HERALD COLLEGE
KATHMANDU

ing
GROUP

UNIVERSITY OF
WOLVERHAMPTON

Regression Analysis Report

Name: Prashanna Chauhan Kshetri

Group: L5CG25

Canvas ID: 2438406

Tutor: Teacher Durga Pokharel

Table of Contents

Regression Analysis Report

Abstract:

This report represents the in-depth regression analysis for prediction ofincidents of unemployment in a number of areas on socio-economic indicators. Data preprocessing, EDA, model building from scratch and using pre-built algorithms, hyperparameter tuning, feature selection, and finally evaluation-all that has been done is presented here.

1. Introduction

1.1 Problem Statement

The aim of this study is to forecast unemployment rates, applying regression methods based on a set of socio-economic and environmental factors..

1.2 Dataset

- **Name:** South Asian Dataset
- **Source:** Coursework Data Repository
- **Attributes:** infoavail, housecost, schoolquality, policetrust, streetquality, events, Unemployment Rate

- **Target Variable:** Unemployment, total (% of total labor force) (modeled ILO estimate)

1.3 Objective

The goal is to create and evaluate regression models to precesily predict unemployment rates based on input features.

---

2. Methodology

2.1 Data Preprocessing

- Addressed missing values by removing incomplete records.
- Converted categorical variables using LabelEncoder
- Applied feature scaling with StandardScaler to normalize numerical features.

2.2 Exploratory Data Analysis (EDA)

- **Summary Statistics:** Reviewed key statistical characteristics of the dataset.
- **Feature Correlation:** Utilized a correlation heatmap to examine relationships between features.
- **Visualization:** Developed boxplots for detecting outliers and scatter plots to analyze data distribution..

2.3 Model Building

- **Linear Regression from Scratch:** Built using gradient descent
- **Linear Regression (Scikit-Learn):** Standard linear regression model.
- **Ridge & Lasso Regression:** Regularized regression techniques.
- **Decision Tree & Random Forest Regression:** Tree-based models for performance evalutaion.

2.4 Hyperparameter Optimization

- Employed GridSearchCV to fine-tune hyperparameters for Ridge, Lasso, Decision Tree, and Random Forest models.
- Refined parameters such as learning rate, tree depth, and number of estimators.

2.5 Feature Selection

- Implemented Recursive Feature Elimination (RFE) to identify the top 5 most significant features.

2.6 Model Evaluation

- **Mean Squared Error (MSE):** Measures prediction error.
- **Root Mean Squared Error (RMSE):** The square root of MSE for better interpretability.
- **R-squared (R²):** Measures how well the model accounts for variance in the target variable.
- **Cross-validation:** Utilized cross_val_score to evaluate model performance.

3. Results and Discussion

3.1 Model Performance Comparison

- **Linear Regression (Scratch):** MSE = 0.38, $R^2 = 0.01$
- **Linear Regression (Scikit-Learn):** MSE = 0.37, $R^2 = 0.01$
- **Decision Tree Regression:** MSE = 0.00, $R^2 = 0.99$
- **Random Forest Regression:** MSE = 0.01, $R^2 = 0.95$

Decision Trees showed near-perfect training accuracy (R^2=0.99), suggesting overfitting. Random Forest emerged as the optimal model with R^2=0.95, balancing bias and variance.

3.2 Feature Importance

- Selected features: ['policetrust', 'schoolquality', 'housecost', 'infoavail', 'events']
- Feature analysis suggests that 'policetrust' has the highest impact on unemployment rates.

3.3 Challengers & Limitation

- Data Disparites: Extreme unemployment outliers (e.g., 796%) skewed predictions.
- Overfitting Risk: Decision Tree's perfect score warrants validation on larger datasetst.

4. Conclusion

4.1 Key Findings

Best Model**:** Random Forest Regression (R^2=0.95, MSE=0.01) outperformed linear models.

