Analysis of Poverty Probability

Mathew Ayotte, July 2019

Executive Summary

This document presents an analysis of data concerning socioeconomic indicators of individuals and the probability that those individuals live below the poverty line. The analysis is based on data for 12,600 individuals. The probability of being in poverty was calculated by the Poverty Probability Index, which estimates an individual's poverty status using 10 questions about a household's characteristics and asset ownership. The remaining data comes from the Financial Inclusion Insights household surveys conducted by InterMedia.

After exploring the data by calculating summary and descriptive statistics, and by creating visualizations of the data, several potential relationships between socioeconomic indicators and poverty probability were identified. After exploring the data, a regression model was created to predict an individual's poverty probability.

After performing the analysis, the author presents the following conclusions:

While many of the 58 factors can help indicate an individual's probability of being in poverty, significant features found in this analysis were:

- **Is_Urban** - whether or not the individual lives in an urban area. The poverty probability for urban dwellers tends to be lower than non-urban dwellers.
- **Education_Level** - the highest level of education has the individual had. There appears to be a negative correlation between education level and poverty probability. Individuals with more education tend to have a lower poverty probability.
- **Country** - which country the individual lives in. Individuals living in certain countries tend to have a higher poverty probability than those living in other countries.
- **Phone_Technology** - the sophistication of the individual's type of phone. There appears to be a negative correlation between phone technology and poverty probability. Individuals with more sophisticated phone technology tend to have lower poverty probability.
- **Can_Use_Internet** - whether or not the individual is able to use internet on their phone. The poverty probability for individuals who can use the internet on their phone tend to be lower than those who cannot.
- **Formal_Savings** - whether or not the individual has savings at a formal institution. The poverty probability for individuals who have a formal savings account tend to be lower than those who do not.

Initial Data Exploration

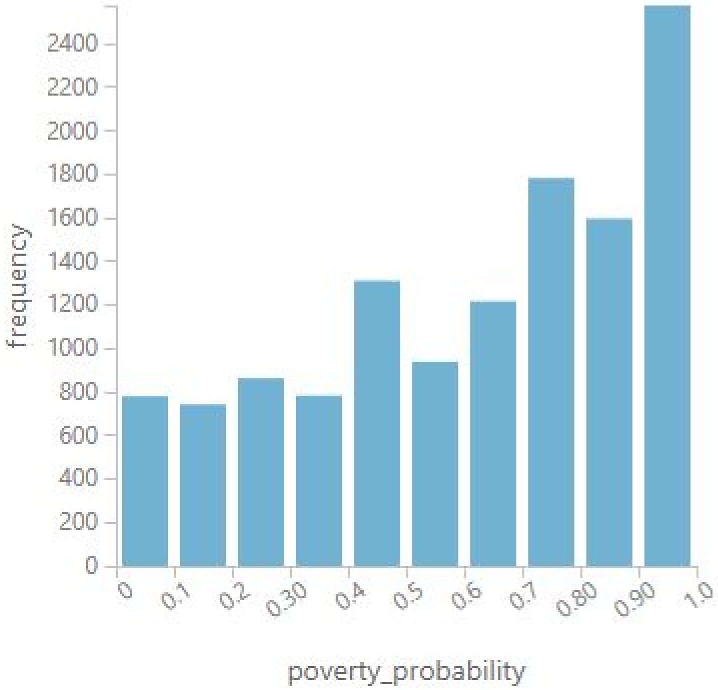
The initial exploration of the data began with some summary and descriptive statistics.

Individual Feature Statistics

Summary statistics for mean, median, minimum, maximum, standard deviation, unique values, and missing values were calculated for numeric columns, and the results taken from 12,600 observations are shown here:

Column	Mean	Median	Min	Max	Std Dev	Unique	Missing
age	36.2807	33	15	115	15.1459	84	0
num_times_borrowed_last_year	0.6577	0	0	3	0.9246	4	0
bank_interest_rate	9.8431	7	0	100	15.0331	30	12311
mm_interest_rate	9.021	7	0	100	13.6202	33	12449
mfi_interest_rate	10.9092	10	0	100	10.3533	33	12399
other_fsp_interest_rate	8.2167	6	0	100	10.6495	25	12361
num_shocks_last_year	1.1002	1	0	5	1.1901	6	0
average_shock_strength_last_year	2.1128	2	0	5	2.0192	41	0
num_formal_institutions_last_year	0.7141	1	0	6	0.8059	7	0
num_informal_institutions_last_year	0.189	0	0	4	0.4737	5	0
num_financial_activities_last_year	1.5597	1	0	10	2.0438	11	0
poverty_probability	0.6113	0.633	0	1	0.2915	166	0

Since poverty probability is of interest in this analysis, it was noted that the mean is slightly lower than the median, but that the relatively small difference between these values indicates that there is no considerable variance in the poverty probabilities. A histogram of the poverty probability column shows that the values are only slightly left-skewed – in other words, the number of individuals with a higher poverty probability is slightly higher than those with a lower poverty probability, as shown here:



The author determined that the columns with an extremely high percentage of missing values should be removed from consideration for further analysis. This includes the following columns:

- bank_interest_rate
- mm_interest_rate
- mfi_interest_rate
- other_fsp_interest_rate

In addition to the numeric values, the socioeconomic indicators include categorical features, including:

DEMOGRAPHICS

- **country** - Unique identifier for each country
- **is_urban** - Urban vs. rural area of residence
- **female** - Sex (True=female, False=male)
- **married** - Marital status
- **religion** - Unique identifier for religion
- **relationship_to_hh_head** - Respondent's relationship to the head of the household

EDUCATION

- **education_level -** Highest level of education (0=no education, 1=primary education, 2=secondary education, 3=higher education) (some missing values found)
- **literacy** - Ability to read and understand
- **can_add -** Ability to add
- **can_divide -** Ability to divide
- **can_calc_percents** - Ability to calculate percentages
- **can_calc_compounding** - Ability to calculate compounding interest

EMPLOYMENT

- **employed_last_year** - Whether the respondent was employed in the last year
- **employment_category_last_year** - Category of employment last year (e.g. employed, retired)
- **employment_type_last_year** - Type of employment last year (e.g. salaried, seasonal)
- **share_hh_income_provided -** Share of household income provided (some missing values found)
- **income_ag_livestock_last_year -** Whether the respondent received income from agriculture or livestock in the last year
- **income_friends_family_last_year** - Whether the respondent received income from friends or family in the last year
- **income_government_last_year -** Whether the respondent received income from the government in the last year
- **income_own_business_last_year -** Whether the respondent received income from their own business in the last year
- **income_private_sector_last_year -** Whether the respondent received income from the private sector in the last year
- **income_public_sector_last_year -** Whether the respondent received income from the public sector in the last year

ECONOMIC

- **borrowing_recency** - Recency of last borrowing activity
- **formal_savings** - Has savings at a formal institution
- **informal_savings -** Has savings at an informal institution
- **cash_property_savings** - Has savings in cash or property
- **has_insurance** - Has at least one form of insurance
- **has_investment** - Has at least one form of investment
- **borrowed_for_emergency_last_year -** Borrowed money for an emergency in the last year
- **borrowed_for_daily_expenses_last_year -** Borrowed money for daily expenses in the last year

- **borrowed_for_home_or_biz_last_year** - Borrowed money for home or business expenses in the last year

PHONE

- **phone_technology** - Sophistication of phone type (0=no phone, 1=basic phone, 2=feature phone, 3=smartphone)
- **can_call** - Ability to make a phone call
- **can_text** - Ability to text
- **can_use_internet** - Ability to use internet on one's phone
- **can_make_transaction** - Ability to make a financial transaction on one's phone
- **phone_ownership** - Phone ownership (0=no phone, 1=shares phone, 2=owns phone)
- **advanced_phone_use** - Ability to do advanced tasks on a phone

