Predicting the Success Index of Second-Generation Immigrants in the United States

December 20, 2019

Abstract

This paper aims to determine the factors that lead to the success of second-generation immigrants in the US. The data set contains the response from 5262 second-generation immigrants on 665 questions over the span of 15 years. The data was collected throughout 3 different surveys since the respondents were in their early adolescence until they were at the beginning of their career. A success index was then generated to identify the success of each respondent from the perspective of US society using the information collected in last survey. A few selected variables from the first two surveys are used as predictors for the success index, ranging from family background to educational attainment. LASSO regression was chosen as our predictive model due to its flexibility, interpretability, variable selection feature, and outstanding low RMSE compared to other models. The study came to a conclusion that family-related variables do not have much effect on the success of the children, but personal development over high school years matter most. It also pointed out the bias for women toward higher success index compared to men. The study successfully achieved its goal in identifying factors and their level of importance toward the success of second-generation immigrants in the United States.

1 Introduction

1.1 Background and Analysis Goal

According to a Pew Research Center analysis of U.S. Census Bureau data in 2012, 24.5% of the U.S. population are first and second-generation immigrants. This number is projected to increase up to 36.9% percent in 2050. Second-generation immigrants are defined as either US-born children of at least one foreign-born parent, or foreign-born children but were then naturalized. [2] These children of immigrants are reported to be doing better than the average of the general U.S. populations, with higher income, college graduation and home-ownership rate, and lower poverty [3], and given current immigration trends and birth rates, virtually all (93%) of the growth of the nation's working-age population between now and 2050 will be accounted for by immigrants and their U.S.-born children. [4] Although these factors do not entirely measure an individual's success, they are undeniably important indicators of how we are doing in life. In this project we explore how to measure success and seek to understand the factors which predict it. Specifically, we considered a variety of factors in the development of a person from a early adolescence to early adulthood, ranging from family background, education attainment to adaptation process in analyses.

1.2 Data Set

Our analysis uses data from Children of Immigrants Longitudinal Study (CILS), which contains of 665 variables with 5262 observations, representing the responses for questions surveyed through the whole longitudinal study.[2] The variables can be analyzed by splitting into 3 different groups, corresponding to 3 different surveys that were conducted in total. The timeline of how data was collected through the entire study is summarized in Appendix 5.1.

2 Methodology

2.1 Data Processing

We created 2 separate subsets of the data: one contains all variables from the first 2 surveys, namely `predictors` and one contains variables from the last survey, namely `results`. The first and second survey were conducted in the neighborhood where attendees went to school, while the last survey was done by individual contacts, leading to a noticeable decline in participation.

The first survey started in 1992, getting 5262 responses from 8^{th} and 9^{th} grades children, with a diversity of 77 (original) nationalities. This survey aimed at baseline information on families, demographic characteristics, language use, self-identities, and academic attainment of the attendees. The second survey was conducted 3 years later (1995), when the attendees finished high schools, and retrieved 4288 responses (81.5% of the first survey). The goal of this follow-up was to examine the evolution of key adaptation outcomes including language knowledge and preference, ethnic identity, self-esteem, and academic attainment over the adolescent years. 43 variables from these 2 surveys were then selected for further consideration, spanning from the reasons parents came to US, socioeconomic situations of the family, living situation, to the person's perspectives of school, grades, and future careers, etc. (see Appendix 5.2 for distributions). We then proceed by examining the pairwise correlation between these variables and running Principal Components Analysis to understand relationships between variables and explore the possibility for dimension reduction. The scree plots summarizing efficiency of the principal components can be found in the Appendix 5.4.

The final survey was conducted in 2001-2003 via mail, with 3613 respondents (68.9% of the first survey). The questionnaires were mainly focused on the outcome of adaptation process, measured by educational attainment, employment and occupational status, income, civil status and ethnicity of spouses/partners, political attitudes and participation, ethnic and racial identities, delinquency and incarceration, attitudes and levels of identification with American society, and plans for the future. 14 variables, ranging from job prestige, income, education attainment to living situation, health condition, marriage status, etc, are then chosen to build the outcome for our model, "success index" - a numeric value that we developed specifically for this project, representing holistically how the society perceives success of a person at the age of 24-26. Section 2.3 will further explain this process.

