Report: Predicting Gross Domestic Household Income

Author: - Taritro Ghoshal

1. Introduction

This report provides an in-depth overview of the methodology applied to predict the gross domestic household income using diverse socio-economic factors. Leveraging extensive data analysis, preprocessing, feature engineering, and cutting-edge machine learning techniques, I unveiled valuable patterns and correlations within the data.

2. Approach

2.1 Data Preprocessing and Feature Engineering

The foundation for any reliable predictive model is rooted in its data quality and the features it utilizes. Here's a breakdown of the preprocessing and engineering efforts:

1. **Ratio Calculations**: By determining the male-to-female ratio, I obtained a nuanced understanding of the gender demographics in the dataset, which can sometimes correlate with income variations.
2. **Age Group Proportions**: I processed age-specific columns to yield proportions of each age group against the total population, revealing potential age-related trends in income.
3. **Educational & Employment Metrics**: Several educational columns were amalgamated to provide broader categorizations. This helped delineate the role of education in influencing household incomes. Additionally, the inferred economic strength combined homeownership and higher-level qualifications, providing a multi-dimensional understanding of wealth and economic stability.
4. **Housing Metrics**: The dataset's housing-specific columns were translated into proportions against the total households. This was pivotal in understanding housing trends and their correlation with household income levels. An intriguing metric that emerged was the ratio of single-person households to family households, shedding light on household composition's role in income determination.
5. **Industry Proportions**: By evaluating different industries' influence on the overall population, I could gauge dominant sectors and their potential correlation with household income.

2.2 Model Selection and Evaluation

A diverse set of models was employed, ensuring a broad exploration of potential algorithms. Our ensemble approach was aimed at leveraging the strengths of individual models to improve overall prediction accuracy:

- **Baseline Models** (RandomForestRegressor, LGBMRegressor, XGBRegressor, ExtraTreesRegressor, DecisionTreeRegressor, KNeighborsRegressor, CatBoostRegressor, Ridge, ElasticNet, GradientBoostingRegressor) – These models were chosen as baseline models to estimate a baseline performance to improve upon. Tree based models were primarily chosen due to their resilience to outliers and scale and ability to learn complex relationships.

- **Gradient Boosting, CatBoost, XGBoost :** These models were chosen for their ability to handle complex non-linear data relationships. They offer robustness against overfitting and can automatically handle missing values.
- **Ensemble Techniques - Stacking and Voting Regressors:** These methods were used to amalgamate individual model predictions, aiming for better accuracy and reduced variance.

2.3 Hyperparameter Optimization

Hyperparameters play a pivotal role in defining a model's performance. Instead of traditional methods like grid search, I employed Bayesian Optimization for its efficiency in hyperparameter tuning. This approach uses prior results to make intelligent decisions about which hyperparameters to try next.

3. Key Findings

1. **Correlation Insights**: A heatmap provided an initial understanding of feature relationships. Notably, several proportion features displayed high inter-correlation, suggesting potential redundancy. This insight drove the decision to perform feature selection during model training.
2. **Dominant Sectors**: The industry, particularly professional, science, and technical sectors, had a pronounced impact on household income. This emphasizes the role of these industries in driving regional incomes.
3. **Education's Role**: The educational spectrum, notably the high-to-no-qualification ratio, turned out to be a key determinant of household income. It underscores education's importance in economic upliftment.
4. **Efficacy of Feature Engineering**: Newly crafted features, such as industry proportions for finance, insurance, and property sectors, and the combined metric of economic strength, emerged as powerful predictors, vindicating our feature engineering efforts.

4. Model Performance

1. **Stacking Regressor**: This ensemble model, integrating GBR, XGBoost, and CatBoost, achieved an average error score of approximately 0.1463 across all folds, indicating its comprehensive predictive capability.
2. **Voting Regressor:** Leveraging a consensus-based mechanism, this model combined predictions from the three aforementioned models. It exhibited an improved average score of around 0.1190, demonstrating the power of ensemble methods.

5. Visualizations

Following are some of the visualizations along with their explanations to help gain a visual understanding of the data and unlock relevant insights. Refer to the notebook for more similar visualizations.

