ANALYSIS OF BILLIONAIRES STATISTICS

ABHILASH SAMPATH(50485796)

HARSHAVARDHAN BAIRA REDDY(50468279)

NIKITHA SADANANDA(50471079)

Github Repository:

https://github.com/AbhilashBharadwaj/statistical-data-mining

1 ABSTRACT

The dataset provides a comprehensive analysis of the world's billionaires, encompassing a wide range of metrics that paint a detailed picture of global wealth distribution. Key elements include individual rankings, net worth, business sectors, and personal details of each billionaire. This data facilitates in-depth examinations of wealth distribution across various industries and geographies, as well as insights into the demographics of the world's richest individuals.

Additionally, the dataset includes demographic details like age, gender, and birthplace, and whether the wealth was self-made or inherited. It also correlates billionaires' wealth with economic indicators such as GDP, CPI, tax rates, and education enrollment levels in their respective countries. Furthermore, the dataset offers a geospatial perspective, illustrating the geographical spread of billionaires globally.

This rich dataset serves as a valuable resource for analyzing trends in billionaire demographics and wealth over time, providing critical insights into the dynamics of global wealth and economic power.

2 INTRODUCTION

The objective of this project was to leverage advanced machine learning techniques to analyze and predict key features related to the world's billionaires, using a comprehensive dataset encompassing various metrics. The primary goal was to understand the relationships between the final worth of billionaires and other significant factors, such as country-specific tax rates, population sizes, educational enrollment levels, and the Consumer Price Index (CPI).

The project followed a systematic workflow:

1. **Data Collection and Preprocessing**: The initial step involved gathering a robust dataset containing detailed information about the world's billionaires, including their final worth, demographic details, and country-specific economic indicators. The data was then

preprocessed to ensure quality and consistency, which involved cleaning, normalization, and handling missing values.

2. **Exploratory Data Analysis (EDA)**: An extensive EDA was conducted to understand the data's characteristics, uncover patterns, and identify potential correlations and anomalies. This step was crucial for gaining insights and informing the subsequent modeling process.
3. **Feature Selection and Engineering**: Based on the EDA insights, relevant features were selected and engineered to effectively capture the complexities within the data. This included creating new variables and transforming existing ones to enhance the models' predictive capabilities.
4. **Model Development**: A variety of machine learning models were employed, including XGBoost, Random Forest, and Linear Regression. Each model was chosen for its unique strengths in handling different aspects of the data and its predictive performance.
5. **Model Training and Evaluation**: The models were trained on a subset of the data, and their performance was rigorously evaluated using appropriate metrics. This step involved tuning hyperparameters and applying cross-validation techniques to ensure robustness and generalizability.
6. **Insight Generation and Prediction**: The final models were used to generate insights into the relationships between billionaires' net worth and other variables like tax rates, population, and CPI. Predictive analyses were also conducted to estimate these features based on the learned patterns.
7. **Validation and Interpretation**: The last phase involved validating the models' predictions and interpreting the results in the context of the initial research questions. This step was essential for drawing meaningful conclusions and identifying areas for further research.

3 DATA DESCRIPTION

This dataset offers a multifaceted view of the world's billionaires, combining personal details with broader economic and demographic data. Each of the 2641 rows represents a unique billionaire, and the dataset includes a wide range of attributes that shed light on various aspects of their wealth, background, and the economic environment of their country. Let’s delve deeper into each attribute:

1. **Rank**: Provides a global wealth ranking of each billionaire, offering a comparative perspective on their financial standing relative to others in the dataset.
2. **FinalWorth**: This is the net worth of the billionaire in U.S. dollars, reflecting their current financial valuation and serving as a key measure of their wealth.
3. **Category**: Indicates the primary business sector or industry in which the billionaire's wealth was generated, such as technology, finance, or manufacturing.

4. **PersonName**: The full legal name of the billionaire, enabling individual identification within the dataset.
5. **Age**: This shows the age of the billionaire, which can be used to analyze wealth accumulation over a lifetime.
6. **Country**: The country where the billionaire currently resides, providing a geographical context for their wealth.
7. **City**: This gives a more specific location of residence within the country, offering insights into urban versus rural distributions of wealth.
8. **Source**: Identifies the primary source of the billionaire's wealth, such as a specific company, investment, or inheritance.
9. **Industries**: Lists the industries associated with the billionaire's business interests, which could be diverse and span multiple sectors.
10. **CountryOfCitizenship**: Shows the billionaire's nationality, which may differ from their country of residence.
11. **Organization**: The name of the company or organization primarily associated with the billionaire, indicating their main business affiliation.
12. **SelfMade**: This binary indicator (True/False) specifies whether the billionaire's wealth was self-made or inherited.
13. **Status**: A categorization of the billionaire's wealth origin: "D" for self-made (founders/entrepreneurs) and "U" for inherited or unearned wealth.
14. **Gender**: The gender of the billionaire, which is vital for gender-based wealth distribution analysis.
15. **BirthDate**: Provides the date of birth, which, combined with age, can be used for age-related analyses.
16. **LastName and FirstName**: These fields offer a breakdown of the billionaire's full name into last and first names for sorting or categorization purposes.
17. **Title**: Any formal title or honorific associated with the billionaire.
18. **Date**: Indicates the date on which the data was collected, essential for understanding the temporal context of the information.
19. **State and ResidenceStateRegion**: These provide further details on the billionaire's location within their country of residence.
20. **BirthYear, BirthMonth, BirthDay**: Break down the birth date into year, month, and day for more granular age-related analysis.
21. **Cpi_country and Cpi_change_country**: These indicators provide the Consumer Price Index and its change over time in the billionaire's country, reflecting inflation and cost of living.
22. **Gdp_country**: The Gross Domestic Product of the billionaire's country, offering a measure of the nation's economic size and health.
23. **Gross_tertiary_education_enrollment and Gross_primary_education_enrollment_country**: These metrics indicate the levels of

higher and primary education in the billionaire's country, useful for correlating wealth with educational attainment.

24. **Life_expectancy_country**: This is a measure of average life expectancy in the billionaire's country, providing a health and development indicator.
25. **Tax_revenue_country_country**: Shows the total tax revenue in the billionaire's country, relevant for understanding fiscal policies.
26. **Total_tax_rate_country**: This represents the overall tax rate in the billionaire's country, which can impact wealth accumulation and distribution.
27. **Population_country**: The total population of the billionaire's country, essential for contextualizing their wealth on a national scale.
28. **Latitude_country and Longitude_country**: Geographic coordinates of the billionaire's country, useful for spatial analysis and mapping.

This dataset is a rich source for analyzing global wealth distribution, understanding the interplay between personal wealth and broader economic factors, and exploring demographic patterns among the world's wealthiest individuals.

4 MATERIALS AND METHODS

4.1 Exploratory data analysis :

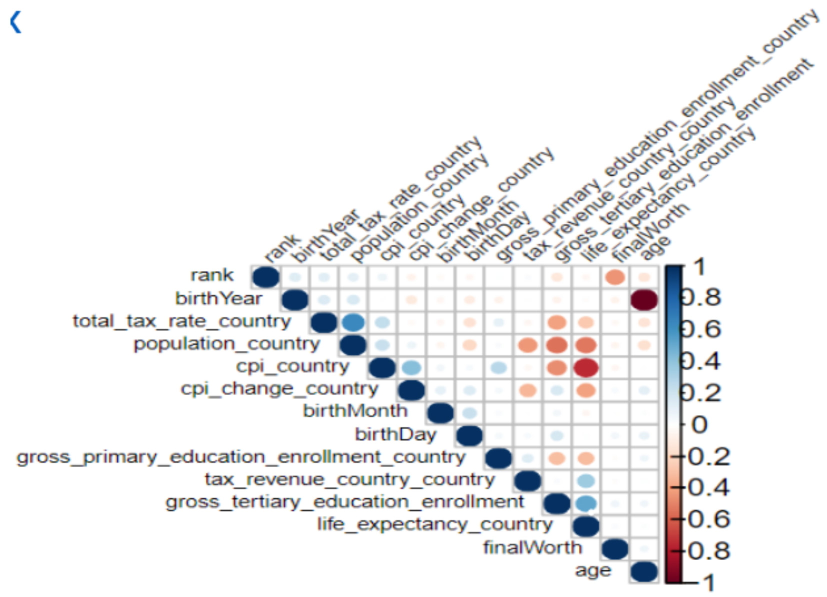


Figure 1: Correlation Matrix

- **finalWorth** has a moderately positive correlation with **age** (the square is moderately large and red), which might suggest that as billionaires get older, their net worth tends to increase.
- There are a few variables such as **rank, birthYear, and total_tax_rate_country** that have a negative correlation with **finalWorth**. This could imply that higher tax rates or a younger birth year might be associated with lower net worth, or that as the rank increases (which usually means a lower position on the billionaire list), the final worth decreases.
- Many squares are white or have very small dots, indicating very low or no correlation between those variables.
- It's important to note that correlation does not imply causation. Even if two variables show a strong correlation, it does not mean one causes the other to change.

Without the exact context or data description, this analysis is based purely on the visualization provided. The actual implications of these correlations would depend on a more in-depth statistical analysis and understanding of the data

Correlation of Tax Rates and Population with Other Economic Indicators: The variables total_tax_rate_country and population_country show correlations with other variables like gross_primary_education_enrollment_country and cpi_country (Consumer Price Index). A correlation between tax rates and education enrollment could suggest that higher tax revenues might be associated with higher investment in education, although this would require further analysis to understand causality or the direction of the relationship. Similarly, a correlation between population size and CPI could indicate that larger populations might be related to higher levels of consumer prices, which could be a result of demand pressures or other economic factors.

Positive Correlation between Different Levels of Educational Enrollment: There is a visible positive correlation between **gross_primary_education_enrollment_country** and **gross_tertiary_education_enrollment_country**, as indicated by the red square where these two variables intersect. This suggests that countries with higher enrollment rates in primary education also have higher enrollment rates in tertiary education. This relationship could be due to a variety of factors, including the overall value placed on education within a country, educational policies that encourage continued education, or socioeconomic factors that enable individuals to pursue higher levels of education.

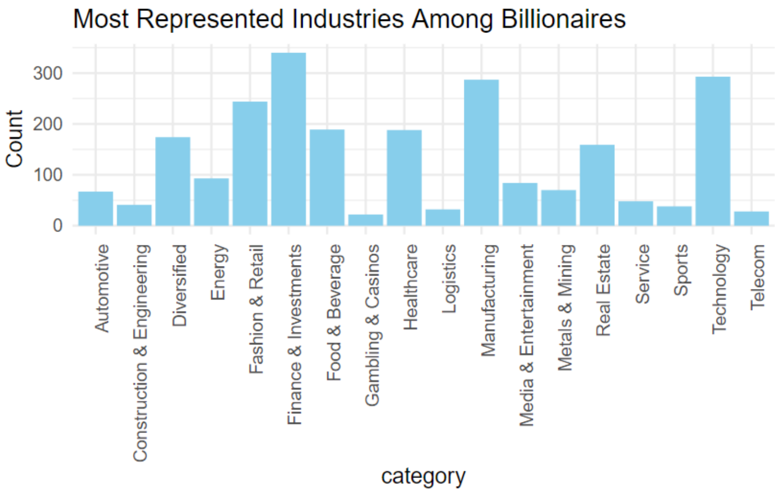


Figure 2: Most Represented Industries among Billionaires

The graph shows different industries on the x-axis, labeled "Category", and a "Count" on the y-axis, which represents the number of billionaires in each industry.