2.2 Exploratory Data Analysis (EDA)

Due to a large number of variables, it is highly likely to exists potential redundancies in predictors. Therefore, we used multiple EDA methods, such as pairwise correlation and Principal Components Analysis (PCA), to explore possible multicollinearity and other distributional challenges before embarking on any model building (see Appendix 5.3). Calculating the correlation, we noticed that most variables are independent, therefore we can be confident that the co-variance problem is not likely to occur and these variables would potentially helpful predictors. However, a few variables still strongly correlate with others, thus we performed PCA on the data set and examined the scree plot (see Appendix 5.4) to see if we can reduce the number of dimensions to consider. We also examined the correlations between variables in the `Result` data set and didn't notice any strongly-correlated relationship. This implies that using dimensions reduction method would not be the most efficient, therefore we proceed with an unique measurement of success, which is further explained in the below section.

2.3 Success Index Formula

Our success index formula is based on a study conducted by Populace, an individuality-focused think tank. The study was conducted through web survey in the July of 2019 with a random sample of 5,242 U.S. adults age 18 and older. Two formulas were then generated: how an individual define success and how they think society defines success. [1]. As we want to see how successful the second-generation immigrants are in the eyes of others, we used the perceived societal measurement with the equation below:

$$\text{swSuccess Index} = 45.9\%\ \text{Social Status} + 19.8\%\ \text{Education} + 8.8\%\ \text{Work} + 4.9\%\ \text{Character} + 4.7\%\ \text{Relationships} + 4.1\%\ \text{Life Quality} + 3.5\%\ \text{Health}$$

A detailed table of how we fit our variables in this data set into different categories above is clarified in Appendix 5.5. Since we are using numeric outcome, root mean squared error (RMSE) would be used as the metric to evaluate the accuracy of our model.

2.4 Model Building: LM, LASSO, GAM, and PCR

Based on what we learned of the data, four regression models are taken into account, including basic linear regression (LM), least absolute shrinkage and selection operation (LASSO), Generalized Additive Model (GAM) and Principle Components Regression (PCR). Considering the pros and cons of each model and analyze the results from our data exploration process, we noticed that top principal components (from PCA) can only account for 33.2% of the variance and since multicollinearity is not one of our concern, it less likely that the PCR would perform well relatively. Moreover, with a large number of variables, LM would raise a problem of over-fitting while GAM has to bear a noticeable slowdown in the run time. Therefore, we predicted that LASSO model would be the most efficient one.

3 Results and Discussion

As we expected, LASSO model stands out with lowest RMSE across all quartiles (see Appendix 5.6 for details and more comparisons between 4 models). LASSO model also aligns with our analysis goal with its flexible variable selection features and high interpretability, which would help us identify the factors that matter most to the future success of second-generation immigrants, as well as get the most accurate predictions. Thus, we chose LASSO as our primary model.

Comparing the variables before and after training the model, most of those related to family such as whether parents divorced, etc. were then ruled out (see Appendix 5.7). In addition, information about the respondent themselves from the first survey are either ignored or has much less effect than the similar variables in the second survey, for instance, `Depression.1991` are ignored while `Depression.1995` is taken into account. Similarly, GPA in both junior high school and high school are considered, however the effect of the latter is twice as much as the former (see Figure 1). Another interesting pattern is that if between a male and female respondents with the same statistics, the success index of the women would be 2.901105e-02 higher. Most important, the percentage of daily attendance and the hours spent on studying of the respondent in high school are negatively affect their success index holding other variables constantFor example, between 2 people who have the same GPA, the one with lower attendance rate and studying hours is likely to have higher success index. This might be because the respondents spent

more time on part-time jobs, volunteering, club events, attending conferences, etc., the activities which are more practically beneficial for them in the long run.

