Exploring global happiness scores and its drivers

Image source: https://globalwellnessinstitute.org/

For Data Science Course: By Irina Nizhnik

What drives a nation's level of happiness?

About the dataset:

Target variable (Y)

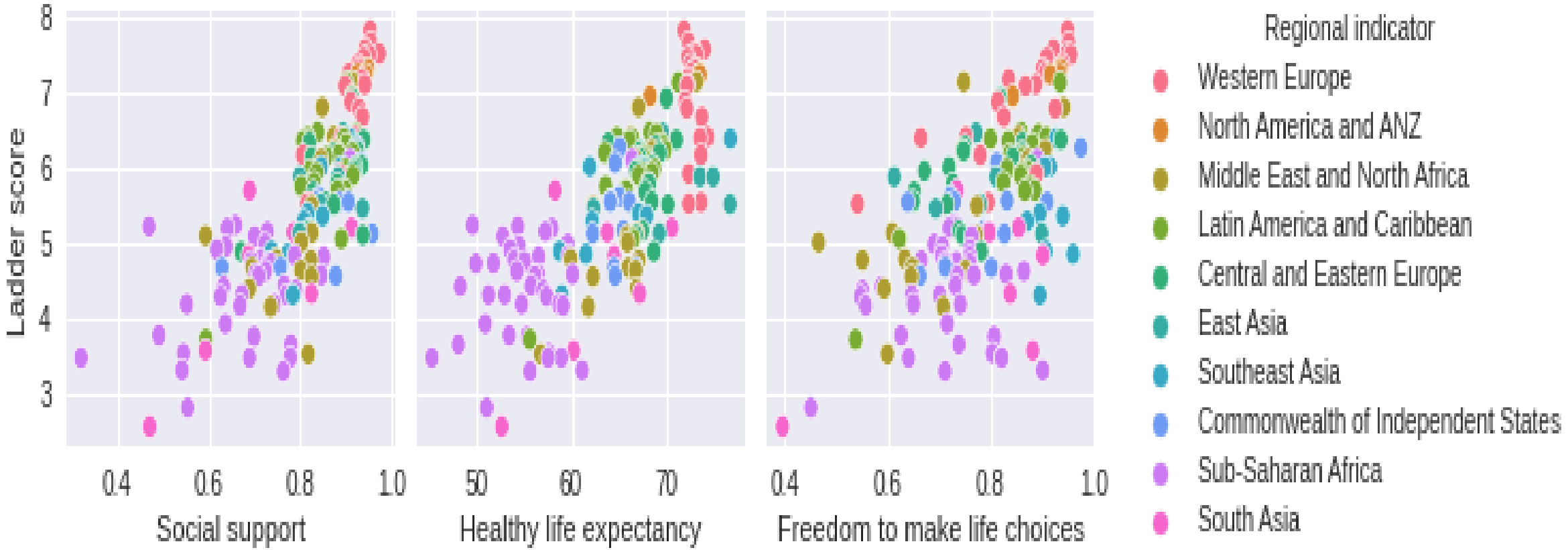
- Ladder score:
  - Stands for overall level of happiness in each country
  - Continuous variable from 1-10
  - Mean: 5.47
  - Median: 5.15
  - Standard Dev: 1.11
  - Max: 7:80
  - Min: 2.56

Independent variables (Xs)

- Includes features that could impact happiness levels, for instance
  - GDP/per capita
  - Social support
  - Healthy life expectancy
  - Freedom score
  - Corruption score
  - Generosity score

The dataset contains 153 rows – one for each country in the dataset

Initial exploration demonstrated high correlation between social support, healthy life expectancy, freedom levels with the level of happiness



Part I: Four models were used to examine the drivers of happiness in more detail

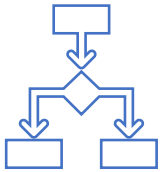
Linear Regression

- Training: 49%
- Testing: 56%

- **Key takeaway:**
  - **Linear Regression** provided a modest of level of predictability
  - **Target feature:** Ladder Score (Happiness)
  - **Features:** Social support, Healthy life expectancy, Freedom to make choices

