# A Guide to Elevated Global Happiness

Nicole Huang and Aiden Kim

# True or False

"Happiness is directly related to economic growth/wealth"

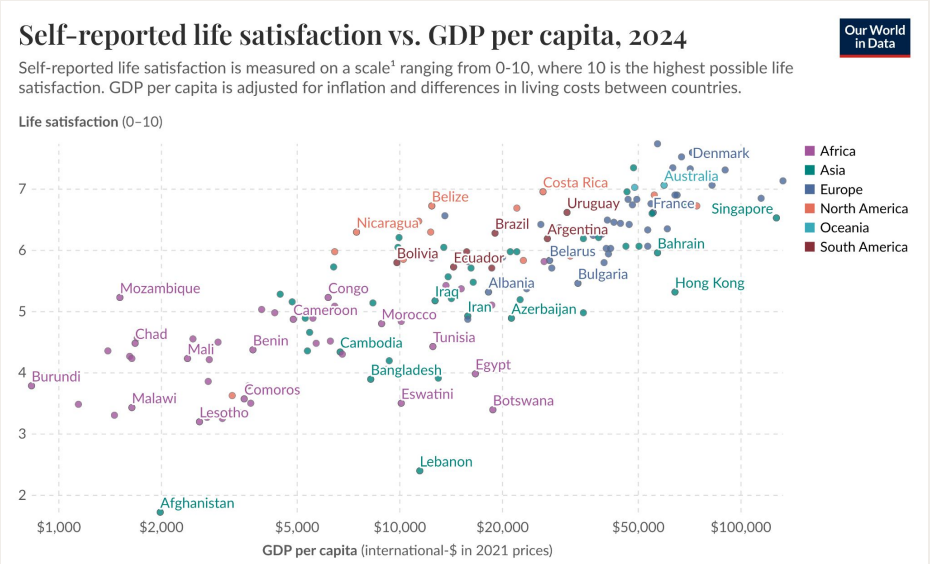



Figure created by OurWorldinData.org based on data collected by Wellbeing Research Center (2025), compiled by the World Bank (2025)

# Our Motivation

- Is it really the most definitive feature for happiness levels?
- Happiness is subjective (hard to measure and define)
- Global happiness provides insight to create informed policy decisions that cater to enhance people's lives

# Scientific Question:

What feature is most directly predictive of high happiness levels?

---

**We Predicted** perceived freedom closely associated with higher global happiness levels – live their lives however they want.

# Table of Contents



# 01

# Data

# Dataset: “World Happiness Report”

Dataset of reported global happiness score with 21 associated features by various countries between 2005-2024.

- Source: kaggle.com
- Size: 4000x24 (n=4000 p=24)
- **Training data**: selected dataset from years 2005 (earliest)

# Dataset included the following features:

y-label

```
Country,Year,Happiness_Score,GDP_per_Capita,Social_Support,Healthy_Life_Expectancy,
Freedom,Generosity,Corruption_Perception,Unemployment_Rate,Education_Index,
Population,Urbanization_Rate,Life_Satisfaction,Public_Trust,Mental_Health_Index,
Income_Inequality,Public_Health_Expenditure,Climate_Index,Work_Life_Balance,
Internet_Access,Crime_Rate,Political_Stability,Employment_Rate
```

Features

**Important Feature Data Types**

- **Happiness_Score**: float between the scale of 1-10 **(label)**
- **Statistical Data** ie GDP_per_Capita, Health_Life_Expectancy, Population
- **Scaled ratings** (between 0-1) ie Social Support, Freedom, Public Trust

# Data Cleaning

## 01
## Removing Data

Removed all datasets that had null values. (not much change)

## 02
## Binary conversion

- Wanted binary labels (low/high happiness levels)
- Threshold: 0.5

## 03
## Features excluded

- Country
- Year
- Happiness_Score (label)

Finalized dataset sorted into dictionary organized by year.

2005 dataset used as training data.