From the plot, Technology and Real Estate are the industries with the highest number of billionaires, with Technology being the most represented. Other industries shown include Finance & Investments, Food & Beverage, Retail, and Healthcare, among others. Automotive, Construction & Engineering, and Telecom have the fewest billionaires represented among the industries listed.

The graph provides an overview of which sectors the wealthiest individuals are most commonly associated with, indicating where the most significant capital accumulation is occurring across industries.

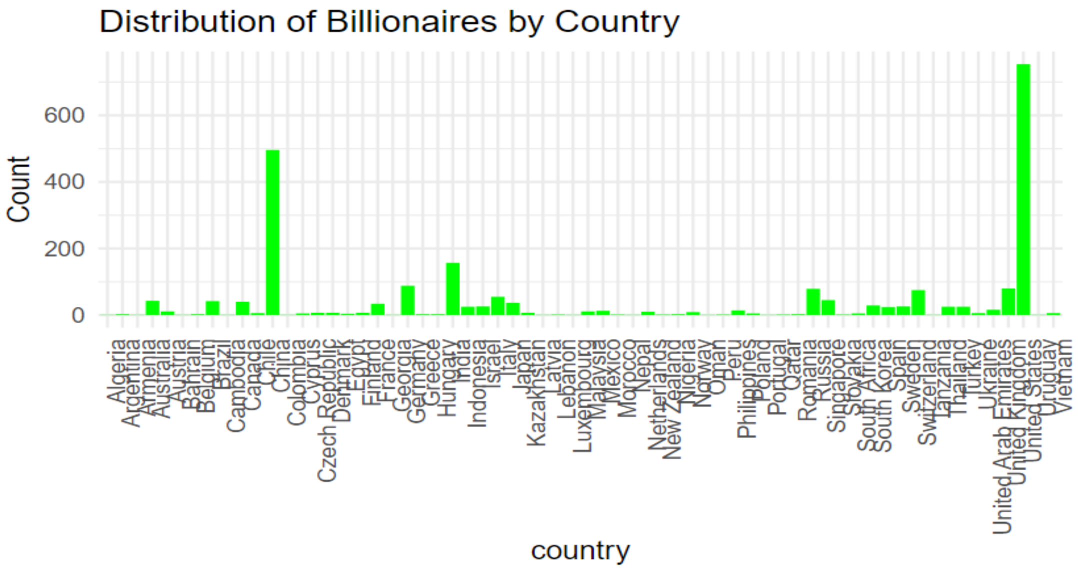


Figure 3: Distribution of Billionaires by Country

The graph displays various countries on the x-axis, and the "Count" of billionaires on the y-axis. The bars represent the number of billionaires in each country.

The bar graph shows that the United States has the highest number of billionaires, with a count far exceeding that of any other country on the chart. There appears to be another significant peak in the graph for China, indicating a high number of billionaires as well, although substantially less than the United States.

Most other countries have relatively few billionaires in comparison, with many countries having counts so low that the bars are barely visible above the x-axis.

This distribution suggests a significant concentration of extreme wealth within the United States, followed by China, with much lower levels of billionaire representation in other countries. This could reflect broader economic patterns, such as market size, the level of industrialization, economic policies, and opportunities for wealth accumulation in different nations.

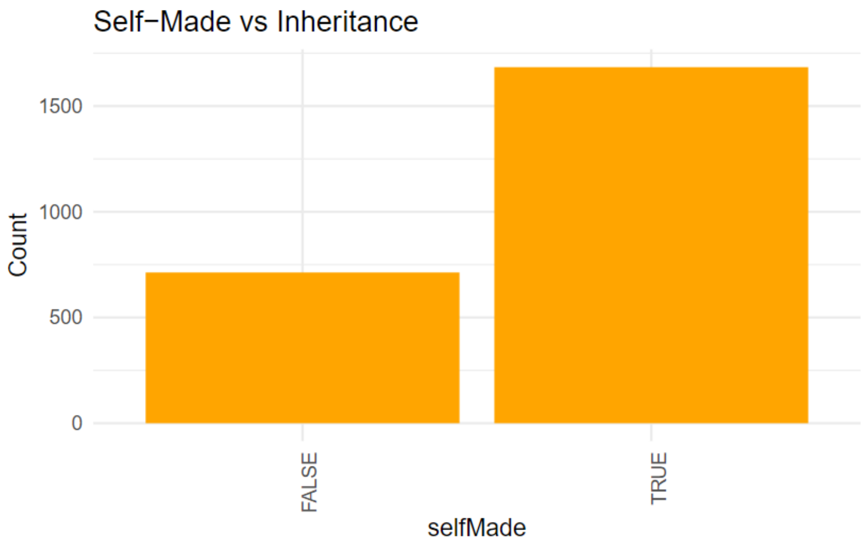


Figure 4: Self made vs Inheritance

The graph compares the count of billionaires who are self-made versus those who inherited their wealth. On the x-axis, there are two categories labeled "FALSE" and "TRUE", which corresponds to whether the billionaires are self-made or not. The y-axis represents the count of individuals.

The "FALSE" category, which presumably represents those who inherited their wealth, has a significantly lower count compared to the "TRUE" category. The "TRUE" category, indicating self-made billionaires, has a count that is more than double that of the "FALSE" category.

The graph suggests that among billionaires, a larger proportion have created their wealth themselves rather than having inherited it. This could reflect trends in global wealth creation, entrepreneurship, and the opportunity for individuals to amass significant wealth through their endeavors.

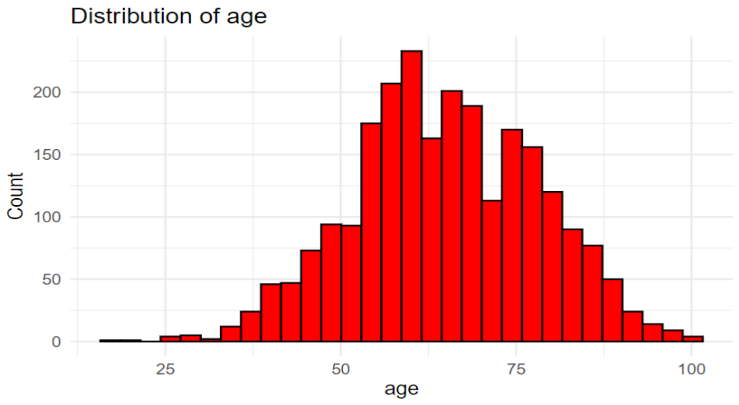


Figure 5: Distribution of Age

The x-axis represents age, and the y-axis represents the count of individuals within each age bracket.

The distribution forms a bell-shaped curve, which is slightly skewed to the right. The peak of the distribution appears to be between the ages of 50 and 60, indicating that the majority of individuals in this dataset are in that age range. There are very few individuals below the age of 40 and above the age of 80.

This suggests that in this particular group, becoming a billionaire might typically occur in middle age or later, possibly after one has had the opportunity to accumulate wealth over time. There's a gradual decline in the count of individuals as age increases past the peak, which could be attributed to a combination of factors such as retirement, passing of wealth to the next generation, or the smaller number of individuals reaching older ages.

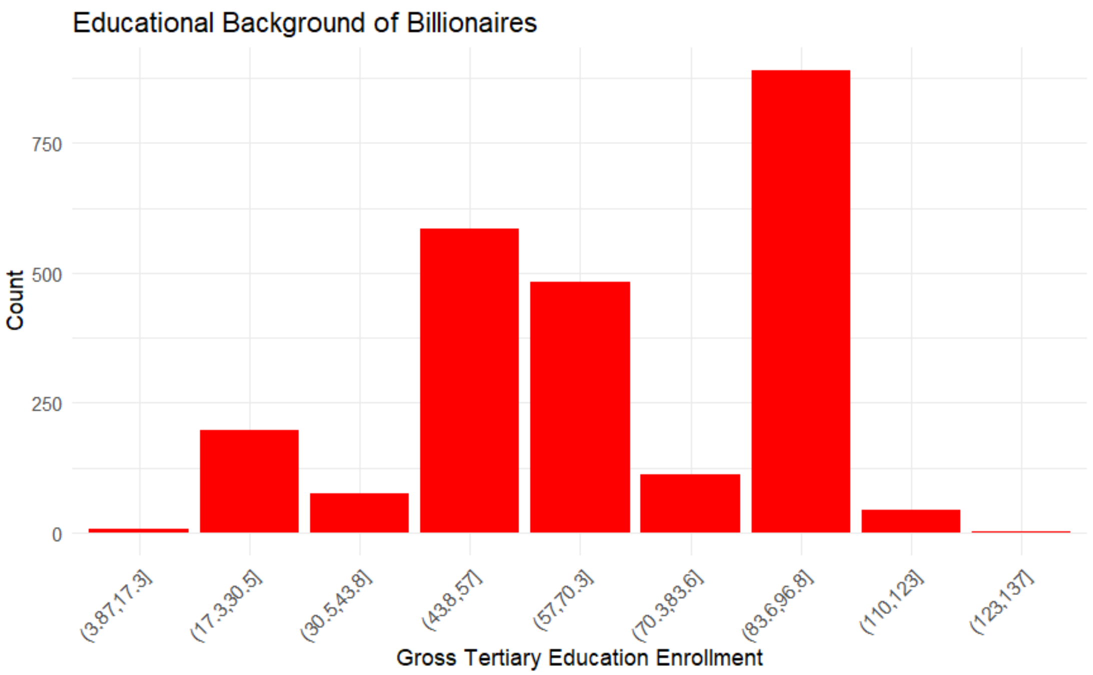


Figure 6: Education background of billionaires

The graph represents the correlation between billionaires and their levels of tertiary (higher) education enrollment. The x-axis is labeled "Gross Tertiary Education Enrollment" with various categories, likely indicating different levels or rates of enrollment, while the y-axis shows the count of billionaires.

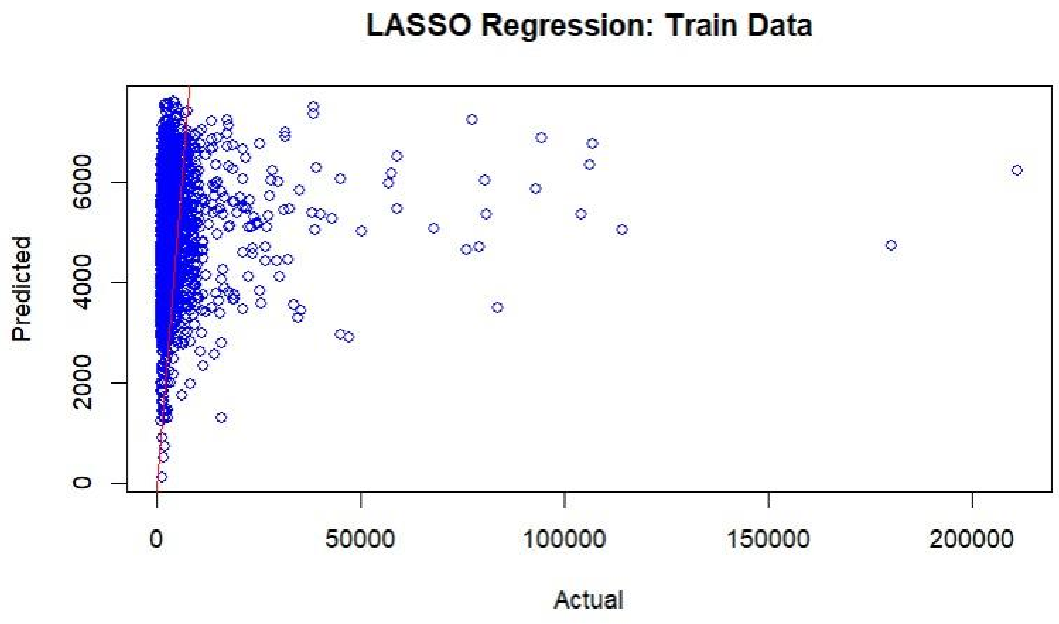
The bar heights vary, with the highest bar indicating that a large number of billionaires fall within the category labeled "([10-13[)", which might suggest that this group has some college education but not necessarily a four-year degree. The second-highest category is "([13-16[)", possibly indicating those who have completed a traditional undergraduate degree or equivalent.