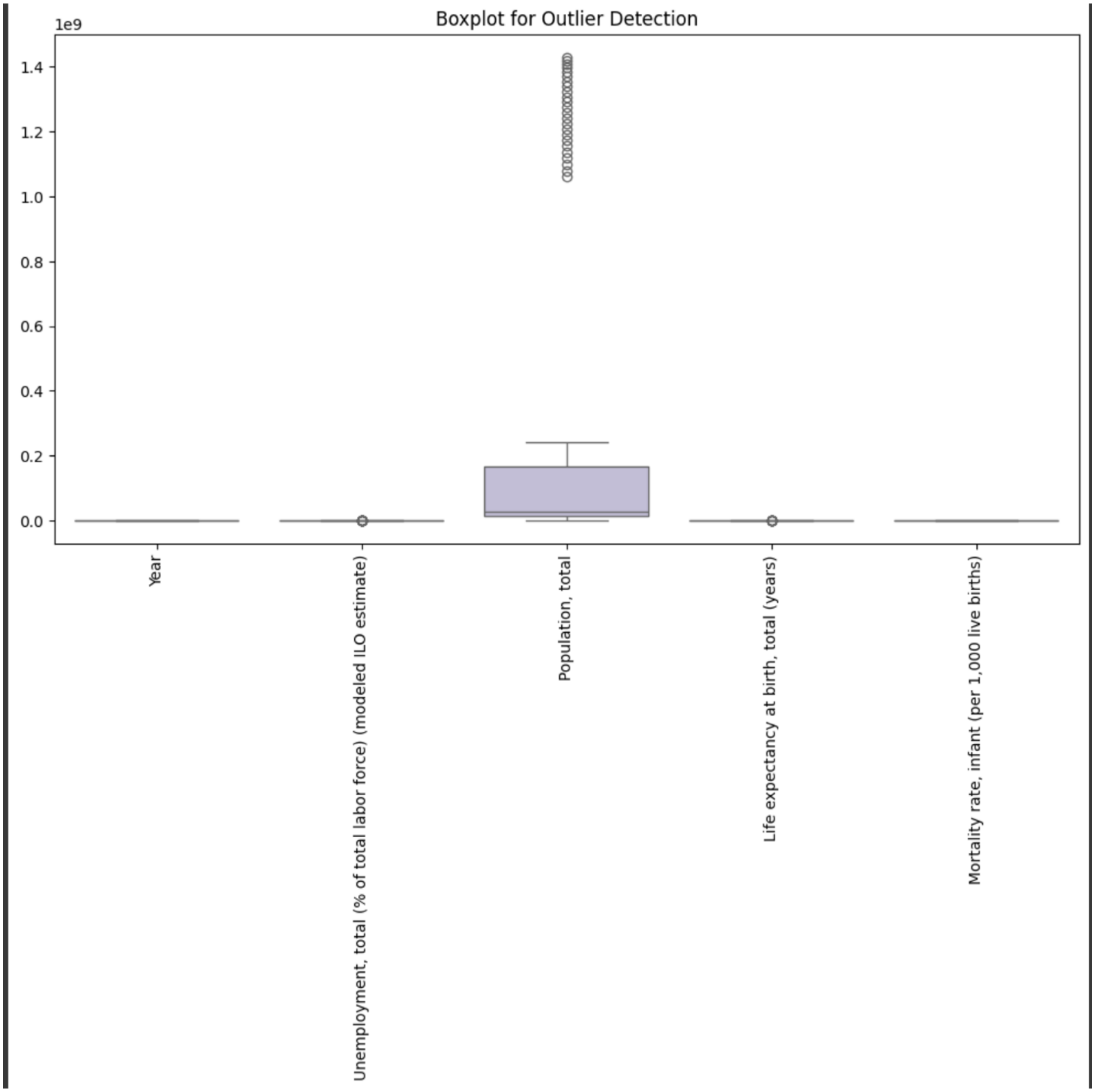
Feature Impact**:** Public trust in institutions and education quality significantly influence unemployment.

