PREDICTION OF POVERTY: A MACHINE LEARNING APPROACH WITH SOCIOECONOMIC INDICATORS

Tin Hang Johnny, Yiu

Professional Program, Microsoft Corporation

DAT102x: Microsoft Professional Capstone: Data Science

Graeme Malcolm

July 17, 2019

1. EXECUTIVE SUMMARY

1.1. BACKGROUND

This report presents an analysis on data concerning the Poverty Probability Index (PPI) of individuals. By exploring the relationships between characteristics of the individuals and their PPI, we extract useful information that would allow us to make prediction of an individual's probability of poverty based on its socioeconomic indicators.

The dataset retrieved from the PPI website[1] and Financial Inclusion Insights household surveys conducted by InterMedia[2] contains PPI data along with 58 features of 12,600 individuals across 7 different countries. Original sources of the dataset include data from The World Bank[3]. The PPI is a measure to identify an individual's probability of living below the poverty line at the $2.50/day threshold, calculated with answers from 10 questions on a household's characteristics and asset ownership statuses[4].

1.2. METHOD

The first part of the report is an exploratory data analysis, where we calculate summary statistics and create data visualizations of the data that identify potential relationships between individuals' characteristics and their PPI. In the second part, we create a predictive machine learning model that predicts the PPI of individuals, based on the information extracted in the first part.

1.3. RESULTS

In summary, the following features are indicators with a relatively higher coefficient of determination (r-squared) against an individual's PPI (in no particular order) –

- **avg_shock_strength_last_year** – average strength of shocks experienced the past year
- **num_financial_activities_last_year** – number of financial activities conducted the past year
- **country** – unique identifier for country of residence (masked)
- **can_calc_percents** – ability to calculate percentages
- **religion** – unique identifier for religion (masked)
- **is_urban** – residence in the urban area (vs. rural)
- **age_group** – age group
- **female** – sex (true = female, false = male)
- **married** – marital status
- **relationship_to_hh_head** – role in the family

Due to high variability and noise around the data, our machine learning model could only make a prediction on an individual's PPI with a low effect size. The adjusted r-squared for our model is 0.443.

2. EXPLORATORY DATA ANALYSIS

2.1. DATA PRE-PROCESSING

Before calculating summary statistics and creating visualizations for our data, the dataset was prepared into suitable formats for analysis. We have

- removed duplicates;
- dropped uninformative features with largely missing values, they include:
  - *bank_interest_rate;*
  - *mm_interest_rate;*
  - *mfi_interest_rate;*
  - *other_fsp_interest_rate; and*
- substituted missing values for *education_level* and *share_hh_income_provided* with a category of their own.

2.2. EXPLORING PPI

The summary statistics for PPI is shown in the table below.

	Count	Mean	Std	Min	25%	Median	75%	Max
PPI	12600	0.611272	0.291476	0	0.394	0.633	0.879	1

Table 1. Summary Statistics of PPI

It is observed that the median is greater than the mean and that the standard deviation is about half of the mean. The distribution should be left-skewed and has some variance, which is verified by the following histogram.