Decision Trees

- Training: 82%
- Testing: 74%



- **Key takeaway:**
  - After testing **different max depths**, the model provided the best results with **max depth of 3**
  - **Max depth in default:** 100% & 50%,
  - **Max depth 2:** 72% & 62%

Bagged Trees

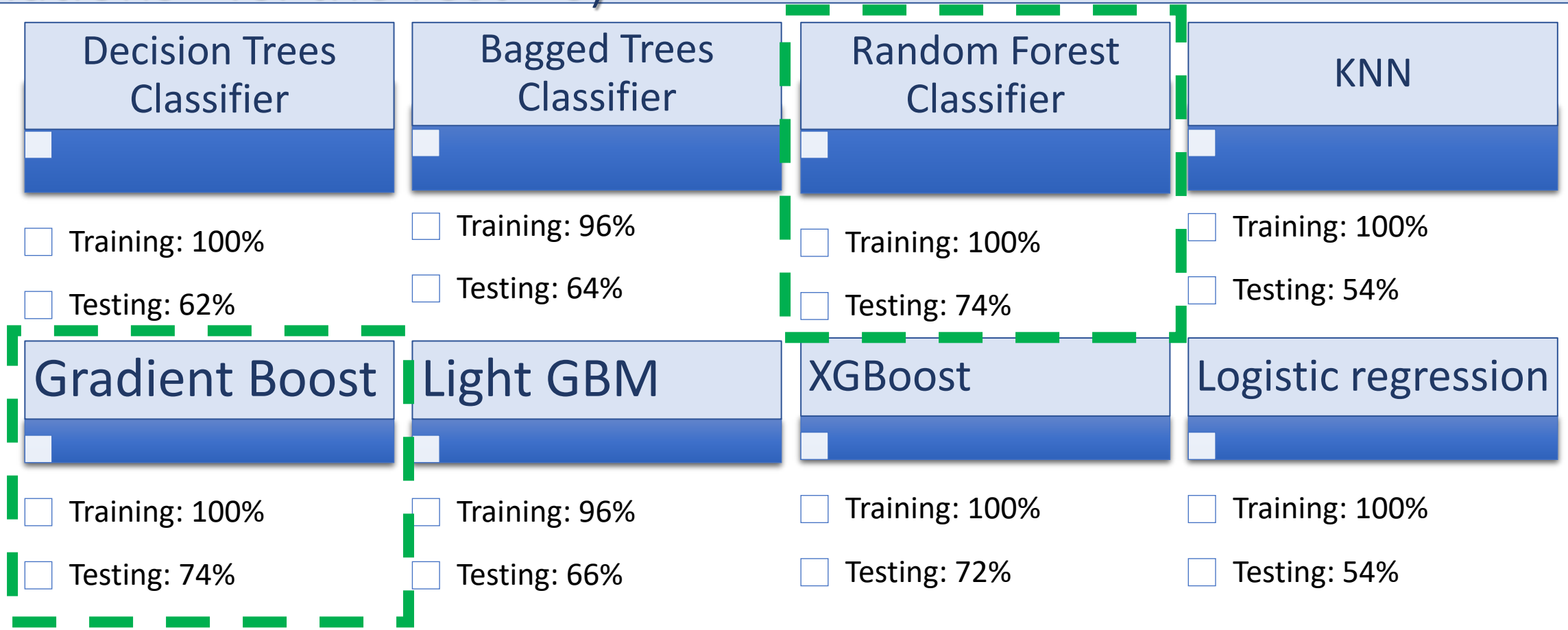
- Training: 95%
- Testing: 73%
- **Key takeaway:**
  - Bagged trees model performs very close to the Decision Tree **in testing**, but not in training
  - The model improved a little with changing n_estimators from 73% on test to 76%

Random Forests

- Training: 97%
- Testing: 77%
- **Key takeaway:**
  - The model improved significantly in running Random Forests: the accuracy is 97% in training & 76% in testing
  - Next, let's look at the most important features of the model

