Using Soft Skills To Predict Labor Market Outcomes

Ben Cahill
(bjcahill@calpoly.edu)

Sahil Bobba
(sbobba@calpoly.edu)

Mason Ogden
(msogden@calpoly.edu)

Jay Ahn
(jaahn@calpoly.edu)

World Bank Group

DATA 452 - Data Science Capstone II

Dr. Dennis Sun and Dr. Jonathan Ventura

June 10, 2021

1. Introduction

In many developing countries, particularly across Africa, there is a high demand for highly-skilled, educated workers, and a surplus of unskilled, minimally-educated workers. These conditions perpetuate a cycle of poverty among youth who cannot find work. Research has shown that labor markets are changing and that soft skills – also called "non-cognitive skills" – are increasing in demand [1]. This suggests that underprivileged youth lacking equitable educational opportunities can develop soft skills to increase their employability and success in the workforce.

Although soft skills offer a promising way to reduce youth unemployment, very little research has been conducted on which specific soft skills are most relevant within different countries and industries. Furthermore, without concrete data to inform public policy, it is difficult for government agencies or organizations like the World Bank to develop programs to help youth find work based on their soft skills.

One of the major difficulties inhibiting research in this field is data availability. The most common sources of data in labor market studies are longitudinal surveys that follow youth into adulthood. These data typically lack explicit measures of soft skills in favor of technical measures like reading comprehension and math skills. Although many surveys include a small section on psychometric questions or personality traits, these non-cognitive questions are not the main focus of the survey and are often limited in scope.

Despite these limitations, longitudinal datasets contain implicit information about soft skills. With the right pre-processing methods and data analysis, it is possible to extract information about specific soft skill aptitudes from the psychometric questions that are present. In this project we use this insight to develop a data-agnostic approach to extract and measure various soft skills. Then, we use these soft skills to predict the success of youth in the labor market across different countries and industries. In particular, we focus on a comprehensive dataset from the United States to construct a generalizable model that finds associations between specific soft skills and market success.

2. Initial Goals

In order to link soft skills to market success, we broke down our project into three distinct sub-tasks. *Figure 1* describes the overall pipeline of our project.

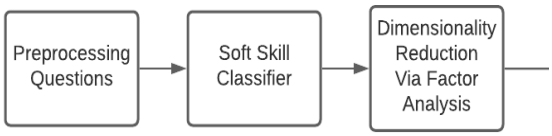
First, we needed to process and extract implicit information about soft skills from the psychometric questions available in the U.S. dataset. This could be achieved by grouping each question into a specific soft skill category and then reducing the dimensionality of each category.

Second, we needed to create a model to predict market success using explicit soft skill measurements and demographic

variables as predictors. If certain soft skills were strongly associated with market success, then these skills could be linked to employability.

Finally, our third task was to subset our data by industry. This would allow us to examine how important soft skills differed by occupation, allowing organizations like the World Bank to develop programs that specifically target these industries.

Figure 1: Model pipeline



3. Extracting Soft Skills

Throughout the course of the project, we tried different approaches to match questions with the soft skills they are measuring. Initially, we manually classified questions with our own intuition. However, this method proved too time consuming and introduced a significant amount of human bias. Instead, we decided to use machine learning to classify questions in order to reduce bias and streamline the process.

3.1 Psychometric Questions

In the US dataset, psychometric questions typically consisted of open-ended questions that asked participants how much they identified with certain prompts. For example, *"Are you relaxed during stressful situations?"* Questions involving a simple statement were also common, such as *"I finish whatever I begin."* Participants would then respond to these questions using an ordinal scale (e.g. *almost never, some of the time, most of the time, always*) that we converted to a numerical scale for analysis.

3.2 Sentence Transformer

Next, we converted the text of each question to a numeric vector representation. Natural language processing embeddings like Glove, Word2Vec, ELMo, or BERT are commonly used to do this. In our project, we utilized a pre-trained sentence transformer [2] based on the vanilla BERT model and fine-tuned it to our dataset. Unlike word-level representations such as Glove and Word2Vec, BERT captures the meaning of

words while taking into account the context of the sentence. Since our psychometric questions consisted of complete sentences, BERT was a good fit for the project.

Once we finished vectorizing each question, we tried unsupervised learning, supervised learning, and a hybrid approach to match questions to soft skills.

3.3 Unsupervised Learning

Unsupervised learning felt like a promising approach for extracting soft skills because it would let the data speak for itself. Using word similarity, the model could cluster questions into soft skills without needing a predefined set of labels going in, making the model more flexible. In our implementation, questions were clustered into groups using K-means clustering. We extracted the five most common words in each cluster, and tried to discern what exactly each cluster was measuring (shown in Table 1).