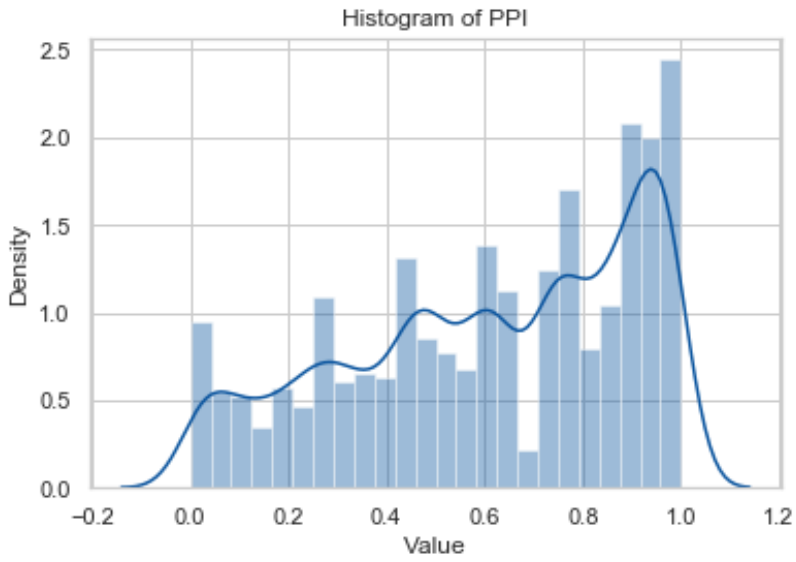


Figure 1. Histogram of PPI

Our PPI histogram shows that there are more people having a higher probability of poverty. The multimodal characteristic entails peaks in 6 different PPIs. To make the distribution more symmetric, transformations are performed. They include natural log (the natural log of 0 was converted into 0), cube root, and square transformations. However, these transformations did not make the distribution more normal. They also did not improve the performance of our machine learning model.

2.3. EXPLORING NUMERIC FEATURES

There are 3 numeric features in the dataset, namely *age*, *avg_shock_strength_last_year* and *num_financial_activities_last_year*. The respective summary statistics and histograms are shown below.

	age	**avg_shock_strength_last_year**	**num_financial_activities_last_year**
Count	12600	12600	12600
Mean	36.28071	2.112765	1.559683
Std	15.14594	2.019239	2.043831
Min	15	0	0
25%	25	0	0
Median	33	2	1
75%	45	4	3
Max	115	5	10

