead, the connections between the learned distribution (representing the latent skills) and the output layer (a reconstruction of students' exam results) are determined by the Q-matrix. This allows for interpretation of the network.

In Curi et. al.'s paper, the response set of students was simulated from the ML2P model after fixing latent skills. It is important to note that the VAE was trained to reconstruct these responses, not minimize the difference between predicted and actual latent traits. Nonetheless, the predicted latent traits serve as highly correlated estimates to the true values. Additionally, after equipping the output layer with a sigmoidal activation function, the weights in the decoder can be interpreted as discrimination parameters, and the biases of the output layer as difficulty parameters. So after training, the decoder of the VAE essentially becomes the ML2P model. At test time, test scores are sent through the encoder to obtain latent skill estimates.

The sports analytics research described in our paper is inspired by Curi et. al.'s work in applying VAE to psychometrics. After determining which baseball skills correlate with measurable statistics, we construct an analogue of the Q-matrix, then use this to determine the connections in the decoder of a VAE. Since there is no analogue in sports analytics to the ML2P model, the weights/biases in the decoder can not be interpreted as parameters from another function. Additionally, the values in the learned low-dimensional representation aren't estimates to any known particular measure. Rather, they will represent a *new* way to measure the underlying skills of athletes.

3 Data Collection and Pre-processing

We gathered our data from FanGraphs.com [1], which is a public website that collects and stores decades of professional baseball statistics. Our model uses all offensive statistics from players between the years of 1960 and 2018, split by player and season. In all, we have 8,604 data points.

When looking through the data, we got rid of the variables that we did not find important to offensive production, such as Times on Base (TOB) and Hit by Pitch (HBP). These statistics are not closely related to the latent skills we are trying to measure. We then divided all counting statistics by Plate Appearances, so that our modl wouldn't favor players who had more opportunities. Next, we normalized all chosen variables using Gaussian normalization to keep all statistics on the same scale [10]. This is a common technique in data science, but is especially important in neural networks so that each input node is considered equally by the network. For a data point $x = (x_1, ..., x_M)^T$ with M features, we normalize this element-wise by

$$z_m = \frac{x_m - \mu_m}{\sigma_m}$$

where μ_m and σ_m are the mean and standard deviation of feature m over all samples. We then write the normalized data sample as $z = (z_1, ..., z_m)^T$. This normalization is often called a z-score, which gives the number of standard deviations from the mean for each data point.

For certain statistics, a large number indicates a negative consequence for the offensive player. We formatted the statistics so that bigger z-scores correlate with better performance and higher skill. For example, a high number of strikeouts (K) is bad for a hitter, so we consider $(-K)$ instead. Because of this adjustment, we can make the assumption in our network that higher skill values in the latent distribution will produce larger numbers in the output layer, interpreted as better stats.

4 Model Description

4.1 Latent Skill Prediction

Our work aims to find connection between a baseball player's batting statistics and offensive skills through a variational autoencoder, an unsupervised learning method. The skills we wish to measure are Contact, Baserunning, Power, and Pitch Intuition. We define Pitch Intuition to measure whether a player is swinging at the pitches that he should be swinging at (strikes), and refraining from swinging on those he shouldn't (balls). The other skills are defined exactly as one would expect. As mentioned earlier, we develop a sort of Q-matrix, where $Q_{ij} = 1$ if skill j influences the statistic i. This matrix is shown in Table 1, and also specifies the statistics we used as inputs/outputs of the neural network.

