Report on Predicting School Requirements in Pakistan

FINAL PROJECT
MAHNOOR AWAN
BYTEWISE DATASCIENCE FELLOWSHIP

Abstract

This report addresses the critical issue of educational resource allocation in Pakistan by predicting the number of schools required in each district based on population data and other related factors. Utilizing both Linear Regression and Random Forest models, the study demonstrates that the Random Forest model significantly outperforms Linear Regression in capturing the complexities of the data, providing more accurate predictions. The findings can inform policymakers to better align school distribution with the population's needs.

Introduction

Education plays a pivotal role in the development of any nation, with equitable access to educational resources being a cornerstone of societal growth. In Pakistan, disparities in the distribution of schools across districts can hinder educational access for many communities. This study aims to predict the number of schools required in various districts based on population size, growth rate, and population density. By identifying districts where the number of schools does not match the population's needs, this research seeks to provide actionable insights for improving educational infrastructure.

Problem Statement

Predicting the required number of schools in each district of Pakistan based on the population and current number of schools.

Data Collection

The dataset used in this project is on kaggle:

- Population in 2023** (`population_2023`)
- Population in 2017** (`population_2017`)
- Growth Rate** (`growth_rate`)
- Area of Districts in km^2** (`area_km2`)
- Population Density in 2023** (`density_people_2023_km2`)
- Number of Households** (`households`)
- Average Household Size** (`average_household_size`)
- Total Number of Schools** (`total_schools`)

Methodology

Exploratory Data Analysis (EDA)

- **Data Distribution Analysis**: Understanding the distribution of the population, area, and number of schools across districts.
- **Correlation Analysis**: Examining the relationships between variables like population, growth rate, and school count to identify key predictors.
- **Visualization**: Using scatter plots, histograms, and correlation matrices to visualize patterns and potential outliers.

- A positive correlation between population size and the number of schools, indicating that more populous districts tend to have more schools.
- A strong correlation between population density and the number of schools, suggesting that densely populated areas may require more educational facilities.

Model Selection and Training

Linear Regression

Linear Regression was used as a baseline model to predict the number of schools based on population_2023, growth_rate, and density_people_km2. The model provided the following results:

- **Mean Squared Error (MSE):** 469494.95
- **R-squared (R^2):** 0.304
- **Coefficients:** [295.135, -222.059, -131.710]
- **Intercept:** 1055.009

The relatively high MSE and low R^2 indicated that Linear Regression did not adequately capture the complexities of the data, leading to suboptimal predictions.

Random Forest Regression

Given the limitations of Linear Regression, a Random Forest model was employed to better account for non-linear relationships in the data. The model was trained on the same features and yielded significantly better performance:

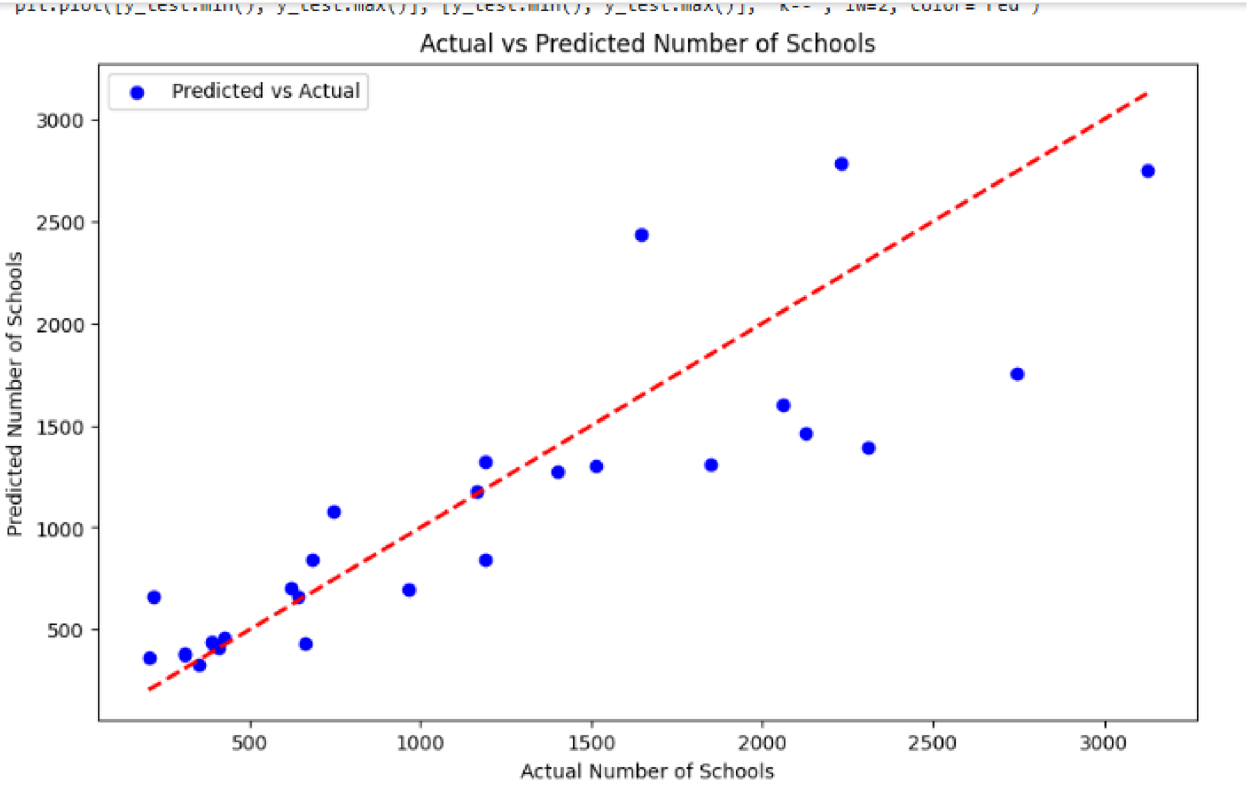
- **Mean Squared Error (MSE):** 162068.29
- **R-squared (R^2):** 0.760

The Random Forest model's lower MSE and higher R^2 demonstrate its superior ability to model the data, resulting in more accurate predictions.