Table 2. Summary Statistics of Numeric Features

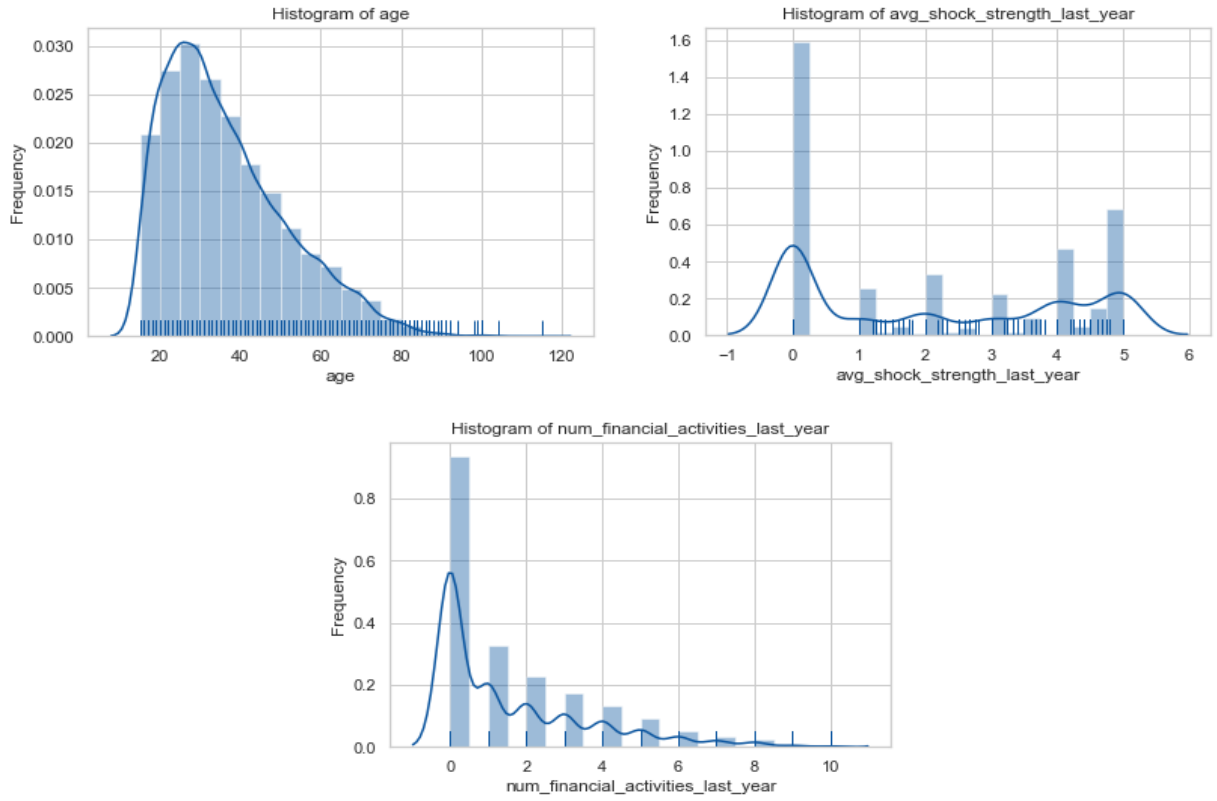


Figure 2. Histogram of Numeric Features

In summary,

- *most respondents are in the age range of 20-40;*
- *the majority of respondents experienced a shock strength of 0 in the past year, and a; significant number of people experienced a 4 or above;*
- *most respondents were not involved in any financial activities last year;*
- *having 6 financial activities or above is uncommon.*

A pair-plot and a correlation matrix of the numeric features are created to study the relationships within features, and between the label and the features.