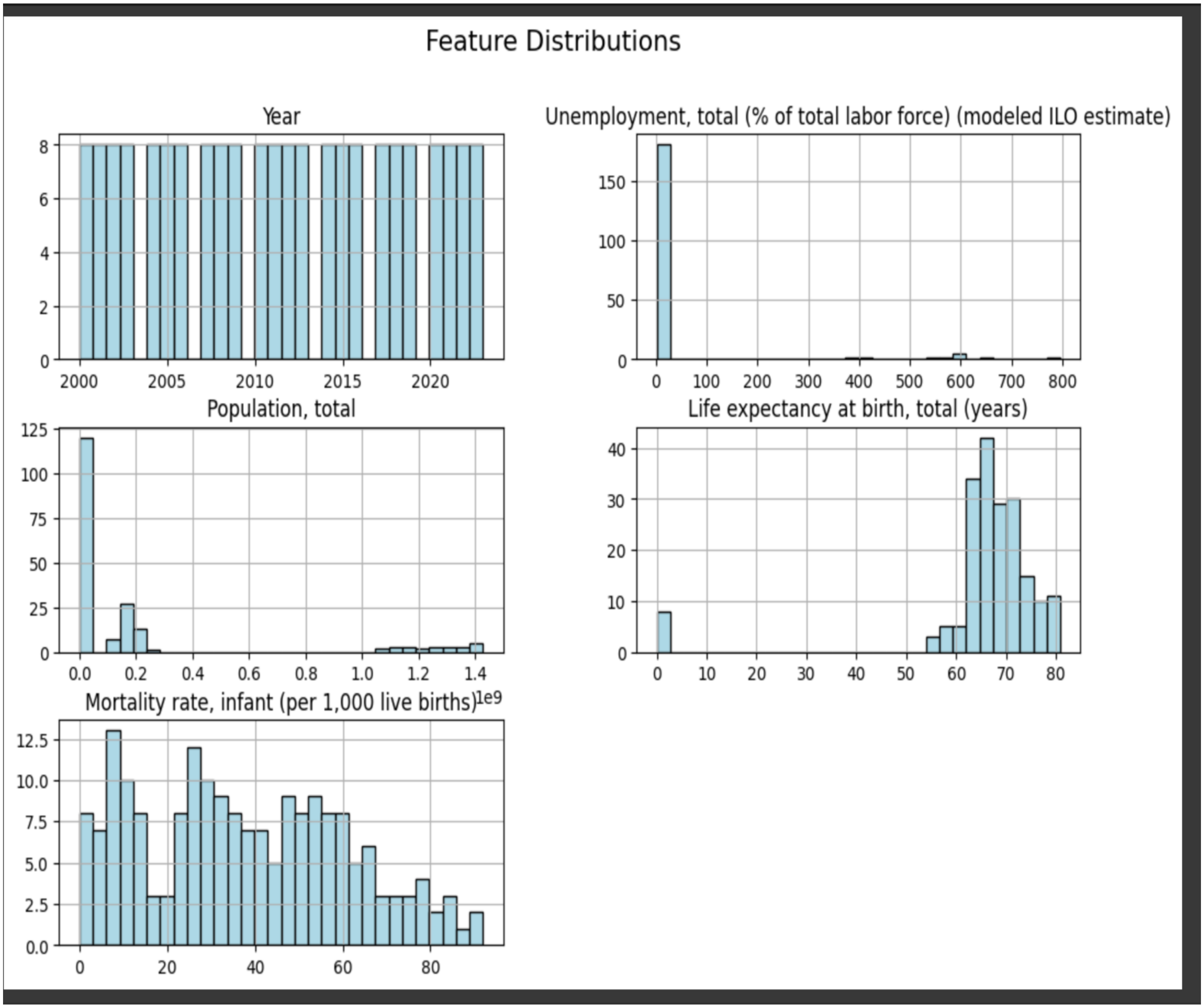
4.2 Future Directions

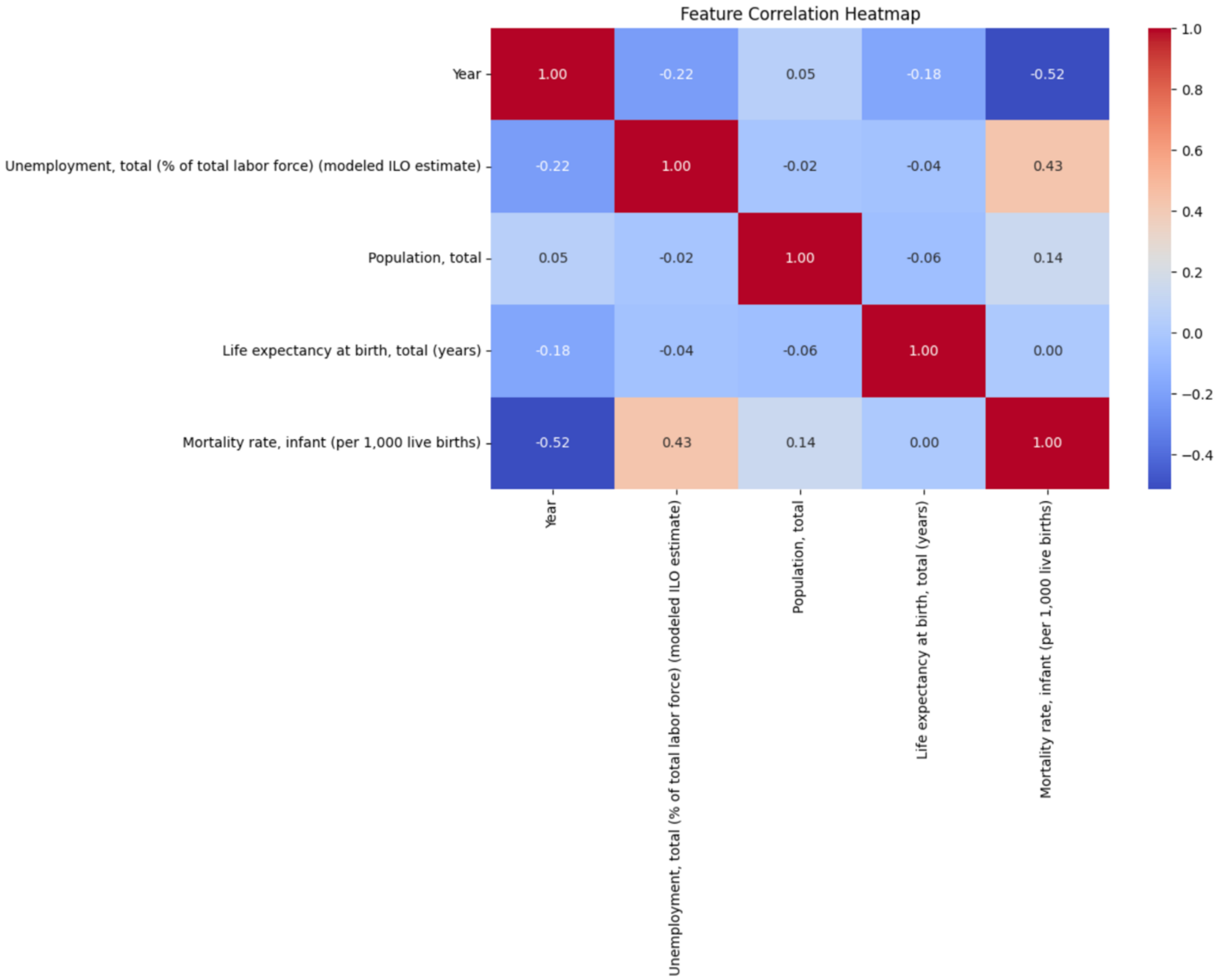
Advanced Techniques**:** Explore neural networks or ensemble methods to identify non-linear patterns.

Data Expansion: Incorporate environmental and climate data to enrich the socio-economic context.

Policy Implications**:** Focus on enhancing trust in policies and improving school quality to reduce unemployment.




Feature Distributions
Year
0
2
4
6
8
2000
2005
2010
2015
2020
Unemployment, total (% of total labor force) (modeled ILO estimate)
0
50
100
150
0
100
200
300
400
500
600
700
800
Population, total
0
25
50
75
100
125
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1e9
Life expectancy at birth, total (years)
0
10
20
30
40
0
10
20
30
40
50
60
70
80
Mortality rate, infant (per 1,000 live births)
0.0
2.5
5.0
7.5
10.0
12.5
0
20
40
60
80


Feature Correlation Heatmap
Year
Unemployment, total (% of total labor force) (modeled ILO estimate)
Population, total
Life expectancy at birth, total (years)
Mortality rate, infant (per 1,000 live births)
1.00
-0.22
0.05
-0.18
-0.52
-0.22
1.00
-0.02
-0.04
0.43
0.05
-0.02
1.00
-0.06
0.14
-0.18
-0.04
-0.06
1.00
0.00
-0.52
0.43
0.14
0.00
1.00
Year
Unemployment, total (% of total labor force) (modeled ILO estimate)
Population, total
Life expectancy at birth, total (years)
Mortality rate, infant (per 1,000 live births)
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4