The other categories, which could represent different ranges of educational attainment (perhaps graduate and post-graduate education, as well as those with less than tertiary education), have fewer billionaires compared to the two aforementioned categories.

Overall it indicates that the majority of billionaires have some level of tertiary education, with significant numbers having attended but not necessarily completed a full undergraduate program, and a substantial number completing what may be equivalent to an undergraduate degree. Fewer billionaires are in the categories that might represent higher levels of education, such as master's or doctoral degrees, or lower levels, such as high school or no college education.

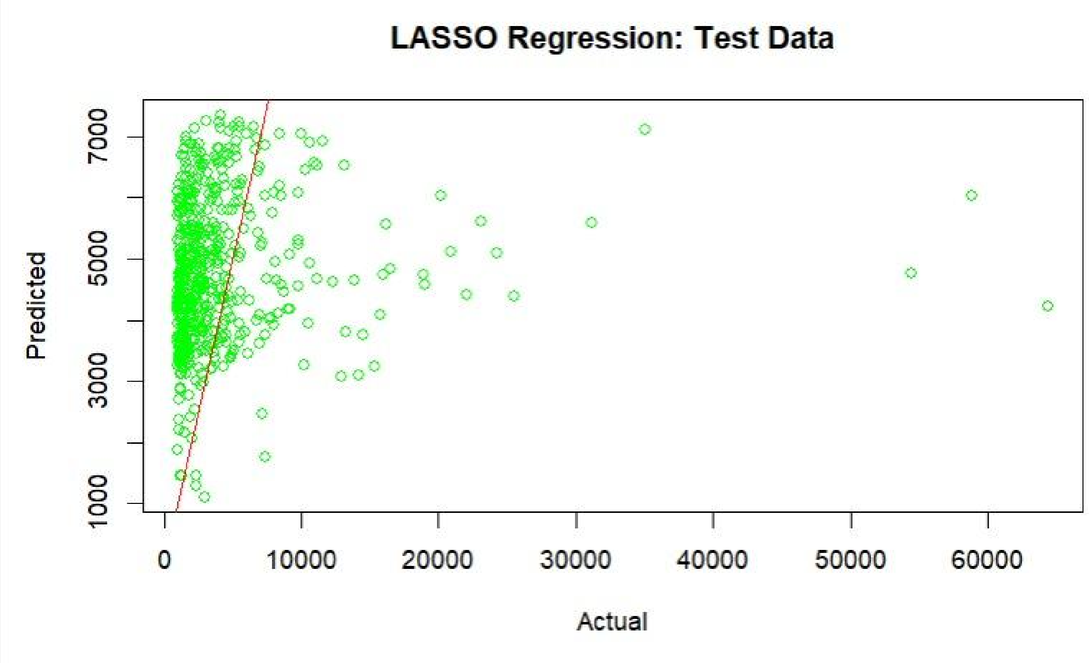
4.2 Data Modelling and Machine Learning algorithms

Lasso Regression to predict the final Worth



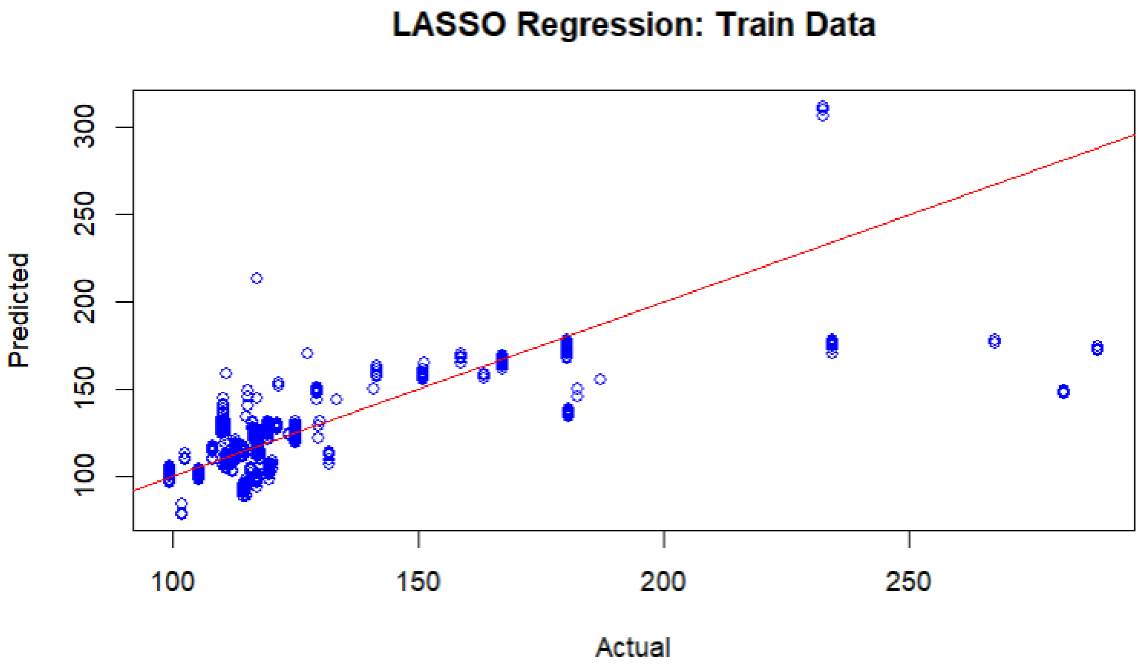
- Density of Points Near Zero: The model clusters many predictions near zero, indicating it may accurately predict lower actual values, though further analysis is needed to confirm this trend.
- Spread Along the Actual Axis: Predictions disperse as actual values rise, suggesting decreased model precision for higher value predictions.
- Potential Overfitting: A notable number of predictions underestimate high actual values, hinting the model might be overfitting and not generalizing well for the entire data range.
- Variability in High-Value Range: The model's variability in predicting higher actual values could reflect conservative estimates or insufficiently captured features for these instances.
- Outliers: Points significantly distant from the main data cluster, particularly where predictions fall short of high actual values, indicate potential outliers or areas where the model's accuracy is limited.

Lasso Regression to predict the final Worth



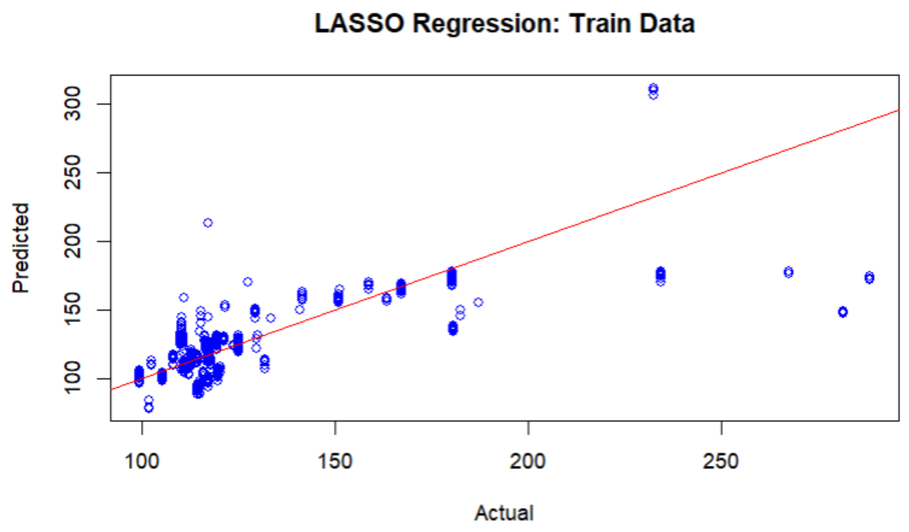
- Concentration at Lower Values: There's a dense concentration of predictions at the lower end of the scale, similar to the training data. This suggests the model predicts lower values with higher frequency.
- Model Predictive Behavior: The predictions appear to fan out as the actual values increase, indicating more variability in the model's accuracy for higher actual values.
- Accuracy at Higher Ranges: The model rarely predicts values at the higher end of the actual value range. This could indicate limitations in the model's ability to predict higher outcomes accurately.
- Outliers: As with the training data, the test data shows some potential outliers, with a few actual values not matched by corresponding predicted values.
- General Model Fit: The overall pattern suggests the model is not overfitting since it shows a consistent trend between the training and test data. However, it may be underfitting, especially at higher actual values, due to a conservative prediction trend.

Lasso Regression to Predict the CPI



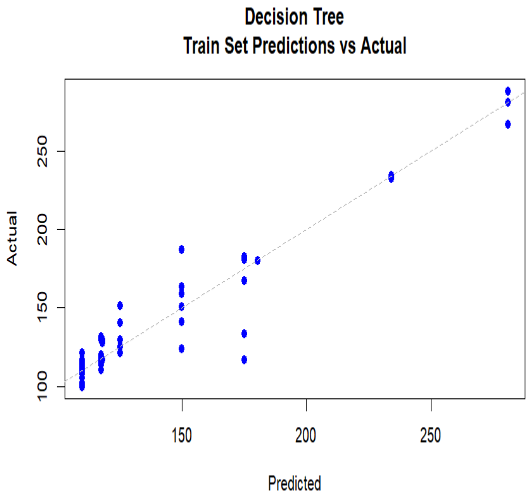
- **Model Overview**: LASSO regression predicts CPI with feature selection to prevent overfitting, visualized by a scatter plot.
- **Predictive Accuracy**: Moderate accuracy is suggested by the clustering of predicted CPI values around the best fit line.
- **Line of Best Fit**: The red line shows the model's CPI predictions, with deviations representing prediction errors.
- **Variance in Predictions**: Points spread around the best fit line indicate varying prediction accuracy for CPI.
- **Potential Outliers**: Notably distant points from the line may be outliers, warranting further investigation.
- **Model Fit**: The model reasonably captures the CPI trend, yet indicates possible improvements for distant points.

