Income Qualification

Course-end Project 2

Problem Statement Scenario:

Many social programs have a hard time ensuring that the right people are given enough aid. It’s tricky when a program focuses on the poorest segment of the population. This segment of the population can’t provide the necessary income and expense records to prove that they qualify.

In Latin America, a popular method called Proxy Means Test (PMT) uses an algorithm to verify income qualification. With PMT, agencies use a model that considers a family’s observable household attributes like the material of their walls and ceiling or the assets found in their homes to classify them and predict their level of need.

While this is an improvement, accuracy remains a problem as the region’s population grows and poverty declines.

The Inter-American Development Bank (IDB)believes that new methods beyond traditional econometrics, based on a dataset of Costa Rican household characteristics, might help improve PMT’s performance.

Summary:

This write-up presents a machine learning project focused on predicting poverty levels in Costa Rican households. The objective of the project is to develop a model that can accurately classify households into different poverty categories, enabling targeted interventions and assistance programs to uplift vulnerable communities.

Data Exploration:

We started with a data preprocessing and exploratory data analysis phase to gain a deeper understanding of the dataset. The dataset was examined for missing values, outliers, and data distributions. The outliers in the target variable were not altered as it was a part of the project classification. The missing values in variables like ‘v18q1', ‘v2a1'were analyzed and handled. Also some variables with inappropriate data like 'tipovivi3', 'v18q','rez_esc' etc were removed from the dataset.

Feature Engineering:

During the feature engineering stage, relevant features were selected and engineered to enhance the predictive power of the model. Techniques such as one-hot encoding in the categorical data, feature scaling, were applied to improve the quality and efficiency of the input data.

Model Development:

Multiple machine learning algorithms are considered to build a predictive model for poverty level classification. Algorithms such as decision tree algorithm, KNeighbours algorithm, random forests classifier along with bagging classifiers(with and without replacement) methods were explored to identify the most effective approach for this specific problem.

Model Evaluation and Validation:

The performance of the developed models is assessed using appropriate evaluation metrics, including accuracy, precision, recall, and F1-score. Cross-validation techniques, such as k-fold cross-validation, are employed to ensure the models' reliability and generalization capability.

Model Selection:

Based on the evaluation results, Randomforest Classifier with replacement was identified as the most accurate and robust model.

The final model can be trained on the entire dataset to maximize its predictive capabilities. The model is saved and prepared for real-world predictions on new, unseen data.

Conclusion:

This project helped me to leverage machine learning techniques to predict poverty levels in Costa Rican households in the given dataset. By accurately identifying households at risk of poverty, policymakers and aid organizations can develop targeted interventions and policies to alleviate poverty and improve the well-being of vulnerable communities. The insights gained from this project can contribute to the design and implementation of effective poverty alleviation strategies in Costa Rica and potentially in similar contexts worldwide.