In our VAE, the encoder has one hidden layer that consists of ten nodes. This connects to the latent distribution, which has two sets of four nodes, representing

Contact	Baserunning	Power	Pitch Intuition	Statistic
1	0	0	0	Singles (1B)
1	0	1	0	Doubles (2B)
0	0	1	0	Homeruns (HR)
0	1	0	0	Runs (R)
0	0	1	0	Runs Batted In (RBI)
0	0	0	1	Walks (BB)
0	0	1	0	Intentional Walks (IBB)
1	0	0	1	Strikeouts (K)
1	0	0	0	Sacrifice (SAC)
0	1	0	0	Grounded into Double Play (GDP)
0	1	0	0	Stolen Bases (SB)
0	1	0	0	Caught Stealing (CS)
0	0	0	1	Walk/Strikeout Ratio (BB/K)

Table 1: Q-matrix relating skills and measurable statistics.

the mean and standard deviation for each of the four skills. After sampling from this distribution, data points are sent through the decoder, which has no hidden layers. Instead, we use the Q-matrix to determine connections between the four nodes in the distribution layer to the thirteen output nodes. This requires all weights which connect unrelated skills/stats to be zero. In addition, we force all weights in the decoder to be non-negative. This helps reinforce our assumption that high skill values will produce better statistics.

To implement the Q-matrix in the decoder, we create a custom kernel constraint function [11]. After each weight update when training, we consider the weight matrix $W \in \mathbb{R}^{13\times 4}$ in the decoder and set $\overline{W} = W \odot Q$. Here, $\odot$ represents element-wise multiplication. Note that this operation simply zeros out weights that we do not want in the network, while not changing anything else. If an entry in $\overline{W}$ is negative, we set it to zero. Then the weights of the decoder are set to be $\overline{W}$.

4.2 Composite Talent Prediction

In addition to training a network to learn the four baseball skills, we also created a network which would give an overall player rating. We built a VAE that uses the same baseball statistics as before to give a new Composite Talent score, which we compare to an existing statistic. For this evaluation stat, we considered two measures, WAR and WRC+, which are both popular among baseball analysts. WAR (Wins Above Replacement) considers not only the batting and baserunning statistics that we have discussed in this paper, but also pitching and fielding stats. As such, we do not wish to compare our Composite Talent to WAR, but Weighted Runs Created Plus (WRC+) instead. WRC+ is used as a relatively reliable measure to quantify the player's offensive value [1]. It aims to determine how many runs a player contributes to their team over the course of a season. All of the same statistics described earlier are used in computing WRC+.

Since we now only wish to obtain a single skill, our VAE only has one dimension in the learned normal distribution. In this model, we require a deeper architecture in the encoder: 4 hidden layers with 10, 5, 10, and 5 nodes respectively. This complexity is required as it is difficult to map 13 statistics to $\mathcal{N}(0,1)$ with a shallow architecture. Again, the decoder has no hidden layers, and we require the decoder weights to be non-negative. Similar to before, the assumption here is that more skill will always correlate with better statistics. However, we do not use a Q-matrix - the weights in the decoder are densely connected, as overall skill relates to all measurable statistics.

5 Experimental Results

5.1 Latent Skill Prediction

Using the previously described modified VAE with our Q-matrix in the decoder, we are able to obtain new evaluation measures for four skill areas: Contact, Baserunning, Power, and Pitch Intuition. In order to evaluate our model's performance, we compare these skills to the commonly used statistics Batting Average (AVG), Speed Score (SPD), Isolated Power (ISO), and On-Base Percentage (OBP), respectively. AVG gives the percentage of hits per at-bat. SPD is a composite baserunning statistic developed by Bill James [12]. A common measure of a hitter's power is ISO, which can be interpreted as the number of extra bases per at-bat. And OBP is a similar metric to AVG, but it also takes into account the number of times a batter is walked or hit by a pitch.

It is important to notice that we did not input any of these statistics into our VAE, though each can be calculated with formulas involving stats which are set as input nodes. Three of them are easy to compute from basic recorded stats:

$$\begin{aligned} AVG &= \frac{(H)}{(AB)} \\ ISO &= \frac{(2B) + (2 \cdot 3B) + (3 \cdot HR)}{AB} \\ OBP &= \frac{(H) + (BB) + (HBP)}{(AB) + (BB) + (HBP) + (SAC)} \end{aligned}$$

SPD is significantly more complicated to compute, and we omit its formula. It takes into account stolen base percentage, frequency of stolen base attempts, percentage of triples, and runs scored percentage [12]. All of these evaluation statistics are used to quantify the competency of baseball players in each skill.