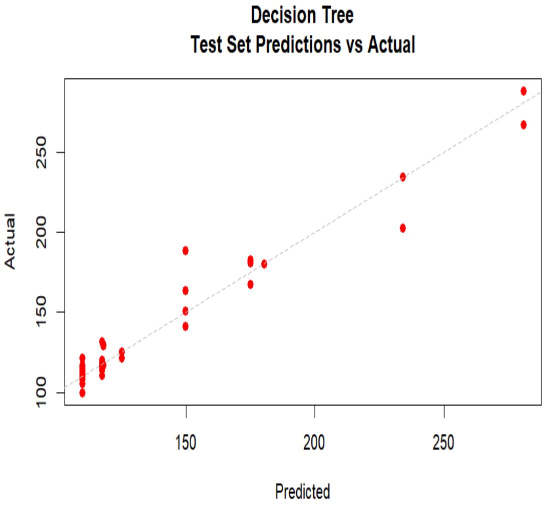
Lasso Regression on test Dataset to predict the CPI



- **Model Overview**: This scatter plot demonstrates the LASSO regression model's predictions on test data for forecasting the Consumer Price Index (CPI). The model leverages regularization to enhance prediction accuracy and feature selection.
- **Predictive Accuracy**: The distribution of green dots around the red line of best fit indicates the model's predictions are closely aligned with the actual CPI values on the test set, signifying a high level of accuracy.
- **Line of Best Fit**: The red line, which should ideally match with all green dots if predictions were perfect, reveals only minor deviations for most points, suggesting the model's effective performance on test data.
- **Variance in Predictions**: The tight clustering of the points around the line of best fit points to a consistent and reliable predictive performance across the range of CPI values, with limited variance.
- **Potential Outliers**: Compared to the training data, the test data shows fewer significant outliers, implying that the model's predictions are robust to overfitting and likely to be more generalizable.
- **Model Fit**: The general pattern on the plot suggests that the model maintains a strong predictive relationship between the features and the CPI on unseen data, which is crucial for practical forecasting applications.

Decision tree for predicting CPI

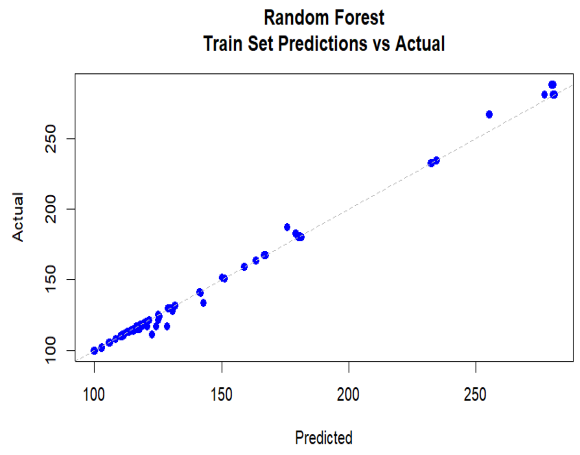


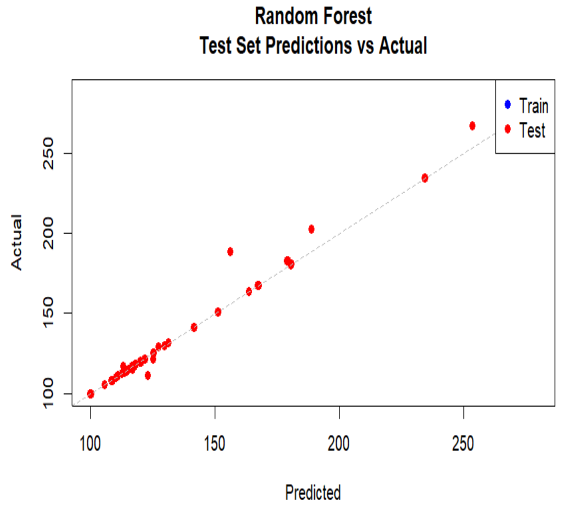


1. **Train Set Predictions vs Actual :** The blue points represent the relationship between the predicted and actual values of the training dataset. The closer these points are to the dashed gray line, which represents perfect prediction, the better the model's predictions. The concentration of points along the line suggests the model fits well to the training data.
2. **Test Set Predictions vs Actual :** The red points illustrate the predicted versus actual values for the test dataset. This graph assesses the model's generalization to unseen data. The spread of points indicates the variance in the model's predictive accuracy on the test data, with some points deviating from the line, showing areas where the model's predictions were less accurate.

Both graphs are crucial for evaluating the model's performance; the training graph for overfitting and learning in the training phase, and the test plot for predictive power and generalization in the testing phase.

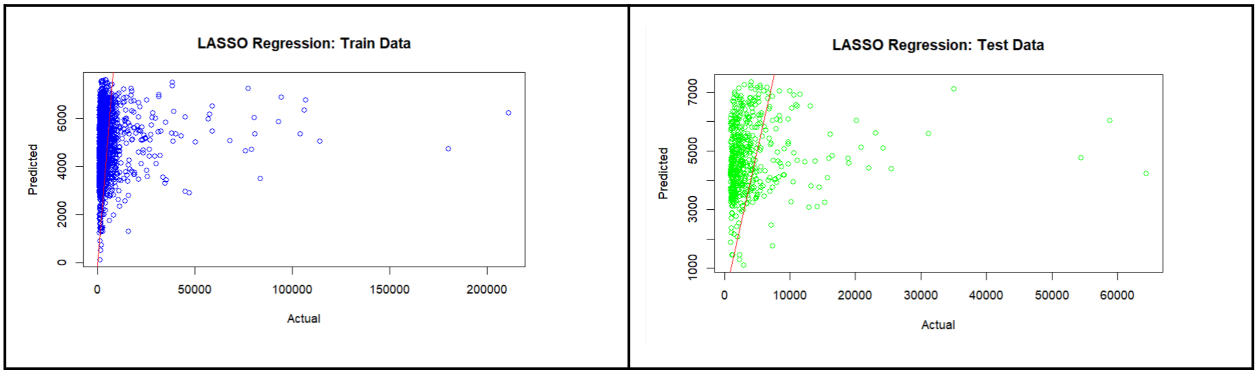
Random Forest for predicting CPI





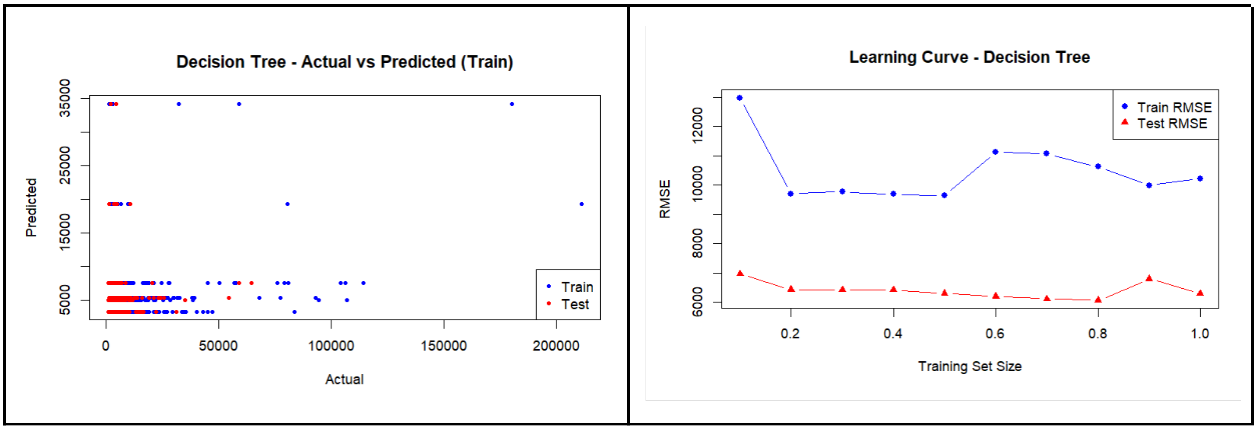
1. Random Forest Train Set Predictions vs Actual : The graph shows a scatter plot of actual versus predicted values for the training set, depicted by blue points. The points are closely aligned along the dashed gray line, which indicates the line of perfect prediction. This close alignment suggests the model has learned the training data well, displaying a high degree of accuracy in predictions.
2. Random Forest Test Set Predictions vs Actual: The scatter plot represents the model's predictions on the test set, indicated by red points. The points are well-aligned with the line of perfect prediction but show a slight spread, indicating variability in the model's performance on unseen data. The graph indicates a strong predictive performance, although not as tight as the training set, which is common in model evaluations due to the test set featuring previously unseen data.

Lasso regression for predicting the finalWorth



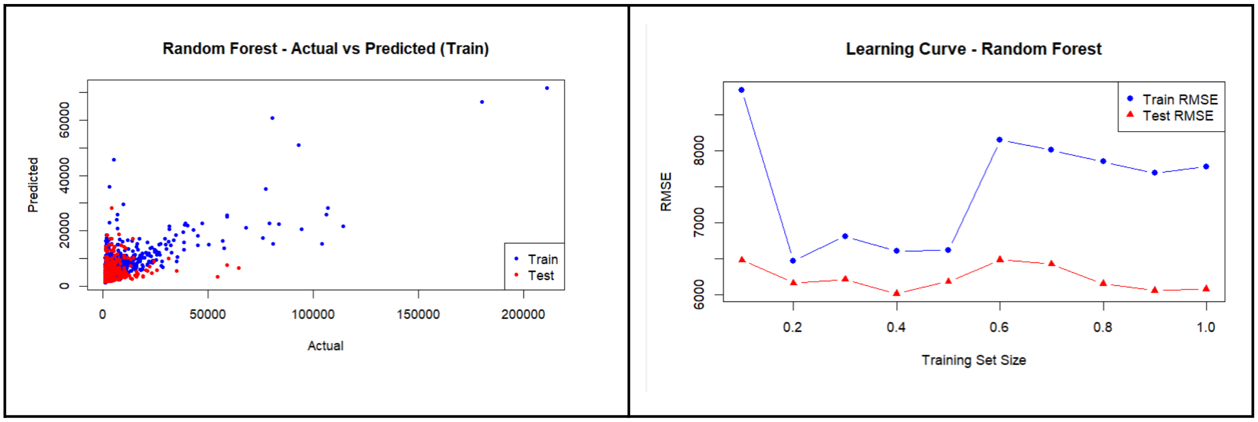
The scatter plots for the LASSO regression model predicting `finalWorth` show a better fit for lower values with a consistent performance across training and test sets, suggesting no overfitting. However, the model underpredicts at higher `finalWorth` values, indicating potential underfitting, likely due to LASSO's regularization effect. This uniform pattern of underprediction implies that the model may benefit from reevaluation of feature selection, engineering, or even reconsideration of the regularization strength to improve predictions across the full range of `finalWorth`. The model's current limitations and the error distribution call for a deeper analysis, potentially looking into residual patterns, to refine the model further and enhance its predictive accuracy, especially for higher `finalWorth` values.