# 02

# Methodology

## + Initial Results

**Our goal** Identify the **best feature** that can closely **predict** high happiness levels.

**Method:** We want to find

**lowest entropy -> highest gain = best feature**

**Then what?** Compare predicted and actual gain values to determine predicted accuracy.

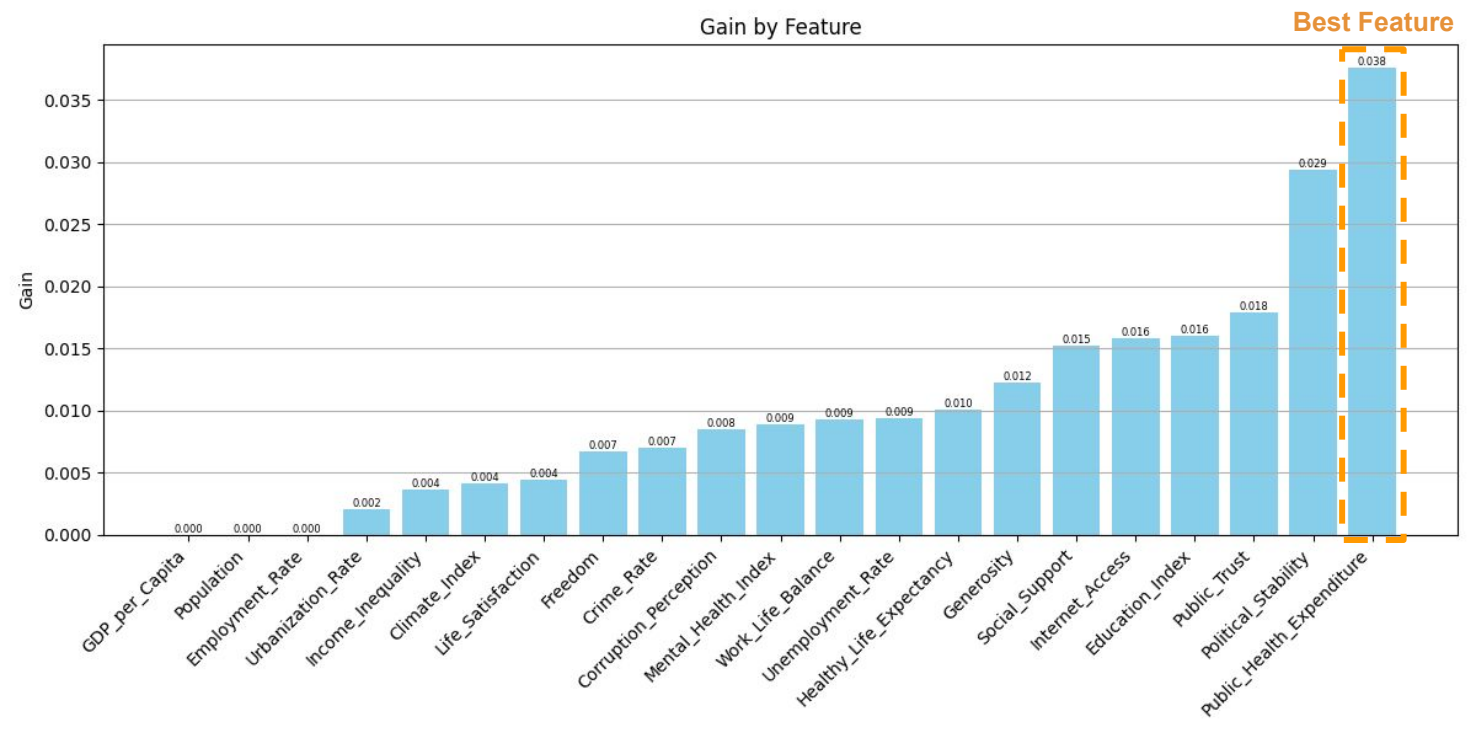
# "Training Process"

**Gain = total entropy - feature entropy**

- For each feature, we calculated the gain value.