Table 1: Clusters and their most common words

Cluster	**5 Most Common Words**
1	time, term, short, project, previous
2	undependable, uncreative, trustful, thorough, finish
3	warm, upset, sympathetic, stable, self
4	tradition, support, setback, set, school
5	worker, work, standards, standard, require

While this approach did not involve any human bias or predetermined constructs, it did not provide fruitful results. The meaning of each group was either unclear or nonsensical in the context of soft skills. For example, cluster 1's most common words were 'time, term, short, project, previous', which are all very similar words, but not related to soft skills in any way.

3.4 Hybrid Approach

As illustrated in Table 1, a significant weakness of unsupervised learning is cluster interpretation. Without predefined labels, it is difficult to understand what each group represents. But, what if, as in a supervised approach, we knew the labels in advance? What if the model assigned labels to clusters based on the similarity between the label description and the question text in each cluster? This way, we could continue to let the data speak for itself, but also put some limiting parameters on the process to make each group more interpretable.

To implement this approach, we used a simple set of soft skills (10 for testing purposes), provided a detailed text description of each, and vectorized each description. Then, for each cluster, we concatenated the text from each question, vectorized the result, and compared the similarity between this text and each soft skill description. Using a greedy algorithm (first matching the most similar soft skill with the most similar cluster), we iteratively assigned a unique label to each cluster.

Unfortunately, the results of the hybrid approach were weak. While the clusters that were assigned first had solid interpretations (e.g. Grit/ Work Ethic was assigned to a cluster with work and motivation as key words), the rest of the clusters didn't make much sense. This was because the later the clusters were assigned, the greater chance the cluster label was simply the result of process of elimination. As such, we decided not to pursue unsupervised learning any further.

3.5 Supervised Learning

The biggest issue we encountered with supervised learning was the lack of an existing training set with a comprehensive, pre-defined set of soft skills. To get around this issue, we created our own data set of about 500 psychometric questions compiled from various sources online [5-10]. Next, we manually labeled each question with one of 18 soft skills that we selected with help from a domain expert. These skills consisted of the "Big Five" personality traits [3] as well as 13 other soft skills that were as comprehensive as possible (the fewer the labels, the easier it would be for our model to classify soft skills correctly). While this approach did introduce some human bias, it proved to be a reasonable strategy because its performance was quantifiable and its results were meaningful.

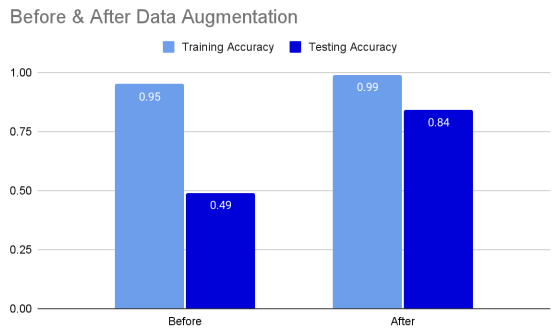
Table 2: Labels and their source

Label	**Source**
Openness	Big Five
Conscientiousness	Big Five
Extraversion	Big Five
Agreeableness	Big Five
Neuroticism	Big Five
Grit/ Work Ethic	Soft Skill
Communication	Soft Skill
Emotional Intelligence	Soft Skill
Teamwork	Soft Skill
Leadership	Soft Skill
Adaptability	Soft Skill
Problem-Solving	Soft Skill
Creativity	Soft Skill
Organization	Soft Skill
Time Management	Soft Skill
Negotiation / Persuasion	Soft Skill
Growth Mindset	Soft Skill
Self-Efficacy	Soft Skill

After testing many different classifiers, a support vector machine proved to be the most effective for labeling soft skills. However, the model was prone to overfitting. To remedy this, we added regularization terms to the model's objective function and performed data augmentation: substituting words in the questions with their

synonyms using NLTK's wordnet and a data augmentation library called nlpaug [4]. Using this method, we increased the size of our training data from 443 to 886 rows and ultimately achieved a test accuracy of 84%. Since our initial accuracies (due to overfitting) were less than 50%, this was a huge improvement. *Figure 2* displays the performance of the classifier before and after the data augmentation.

Figure 2. Model performance after data augmentation



3.6 Approach Selection

Based on the results from each machine learning method, it was clear that supervised learning produced the best results. As such, we decided to move forward with our support vector machine. Running the support vector machine on the U.S. data yielded 9 different soft skills with questions assigned to them. While it would have been best for all 18 soft skills to be represented, it was clear that the psychometric questions present in the U.S. data were not comprehensive enough to implicitly measure each of these soft skills.