Decision tree for predicting finalWorth

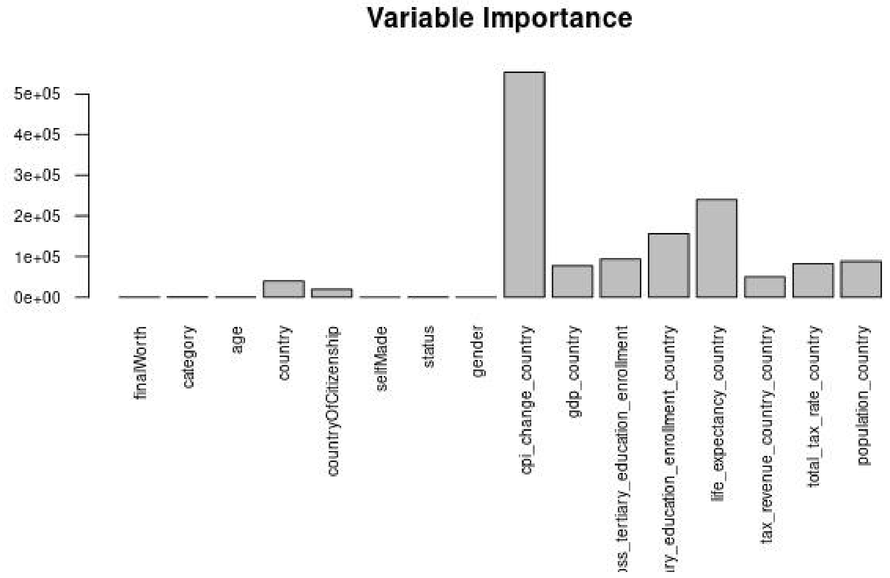


The left graph, an Actual vs Predicted plot, shows that the model has a relatively consistent performance on both training and test data, with a tendency to underpredict at higher values of `finalWorth`. The right graph, depicting the Learning Curve, demonstrates that as more data is used for training, the RMSE for the training set shows some volatility but generally remains lower than the test RMSE. This indicates that the model may not be improving with additional data and might be experiencing high variance. The test RMSE, while higher, shows a downward trend as the training set size increases, suggesting some learning but with potential overfitting as the model does not generalize as well on the test data. Overall, the model's prediction accuracy could potentially be improved, possibly by tuning model complexity or by addressing any data-specific challenges such as outliers or feature engineering.

Random forest for predicting finalWorth



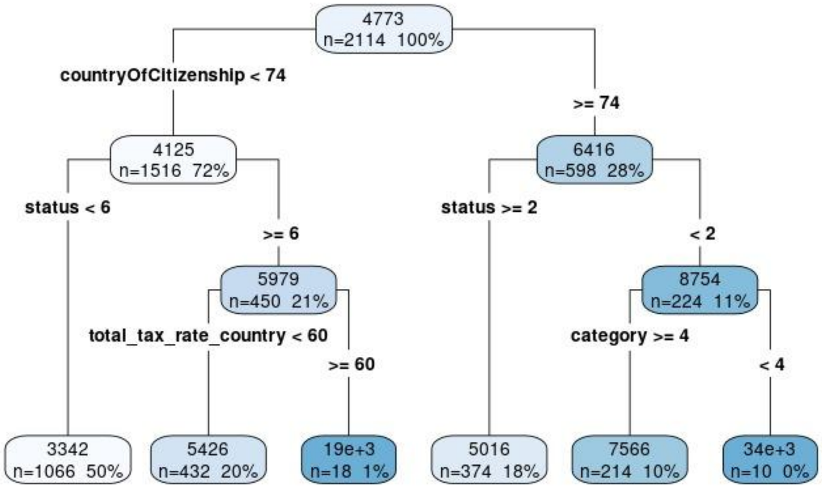
In the Actual vs Predicted plot on the left, the data points suggest that the model's predictions are closely aligned with the actual values for lower `finalWorth`, but there is greater variance in predictions as the value of `finalWorth` increases. This is indicated by the spread of points away from the line of equality (where predicted values would match actual values). The right graph is a Learning Curve that displays the Root Mean Square Error (RMSE) for both the training and test sets across different sizes of the training data. The RMSE for the training data appears relatively high with small training sizes but improves significantly as more data is included. Conversely, the RMSE for the test data decreases as the training size increases, suggesting that the model benefits from more data and is generalizing well. The overall trend in the Learning Curve indicates that adding more training data improves the model's performance, with the test RMSE converging to a lower value, indicative of a good balance between bias and variance.



1. The chart ranks the importance of different variables used in a predictive model.

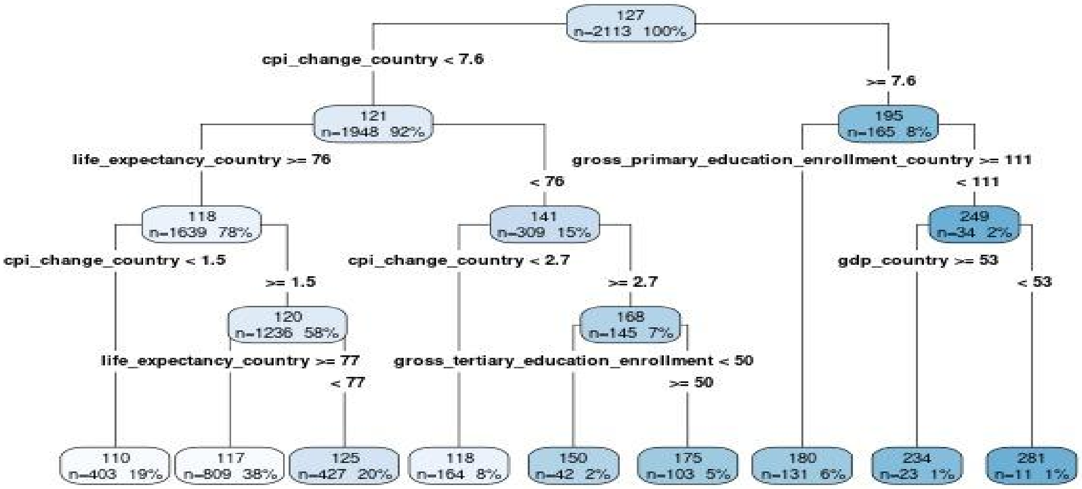
2. The variable cpi_change_country appears to be the most important feature by a significant margin.
3. Other notable variables include age, gender, and status, but they are much less important than cpi_change_country.
4. The variable category has the second-highest importance but is still less than half as important as cpi_change_country.
5. Variables related to the country (countryCitizenship, country, gdppercapita_country, etc.) show moderate importance.
6. The chart uses a scale that likely represents a metric used to quantify importance, such as gain, frequency, or some form of statistical measure.
7. Variables like education_country, taxrevenue_country, and population_country seem to have the least importance among those shown.

Decision tree predicting finalWorth



The decision tree diagram represents a model for predicting the `finalWorth` variable, beginning with an initial split on `countryOfCitizenship` and further partitioning based on `status`, `total_tax_rate_country`, and `category`. The tree's branches conclude with predicted `finalWorth` values, each branch reflecting a specific path of decision criteria. Notably, the tree suggests `status` is a significant predictor, as it appears in two separate splits, with the final predictions ranging from 3342 to a high of `34e+3`, depending on the path taken through the tree. The structure of the tree, along with the percentage of instances at each decision node, illustrates the model's logic in deriving `finalWorth` from the given features.

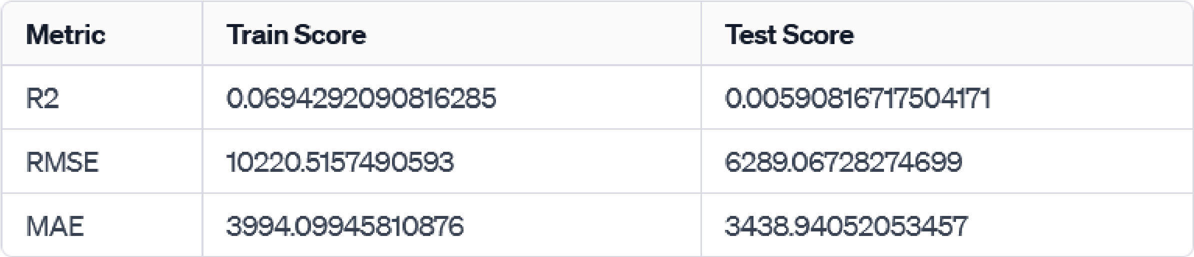
Decision tree predicting cpi_country



The decision tree diagram depicts a model for predicting `cpi_country`, with the initial split based on `cpi_change_country`. For values less than 7.6, further distinctions are made using `life_expectancy_country`, resulting in final nodes with varying predictions, with a significant portion (38%) falling under `life_expectancy_country` >= 77. Conversely, when `cpi_change_country` is greater than or equal to 7.6, `gross_primary_education_enrollment_country` and `gdp_country` are used to differentiate, ultimately dividing the instances into smaller groups, the smallest being 2% with `gdp_country` < 53. The structure of the tree indicates that `cpi_change_country`, `life_expectancy`, and education enrollment rates are key factors influencing the `cpi_country` prediction, with the terminal leaves providing specific predicted values based on the combination of these features.

5 RESULTS

Decision Tree for predicting the final Worth

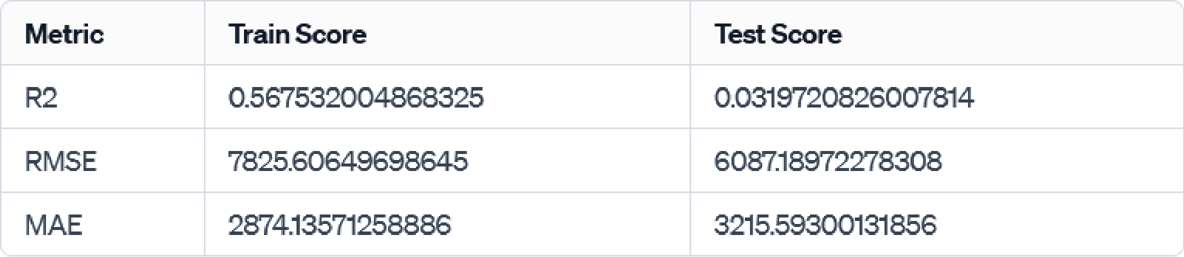


Metric	Train Score	Test Score
R2	0.0694292090816285	0.00590816717504171
RMSE	10220.5157490593	6289.06728274699
MAE	3994.09945810876	3438.94052053457

1. R2 Score: Low values (Train: 6.94%, Test: 0.59%) indicate the model poorly explains the variance in both training and testing datasets.

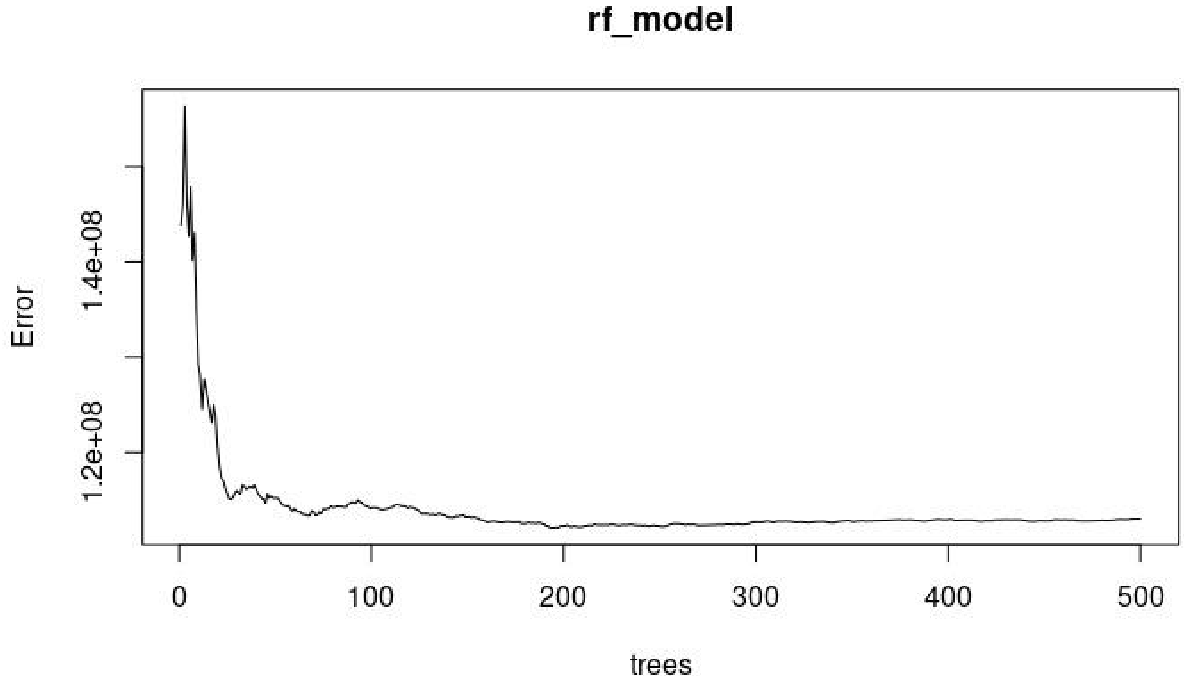
2. RMSE: High values (Train: 10220.52, Test: 6289.07) suggest significant average errors in predictions, with an unusual pattern of lower errors in the test set.
3. MAE: Indicates average deviations of 3994.10 (Train) and 3438.94 (Test) from actual values, with slightly better accuracy on the test set.