After training our network on 7,600 data points (selected randomly), we run the remaining 1,003 player/seasons through the encoder to obtain their latent skill ratings. We calculate the correlation between our new skill measures and their respective evaluation statistics. We see high correlation in Power/ISO

(0.9033), and significant correlation in Baserunning/SPD (0.7953) and Pitch Intuition/OBP (0.6841). There is little to no correlation between Contact/AVG (0.2138). However, this alone does not necessarily mean that our Contact measure is bad. It's possible that we are simply not using the best evaluation statistic (i.e. AVG).

It was previously mentioned that the evaluation stats are computed using stats that serve as input/output nodes in the VAE. We should further remark that the nonzero entries in the Q-matrix seen in Table 1 often line up with these same stats. For example, Power is related to the stats 2B, HR, RBI, and IBB; two of these are found in the formula for ISO. So intuitively, it makes sense that our Power measure is so highly correlated with ISO.

Note that while high correlation is good, we are not necessarily interested in maximizing these quantities. We do desire our skill measures to be similar to the evaluation statistics, but we don't want them to be *exactly* the same. Our goal is not to reproduce the evaluation statistics - rather, our model provides brand-new measures of these skill areas. As such, we are not concerned with error terms like root mean squared error. The evaluation statistics really just give assurance in the validity of our measures.

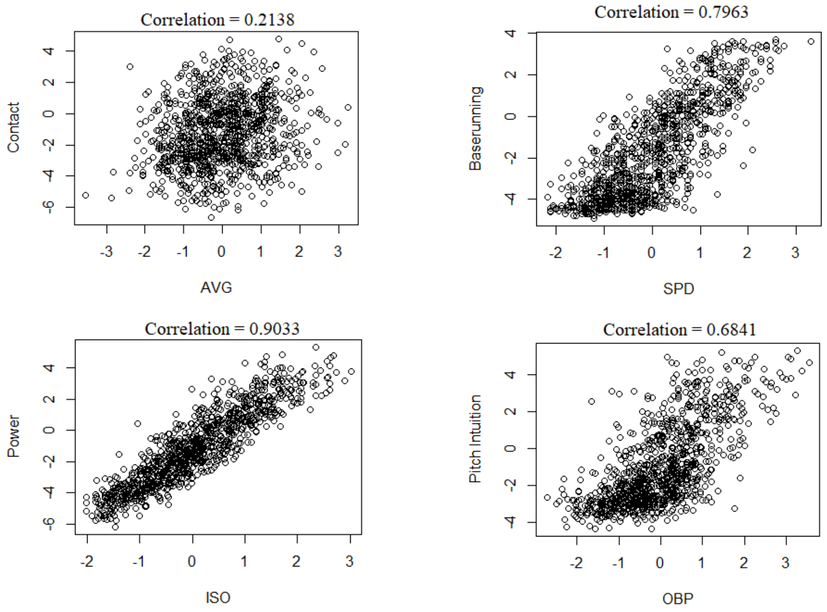

Fig. 2: Each latent skill plotted against its evaluation statistic.

We can learn more about our results by looking at the correlation plots, seen in Figure 2. There is a stark linear trend in Power/ISO, and a similar pattern in Baserunning/SPD and Pitch Intuition/OBP. Of course, the Contact/AVG plot is not very telling at all. But also notice that in the three relevant plots, the top of the graph (representing the highest scoring players) is relatively sparse. Star players are typically statistical outliers, and our new measures are able to identify them.