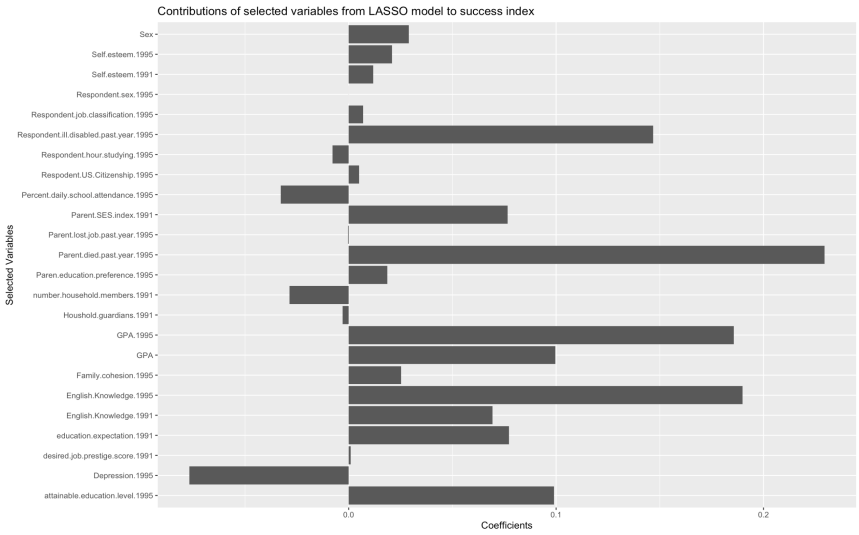

Figure 1: Model effects for 1 standard deviation increase

4 Limitations and Future Work

4.1 Limitations

Since the data was generated by a longitudinal study that lasts for 15 years, many of the observations contain missing values which must be omitted when fitting into the models. Moreover, it has been 3 decades ago since the data was first collected, so it failed to reflect the patterns in current days. Immigration has always been a matter of concern due to its constant changes in trends and targets, for example the reasons and socioeconomic situations of immigrants can differ greatly compared to now. Therefore, using data that is not up-to-date and noisy can easily create bias and flaws in our model.

In addition, although the source we based our formula on is primary, more researches are still needed in order to find the most appropriate method to particularly measure success of children of immigrants since it should take into account extra factors compared to the formula for the general. Additionally, human's perspective is very subjective and constantly changes from generations to generations, so using a formula built in 2019 for data collected in 1991 will need more validation on the results.

4.2 Future Work

Future research should consider finding better and tidier data sets to train the model, as well as taking initiative for a pilot study to build its own data. Additionally, further investigations are needed to validate the hypotheses that are drawn in previous sections of this study. Moreover, as we mentioned above, using different metrics to evaluate success is a good starting point for discussion and follow-up study.

Other recommendations to expand on this topic could be comparing the statistics second-generation immigrants with the average of all Americans and detecting different factors that drive the success of each type. It is also possible to extend this idea to different races and ethnicity within the second-generation immigrants community.

