Determinants of Happiness in Rural Bangladesh

Ahmed Mumin, John Brown, Vincent Chen, Clay Cohen, Felipe Dantas

The University of Texas at Austin

SDS 322E: Elements of Data Science

Dr. Arya Farahi

December 6, 2022

Introduction

We set out to examine what factors influence life satisfaction and overall happiness in developing countries. Current development economics and international development circles overly concern themselves with GDP per capita and do not take into account what makes people happier. The main reason for this is that current academics have trouble pinpointing predictors of happiness and what the goals of development programs should be. This study seeks to provide direction for the international development community by incorporating results of a comprehensive survey of residents of 3 villages in rural Bangladesh. We hope to shed light on what improves people's quality of life beyond improved financial stability.

Data

The data we are using is from The Tangail Survey, a series of interviews organized by Dr. Muhammad Bhuiyan and Dr. Radek Szulga. Each interview was conducted at the household level with one person representing each household and providing information on the other members of the household. For this project we will only be considering the people who were interviewed since they were the only ones asked about their happiness and life satisfaction. They were also asked about household income, immigration status of household members, education level in schools and Madrasas, health, disability, and if they owned various things like animals, appliances, and vehicles. This original dataset has 1430 households represented.

To clean and prepare the data we had to deal with the different kinds of NAs, assign factor values, and make aggregate dependent variables and classes. The dataset contains different kinds of NAs for cases of not being applicable and cases of being missing. Oftentimes not applicable meant 0, such as the case for non-wage income. The value itself was assigned ".n NA" if it was not applicable, but assigned ".m NA" if it was missing. These different kinds of NA entries come from Stata, a statistical programming language widely used by economists. After assigning proper values to specific NAs, we filtered out all observations with remaining missing data NAs for variables that were included in the final dataset. We added the monthly income variables in order to get total monthly income so that we could manipulate a single variable in order to observe the effects of income.

The life satisfaction variables were in categories solely as characters, so we transformed them into proper categories with assigned numeric values using the "forcats" package. Lastly we used 11 measures of satisfaction in various areas of life - from people's satisfaction with their jobs, living standards, to their leisure time - to create an aggregate measure of happiness on a 40 point scale. The other seven satisfaction variables were about general happiness, achievements, health, safety, community, financial security, social environment, and family. When creating this aggregate measure we functionally shifted the scales from 1-5 to 0-4 because the response on the survey corresponding to the lowest level of satisfaction is "Not at all Satisfied". We chose to interpret this response as a 0. It would also make less sense to assign someone who gave this response to every question an aggregate happiness score of 10 instead of 0. We then used this aggregated measure to

create factored class variables that split the observations into 2 classes. This allowed us to perform classifications to determine the best predictor variables during our exploratory analysis.

Exploratory Analysis

For our exploratory analysis we created four different visualizations. Our first two visualizations focused on the Income and Age variables to gather preliminary hypotheses about the variables' relationship with the aggregated happiness variable. We then constructed a random forest model exploring happiness as a factor variable to look at different variables and their importance to the model. Lastly, we constructed a boxplot examining employment status and its effect on happiness.