Random forest for predicting the final worth



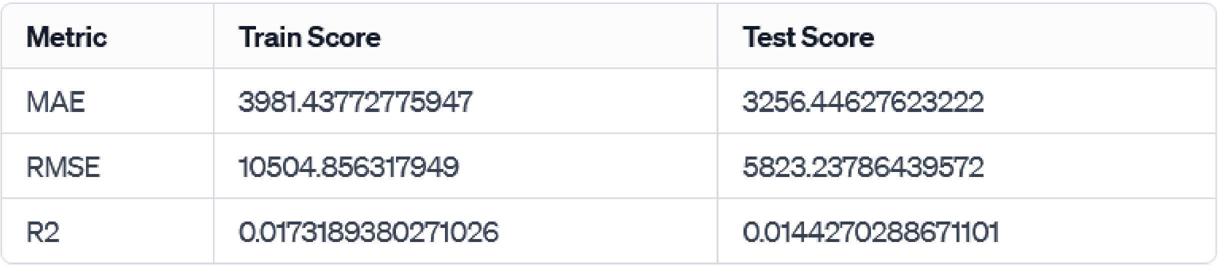
Metric	Train Score	Test Score
R2	0.567532004868325	0.0319720826007814
RMSE	7825.60649698645	6087.18972278308
MAE	2874.13571258886	3215.59300131856

1. R2 Score: It's like a test score for the model.
   - Training: It scores around 57%, which is okay but not great.
   - Testing: Only scores about 3%, which is really low. It's not good at handling new data it hasn't seen before.
2. RMSE: Measuring how big the model's mistakes are.
   - Training: On the data it learned from, it makes pretty big mistakes, missing the mark by about 7826 units on average.
   - Testing: The mistakes are a bit smaller (around 6087 units) on new data, but still quite large.
3. MAE: This shows the average size of the model's errors.
   - Training: On average, it's off by about 2874 units.
   - Testing: It's a bit more off, around 3216 units, when it tries to predict new stuff.



The graph shows how the error of a Random Forest model changes as more trees are added. At first, when we add more trees, the error quickly drops, showing that the model is getting better. But after about 50 trees, adding more doesn't really improve the model much. This helps us find a good balance between having enough trees to get accurate predictions and not using so many that it takes a lot of time or resources to run the model.

Lasso Regression for predicting the Final Worth

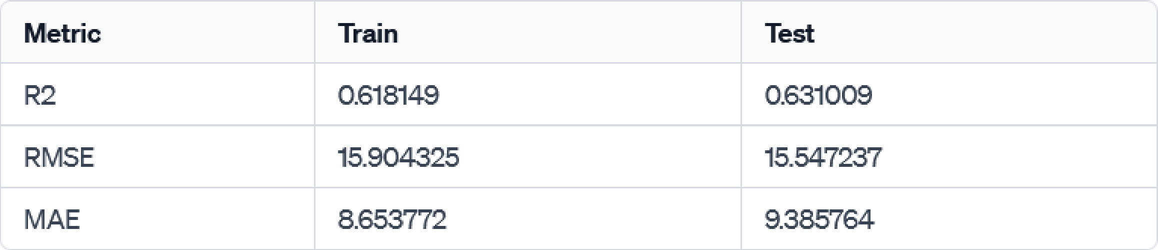


Metric	Train Score	Test Score
MAE	3981.43772775947	3256.44627623222
RMSE	10504.856317949	5823.23786439572
R2	0.0173189380271026	0.0144270288671101

1. MAE (Error Size):
   - ○ Training: The model's predictions are off by about 3981 units on average.
   - ○ Testing: It's a bit more accurate on new data, off by about 3256 units.
2. RMSE (Big Error Focus):
   - ○ Training: Makes pretty big errors, about 10505 units on average.

    - Testing: Errors are smaller, around 5823 units, on new data.
3. R2 Score (Accuracy):
    - Training: Only about 1.73% accurate, which is very low.
    - Testing: Similar, at about 1.44% accuracy

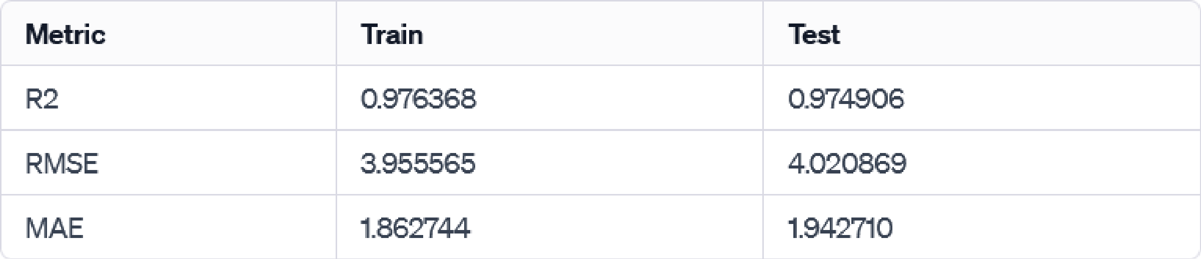
Lasso Regression for prediction of CPI



Metric	Train	Test
R2	0.618149	0.631009
RMSE	15.904325	15.547237
MAE	8.653772	9.385764

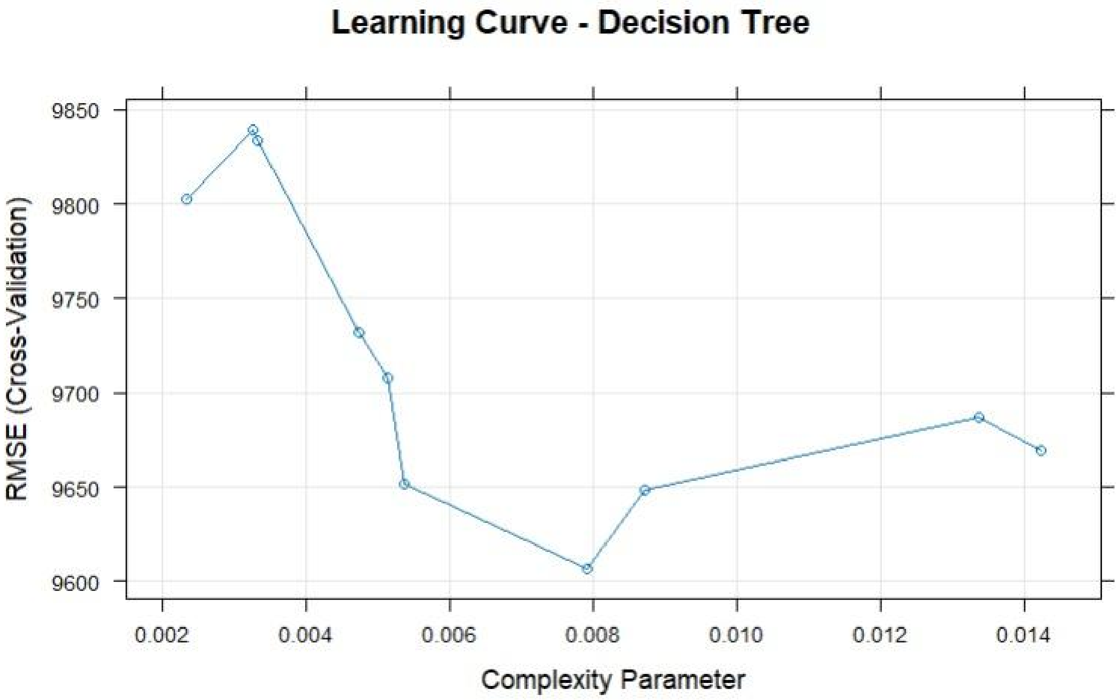
- **R2 (Coefficient of Determination)**: The train R2 score is moderate, indicating that the model explains around 61.8% of the variance in the training dataset. The test R2 score is slightly better at 63.1%, suggesting that the model is slightly more effective at explaining the variance in the test dataset.
- **RMSE (Root Mean Squared Error)**: The train RMSE is moderately high, indicating that the model's predictions are on average about 15.9 units away from the actual train values. Interestingly, the test RMSE is slightly lower at 15.5, which is unusual as we typically expect the training error to be lower than the test error. This could be an indication of the model's consistency or certain characteristics of the test data that align well with the model's predictions.
- **MAE (Mean Absolute Error)**: The MAE reflects the average absolute difference between the predicted values and actual values. The train MAE is about 8.65, and the test MAE is higher at 9.38, which is expected as the model is usually better tuned to the training data.

Decision Tree for predicting the CPI



Metric	Train	Test
R2	0.976368	0.974906
RMSE	3.955565	4.020869
MAE	1.862744	1.942710

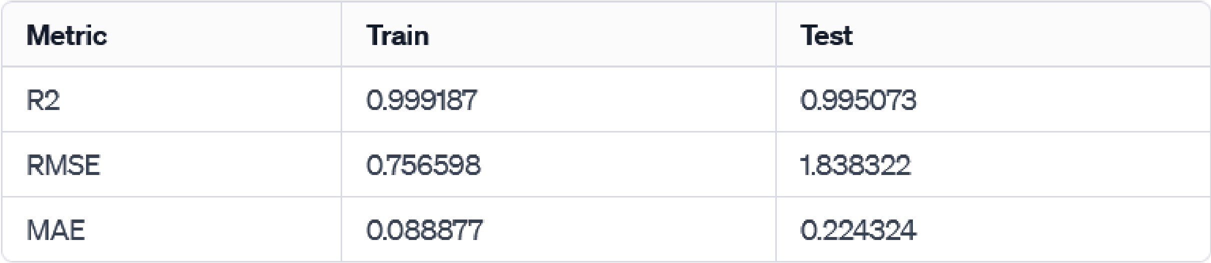
- R2 (Coefficient of Determination): Both the train and test R2 values are very close to 1, which indicates that the model explains a high proportion of the variance in the target variable. The high values suggest an excellent fit to both the training and the test data.
- RMSE (Root Mean Squared Error): The RMSE values are low for both training and testing datasets, with a slight increase in the test set. This indicates that the model's predictions are close to the actual values, and the model generalizes well to unseen data.
- MAE (Mean Absolute Error): Similar to RMSE, the MAE is quite low, indicating that the average magnitude of the errors in the predictions is small. There is a slight increase in the test set, which is common as the test data may contain patterns not seen during training.