  *Highest gain value = best feature*

- Store the **best feature** → **pred_feat_entropy**


Gain by Feature
Best Feature
Gain
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0.000
0.000
0.000
0.002
0.004
0.004
0.004
0.007
0.007
0.008
0.009
0.009
0.009
0.010
0.012
0.015
0.016
0.016
0.018
0.029
0.038
GDP_per_Capita
Population
Employment_Rate
Urbanization_Rate
Income_Inequality
Climate_Index
Life_Satisfaction
Freedom
Crime_Rate
Corruption_Perception
Mental_Health_Index
Work_Life_Balance
Unemployment_Rate
Healthy_Life_Expectancy
Generosity
Social_Support
Internet_Access
Education_Index
Public_Trust
Political_Stability
Public_Health_Expenditure

# Data Collection Process

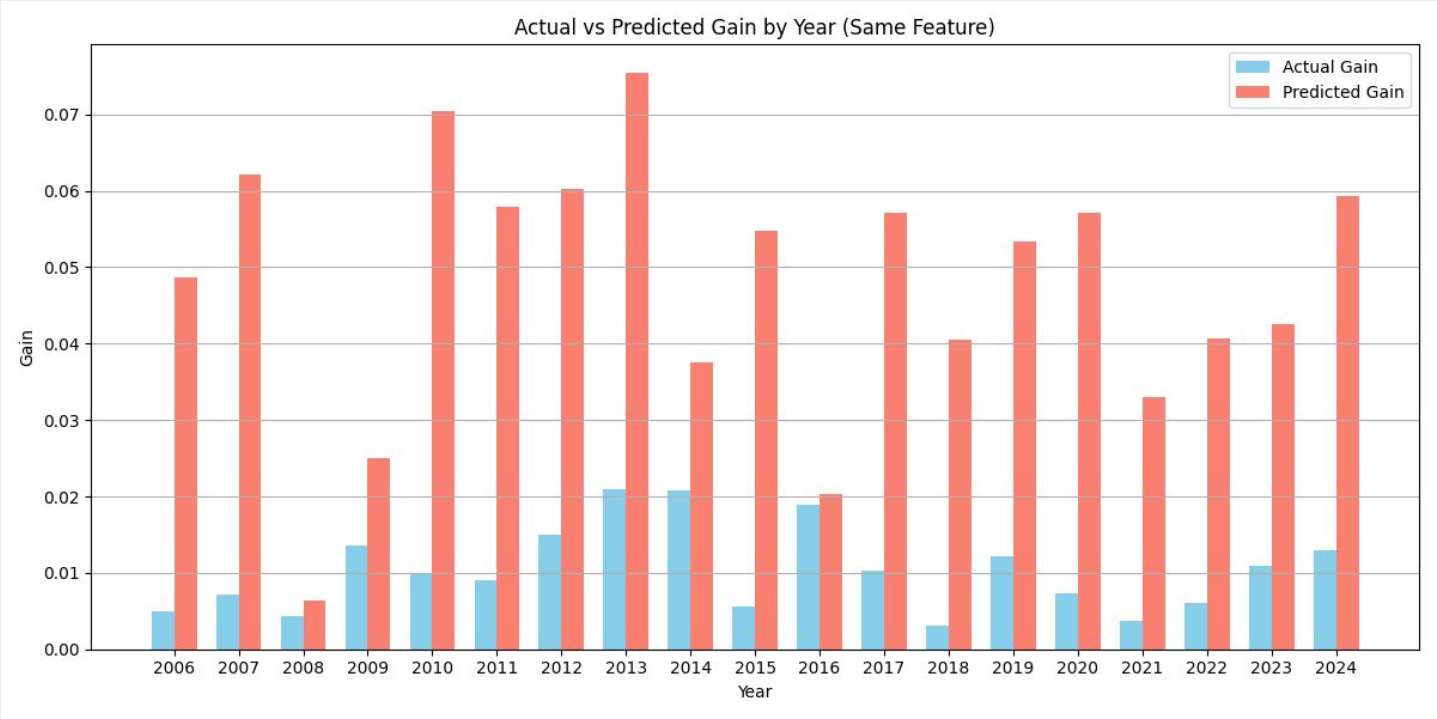
- Assume best_feature consistent throughout all datasets
- Calculate **gain of best_feature for each year**