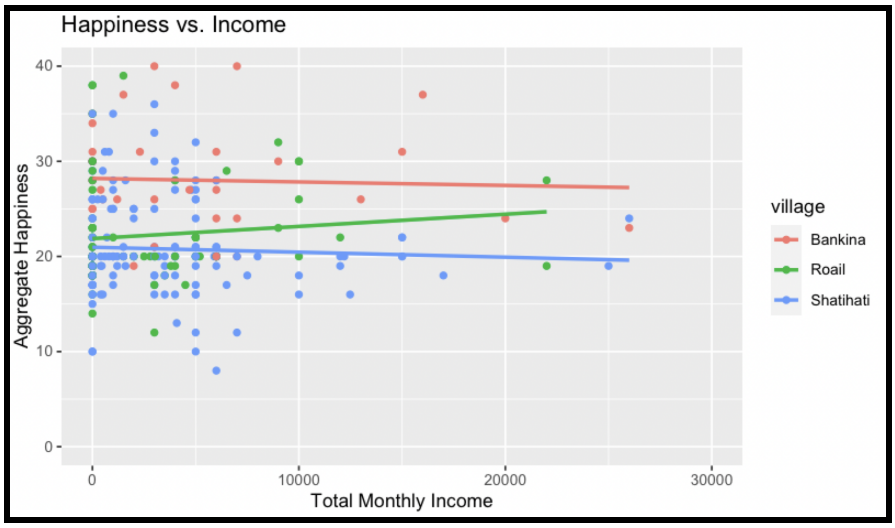

Figure 1. Happiness vs. income

Figure 1 is a scatter plot that graphs the aggregated happiness measure on the y axis and total monthly income on the x axis. The colors represent the different villages and the lines were added using geom_smooth with the method set to "lm" so we would get an idea of what a regression might look like. From this visualization we generated our first hypothesis: Total income has little effect on the happiness of residents in rural Bangladesh. We confirmed this hypothesis by running a t-test and discovering a p-value of .953, showing that income does not have a statistically significant effect on aggregate happiness score.

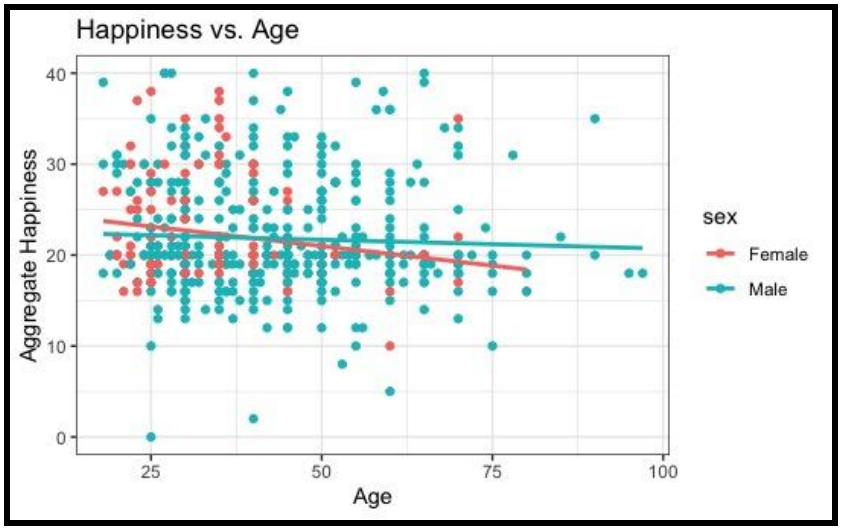

Figure 2. Happiness vs. age

Figure 2 is a scatter plot that again has the aggregated happiness measure on the y axis but has age on the x axis, with color representing sex. We also added regression lines to get a idea about the relationship between the variables stratified by sex. From this visualization we generated our second hypothesis: Age has little effect on happiness in males, but has an effect on female happiness. We confirmed this hypothesis by running a t-test for each sex. The p-value for males was

.2997, showing that age does not have a statistically significant effect on happiness for males. The p-value for females was .0240, which, based on an alpha level of .05, shows that age does have a statistically significant effect on females' happiness.

