1. Intro
   1. Project Purpose: What makes this an interesting setting that would benefit from analysis? What is the anticipated benefit of this modeling exercise?
      1. Everyone might have different answers regarding what is happiness in their personal life. From a national standpoint, some people think we can can find a way to quantify happiness now. This project seeks to answer what can be used to predict happiness and what would be the best model to predict happiness.
      2. anticipated benefit: people will have deeper understanding of happiness.
   2. The importance of Happiness
2. Hypotheses: What is the implicit status quo or a priori claim before the data analysis? What is the research (alternative) hypothesis?
   1. If the GDP of a country is higher, the happiness score will be higher
   2. If the social support of a country is stronger, the happiness score will be higher
   3. If the life expectation of a country is higher, the happiness score will be higher
   4. If the people in a country have a higher degree of freedom to make choice, the happiness score will be higher
   5. If the people in a country are more generous, the happiness score will be higher
   6. If the country is more corrupt, the happiness score will be lower
   7. If the people have more confidence in government, the happiness score will be higher
   8. If the democratic quality of a country is higher, the happiness score will be higher
   9. If the FDI of a country is higher, the happiness score will be higher
3. Analysis Process:
   1. Where did the data set come from?
      1. Survey from Gallup
      2. World bank used the survey to create the Happiness report.
   2. What was the sample size?
      1. There are 157 countries’ happiness score
      2. After removing the missing data, there are 122.
   3. What does the data look like? What are the variables?
      1. Dependent variable is continuous: Happiness.
      2. Independent variable: GDP, social support, life expectation,

Freedom to make choice, generosity, corruption, confidence in government, democratic quality and foreign direct investment.

* 1. What kind of cleaning/pre-processing did the team engage in?
     1. Removing missing data
     2. Separate data to training data & test data by 50/50 to train our model
  2. What techniques were used in the analysis?
     1. Regression and Tree analysis
  3. Why were these modeling techniques employed?
     1. Linear Regression and Tree—because our dependent variable is continuous
     2. Logistic Regression and Classification Tree-We created an indicator variable to represent “high” happiness score

1. Results:
   1. What discoveries or insights resulted from the statistical analysis?
      1. Continuous dependent variable: We choose linear even though tree and prune tree have a lower MSE. (We perform Linear Regression and Tree on the whole data set in the beginning. It has a very nice R square and MSE.
         * 1. Nice R-square from linear regression: 82%
           2. Easy to interpret
           3. Social support, Freedom to make choice and population density is easy to understand for high happiness score.

However, We would like to know how the regression model perform on unseen data. Therefore, we perform the same regression and tree model on the training dataset and test dataset.)

The result shows that Linear regression has the second lowest test MSE.

Why not choose Bagging (the lowest Test MSE)?

Loss of interpretability: the final bagged classifier is not a tree, and so we forfeit the clear interpretative ability of a classification tree

* + 1. Go in deeper to see what happen if we create an indicator variable for high happiness(Categorical Dependent Variable):
       - 1. Why create an indicator variable “high”?

We want to have a closer look at what make happiness score high

* + - * 1. Logistic Regression has the lowest classification error rate

Whole dataset: 7.38%

Test dataset: 6.57%

* + - * 1. Social support and freedom to make choice is the most significant variable in logistic regression
  1. What do the results mean (in human terms)?
     1. Social support and freedom to make choice are the primary factor that have impact on people’s happiness
  2. Are the results intuitive or surprising?
     1. We are surprise that variables like perception of corruption, democratic quality and confidence in government are not significant in terms of predicting happiness score.
        + 1. Limitation on our data source
          2. All data are self-reported. People have different perception of what is corruption. (culture difference between China and US)

1. In order to draw more general or larger conclusions from the research, We make comparison between China and the U.S
   1. Why China and US?
      1. We consider both countries are the world’s leading economic power
      2. China is developing country and US is developed country
         * 1. Developing v.s Developed
           2. Culture different?
2. Conclusion:
   1. What were the key findings?
      1. Which variables are significantly contribute to the change of happiness?
   2. Would additional data/variables have aided the research?
      1. Yes. We can add number of immigrants.
      2. Price of purchasing a house?
   3. What are the practical implications of the conclusions?
      1. People are happier if they have more freedom to choose what to do with their life.
      2. People are happier if they know they can get help when they are in trouble.
   4. What are the lessons learned in this project?
      1. The variables, which we thought are important, do not yield the best prediction at the end.