Wealth Forecast: Exploring Billionaires and Socio-Economic Factors

Sri Sudersan Thopey Ganesh,
Muqi Guo,
Xinmeng Wang

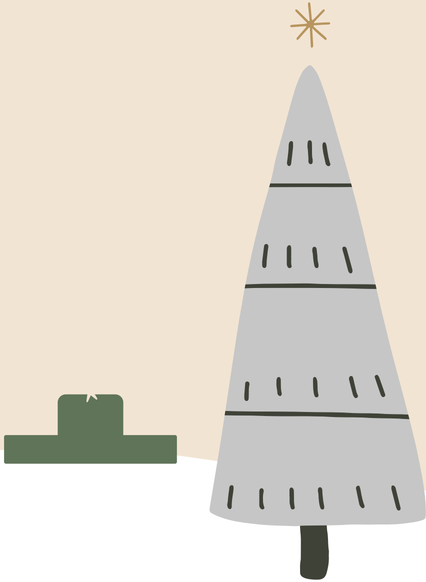
Introduction

Our project delves deep into the intricate relationship between socio-economic factors and individual financial success, with a specific focus on billionaires' wealth distribution.

Research Question

Gather Insights:

- Macroeconomic Factors
- Social Factors
- Personal Factors



Dataset

Acquired from Kaggle

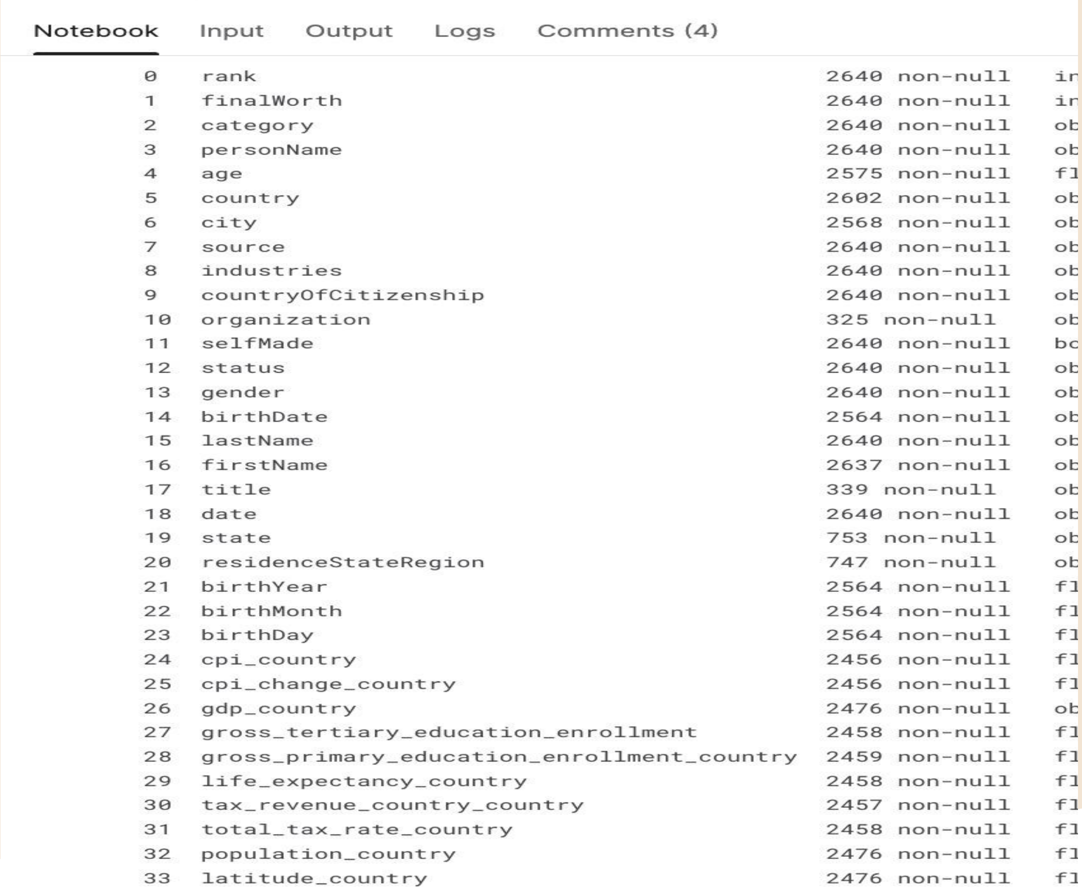
The dataset covers a spectrum of details, including businesses, industries, and personal profiles of billionaires worldwide.
– Wealth Distribution: Offers insights into the distribution of wealth among the global billionaire population.
– Business Sectors: Provides a breakdown of billionaires across various industries.
– Demographics: Reveals demographic trends within the billionaire community.