4. Dimensionality Reduction

Once we assigned a soft skill to each question, we moved on to dimensionality reduction. Our goal was not to measure the relationships between individual questions and market success, but the overall relationships between soft skill/personality constructs and market success. Thus, we had to find a way to condense information from multiple questions into a single column that reflected an underlying soft skill.

We first tried principal components analysis. If multiple questions fell within a certain soft skill group, we extracted a single principal component that contained the maximum amount of variation. With any dimensionality reduction technique, information will be lost. In this case, the proportion of variation contained in just the first component varied by soft skill and the number of questions measuring that soft skill. As would be expected, for soft skills with one question, 100% of the variation was captured with a single component. However, for a Grit/Work Ethic, a soft skill with nine questions, only 24% of the total variance was captured in the first component. On average, a substantial amount of information was being lost in the process.

Another problem with principal component analysis was the fact that a single component could not be interpreted in a meaningful way. Questions within a soft skill could be measuring the skill on a negative scale (for example, a question measuring Organization could ask "how

disorganized are you on a scale from 1 to 5"), but we had no way of programmatically controlling whether questions were assigned a positive or negative loading in the first principal component. As such, we were not able to say "people with higher values of the extracted Organization component were more organized" since we simply didn't know the direction of the overall component.

To solve this issue, we tried to use a pretrained sentiment analyzer to determine if a question was positively or negatively worded. However, the performance of the analyzer depended very heavily on the structure of the input question. For example, if the question had a double negative, it failed to assign the correct sign. Furthermore, we believed that it added too much complexity to our general pipeline, leading us to consider a different dimensionality reduction technique: factor analysis.

Factor analysis proved to be a far superior approach for dimensionality reduction. The primary benefit was its ability to discern positively-worded questions from negatively-worded questions and assign coefficients accordingly. For example, within the "Grit/Work Ethic" skill, factor analysis assigned a positive coefficient to "I am diligent", and assigned a negative coefficient to "I do not work as hard as the majority of people around me." This classification follows our intuition, as not working as hard as other people detracts from a person's Grit/Work Ethic 'score', while being diligent adds to it. We could then say with confidence that people who have higher values of Grit/Work Ethic were harder workers. This allowed us to make real, meaningful interpretations about how these soft skills affected market success.

5. Predicting Market Success

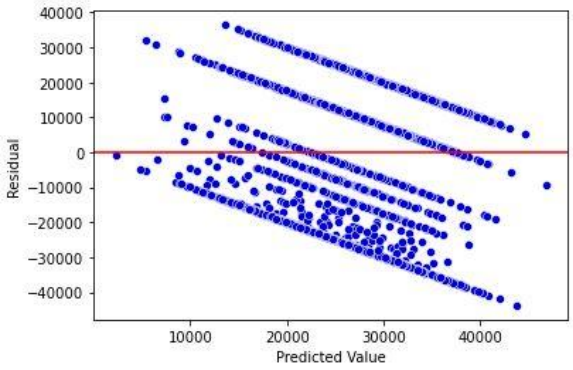
Measuring a person's success in the labor market proved to be difficult. It turns out that market success has many different definitions. The most obvious definition is salary: people who get paid more are more successful. However, this approach lacks nuance about what success means. Other definitions involve taking into account work benefits, while some of the most complex approaches create an "asset index" that quantifies the assets held by a worker.

In practice, the best information (the most generalizable to other datasets) we could use to define market success in the US data was salary. Salary was split into 16 'bins' representing ranges of yearly earnings, starting at $0 and ending with $50,000+. To convert this into a numeric response for regression, we replaced each range with its midpoint. For example, if a participant reported their salary as "$15,000 - $19,999", their response was changed to 17499.5. This resulted in 16 unique values in the response, ranging from 0 to 50,000.

All soft skill scores were then used as predictors of market success in a multiple regression model. Additionally, demographic attributes of the subjects (sex, age, and ethnicity) were added to control for differences in the participants. The resulting

model was fit using data from 3,227 participants, and had an adjusted R^2 of 8.4%. While this R^2 value is quite low, this model did not take into account industry or experience level.

