PROJECT 03 - HOUSEHOLD POVERTY

Author - Max Lan

Objective

To predict whether a household falls under federal and state poverty levels based on census data.

Business Case

Demographic information on household poverty can help identify households at risk of financial hardship or better target features related to poverty for governmental policy purposes. Further analysis can also provide understanding of factors impacting poverty over time and how to preemptively identify risk for lenders.

Data

Anonymized US census microdata on the household and person level will be sourced from the Integrated Public Use Microdata Series (IPUMS) located at IPUMS.org. As the data available includes multiple years and millions of observations, the key focus of data cleaning will be the following:

- Understanding and selecting metrics provided
- Adjusting the data for bias related to census and survey data
- Formatting or grouping the data using SQL.

Model and Metrics

The project will focus on classification methods to determine whether a household falls under specific annual income levels as defined by federal and state regulations. At a minimum, K-Nearest Neighbors and Logistic Regression models will be evaluated for their predicted accuracy. Probability thresholds and metrics, including accuracy, precision, recall, and AUS will be examined for relevance. The author has initially considered the risk lies more with imprecise predictions (guessing poverty incorrectly) than with low recall (not capturing all impoverished households in the nation).

Features

Potential features will be identified throughout the iterative project process. An initial listing of features includes, but is not limited to:

- Locational features: state, city, metropolitan area
- Educational features: level of education received, educational field
- Occupational features: occupation, industry, commute hours, Hauser and Warren Socioeconomic Index
- Demographic features: Age, home-ownership, internet availability