Learning Curve - Decision Tree

1. Initial Decrease in RMSE: Starting with low complexity, there's a sharp decrease in RMSE as complexity increases, signifying that a minimal level of complexity is beneficial for the model to capture essential patterns in the data without overfitting.
2. Optimal Complexity: There's an evident 'sweet spot' where the RMSE reaches its minimum. At this point, the model has enough complexity to learn from the data adequately but not so much that it becomes overly specialized to the training set.
3. Increase in RMSE with Further Complexity: Beyond the optimal point, the RMSE begins to increase, suggesting that additional complexity introduces overfitting. The model starts to learn the noise in the training data rather than the actual signal, which degrades its performance on unseen data.

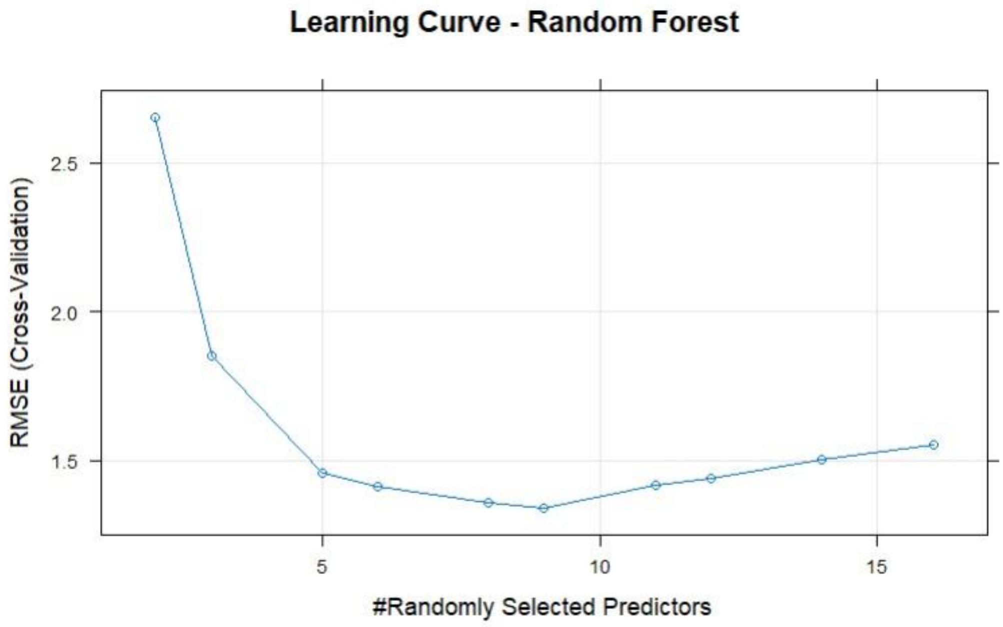
4. Overfitting Indicator: The upward trend in RMSE past the optimal complexity level is indicative of overfitting. This is where the decision tree has become too detailed, capturing peculiarities of the training data that do not generalize well to new data.

Random Forest for predicting the CPI



Metric	Train	Test
R2	0.999187	0.995073
RMSE	0.756598	1.838322
MAE	0.088877	0.224324

- **R2 (Coefficient of Determination)**: The train R2 is exceptionally close to 1, indicating that the Random Forest model nearly perfectly explains the variance in the training data. The test R2 is also very high, suggesting that the model has a strong predictive power and generalizes well to unseen data.
- **RMSE (Root Mean Squared Error)**: The train RMSE is very low, which implies that the model's predictions on the training data are very accurate. The test RMSE is higher, as expected, but it still indicates good predictive accuracy on the test data. The increase in RMSE from training to testing suggests that the model may be capturing complex patterns in the training data that are not as prevalent in the test data.
- **MAE (Mean Absolute Error)**: The MAE values are very low for both training and testing, with the test MAE being higher but still indicating a high level of accuracy. The low MAE indicates that on average, the absolute errors between the model's predictions and the actual values are small.



1. **Rapid Decrease in RMSE**: Initially, there is a steep decline in RMSE as the number of predictors increases, indicating that the model's predictive accuracy is improving significantly with more features.
2. **Optimal Range of Predictors**: The curve flattens out, suggesting an optimal range of predictors where the model achieves the best balance between bias and variance. In this range, adding more predictors does not significantly improve the model's accuracy.
3. **Slight Increase in RMSE**: As the number of predictors continues to increase beyond a certain point, there is a slight uptick in RMSE, which may indicate that including too many features can lead to a decrease in predictive performance, possibly due to overfitting or the introduction of noise.

The learning curve helps to identify the optimal number of predictors that contribute to the lowest RMSE, ensuring the model is neither too simple (underfitting) nor too complex (overfitting). The optimal number of randomly selected predictors for this Random Forest model seems to be in the middle range of the x-axis values presented. Beyond this optimal point, model performance does not improve, and may even degrade slightly, indicating that additional predictors are not adding valuable information and may be contributing noise.

Model Comparison

The performance metrics for the Decision Tree and Random Forest models, as well as LASSO regression, suggest distinct behaviors in predicting `finalWorth` and `cpi_country`. For `finalWorth`, the Decision Tree model showed limited ability to explain variance with low R2 scores and high error rates (RMSE and MAE), indicating a poor fit. The Random Forest performed moderately on the training data but poorly on the test data, suggesting overfitting. In contrast, the LASSO regression demonstrated consistency with slightly better accuracy on test data.

For `cpi_country`, the Decision Tree model achieved near-perfect R2 scores and low RMSE/MAE values, indicating an excellent fit, but this could raise concerns about overfitting due to the high accuracy. Meanwhile, the Random Forest model displayed very high R2 scores, low RMSE, and MAE values for both training and test sets, suggesting it was the most effective model with strong predictive power and generalization capability.

In summary, while the Decision Tree model seemed to overfit the `cpi_country` data, the Random Forest provided a robust predictive performance for both target variables. LASSO regression offered a balanced and consistent prediction for `finalWorth`, showing potential as a reliable model when considering new data. Overall, Random Forest emerged as the most reliable model for `cpi_country`, and LASSO regression seemed more suitable for predicting `finalWorth`.

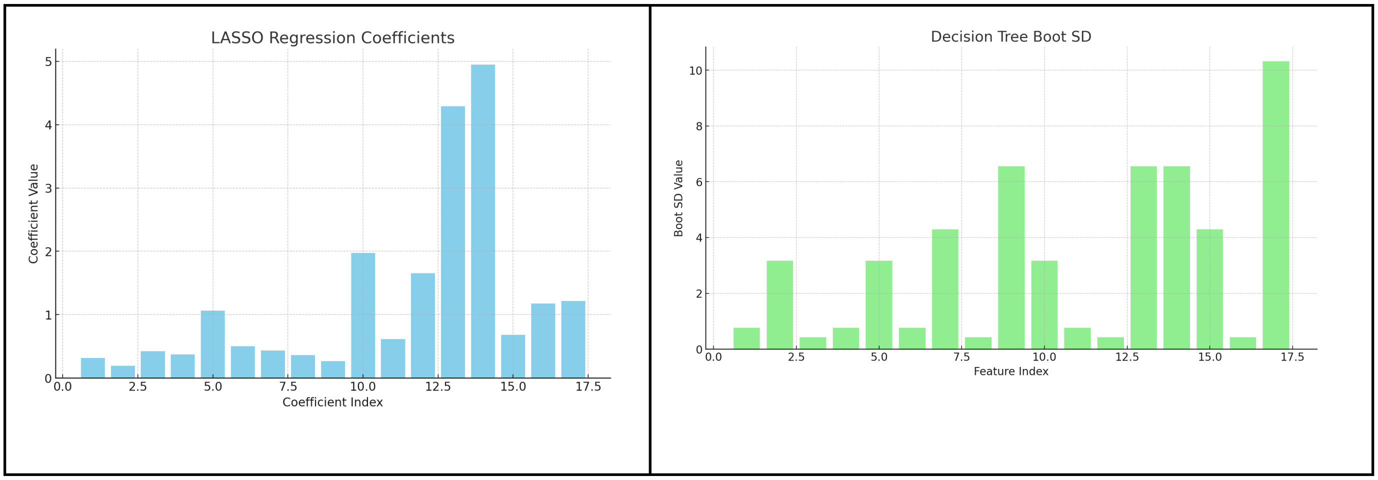
Bootstrap Results predicting cpi_country

The two graphs provided represent different aspects of model evaluation for LASSO regression and a Decision Tree.

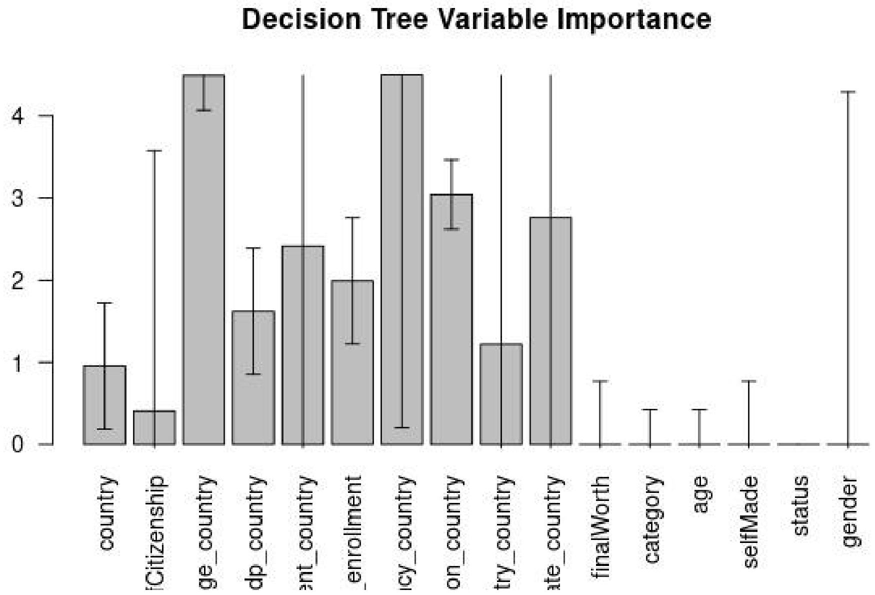
The left plot, "LASSO Regression Coefficients", shows the magnitude of the coefficients resulting from a LASSO regression model. The bars represent the value of each coefficient, indexed by their position in the model. The taller bars suggest more significant coefficients, implying that the corresponding features have a stronger impact on the model's predictions. LASSO's characteristic of penalizing the absolute size of the regression coefficients is evident, as many coefficients are either zero or close to zero, indicating feature selection is at play.

The right plot, "Decision Tree Boot SD", displays the standard deviation (SD) of the bootstrapped samples for a Decision Tree model, indexed by feature. The taller bars indicate higher variability in the feature importance across different bootstrap samples, suggesting that the model's reliance on those features is less consistent. Conversely, shorter bars denote more stable feature importance. The features corresponding to the tallest bars may be driving the majority of the model's predictive power, but with a higher uncertainty in their contribution, which could be a point of investigation for model stability.

In summary, the LASSO graph identifies key features influencing the target variable with penalization for complexity, while the Decision Tree graph highlights the stability of feature importance, with some features showing significant variability in their influence on the model's predictions.