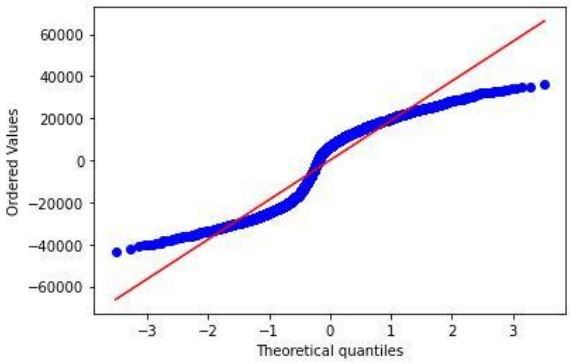
Figure 3: Residual vs. Predicted Plot



Due to the pseudo-categorical nature of the response, checking the assumptions of the model was difficult. Since the response variable could only take one of 16 values, the residual plot (*Figure 3*) contained many parallel lines of points, and was not a random scatter. However, no non-linear pattern appeared, so it can be concluded that the relationship between the predictors and the response is well represented by a linear model. Additionally, the distance between the lowest residuals and the highest residuals stayed constant over every predicted value, so there was not any heteroskedasticity in the errors.

Since this data came from a random sample, we are confident that the responses of the participants are independent from one another. However, problems arose when checking whether the residuals were normally distributed. Instead of a straight line, the normal probability plot of residuals (*Figure 4*) was S-shaped, indicating bimodality, with the dip in between the modes occurring at zero. Despite this departure from normality, we were confident that we could trust the results of the regression, since the substantial sample size (3,227) compensates for the problems in the residuals.

Figure 4: Normal Probability Plot



Once we were able to trust the standard errors of the estimated regression coefficients, we could begin interpreting the p-values to determine what exactly is associated with success in the labor market.

6. Comprehensive Results

Every soft skill in the regression except self-efficacy had a significant p-value. However, since our research question was to determine which specific soft skills were *most* relevant to market success, we wanted a way of identifying which predictors were the most important to our model.

One way was to look at the F-statistic of each predictor in a model with only that predictor. This way, we could rank the soft

skills with the highest F-statistics as the most important overall. However, this approach failed to consider the relationship between soft skills; many soft skills could be statistically significant on their own, but not statistically significant after adjusting for the effect of other important predictors.

As such, we implemented a different technique to identify important skills while taking into account other predictors. We started with a model with only the predictor with the highest F-statistic, and then iteratively found the single predictor that could be added to the model to maximize the model's F-statistic (similar to a forward-stepwise approach). Using this technique, we identified the top three most important soft skills to be Grit / Work Ethic (overall F-stat of 63.22), Teamwork (43.95), and Conscientiousness (33.56). The soft skills Organization and Growth Mindset were found to be collectively the least important.

7. Segmentation Analysis

7.1 Approach

Each country has a diverse set of occupations that require different skills to be successful. Although a soft skill like Teamwork might be predictive of market success across the U.S. as a whole, it might not be predictive of market success within a specific industry like manufacturing. As such, an important next step for us was to segment our results by industry to determine which soft skills were most important in each occupation.

Our process for segmenting by industry was similar to our approach overall. The U.S. data contained information on 17 top-level industries (e.g. construction or public administration), so we split the data by each of these industries and ran a separate regression model on each (the sample size of each model ranged from <10 to >500 participants). To get the most important soft skills from each of these models, we used F-statistics to calculate the top three soft skills for each industry in the same way that we did before.

7.2 Results

Taken together, our industry specific results were consistent with our comprehensive results. Grit / Work Ethic, Teamwork, and Conscientiousness were all among the most frequent soft skills to appear in each specific industry, although each industry varied from one and other. For example, Communication was found to be important along with Grit / Work Ethic and Conscientiousness in the entertainment industry, whereas Grit / Work Ethic Self-Efficacy, and Adaptability were found to be important in the construction industry.

These results were also consistent with intuition about each industry. For instance, in the retail industry, Teamwork was found to be the most important soft skill, reflecting the fact that retail employees have to work frequently with customers and other co-workers to succeed at work.

A plethora of interesting results can be obtained by looking closely at *Appendix Table 1*. One interesting result, for instance, was that Agreeableness appeared the most out of any soft skill in the table (in 9 out of 17 industries), but was never the most significant predictor; instead, it was usually significant alongside other soft skills like Teamwork.

8. Conclusions

Ultimately, we were able to find strong evidence that certain soft skills are associated with labor market success in the U.S. By focusing on one primary country (the U.S.), we were able to map out a process to classify and extract numeric representations of soft skills from psychometric questions and create a model to predict market success both holistically across the country and segmented by industry.

We found Grit/Work Ethic, Teamwork, and Conscientiousness to be the most relevant soft skills across all industries, and other soft skills showed importance sparsely in specific industries. These findings should eventually help organizations like the World Bank develop programs to help underprivileged youth find work based on soft skills.