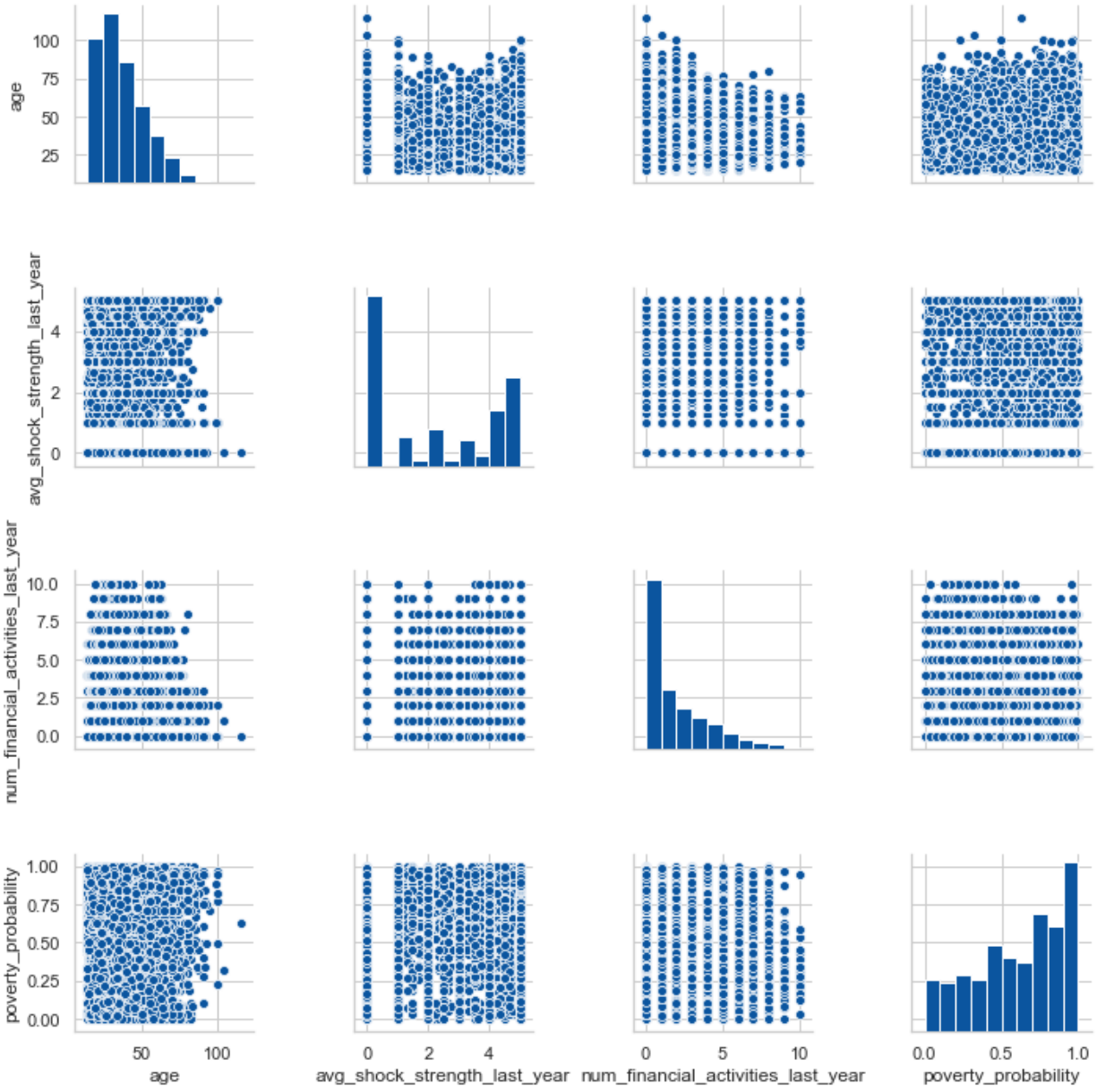


Figure 3. Pair-plot of Numeric Features

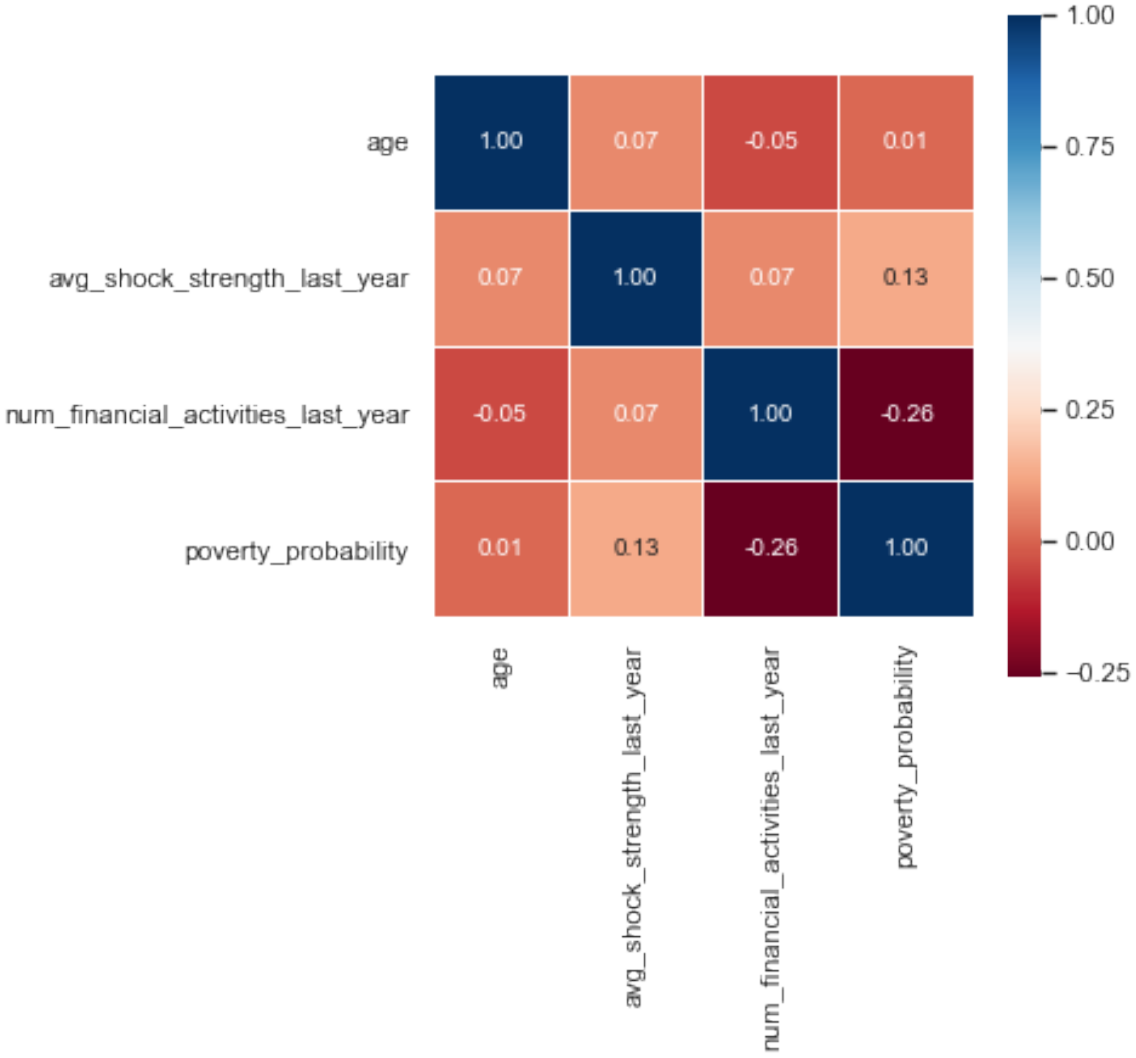


Figure 4. Correlation Matrix of Numeric Features

Little apparent relationships are found, except that

- *on average, the amount of financial activities seems to decrease with age*

2.4. EXPLORING CATEGORICAL FEATURES

After assigning proper discrete variables as categorical features, we ended up with 52 categorical features. Bar charts were used to study the frequency of these features. They indicate the following –

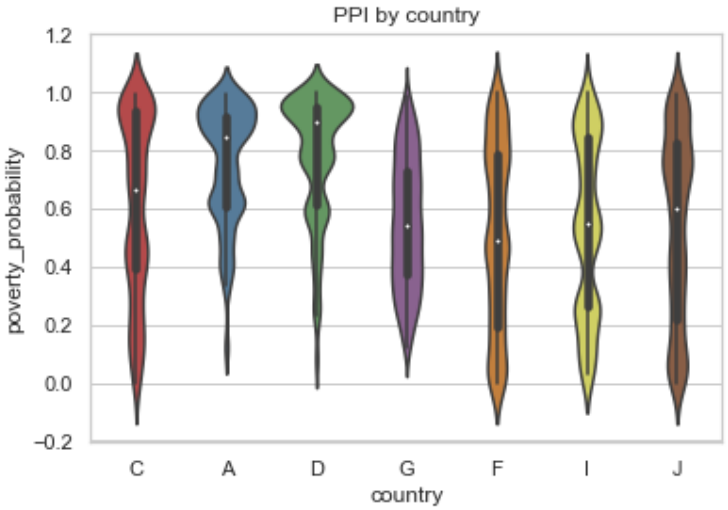
- *86% of the respondents are either in the religion q or x;*
- *over 80% of the respondents can either add or divide;*
- *75% of the respondents are either the head of the family or the spouse;*
- *75% of the respondents do not have the ability to use internet on their phone;*
- *72% of the respondents do not have a bank account under their own name;*
- *70% of the respondents do not have any form of investment;*
