Classification of Poverty Levels

By: Crystal Han

Abstract

The goal of this project was to use the Health Professional Shortage Areas (HPSA) dataset from the google cloud big query to analyze and classify county and states with greater poverty levels. This will help the government and other private companies to be able to allocate workers and resources to certain areas to develop and provide. The HPSA Score is analyzed between 2010 and 2019 to classify poverty levels for the most recent decade. I will use Logistic Regression and KNN to model this classification.

Design

This data for this project is provided by Google Cloud Big Query. It was extracted from the Google Cloud Big Query using SQL commands. This presents data of the county, state, population, designation date, proximity to the US-Mexico border, rural code, Metropolitan Indicator Code, HPSA score, and more. From this data, I encoded rural, proximity to the US-Mexico border, and HPSA scores less than 15 with the preprocessing module from sklearn. Supplementary data from the US Census Bureau was used to fill in the missing population data section of the HPSA dataset. The government could allocate resources and energies to help areas that are in need.

Data

The dataset contains Health Professional Shortage Areas data from 2010 to 2019. The data is categorized by County, State, Source ID, Federal Information Processing Standards (FIPS), Designation Date, Estimated served population, Rural code, Metropolitan Indicator code, HPSA Score, and proximity to the US-Mexico Border. Some of these features can be used to create new features such as year and label encoding with numbers. Supplementary data was provided from the census bureau to fill in the missing population data for counties and states. The data can be used to visualize the relation between score, population, and more. An in-depth analysis of this was undertaken to help visualize HPSA score data and population. Logistic Regression and KNN were used to classify the remaining data as in poverty or not.

Algorithms

Feature Engineering

1. Plotting pair plots, scatter plots, and bar charts to observe the relation between some features and HPSA score
2. Using the date to extract the Year and use it to match up with the census bureau data based on county and state name
3. Label encoding proximity to US-Mexico Border, Rural status code, and HPSA scores less than 15 using sklearn preprocessing
4. Fixing outliers or missing data that the supplementary dataset or HPSA dataset could not fill (as there were only a few it did not affect overall number of datapoints)

Models

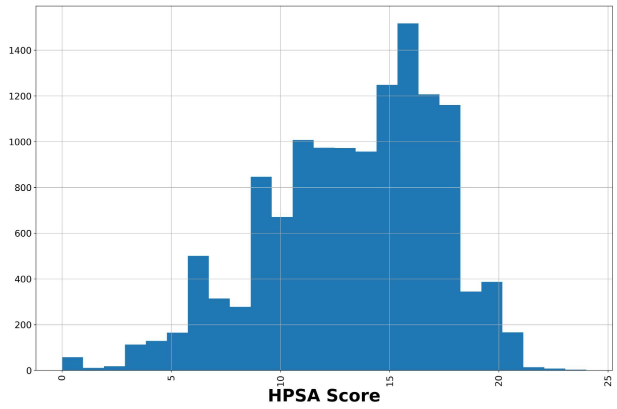
Histograms, scatter plots, and bar plots were used to observe population and HPSA Score. Bar plots were used to visualize HPSA data against population and other features but this did not show much to take from. So scatter plots were used to observe the relationship between population and HPSA score. Then a histogram was done on the HPSA score to observe this feature.

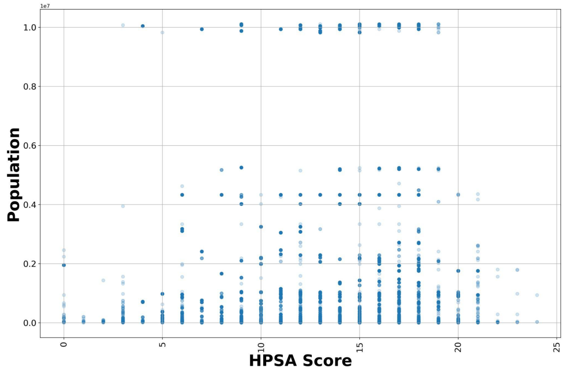
Logistic Regression and KNN were used for classification and regression. Scores for accuracy, precision, recall, and f1 score to compare against each other. KNN showed that it had higher accuracy and f1 score than Logistic Regression.

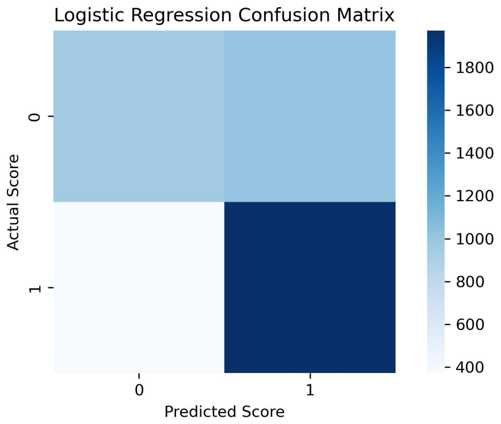
Tools

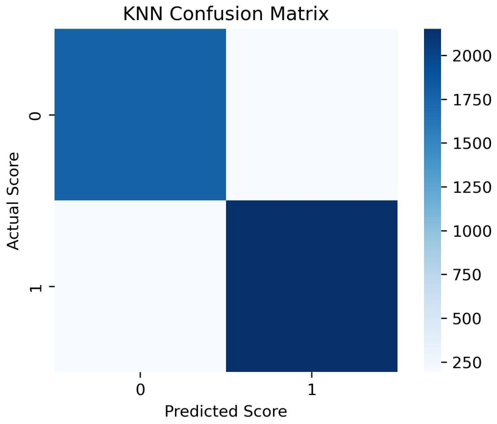
- NumPy and Pandas for data manipulation
- Matplotlib and Seaborn for plotting and visualizations
- SQLAlchemy to write the python dataframe into a database and read from database
- Scikit-learn for modeling, classification, and regression

Communication









Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.6785	0.6899	0.6630	0.6596
KNN	0.9094	0.9087	0.9086	0.9086