9. Future Work

Moving forward, we would like to have our model run on data from other countries such as South Africa, Bulgaria, UK, or Jordan. We believe that this is possible with a minimal amount of data preprocessing before the model pipeline. Furthermore, we would like to compare important soft skills among entry-level, mid-level, and senior-level industry positions.

10. References

[1] Deming, D. J. (2017). The Growing Importance of Social Skills in the Labor Market. *The Quarterly Journal of Economics*, 132(4), 1593–1640. https://doi.org/10.1093/qje/qjx022

[2] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[3] Kendra Cherry. (2021, February 20). *What Are the Big 5 Personality Traits?* Verywell Mind. https://www.verywellmind.com/the-big-five-personality-dimensions-2795422#:%7E:text=The%20five%20broad%20personality%20traits,how%20many%20personality%20traits%20exist.

[4] Edward Ma. (2019). NLP Augmentation.

[5] 20 Interview Questions to ask to Assess a Candidate's Soft Skills. (n.d.). 24Seven. Retrieved June 10, 2021, from https://www.24seventalent.com/blog/2020/09/20-interview-questions-to-ask-to-assess-a-candidates-soft-skills

[6] Buswell, G. (2021, June 3). 100 Soft Skills Assessment and Interview Questions. Toggl Blog. https://toggl.com/blog/100-soft-skills-questions-to-help-you-hire-top-talent

[7] Fabiano, J. (2019, October 11). Why you need to ask these 7 questions to test soft skills during a job interview. Ladders | Business News & Career Advice. https://www.theladders.com/career-advice/test-soft-skills-job-interview

[8] Frame, P. (2000, August). Soft Skills Interview Questions. The Regents of the University of California Agricultural Extension, Stanislaus County. https://nature.berkeley.edu/ucce50/ag-labor/7labor/b003.htm

[9] Matthew Chulaw (2021, April 14). TOP 30 Soft Skills Interview Questions + Sample Answers. InterviewPenguin.Com - Your Best Job Interview Coach. https://interviewpenguin.com/soft-skills-interview-questions/

[10] McNutt, R. (n.d.). 15 Critical Soft Skills Interview Questions for Identifying Leadership. RMI Executive Search. Retrieved June 10, 2021, from https://www.rmiexecutivesearch.com/15-critical-soft-skills-interview-questions-for-identifying-leadership

11. Appendix

Appendix Table 1: Industry Level Results

Industry	1st Soft Skill	F-stat	2nd Soft Skill	F-stat (including previous)	3rd Soft Skill	F-stat (including previous two)	Sample Size
Educational, Health, and Social Services	Grit/Work Ethic	25.900	Teamwork	18.730	Agreeableness	14.710	549
Professional and Related Services	Grit/Work Ethic	6.399	Conscientiousness	6.458	Teamwork	5.285	273
Retail Trade	Teamwork	4.213	Conscientiousness	2.538	Grit/Work Ethic	2.126	200
Manufacturing	Grit/Work Ethic	10.490	Teamwork	6.057	Agreeableness	4.502	190
Entertainment, Accommodation, and Food Services	Grit/Work Ethic	6.062	Conscientiousness	6.187	Communication	5.123	190
Finance, Insurance, and Real Estate	Agreeableness	1.985	Adaptability	1.740	Self Efficacy	1.6060	166
Construction	Grit/Work Ethic	3.725	Self Efficacy	2.233	Adaptability	2.6430	131
Public Administration	Teamwork	5.888	Self Efficacy	3.557	Communication	2.7900	117
ACS Special Codes	Teamwork	5.337	Communication	6.258	Conscientiousness	5.3090	104
Transportation and Warehousing	Growth Mindset	2.584	Communication	1.918	Agreeableness	1.4140	90
Other	Self	3.172	Agreeablenes	3.394	Adaptability	2.8310	87

Services	Efficacy		s				
Wholesale Trade	Communication	4.398	Growth Mindset	3.851	Organization	2.7500	52
Information and Communication	Conscientiousness	3.266	Agreeableness	2.517	Growth Mindset	1.8870	50
Agriculture, Forestry and Fisheries	Teamwork	6.077	Communication	6.646	Agreeableness	7.1640	18
Utilities	Self Efficacy	8.072	Adaptability	9.793	Grit/Work Ethic	13.4200	12
Mining	Organization	1.443	Agreeableness	2.177	Teamwork	5.2520	8
Active Duty Military	Adaptability	-4.000	Agreeableness	-1.500	Communication	-0.6667	6