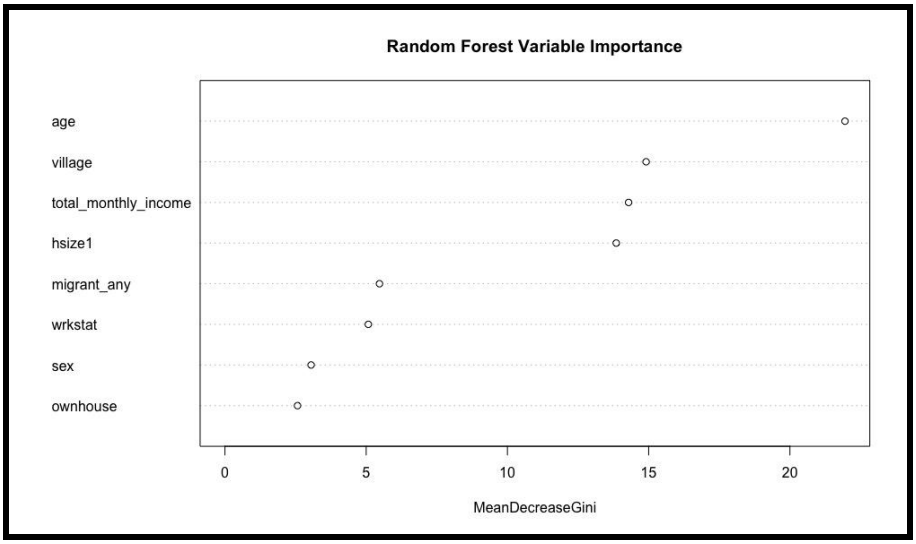

Figure 3. Random forest variable importance

Figure 3 displays the relative importances of our explanatory variables based on a random forest classification using the aggregate happiness score as a binary response variable. From this plot, we generated the hypothesis: Age is the most important variable in determining happiness.

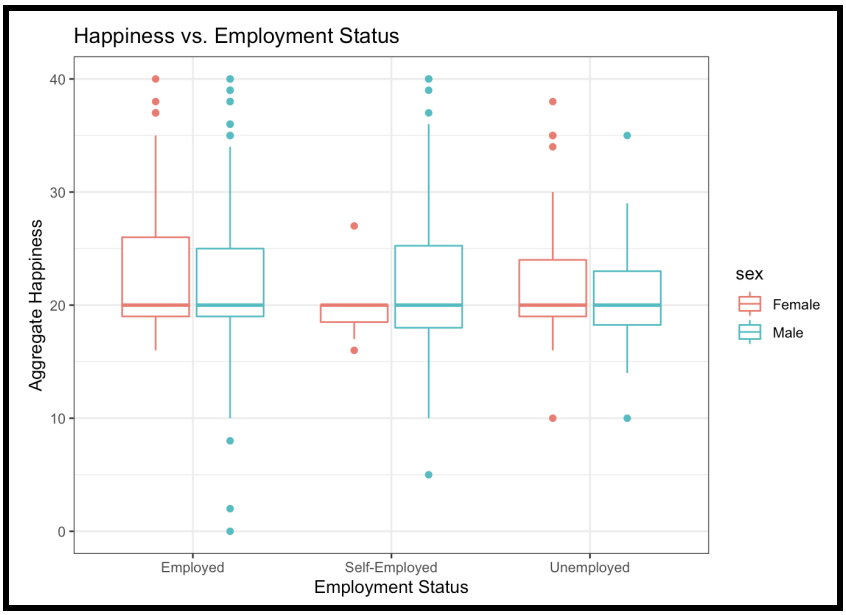

Figure 4. Employment status vs. happiness

Lastly, we examined a boxplot looking at employment status and its relationship with aggregate happiness. From this plot we generated the hypothesis: Work status has little effect on happiness regardless of sex. We confirmed this by running a t-test and discovering a p-value of .7237, showing that work status is not a statistically significant predictor of happiness.

Modeling

For our dataset, we found it appropriate to compare how three different regression models predicted happiness, seeing as our measure of aggregate happiness ranged from zero to forty. The three models we chose were linear regression, multiple regression, and random forest regression. For the linear regression, we trained off of age as a predictor variable, since our exploratory random forest model (Figure 3) showed that it was the single best predictor. For the other two models, we

used the eight best random forest predictors - age, village, total monthly income, house size, migrant status, employment status, sex, and house ownership. 80% of the dataset was randomly sampled to be the training set, and the other 20% was used for testing model accuracy.