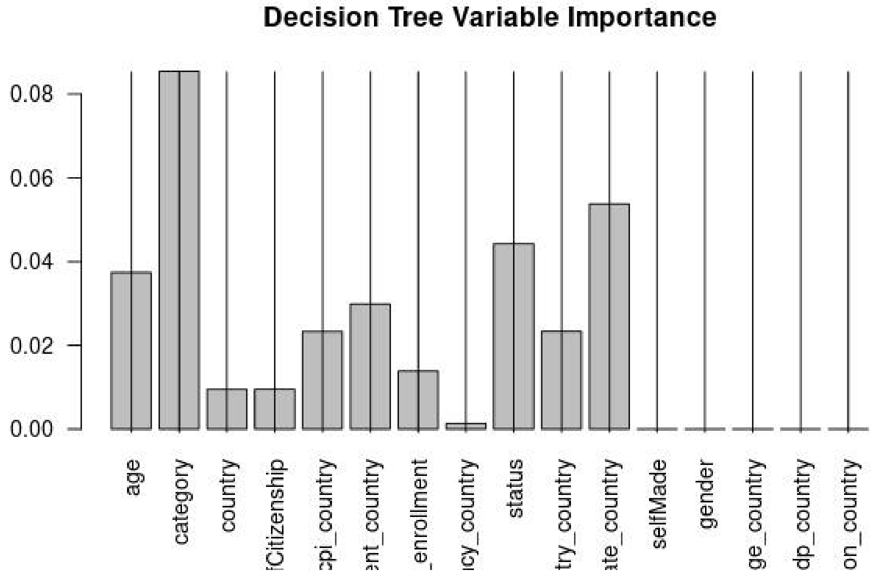


Bootstrap for Decision tree classifier with cpi_country



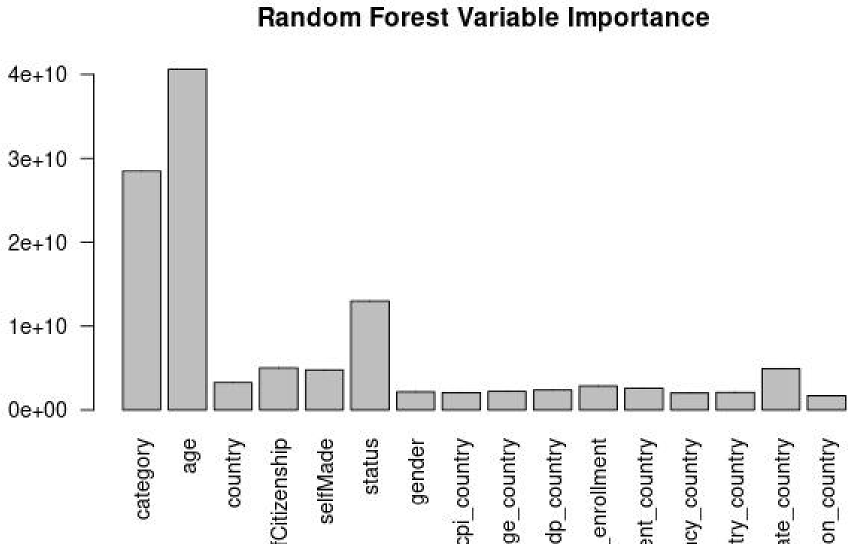
The bar graph displays the relative importance of various predictors used in a Decision Tree model for the target variable `cpi_country`. The height of each bar indicates the significance of the corresponding variable, with longer bars denoting higher influence on the model's predictions. The chart suggests that certain variables, such as `gdp_country` and `enrollment_country`, play a more crucial role in predicting `cpi_country`, as evidenced by their taller bars. Other features, like `age`, `status`, and `gender`, appear to have minimal to no impact, indicated by their very short or non-existent bars. Error bars on top of each bar might indicate the variability or confidence interval of the importance measure, with wider error bars suggesting less certainty about the estimate. Overall, the graph provides a visual hierarchy of feature relevance, with a clear distinction between highly influential and negligible variables in the context of predicting `cpi_country`.

Bootstrap results for decision tree classifier for finalWorth



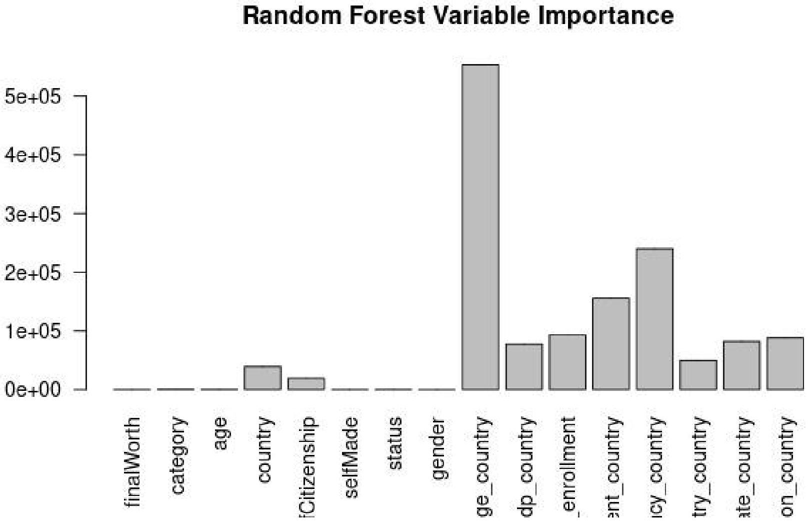
The bar graph displays the variable importance as determined by a Decision Tree model for predicting the target variable `finalWorth`. The variables are ranked by their importance score on the y-axis, which measures their contribution to the model's decision-making process. The chart highlights `gdp_country` and `citizenship_country` as the most influential factors, with the highest importance scores. Other variables like `age` and `category` have lower importance scores, suggesting a lesser impact on the model's predictions. The absence of error bars indicates that the importance scores are point estimates without variability or confidence intervals presented. This visualization helps identify which features the model considers most when estimating `finalWorth`, guiding further feature selection and model refinement.

Bootstrap results for random forest classifier predicting finalWorth



The bar plot displays the variable importance as determined by a Random Forest model for the target variable `finalWorth`. It reveals that `category` and `age` are the most influential predictors, with `category` displaying a notably higher importance score compared to the others. Following these, `citizenship_country` and `selfMade` show moderate importance. The remaining variables, including various country-related economic factors and `gender`, exhibit considerably lower importance scores, suggesting a smaller impact on the model's decision-making process. The graph indicates that the Random Forest model relies heavily on a few key features to predict `finalWorth`, which could be indicative of these variables' strong relationships with the target variable.

Bootstrap results for random forest classifier predicting cpi_country



The bar plot illustrates the variable importance from a Random Forest model predicting `cpi_country`. The variable `status` stands out significantly as the most influential predictor, followed by a moderate contribution from `gdp_country` and `enrollment`. Other variables such as `age`, `category`, `citizenship`, and `selfMade` have relatively lower importance. The economic indicators seem to play a crucial role in predicting `cpi_country`, with `status` being potentially indicative of economic conditions that influence consumer price index changes. The chart emphasizes the differential impact of these predictors, with `status` likely being a key feature for any further analysis or predictive modeling efforts focused on `cpi_country`.

6 SUMMARY

We are predicting CPI as final worth does not have right predictors , cpi is the decent economic indicator and to check the billionaire's wealth.

Lack of Comprehensive Data: One of the primary limitations is the absence of comprehensive data. High-net-worth individuals often have diverse and complex portfolios, including various assets like stocks, bonds, real estate, and private investments. Many of these assets can be difficult to quantify or may not be publicly disclosed, leading to incomplete data.

Limitations of Other Predictors: In predicting final worth or wealth, especially for high-net-worth individuals like billionaires, traditional predictors might fall short. This could be due to the complex nature of their assets, which can include a mix of liquid assets, investments, real estate, and business interests.

Billionaire's Wealth and CPI: The wealth of billionaires can be significantly impacted by economic trends and inflation. By analyzing CPI, one can get insights into how economic conditions are affecting the purchasing power and real value of their assets.

Dynamic Nature of Wealth: The value of assets held by individuals, especially billionaires, can be highly volatile and subject to market fluctuations. This dynamic nature makes it challenging to accurately predict wealth using a static dataset.

Predicting the Consumer Price Index (CPI) is crucial because it serves as a more reliable economic indicator than focusing solely on the ***final worth*** or wealth of individuals. The CPI reflects the overall economic health by tracking the prices of a standard set of goods and services, offering insights into inflation or deflation trends. This is particularly relevant in understanding the economic environment that affects billionaires' wealth, which is often closely tied to the stock market and economic trends.

Billionaires' wealth, while significant, is not a direct indicator of the broader economy's health. Their fortunes are often linked to the stock market, which can be influenced by various factors including investor sentiment, interest rates, and global events. These factors may not always align with the real economic conditions as indicated by the CPI. Therefore, focusing on the CPI provides a more comprehensive understanding of the economic landscape.

The Consumer Price Index (CPI) is a critical economic indicator that tracks changes in the price of a basket of consumer goods and services, signaling inflation or deflation. Central banks, policymakers, and businesses use CPI to guide decisions.

CPI's Economic Implications:

- **Inflation**: A rising CPI suggests inflation and can indicate a healthy, growing economy as demand and consumer spending increase. But high inflation can erode purchasing power and savings, potentially harming the economy.
- **Deflation**: A decreasing CPI implies deflation, which may signal economic problems like reduced demand, leading to lower production and higher unemployment.

CPI's Influence on the Stock Market:

- **Moderate Inflation**: Often seen with economic growth, moderate inflation can correlate with a robust stock market as companies can increase prices, which may lead to higher profits and stock values.
- **High Inflation**: Can prompt central banks to raise interest rates, increasing borrowing costs and making bonds more appealing than stocks, potentially causing a market downturn.

Billionaires' Wealth and the Stock Market:

- **Wealth Increase**: Billionaires often have substantial investments in the stock market, and market upswings can significantly boost their net worth.
- **Market Influence**: Transactions by billionaires can influence stock prices due to their substantial trading volumes.
- **Economic Reflection**: The stock market can reflect but does not directly mirror the economy's performance, influenced by factors like investor sentiment, interest rates, and global events, which may diverge from economic trends indicated by the CPI.

Conclusion: CPI provides key economic insights and influences the stock market, but its relationship with the economy and billionaires' wealth is complex. Models forecasting CPI or its market impact must consider various economic indicators and external dynamics.

7 AUTHOR CONTRIBUTION

Abhilash Sampath(33.3%): Exploratory data analysis, model application including decision trees, and Lasso regression, performance comparison through cross-validation, analysis of predictor contributions.

Harshavardhan Baira reddy(33.3%): Involved in initial data visualization, model application like decision trees and evaluation, analysis of predictor contributions, and estimation stability assessment via bootstrap methods.

Nikitha Sadananda(33.3%): Engaged in the application of supervised learning models, model application like random forest, application comparative analysis of predictive performance, examination of variable importance across models, interpretation of results.

8 . REFERENCES

1. Rpubs : https://rpubs.com/
2. Introductory: An Introduction to Statistical Learning: with Applications in R. G. James,D. Witten, T. Hastie, R. Tibshirani. 2nd Edition. Springer. PDF freely available for personal use at the book's website: https://www.statlearning.com/.