Lead Testing in NYC Children

Introduction

- This report presents the findings of a big data analysis project on child lead testing in New York City. We aimed to examine public health data through the application of statistical and machine learning methods with the objective of gaining a deeper understanding of patterns and disparities in testing rates. The project was undertaken in three primary phases: data cleaning and preprocessing, data visualization, and regression modeling.

  1. Data Preparation and Preprocessing I began with importing and pre-processing the "Children Tested for Lead by Age 3 Years" dataset from the NYC Open Data Portal. The following pre-processing were done:

    - **Column Renaming:** Long and descriptive column names were shortened for conciseness and better readability of code (for example, "Children tested for lead by age 3 years Number" was changed to "Tested_Num").
    - **Column Dropping:** Meta features such as timestamps, IDs, and notes were dropped to eliminate noise and keep meaningful features.
    - **Missing and Zero** Values: Numerical columns that contained missing values were imputed with their respective column means. Zero values, which were probably missing or incorrect data, were also imputed with means.
    - **Data Types:** Numerical columns were all forced to the correct format. Categorical variables such as "GeoType" were changed to categorical data types.
    - **Encoding:** One-hot encoding was used on the "GeoType" column to ready it for regression modeling.

2. Data Visualization The visualization process revealed patterns and relationships in the data:

- **Skewness Analysis:** Skewness was measured for all numeric variables. There was strong right-skew in variables such as "Tested_Num."
- **Correlation Heatmaps:** Both Pearson and Spearman heatmaps were used to discover linear and non-linear correlations between variables. Spearman, specifically, revealed monotonic trends that Pearson did not catch.
- **Histograms, Boxplots, and Scatter Plots:** These plots gave information about the distribution and spread of each variable, presented outliers, and indicated relationships between predictors and response variable (Tested_Pct).

- **Transformations:** To correct skewness, log, square root, and Yeo-Johnson transformations were used. Yeo-Johnson worked best for normalization and was utilized for modeling.

These plots informed preprocessing steps required and directed feature selection for modeling.

3. Regression Analysis For the regression phase, I picked the following as features:

- Tested_Num (numerical)
- Year (numerical)
- GeoType_Citywide (categorical, one-hot encoded)

Target variable was Tested_Pct. We developed a multiple linear regression model after feature scaling. The early performance was poor, with negative R^2 on the test set, so a simple linear model would not be feasible.

To enhance the model, we used polynomial regression (degree 2) to enable the model to pick up on interaction and nonlinear relationships. This substantially boosted the R^2 scores:

- Training R^2: 0.46
- Testing R^2: 0.19

Evaluation Metrics:

- MAE (Mean Absolute Error)
- MSE (Mean Squared Error)
- RMSE (Root Mean Squared Error)

These metrics confirmed improved performance and reduced prediction error.

4. Communication and Interpretation In order to interpret the model visually:

- A 3D regression surface plot was created based on two features (Tested_Num and GeoType_Citywide).
- This plot illustrated how the model modifies its predictions in response to feature inputs.
- Intercept and coefficient values were interpreted in the model equation.

Key Insights:

- Testing rates vary by geographic type and year.
- The categorical variable "GeoType" made a significant difference in prediction results.
- Nonlinear modeling more accurately represented real-world trends in the data.

Conclusion and Recommendations This discussion illustrates the necessity of proper preprocessing, informative visualization, and suitable model selection for public health data analysis. As recommendations for future studies, we suggest:

Adding more detailed location information (e.g., neighborhood or borough level)

Exploring alternative modeling approaches like decision trees or ensemble methods