Discussion

None of the three models proved very reliable in predicting aggregate happiness - the linear regression's mean square predicted error (MSPE) was found to be 25.25 units, the multiple regression's was found to be 24.88, and the random forest regression's was found to be 23.57. Seeing as all of the MSPEs are more than half of the range of the response variable, none of the models would be very appropriate for reliably predicting aggregate happiness. Random forest regression did prove to have the least error however, but relative to the other regression models, it was not a significant improvement. In fact, the correlation coefficient for the random forest regression was computed to be 0.1025, meaning that a mere 10.25% of the variability in aggregate happiness could be predicted by the eight predictor variables. This points to the predictor variables being the issue, and not the models - it is likely that the data collected is insufficient to predict this population's aggregate happiness as measured by self-reported satisfaction levels.

Our initial hypotheses that income and employment status have little effect on happiness seem to be confirmed. Our hypothesis that age is the most important variable in determining happiness stands true, however that's only because it's being compared to other weak predictors.

Some possible limitations are that 57.6% of the data was omitted due to null values. If those null values were MNAR, or missing not at random, then better data collection is advised, as an unobserved variable may have caused the missing values. However, most of the omitted rows were omitted due to missing answers to some of the life satisfaction questions, therefore it is likely that the null values were MCAR, or missing completely at random. Another possible limitation is voluntary response bias. The only surveyed people were those who allowed the interviewers to enter their homes, perhaps inducing the measured happiness to be higher than reality. There could also have been measurement bias introduced by the length of the survey - the survey was quite long and respondents may have been fatigued by the time they got to the satisfaction questions, thereby increasing the uniformity of their responses. Social desirability bias may have also caused respondents to avoid reporting low happiness even if it was true.

Conclusion

In summary, we found that using socioeconomic factors to predict determinants of happiness was much easier said than done, as there are a number of intangible and immeasurable factors that affect subjective measures of happiness. With our data from the Tangail Survey Data, we were limited in what we could explore and the quality of the responses may have been diminished due to immeasurable bias. Overall, the indicators we chose to explore proved to not be good predictors of happiness.

We do believe there are many positive benefits to further exploring determinants of happiness and applying the findings to development programs across the world. Future studies could consider employing longitudinal measures of happiness, allowing normalized comparisons between individuals, especially considering the abstract nature of happiness and the highly variable definition of it by person. Future studies could also consider using respondents' friends' and families' judgements of the respondents' happiness rather than respondents' own self-reported happiness. This could reduce social desirability bias. Most of all though, future studies should make sure to survey a very large breadth of predictor variables, as what was most clear from this study is that the best measures of happiness are likely not what we expect.

Acknowledgment

Mumin Ahmed contributed to the introduction and background sections of the report and powerpoint and conclusion of the report. John Brown contributed code to the data wrangling section as well as prepared that section in the report and powerpoint. Vincent Chen contributed to the discussion and conclusion sections of the report and powerpoint. Clay Cohen contributed code to the exploratory data analysis section as well as prepared that section in the report and powerpoint. Felipe Dantas contributed code to the modeling section as well as prepared that section in the report and powerpoint.

Bibliography

Bhuiyan, M., & Szulga, R. (2021). The Tangail Survey: Household Level Census of Subjective Well-Being, Perceptions of Relative Economic Position, and International Migration: 2013 [Tangail, Bangladesh]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research.. https://doi.org/10.3886/E100177V5

Kushlev, K., Radosic, N., & Diener, E. (2022). Subjective Well-Being and Prosociality Around the Globe: Happy People Give More of Their Time and Money to Others. *Social Psychological and Personality Science*, 13(4), 849–861. https://doi.org/10.1177/19485506211043379

Group Member Name	Percent Contribution
Mumin Ahmed	100%
John Brown	100%
Vincent Chen	100%
Clay Cohen	100%
Felipe Dantas	100%