With a total of 35 features

While there are some missing values, columns with a significant percentage

Data was gathered in 2023

Dataset

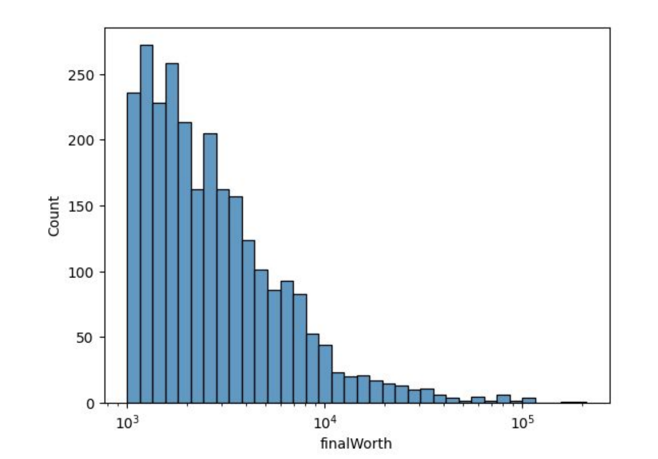


```
 0   rank                                      2640 non-null   in
 1   finalWorth                                2640 non-null   in
 2   category                                  2640 non-null   ob
 3   personName                                2640 non-null   ob
 4   age                                       2575 non-null   fl
 5   country                                   2602 non-null   ob
 6   city                                      2568 non-null   ob
 7   source                                    2640 non-null   ob
 8   industries                                2640 non-null   ob
 9   countryOfCitizenship                      2640 non-null   ob
 10  organization                              325 non-null    ob
 11  selfMade                                  2640 non-null   bo
 12  status                                    2640 non-null   ob
 13  gender                                    2640 non-null   ob
 14  birthDate                                 2564 non-null   ob
 15  lastName                                  2640 non-null   ob
 16  firstName                                 2637 non-null   ob
 17  title                                     339 non-null    ob
 18  date                                      2640 non-null   ob
 19  state                                     753 non-null    ob
 20  residenceStateRegion                      747 non-null    ob
 21  birthYear                                 2564 non-null   fl
 22  birthMonth                                2564 non-null   fl
 23  birthDay                                  2564 non-null   fl
 24  cpi_country                               2456 non-null   fl
 25  cpi_change_country                        2456 non-null   fl
 26  gdp_country                               2476 non-null   ob
 27  gross_tertiary_education_enrollment       2458 non-null   fl
 28  gross_primary_education_enrollment_country 2459 non-null  fl
 29  life_expectancy_country                   2458 non-null   fl
 30  tax_revenue_country_country               2457 non-null   fl
 31  total_tax_rate_country                    2458 non-null   fl
 32  population_country                        2476 non-null   fl
 33  latitude_country                          2476 non-null   fl
```

BACKGROUND/RELATED WORK

- Exploratory Data Analysis : Some Programmers have looked at businesses including manufacturing, banking, and investment that produce a large number of billionaires, as well as which countries are home to the greatest number of billionaires.

- Dataset : Classifies sources of billionaire-wealth, distinguishing between self-made and inherited fortunes.



Motivation

Comprehensive Understanding of Economic Dynamics:

- Motivation stems from a deep interest in unraveling intricate dynamics shaping individual financial success.
- Focus on patterns, correlations, and contributing factors to gain insights extending beyond personal finance.

Crucial Role in Wealth Distribution:

- Understanding implications for wealth distribution is essential for promoting fairness and equity within society.
- Identification of factors contributing to financial success provides insights into wealth accumulation or disparity.

Influence on Economic Policies:

- Insights gained aim to inform future policy formulation and decision-making.
- Nuanced understanding empowers policymakers to craft targeted and effective interventions for a more inclusive resource distribution.

Empowerment for Informed Decision-Making:

- Findings empower individuals to make informed choices about their lives.
- Knowledge about factors contributing to financial success guides decisions in education, career, investments, and lifestyle.

Holistic Approach to Societal Change:

- Motivation encapsulates a holistic approach to understanding the interplay of factors determining financial success.
- Empowerment of individuals and policymakers contributes to the creation of a more equitable and prosperous society.