- *only 13% of the respondents have insurance;*
- *only 6% of the respondents received income from the government last year;*
- *only 5% of the respondents experienced 4 shocks or above last year;*
- *only 4% of the respondents received income from the public sector last year;*
- *less than 3% of the respondents used services of 3 or more formal financial institutions last year;*
- *only 2% of the respondents used services of 2 or more informal financial institutions last year;*
- *respondents in the religion o or n account for less than 1% of the sample.*

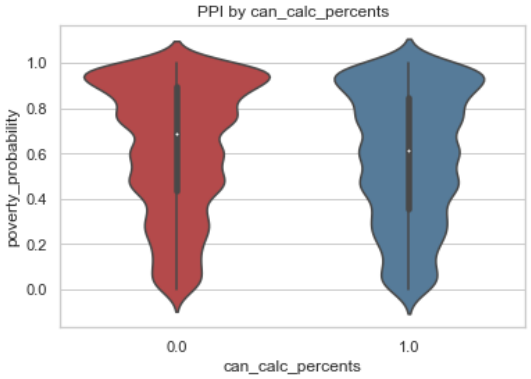
2.4.1. FEATURE ENGINEERING

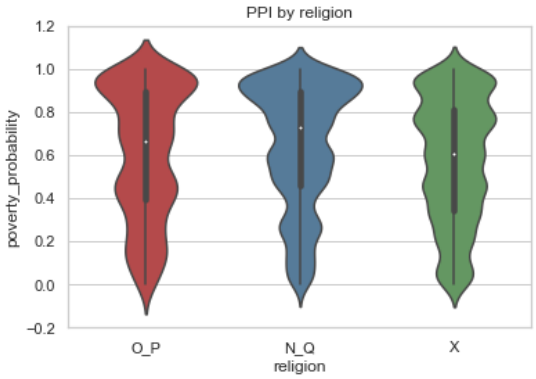
We aggregated categorical features to reduce the number of categories. For discrete variables, rare values are combined to form a range of values. For categorical variables, rare values are combined with common values that share a similar distribution in PPI. For example,