5 Appendix

5.1 Data Collection Process

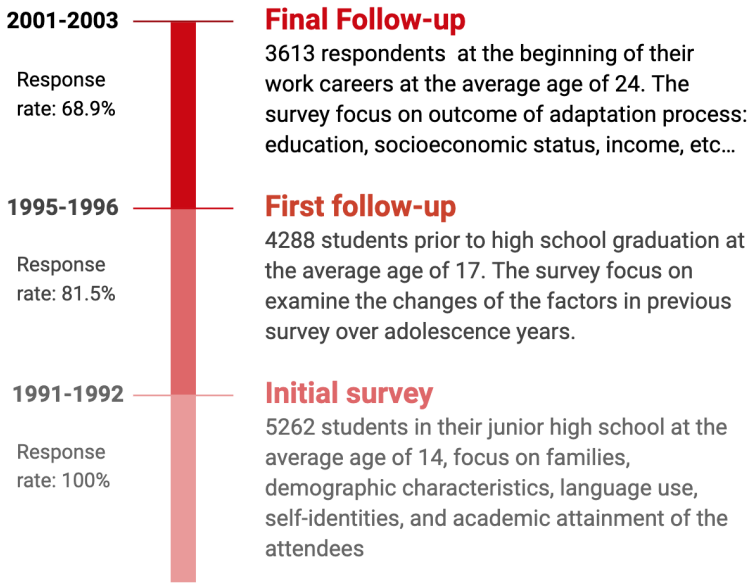

Figure 2: Timeline and statistics of three surveys in the study

5.2 Distributions of variables

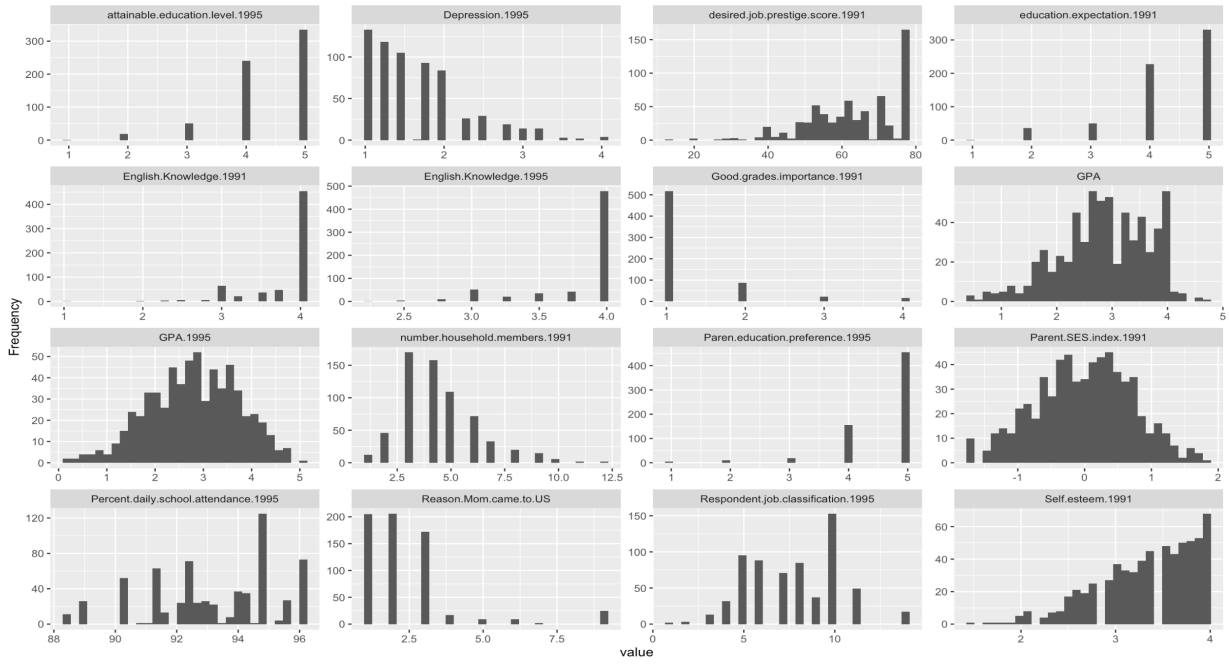

Figure 3: Distributions of predictors selected by LASSO model