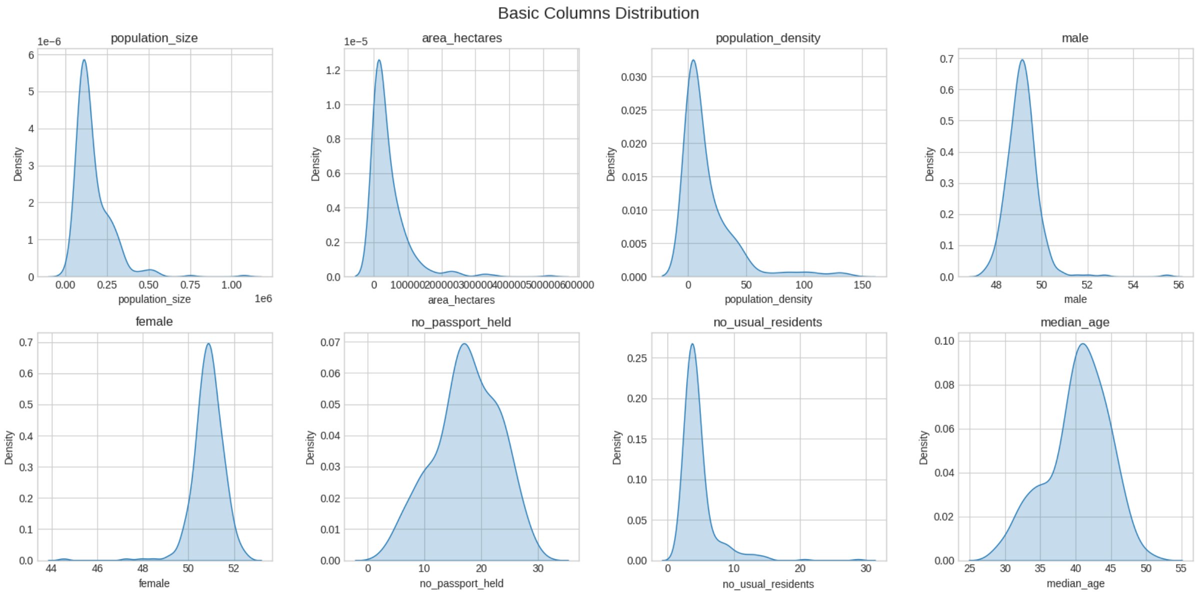

Fig 1: Basic Columns KDE Plots

Features such as population_size, population_density, and area_hectares display a right-skewed distribution. This suggests that a majority of the data points for these features lie towards the lower end, with fewer high-valued outliers. The significant right skewness in area_hectares might hint at a few regions with exceptionally large areas compared to the majority.

On the contrary, columns like median_age and no_passport_held exhibit a more normal or bell-shaped distribution, signifying a balanced spread of data points across their range.