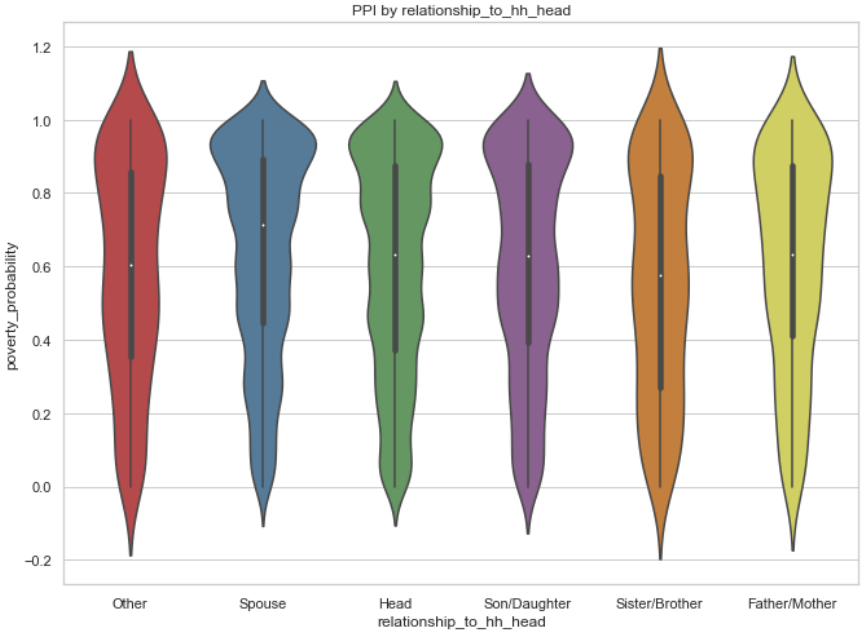
- *religion o is combined with religion p to form religion o_p, and religion n is combined with religion q to form religion 'n_q';*
- *the number of shocks of 4 or above are combined to form '4_5';*
- *the number of formal financial institutions used of 3 or above are combined to form '3_4_5_6';*
- *the number of informal financial institutions used of 2 or above are combined to form '2_3_4';*
- *the number of shocks of 4 or above are combined to form '4_5';*
- *the category 'unknown' for relationship to the head of the household is combined with 'other'.*

Violins plots of the features with a relatively higher r-squared are created to study their relationships with PPI.