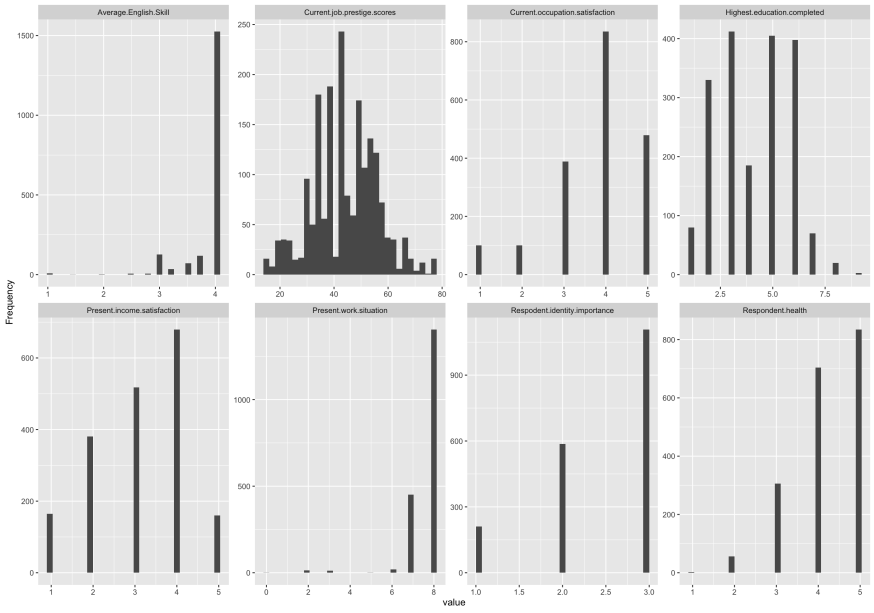

Figure 4: Distributions of components of success index formula

5.3 Pairwise Correlations

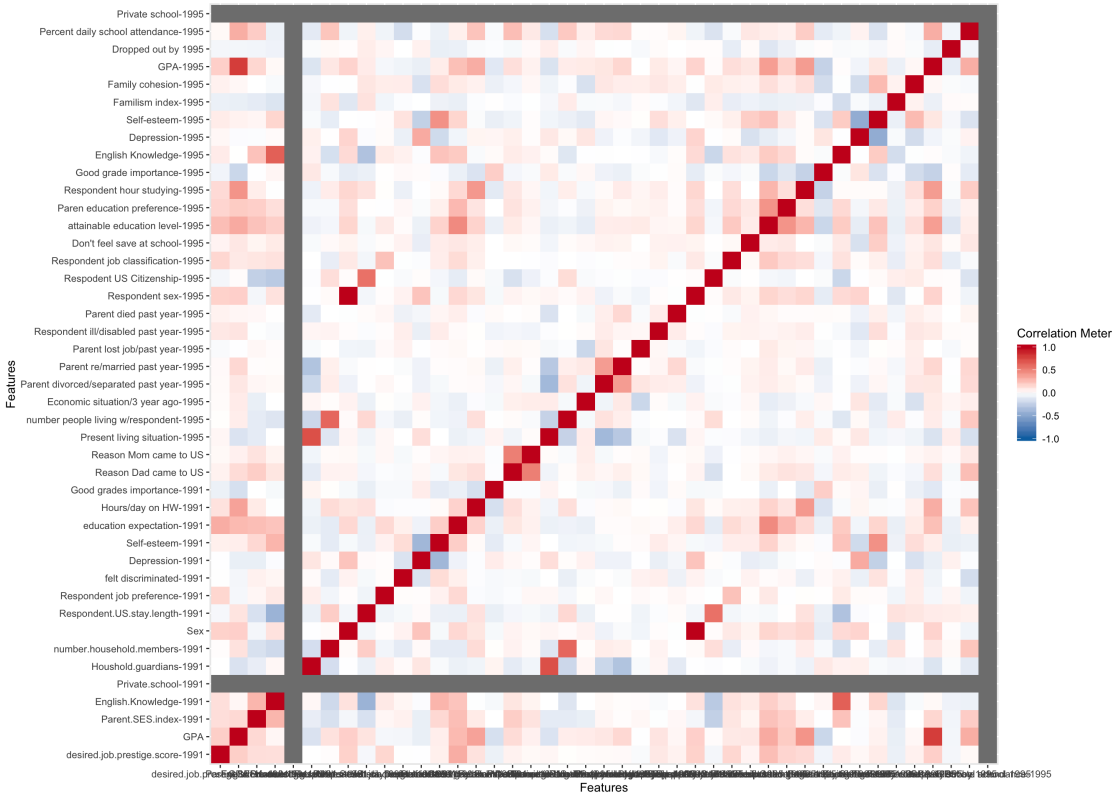

Figure 5: Pairwise Correlations between Predictors