Besides comparing our new skill measures with existing statistics, we can look at the top performers in each skill, and see if those results are consistent with human judgment. This is shown in Table 2. As expected from the lack of correlation in our Contact skill, the top ranking Contact players don't make much of sense. However, the top Power hitters are all players that are well-known for their homerun-hitting ability - Bonds and Sosa famously went head-to-head in the early 2000's to chase the single season homerun record.

As for Baserunning, all of the top five performers were prolific base-stealers throughout their careers. In fact, Rickey Henderson is the MLB's all-time leader in stolen bases. Throughout the developmen of this project, we had thought that Pitch Intuition would be the most difficult skill to capture using the statistics that were available to us. But the best ranking individuals actually make a lot of sense. In his prime, Barry Bonds was one of the most dominant athletes of all time, and was capable at hitting all areas of the strike zone. Ferris Fain has the 13th best on-base percentage of all time. And the top ranking Pitch Intuition player, Elmer Valo, was known for having exceptional "strike zone judgment"[13]; this is exactly what we were trying to capture with this skill.

Contact	Power	Baserunning	Pitch Intution
1981 Tim Foli	2009 Mike Piazza	1980 Ron LeFlore	1952 Elmer Valo
1977 Bert Campaneris	2010 Miguel Cabrera	1982 Rickey Henderson	2004 Barry Bonds
1974 Len Randle	2008 Mike Piazza	1987 Vince Coleman	1953 Ferris Fain
1994 Felix Ferman	2003 Barry Bonds	1974 Lou Brock	2002 Barry Bonds
1979 Craig Reynolds	2001 Sammy Sosa	1981 Tim Raines	1973 Dave Rader

Table 2: The top five ranking player/seasons for each skill.

5.2 Composite Talent Prediction

A question that is debated endlessly on sports talk shows is "who is the G.O.A.T. - the greatest of all time?" This discussion is not unique to baseball - the quest to determine the best ever basketball, soccer, or tennis player is always of interest. Often, this discussion takes into account an athlete's entire career, including the caliber of their team accomplishments, along with individual feats and accolades. It can be especially difficult to compare athletes from different eras, as they may have had different styles of play, rules, and competition. But whether it's Jordan vs. LeBron, Pele vs. Messi, or Federer vs. Nadal, individual statistics are always a topic of discussion.

Our Composite Talent model focuses on a similar task as our latent skill predictions. We use season-long (individual) statistics, in order to rank the overall offensive productivity of baseball players. This has immediate applications for professional teams, as they want to obtain the best players available, and give them a favorable contract.

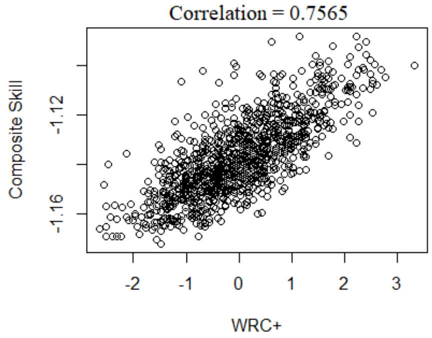

Fig. 3: Composite talent score plotted against weighted runs created plus.

In our experiments, we used the same train/test split of our data as before. On the test data, the correlation between our Composite Skill and WRC+ was 0.7565, which is significant. A clear linear relationship is also shown in Figure 3. This shows that our new overall offensive measure is comparable to those already in use. But how does it answer the question "who is the greatest of all time?"

After obtaining predictions on all collected data, our VAE model ranked the top ten offensive seasons of all time, seen in Table 3. Of the six unique players listed, four have been inducted into the Baseball Hall of Fame. Albert Pujols is still an active player, and thus not eligible for this honor. Barry Bonds has appeared on the voting ballot seven times [14], but has not been inducted due to his affiliation with performing enhancing drugs and steroids. Six of the ten slots are players who won the Most Valuable Player Award in either the American or National League in that season. Frank Thomas and Ted Williams both won this award in different years than listed (each twice).