Results

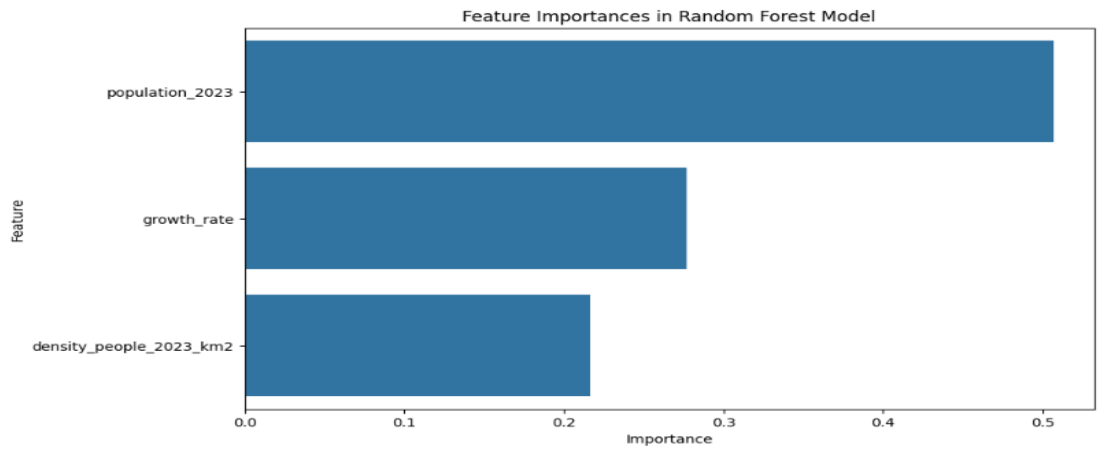
The Random Forest model identified population_2023, growth_rate, and density_people_km2 as the most important features influencing the number of schools required in a district. The model's predictions closely matched the actual number of schools, as indicated by the high R^2 value.

Residual analysis confirmed that the model was well-fitted, with residuals randomly distributed around zero.

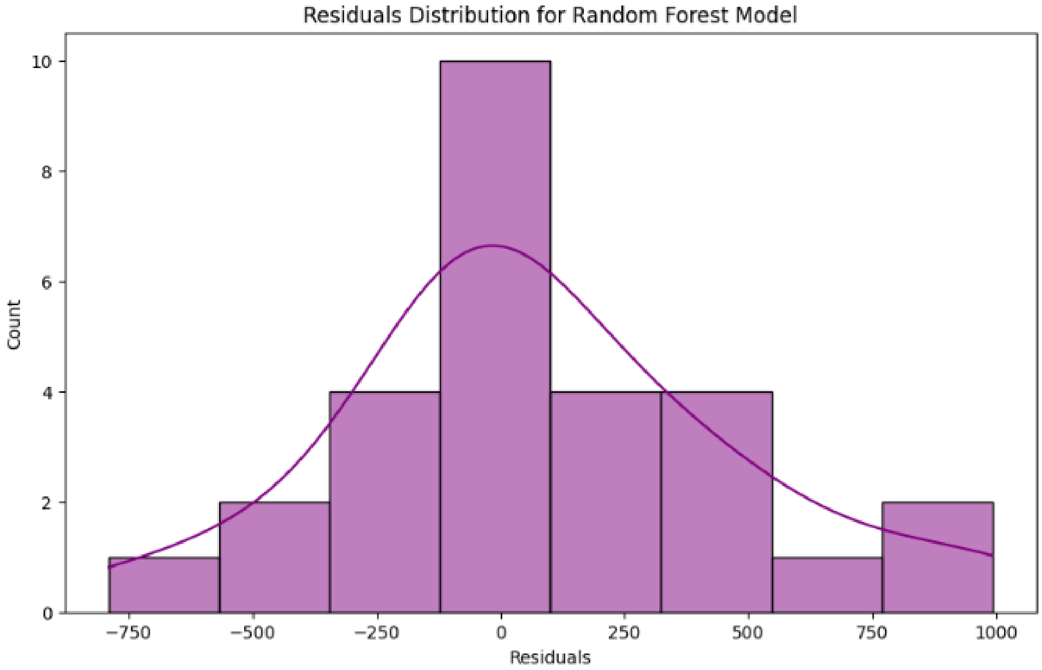


```
plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], 'k--', lw=2, color='red')
```

This figure is a scatter plot comparing the actual number of schools to the predicted number of schools in various districts. The x-axis represents the actual number of schools, while the y-axis represents the predicted number of schools .Each blue dot corresponds to a district, showing where the actual and predicted values align. The red dashed line is the ideal line where predicted values would perfectly match the actual values (y = x).



This figure is a horizontal bar chart displaying the feature importance in a Random Forest model used to predict the number of schools in various districts.



This figure shows the residuals

Conclusion

The study highlights the effectiveness of the Random Forest model in predicting school requirements based on district-level population data in Pakistan. The insights derived from this model can guide policymakers in allocating educational resources more equitably. The comparison with Linear Regression underscores the importance of using more complex models like Random Forest for tasks involving non-linear data.

Recommendations

- **Policy Implications**: Policymakers should prioritize districts with high populations but relatively few schools for additional resources.
- **Future Work**: Incorporating additional variables such as income levels, educational outcomes, and infrastructure quality could further refine the predictions and enhance decision-making.