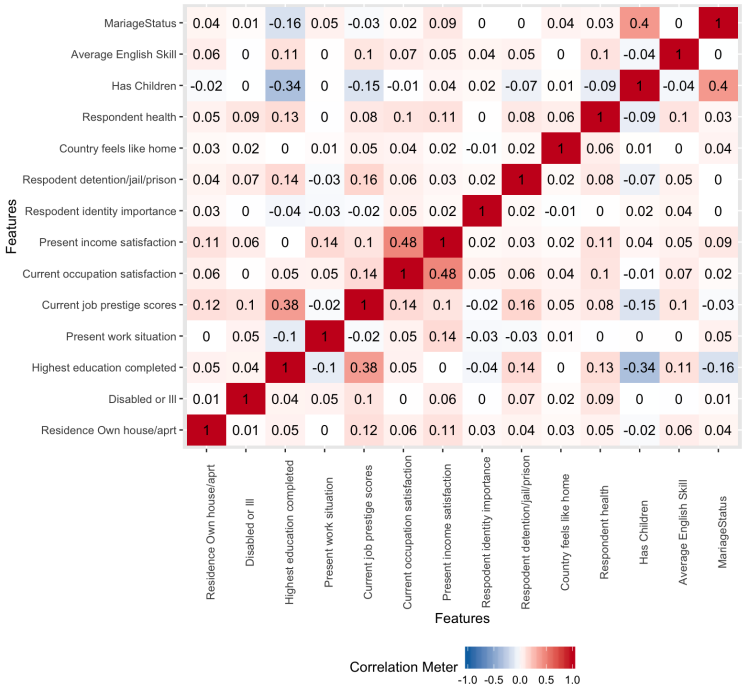

Figure 6: Pairwise Correlations between components in success index formula

5.4 Principle Components Analysis

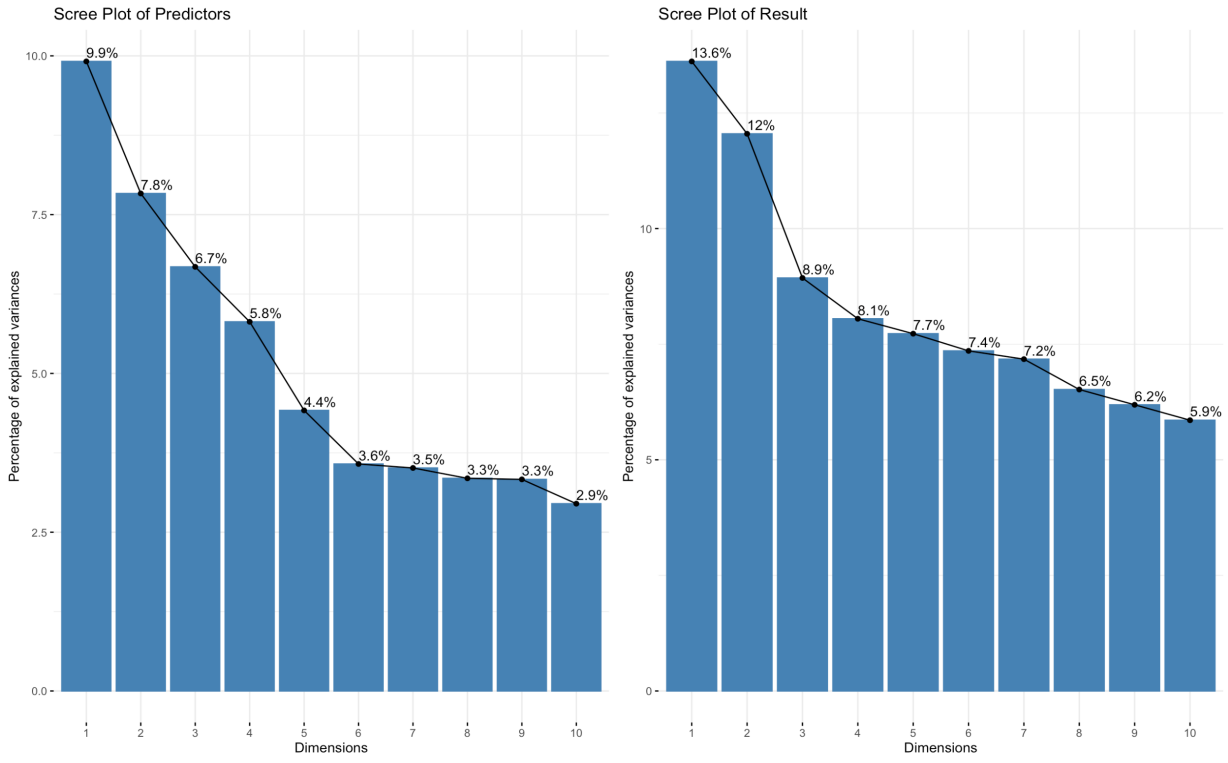

Figure 7: Scree Plots of the Principal Components in both Predictors and Result data sets