Data Preprocessing:

Feature Selection:

- Elimination of irrelevant columns showing little correlation with the target variable ('finalWorth').
- Selected features for the model include demographic information, economic indicators, and country-specific details.

Categorical Data Transformation:

- Use of LabelEncoder to convert categorical columns ('country,' 'industries,' 'gender') into numerical labels.
- Conversion of data types to float to meet machine learning algorithm requirements.

Data Preprocessing:

Cleaning and Converting 'gdp_country':

- Cleaning and converting 'gdp_country' to Decimal type.
- Handling non-numeric or missing values using NumPy's np.nan, pandas' pd.isnull, and Decimal class.
- Application of regular expressions to eliminate non-numeric characters before conversion.

Calculation of GDP per Capita:

- Calculation of GDP per capita for a more meaningful representation of a country's economic output.
- Creation of 'gdp_per_capita' column by dividing 'gdp_country' by 'population_country.'

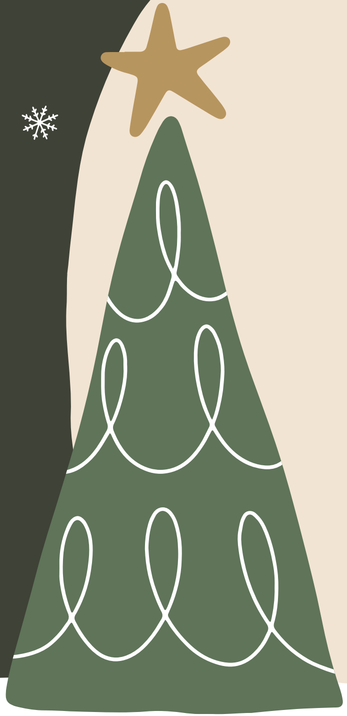
Data Preprocessing:

Quartile Categorization of 'finalWorth':

- Calculation of quartiles (percentiles) for the 'finalWorth' column.
- Categorization into four distinct categories based on quartiles:
  below 25th percentile (0), between 25th and median (1), median to 75th percentile (2), and above 75th percentile (3).
- Use of percentiles for labeling to reduce sensitivity to extreme values or outliers.

Handling NaN Values:

- Acknowledgment of the presence of NaN values in the dataset.
- A more detailed discussion on NaN values planned during the modeling phase.



Models & Approach

Logistic Regression

Sigmoid Function and Non-Linearity:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

- The sigmoid function played a crucial role in introducing non-linearity to the model.
- It enables the model to capture intricate relationships within the data

Cost Function and Gradient Descent:

- The logistic regression model includes a cost function
- Gradient descent optimization is used for iterative parameter adjustment

$$-\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i))$$

Binary Cross-Entropy / Log Loss

Logistic Regression

Handling Missing Values:

- Missing values were addressed during data preprocessing by replacing them with the median of respective columns.

Feature Standardization and Bias Term:

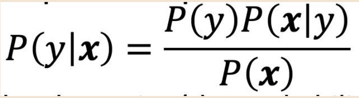
- Features were standardized in the preprocessing pipeline
- Addition of a bias term

One-Hot Encoding for Multiclass Nature:

- The target variable was transformed into one-hot encoding to facilitate training for multiclass classification.
- This ensures the model can effectively classify instances across multiple categories.

Naive Bayes

1. Choice of Algorithm:
   - Reasons for Selection:
     - Computational simplicity and efficiency.
     - Less sensitivity to missing data.
2. Data Preparation:
   - Conversion of continuous features into binary (0 and 1) format, except for gender.
3. Core Computation:
   - Posterior Probability Calculation:



$$P(y|\boldsymbol{x}) = \frac{P(y)P(\boldsymbol{x}|y)}{P(\boldsymbol{x})}$$

4. Implementation Details:
   - Calculation of class prior and generative likelihood using the formulas.
   - Use of dictionaries for storing likelihoods:
     - Enhances efficiency through constant complexity in dictionary look-ups.