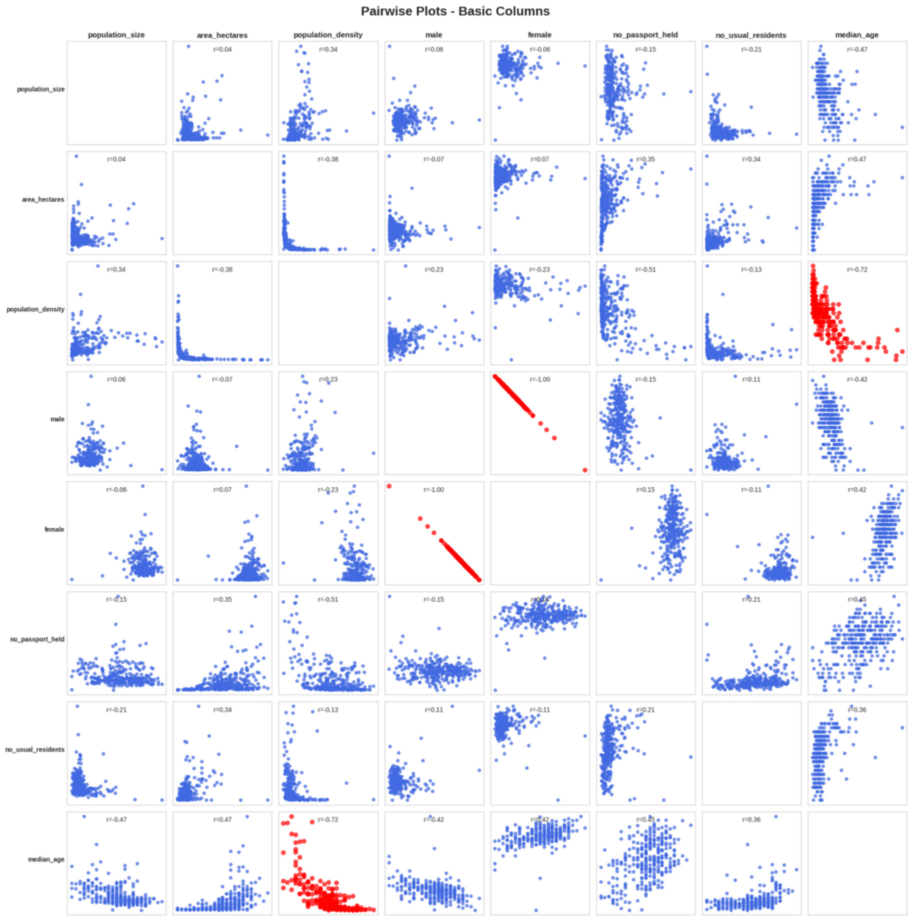

Fig 2: Pairwise Plots – Basic Columns

The pairwise plots for the basic columns provide visual insights into the relationships between different features within the dataset.

1. Population Density and Median Age: There is a noticeable correlation of 0.72 between population density and median age. This suggests that regions with higher population density tend to have a specific median age, possibly implying that certain age groups might be more inclined to live in densely populated areas, perhaps due to factors like job opportunities or urban attractions.
2. Male and Female Correlation: The perfect correlation of 1 between male and female populations is expected. In datasets representing a balanced and broad demographic, as the male population for a specific region increases, so does the female population. This essentially captures the total population of the region and highlights the complementary nature of these two features.

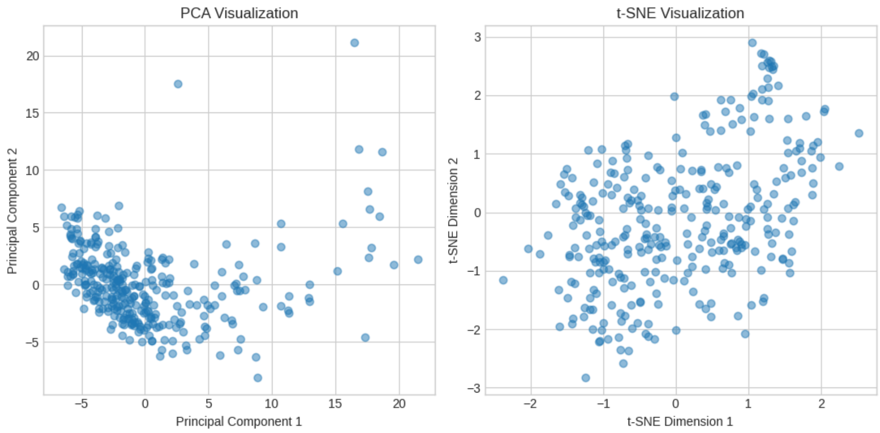

Fig 3: PCA and TSNE Visualizations

The t-SNE and PCA plots give a compressed view of the data. Regions that are closer in this 2D space might have similar characteristics. The lack of distinct patterns in both PCA and t-SNE plots underscores the complexity of our dataset. Such insights imply that relying on simple linear models might not be sufficient to capture the intricacies in the data. Thus, exploring more complex models or ensemble techniques could be beneficial for our predictive tasks.