5.5 Success Index Formula

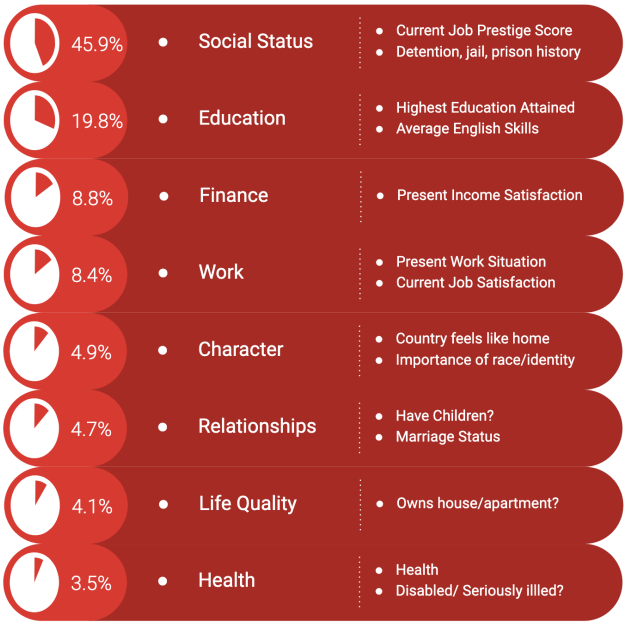

Figure 8: Categorizing variables into different factors of success index

5.6 Different Models Comparison

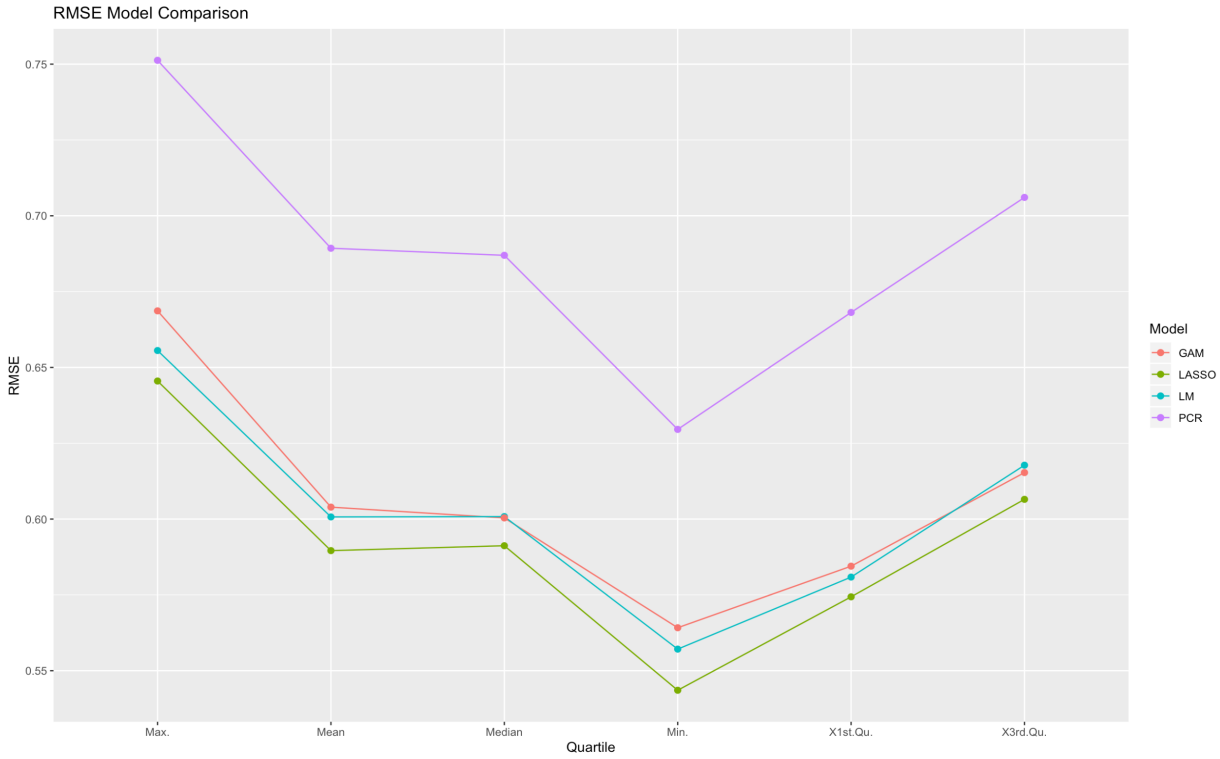

Figure 9: Comparing RMSE of 4 models: LM, LASSO, GAM, PCR across different quartiles