6 Future Work

The work we've done thus far is very useful to offensive baseball analytics. However, there are ways to improve upon and expand our model. The most obvious improvement is to hone our Q-matrix with area experts. Our current Q-matrix is very subjective and relies upon our own opinions of which skills are necessary to achieve good statistics. By working with baseball experts, we expect to have a more precise matrix, and therefore a better functioning model. Another possible way to improve the Q-matrix is to use a VAE to develop it. By starting with a more dense matrix, and removing connections with low weights after repeatedly

Year	Player
2003	Barry Bonds*
2002	Barry Bonds*
1957	Ted Williams
1969	Willie McCovey*
2004	Barry Bonds*
1995	Frank Thomas
2009	Albert Pujols*
1954	Ted Williams
1950	Yogi Berra*
1996	Frank Thomas

Table 3: Top individual offensive seasons based on composite skill score. A * indicates a league MVP award winner that year.

training, we could possibly find a better matrix than the one we decided ourselves. This process is similar to the backwards feature selection method used in other machine learning techniques.

This work could also be enhanced by experimenting with different architectures. Our model is fairly simple, and could possibly benefit from a more complex structure. Also, we only chose 13 statistics to apply our traits to. Adding more could help solidify our latent skills, possibly making them correlate more with other evaluation statistics.

This type of work can also be extended to other sports. Basketball has a large number of latent traits associated with it, and would be a great place to start. Besides the fact that this would be another new application, the data would be interesting, because all work combining VAE with a Q-matrix has been performed on data with a very small number of latent skills.

7 Conclusion

The work we have done thus far is increasingly important in the world of sports analytics. By determining players latent traits and how they match up with certain batting statistics, we are able to evaluate players and determine their usefulness to a team. Using a variational autoencoder in offensive evaluation, we are able to pinpoint these underlying skillset of each player, something that cannot be done through the simple calculation of a WAR or WRC+ statistic. We believe that a player's skill is most likely not a simple linear combination of statistics, but much more complicated. Though uninterpretable, a VAE is capable of learning this complex relationship. Baseball analytics has come a long way since the emergence of the movie "Moneyball," but there is still much more that can be done as well. Baseball fans often prefer interpretability and simplicity over more rigorous methods. We hope that the compelling results of our research opens the door to a wider acceptance of unsupervised machine learning techniques in sports analytics.

References

1. FanGraphs.com. Mlb batting statistics. 2018.
2. Sayar Banerjee. Linear regression: Moneyball - part 1. *Towards Data Science*, Apr 2018.
3. Benjamin S. Baumer, Shane T. Jensen, and Gregory J. Matthews. Openwar: An open source system for evaluating overall player performance in major league baseball. *Journal of Quantitative Analysis in Sports*, 11(2), 2015.
4. Nikolas Ahrendt Furnald and Michael E. O'Hara. The impact of age on baseball players' performance how was this altered during the steroid era. 2012.
5. Carl Doersch. Tutorial on variational autoencoders. *arXiv:1606.05908*, 2016.
6. S Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 1951.
7. D. Kingma and M. Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.
8. M.A. da Silva, R. Liu, A.C. Huggins-Manley, and J.L. Bazan. Incorporating the q-matrix into multidimensional item response theory models. *Educational and Psychological Measurement*, 2018.
9. Curi, Converse, Hajewski, and Oliveira. Interpretable variational autoencoders for cognitive models. In *International Joint Conference on Neural Networks*.
10. James McCaffrey. How to standardize data for neural networks, Jan 2014.
11. Keras. Keras: The python deep learning library, 2019.
12. Steve Slowinski. Spd. *FanGraphs.com*, 2010.
13. Wikipedia. Elmer valo. 2019.
14. Dayn Perry. Baseball hall of fame: Barry bonds and roger clemens might have to wait longer, but they're trending toward induction. *CBS Sports*, 2019.