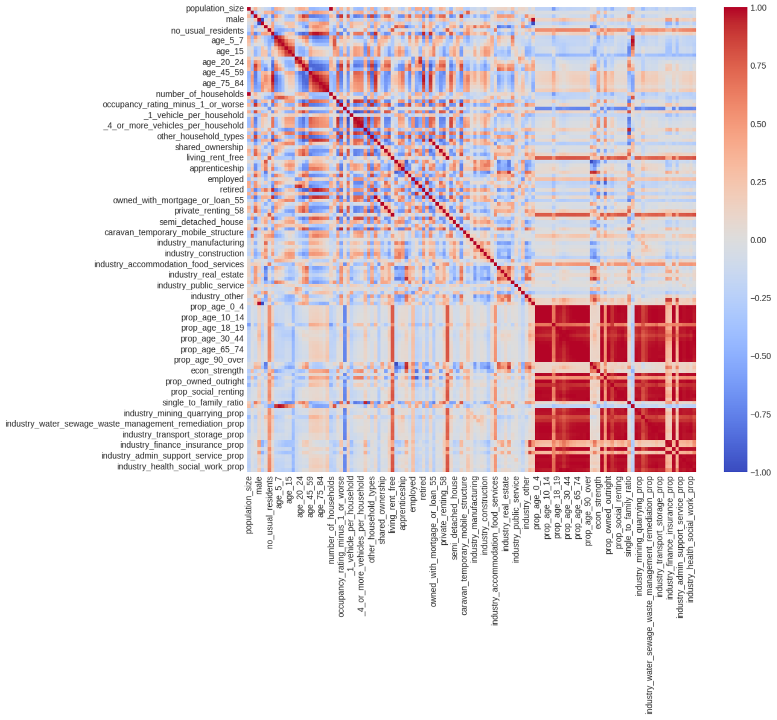

Fig 4: Correlation Heatmap

From our heatmap visualization, it's evident that a good number of proportion features exhibit high correlation with each other. This suggests that some of these features might be providing redundant information. In contrast, many of the other generated features and the standard features don't seem to have a high correlation with each other, implying that they are providing diverse information to our models. It would be a good idea to remove the highly correlated features as a part of feature selection. I am doing this during KFold cross validation during the modelling stage.

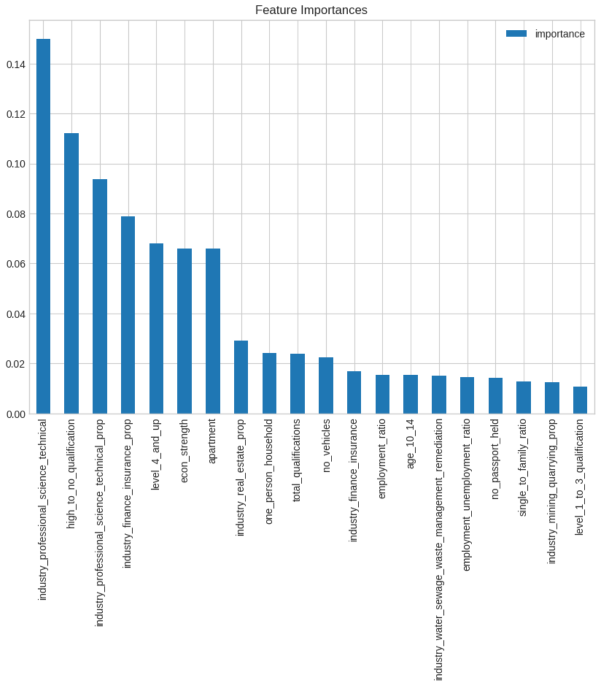

Fig 5: Feature Importance (GBRegressor)

The feature importance visualization derived from the model offers intriguing insights. At the forefront is the industry_professional_science_technical feature, which suggests a notable contribution of the professional, science, and technical sectors to the gross domestic household income. This is followed by high_to_no_qualification, indicating a significant relationship between household income and the educational spectrum, from the highest qualifications down to none.

Interestingly, the feature engineering efforts proved valuable. The generated features like industry_finance_insurance_prop and econ_strength are among the top determinants of the target variable, reinforcing the importance of the finance, insurance, and property sectors, as well as the overall economic strength in predicting household income. Additionally, features such as level4andup and apartment further reflect the role of advanced education and property types in determining income levels.

6. Conclusion

Through meticulous data analysis, feature engineering, and iterative model optimization, I created a robust predictive model for household income. Our systematic approach underscores the value of a comprehensive, iterative analysis in tackling complex prediction challenges.