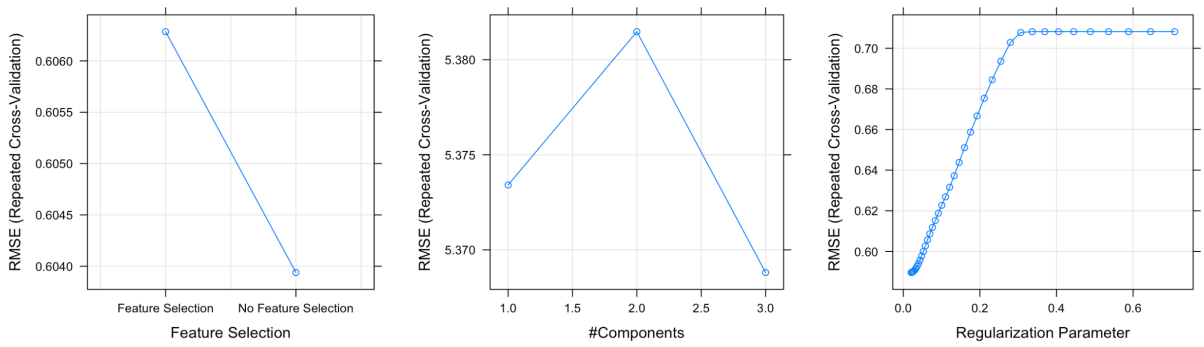

Figure 10: RMSE plots of GAM, PCR, and LASSO respectively

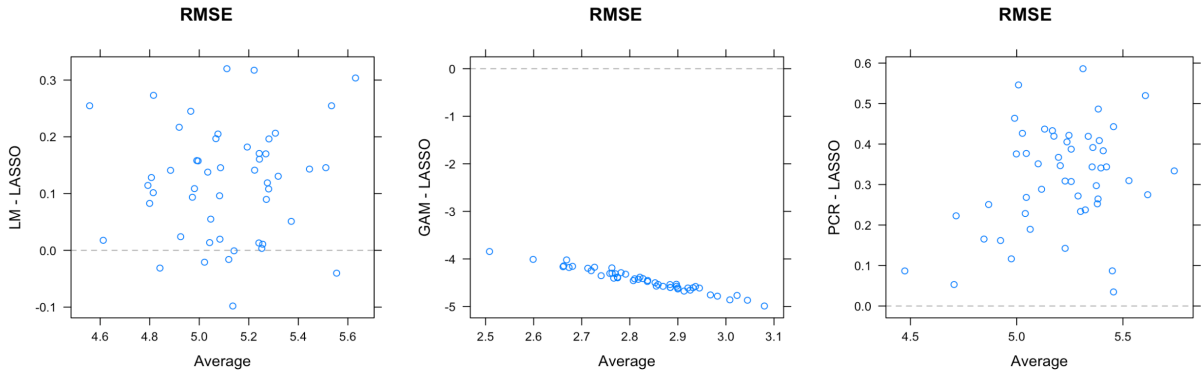

Figure 11: Comparing different models using BlandAltman plots

5.7 LASSO Not-selected Variables

'Private.school.1991'	'Private.school.1995'
'Good.grades.importance.1991'	'Good.grade.importance.1995'
'Reason.Dad.came.to.US'	'Reason.Mom.came.to.US'
'felt.discriminated.1991'	'Don.t.feel.save.at.school.1995'
'Hours.day.on.HW.1991'	'Dropped.out.by.1995'
'Respondent.US.stay.length.1991'	'Economic.situation.vs.3.year.ago.1995'
'Respondent.job.preference.1991'	'Familism.index.1995'
'Depression.1991'	'Parent.divorced.separated.past.year.1995'
'Present.living.situation.1995'	'number.people.living.w.respondent.1995'
'Parent.re.married.past.year.1995'	

Table 1: Ruled-out variables that have zero coefficient

References

[1] Success index. *Populace*, 2019. Retrieved from https://static1.squarespace.com/static/59153bc0e6f2e109b2a85cbc/t/5d939cc86670c5214abe4b50/1569955251457/Populace+Success+Index.pdf.

[2] Alejandro Portes and Rubén G. Rumbaut. Children of Immigrants Longitudinal Study (CILS), San Diego, California, Ft. Lauderdale and Miami, Florida, 1991-2006. 2007. Retrieved from http://dx.doi.org/10.3886/ICPSR20520.V3.

[3] Second-Generation Americans. A portrait of the adult children of immigrants. *Washington, DC: Pew Research Center*, 2013. Retrieved from https://www.pewsocialtrends.org/2013/02/07/second-generation-americans.

[4] Jeffrey S Passel and D D'Vera Cohn. Us population projections, 2005-2050. *Pew Research Center Washington, DC*, 2008. Retrieved from https://www.pewresearch.org/hispanic/2008/02/11/us-population-projections-2005-2050.