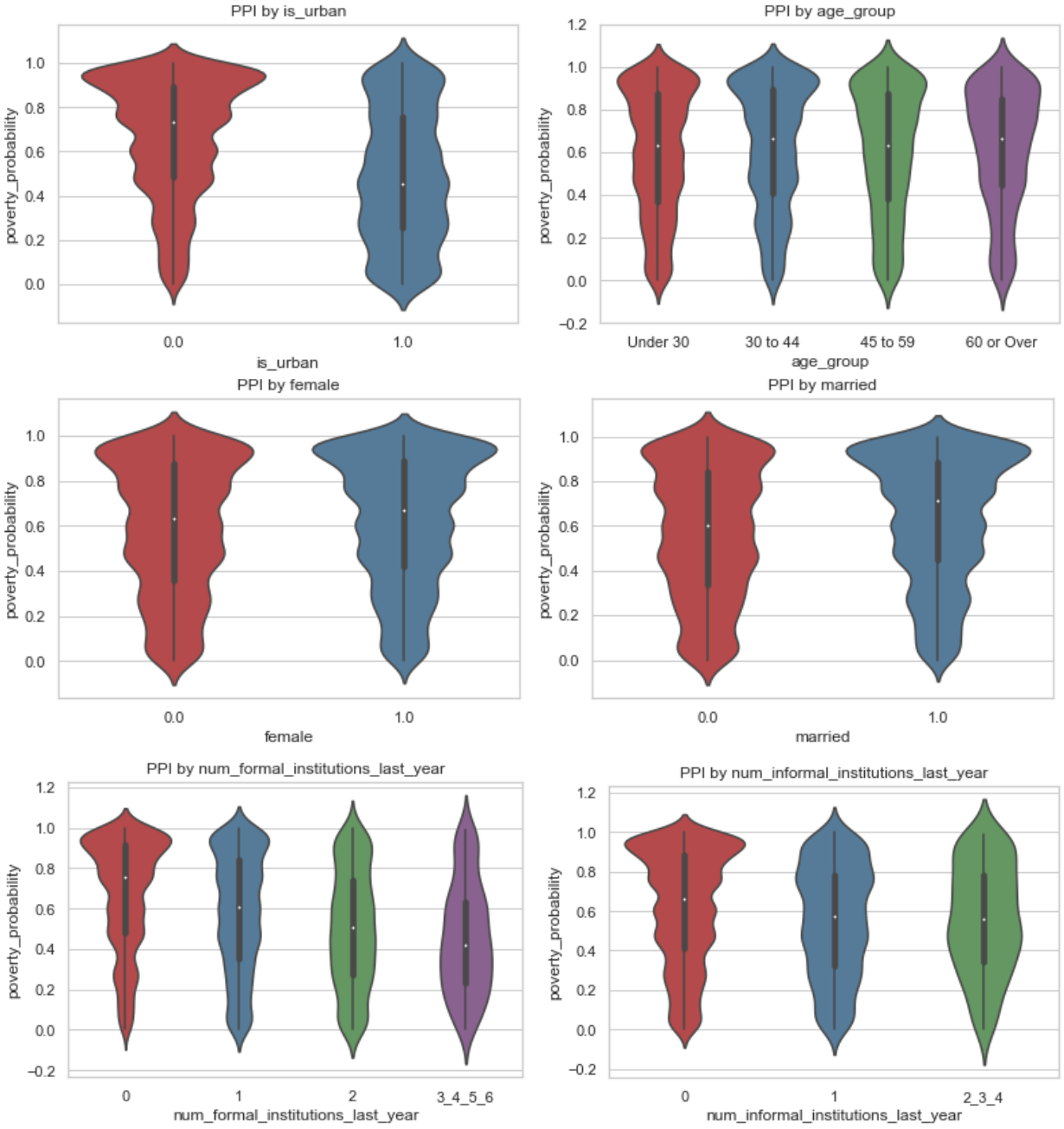


Figure 5. Violin Plots of Categorical Features

They indicate that on average,

- *country D has the highest probability of poverty;*
- *respondents that cannot calculate percentages have a higher probability of poverty;*
- *respondents in the religion N or Q have a higher probability of poverty*
- *spouses have a higher probability of poverty;*
- *rural residents have a much higher probability of poverty;*
- *females have a higher probability of poverty;*
- *married persons have a higher probability of poverty;*
- *respondents that do not use services of a financial institution have a higher probability of poverty.*

After aggregating the categorical features, we eliminated features with low variance. For the variance threshold $Var(x)=p(1-p)$, we have used a $p = 0.95$ and eliminated any features under the threshold. We selected informative features with Recursive Feature Elimination using cross validation, and listed those with the highest feature importance scores in the Executive Summary.

3. POVERTY PREDICTION

Based on the analysis of the probability of poverty, we created a regression model to predict PPI.

We created the model based on a boosted decision trees algorithm and trained with 80% of the data. Testing the model with the remaining 20% of the data yielded the following results.

Mean square error	0.0469
Root mean square error	0.217
Mean absolute error	0.174
Median absolute error	0.147
Adjusted r-squared	**0.443**

Table 3. Prediction Model Test Results

The result of our test shows that the model only explains 44.3% of the variability of the response data around its mean. Due to high variability and noise around the data, our model could only make a prediction on an individual's PPI with a low effect size. Plot on residuals are shown below.