Random Forest

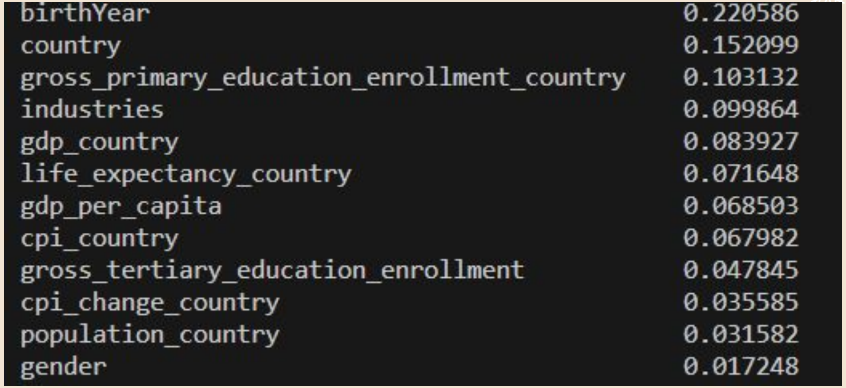
- Ensemble Learning: Using many decision trees, the Random Forest Classifier is an ensemble learning technique that produces predictions that are more reliable and accurate.
- Decision Trees: Every ensemble tree is a decision tree, a structure resembling a flowchart in which nodes stand in for choices made in response to certain features.
- Model Pattern : Birthyear, gender, country, industries
- Versatility, Feature Importance

$$Entropy = \sum_{i=1}^{C} -p_i * \log_2(p_i)$$

Conclusions/Analysis

Accuracy:

- Logistic Regression: Training Accuracy (30.91%), Testing Accuracy (31.93%)
- Naïve Bayes: Validation Accuracy (31.3%)
- Random Forest: Validation Accuracy (32.45%)



```
birthYear                                    0.220586
country                                      0.152099
gross_primary_education_enrollment_country   0.103132
industries                                   0.099864
gdp_country                                  0.083927
life_expectancy_country                      0.071648
gdp_per_capita                               0.068503
cpi_country                                  0.067982
gross_tertiary_education_enrollment          0.047845
cpi_change_country                           0.035585
population_country                           0.031582
gender                                       0.017248
```

Random Forest shows a slightly higher accuracy compared to Logistic Regression and Naïve Bayes.

Precision:

- Logistic Regression: 31%
- Naïve Bayes: 26.8%
- Random Forest: 30.492%

Logistic Regression and Random Forest have similar precision, with Naïve Bayes having a slightly lower precision.

Conclusions/Analysis

Random Forest achieves the highest F1 score.

Logistic Regression follows with a moderate F1 score.

Naïve Bayes has the lowest F1 score.

Random Forest generally outperforms Logistic Regression and Naïve Bayes in terms of accuracy, precision, recall, and F1 score.

However, the performance of each model is contingent on the dataset's specific characteristics and requirements.

Further tuning and feature exploration may be necessary to enhance overall model effectiveness..

Future Work/Extensions

1. Refining Feature Selection:
    a. Conduct a more rigorous analysis to select more predictive features.
    b. Techniques for Improvement:
        i. Advanced feature engineering, e.g., Principal Component Analysis (PCA).
        ii. Emphasis on reducing redundancy and extracting informative features.
2. Reformulating Research Questions:
    a. Current results indicate a misalignment between our questions and methodologies.
    b. Future work:
        i. Reassess the research objectives to align better with our methods.
        ii. Consider focusing on different dataset aspects or redefining scope to fit our model capabilities.
3. Explore Regression Models:
    a. Converting continuous features into ordinal ones could result in a loss of information.

Bibliography

Burlick, M. (2023). Probabilistic/Statistical Classification. [PowerPoint slides]. Drexel University. Retrieved from BBlearn.

Webb, G.I. (2011). Naïve Bayes. In C. Sammut & G.I. Webb (Eds.), Encyclopedia of Machine Learning (pp. [pages]). Springer. https://doi.org/10.1007/978-0-387-30164-8_576

Freund, C. (2016). The Origins of the Superrich: The Billionaire Characteristics Database. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2731353

mawro73. (2023). Regression on Millionaire Data. Kaggle. Retrieved [Access Date], from https://www.kaggle.com/code/mawro73/regressionon-millionairedatas

Nelgiriye Withana. (2023). Billionaires Statistics Dataset. Kaggle. Retrieved [Access Date], from https://www.kaggle.com/datasets/nelgiriyewithana/billionaires-statistics-dataset
codeRuslan. (2023).
Forbes Classification of Billionaires. GitHub. Retrieved [Access Date], from https://github.com/codeRuslan/Forbes-classification-of-billionaires

Shivansh R. (2023). Exploratory Analysis - Billionaires Dataset. Kaggle. Retrieved [Access Date], from https://www.kaggle.com/code/shivanshr12/exploratory-analysis-billionaires-dataset

Happy Holidays

Thank You!