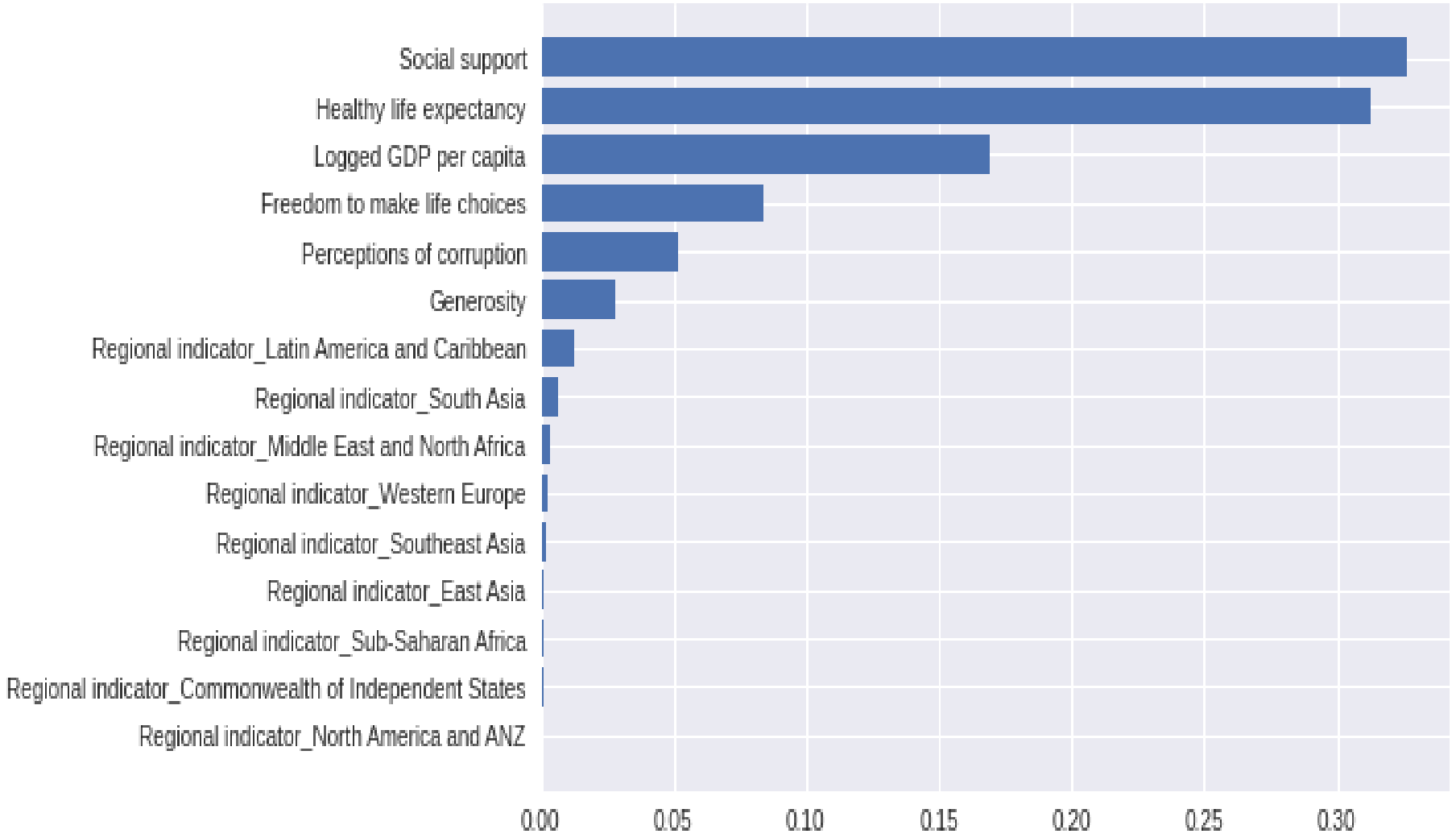
Four models were run for numerical continues variable, Ladder Score (Happiness)

Part II: Eight more models were tested by turning Happiness into a categorical variable (>6 ladder score as “Happy nations” vs. the rest <6)



- Of all the models that were tested (Decision trees classifier, bagged trees classifier, random forests classifier, KNN, Logistic regression, and variations of Gradient Boost), Radom Forest and Gradient appear to provide the best results.

What drives a nation's level of happiness?



Random forests provided the highest accuracy on the test data. As such, it is picked as the MAIN MODEL. Looking at which features have the highest impact on happiness score, social support and healthy life expectancy followed by GDP/Capital