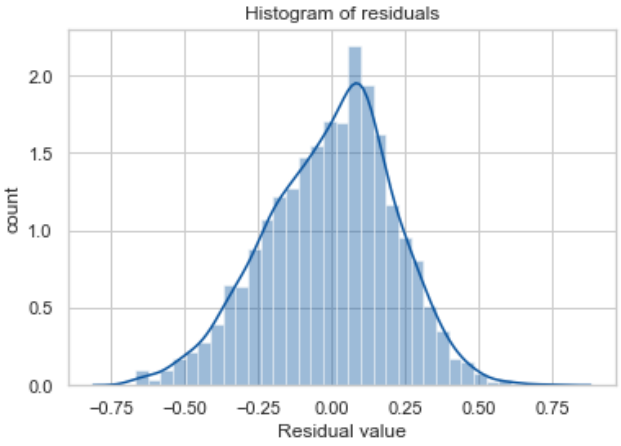


Figure 6. Histogram of Residuals

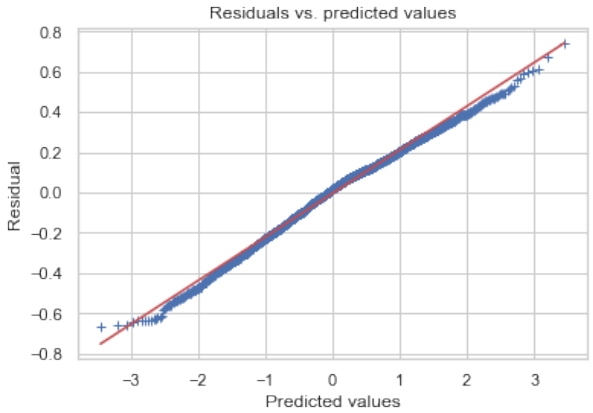


Figure 7. Quantile-Quantile (Q-Q) Plot of Residuals vs. Predicted Values

The above plots show a wide range of residuals across predicted values. However, the distribution of residuals is close to normal, which indicates that the model's predictive ability is consistent across the full range of the label (PPI).

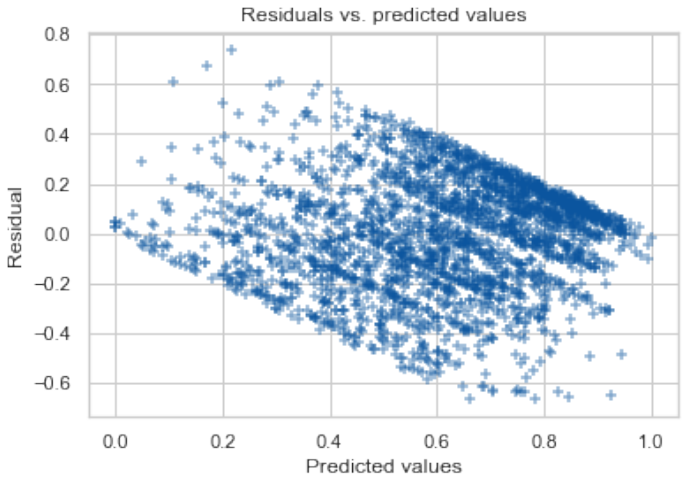


Figure 8. Scatter Plot of Residuals Vs. Predicted Values

The scatter plot of above shows that the model performs well when predicting individuals with a PPI close to 0 or 1, but is less accurate when predicting individuals with a medium probability of poverty.

4. CONCLUSION

This report has extracted socioeconomic indicators with better explanatory power on poverty from the rest. They are –

- *the average strength of shocks experienced the past year,*
- *number of financial activities conducted the past year,*
- *country of residence,*
- *the ability to calculate percentages,*
- *religion,*
- *residence in the urban area,*
- *age group,*
- *sex,*
- *marital status, and*
- *the role in a family*

It is concluded that an individual's PPI can only be roughly predicted from their socioeconomic indicators, especially when they have a PPI towards the middle.

5. REFERENCES

1. Getting Started with the PPI. Innovations for Poverty Action. https://www.povertyindex.org/get-started-ppi
2. Data Fiinder. Financial Inclusion Insights by InterMedia. http://fii-website.staging.interactive.columnfivemedia.com/data_fiinder.php
3. Financial Inclusion Insights Survey 2014. The World Bank. https://microdata.worldbank.org/index.php/catalog/2730
4. How does the PPI Work. Innovations for Poverty Action. https://www.povertyindex.org/about-ppi#How%20does%20the%20PPI%20work