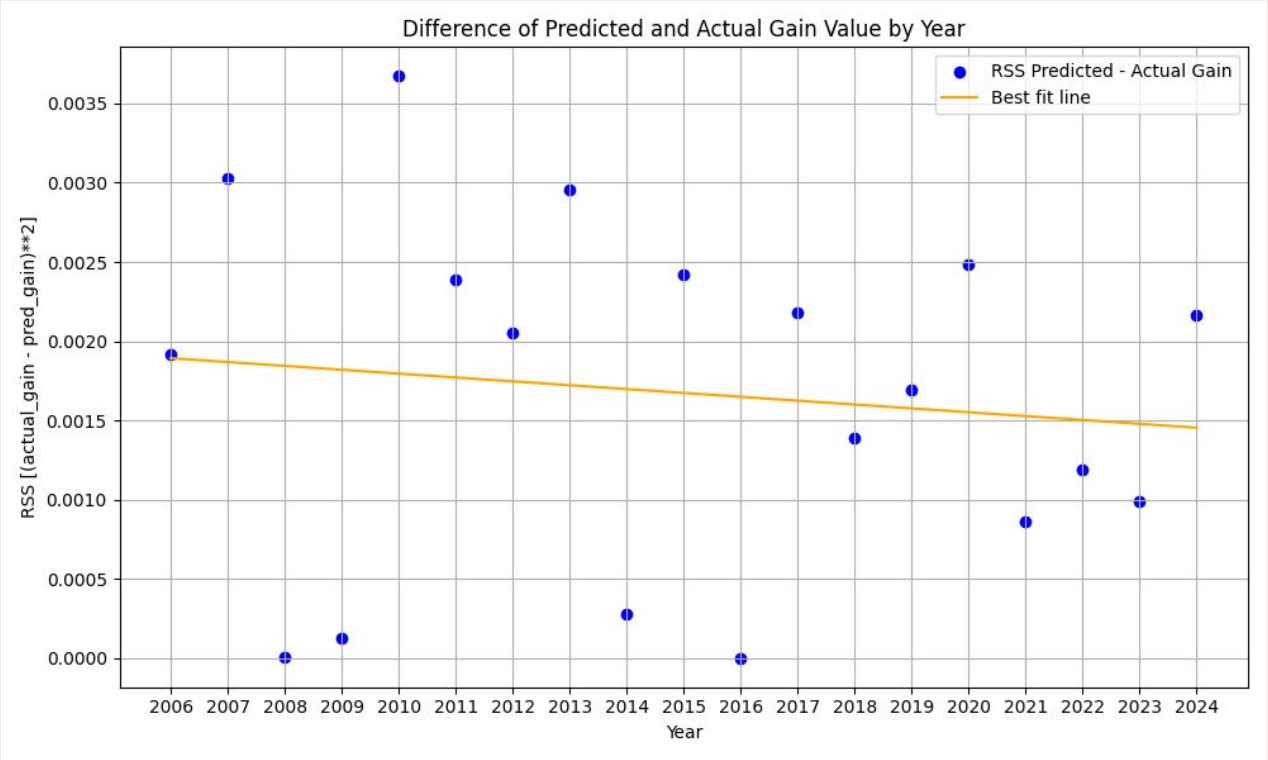
Predicted value

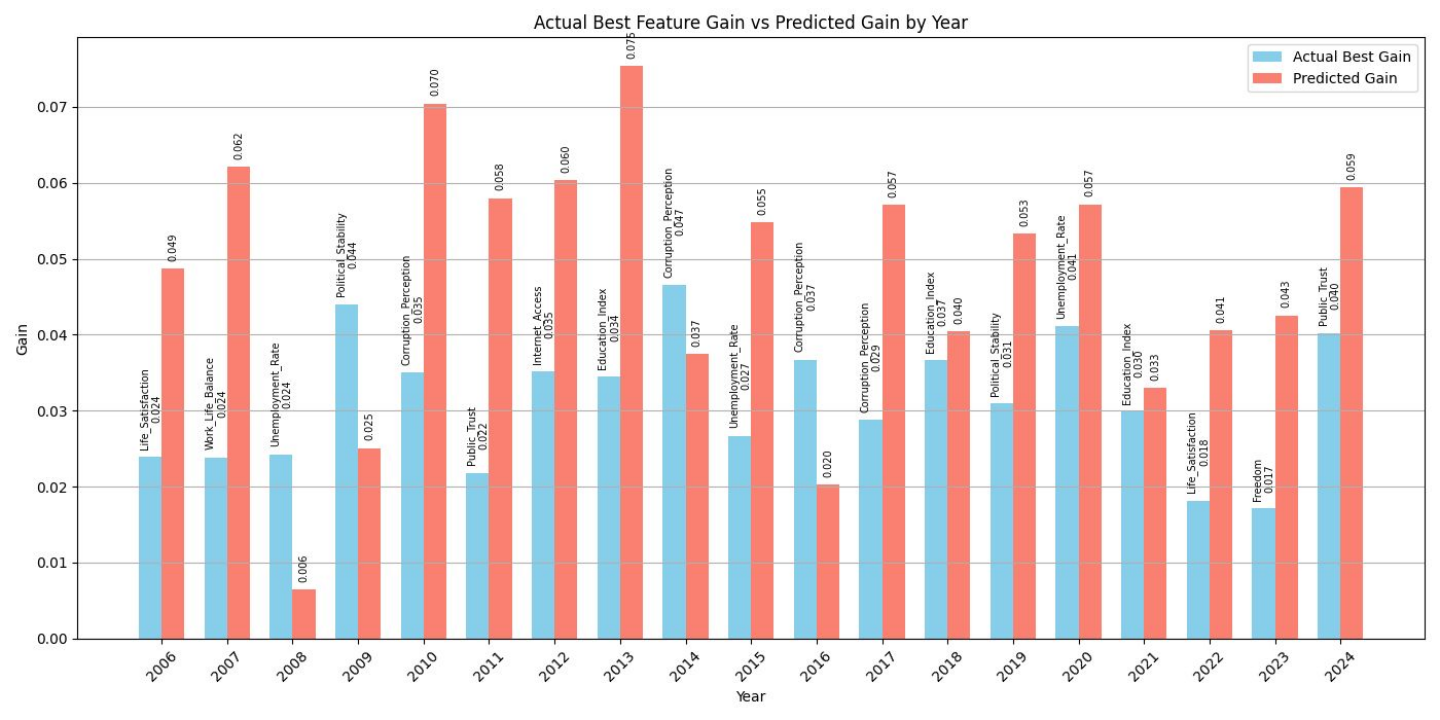
**pred_gain = total_entropy - feature_entropy**

- Calculate for all years

**difference = (actual_gain - pred_gain)**2**


Actual vs Predicted Gain by Year (Same Feature)
Actual Gain
Predicted Gain
Gain
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
Year


Difference of Predicted and Actual Gain Value by Year
RSS Predicted - Actual Gain
Best fit line
RSS [(actual_gain - pred_gain)**2]
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024
Year


Actual Best Feature Gain vs Predicted Gain by Year
Actual Best Gain
Predicted Gain
Gain
Year
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
Life_Satisfaction 0.024
0.049
Work_Life_Balance 0.024
0.062
Unemployment_Rate 0.024
0.006
Political_Stability 0.044
0.025
Corruption_Perception 0.035
0.070
Public_Trust 0.022
0.058
Internet_Access 0.035
0.060
Education_Index 0.034
0.075
Corruption_Perception 0.047
0.037
Unemployment_Rate 0.027
0.055
Corruption_Perception 0.037
0.020
Corruption_Perception 0.029
0.057
Education_Index 0.037
0.040
Political_Stability 0.031
0.053
Unemployment_Rate 0.041
0.057
Education_Index 0.030
0.033
Life_Satisfaction 0.018
0.041
Freedom 0.017
0.043
Public_Trust 0.040
0.059

# 04

# Conclusions

# To Summarize...

## Main Takeaways

- Public Health Expenditure most directly impact happiness level for **2005 ONLY**
  - Gain values not really high
- Predictions overemphasize the gain value of the feature.
- We **should not** use gain and entropy to **predict** best features.

## Future Works

- Implement **bootstrapping or random sampling**
  (randomly select predictive data, select smallest RSS iteration)
- Including a **larger dataset** (more years)

# Questions?