FINANCIAL INCLUSION

- **reg_bank_acct** - Has a bank account in their own name
- **reg_mm_acct** - Has a mobile money account in their own name
- **reg_formal_nbfi_account** - Has an account at a non-banking financial institution (NBFI) in their own name
- **financially_included** - Financially included, which is defined as having at least one of the following: a registered bank account, a registered mobile money account, or a registered NBFI account
- **active_bank_user** - Has used their bank account in the last 90 days
- **active_mm_user** - Has used their mobile money account in the last 90 days
- **active_formal_nbfi_user** - Has used their formal NBFI account in the last 90 days
- **active_informal_nbfi_user** - Has conducted financial activity at an informal NBFI in the last 90 days
- **nonreg_active_mm_user** - Has used a mobile money account in someone else's name in the last 90 days ('over-the-counter' use)

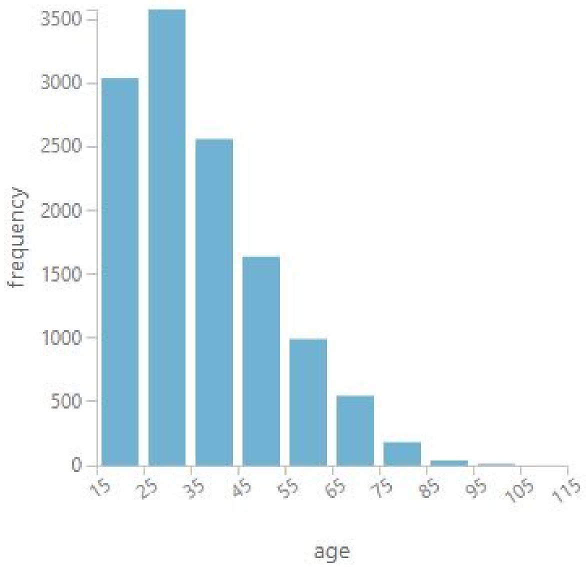
Bar charts were created to show frequency of these features, and indicate the following:

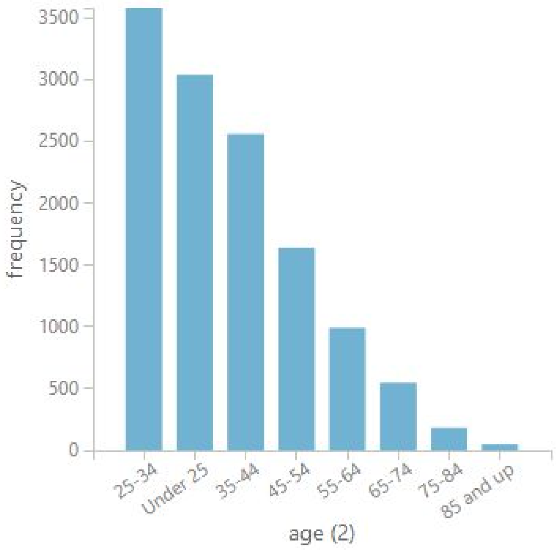
- Urban dwellers are less common than non-urban dwellers.
- Religions Q and X are most common, followed by religion P; religions O and N are relatively uncommon.
- Primary and secondary education levels are most common, followed by no education; higher education was the least common.
- Individuals that can_add and can_divide are much more common than those who cannot add or divide.
- Individuals receiving income from government sources last year is highly uncommon.

- Individuals receiving income from public sector sources last year is extremely uncommon.
- Most individuals had no formal or informal savings.
- It is highly less common for individuals to have insurance.
- Individuals with no phone are most common, followed by basic and feature phones; it is least likely for individuals to own a smartphone.
- Most individuals are unable to use the internet or make a transaction on their phone.
- It is less likely for individuals to have a bank account in their own name.

It was decided by the author that the missing values in education_level and share_hh_income_provided should be replaced by the median of those columns respectively.

The age column was converted to a categorical column and was then grouped by the following categories: Under 25, 25- 34, 35 - 44, 45 - 54, 55 - 64, 65 - 74, 75 - 84, 85 and up





It was also decided to convert the following numeric columns to categorical features for the purposes of assisting the efficiency of the regression algorithm:

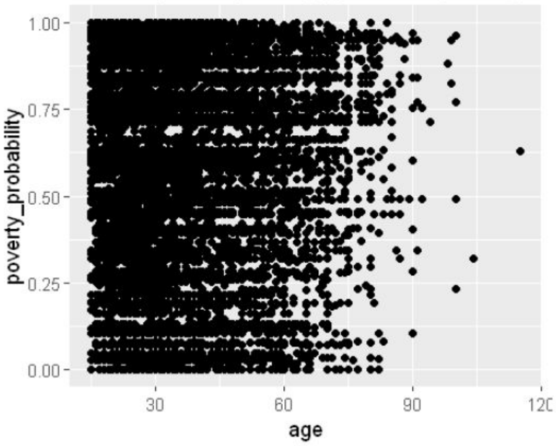
- num_times_borrowed_last_year
- num_shocks_last_year
- avg_shock_strength_last_year
- num_formal_institution_last_year
- num_informal_institution_last_year
- num_financial_activities_last_year

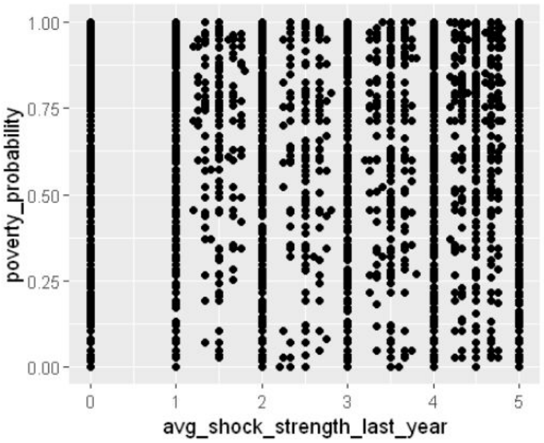
Correlation and Apparent Relationships

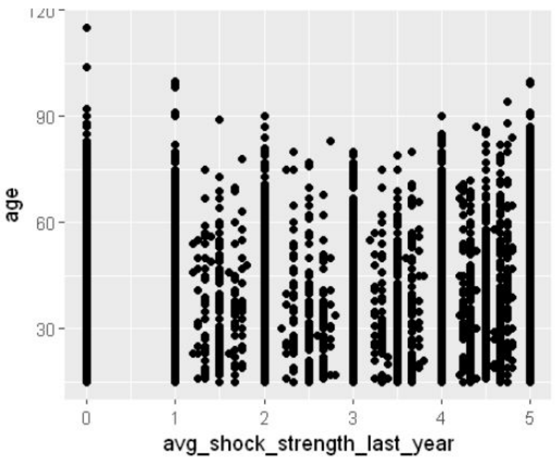
After exploring the individual features, an attempt was made to identify relationships between features in the data – in particular, between poverty probability and the other features.

Numeric Relationships

The following scatter-plots were generated initially to compare the two numeric features with the poverty probability label.





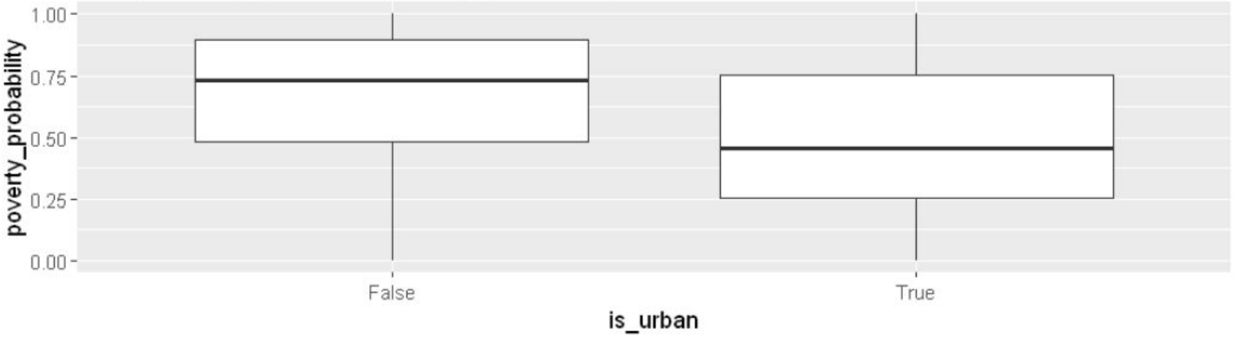


Based on these scatter plots, no apparent relationship is noticed between the numeric columns.

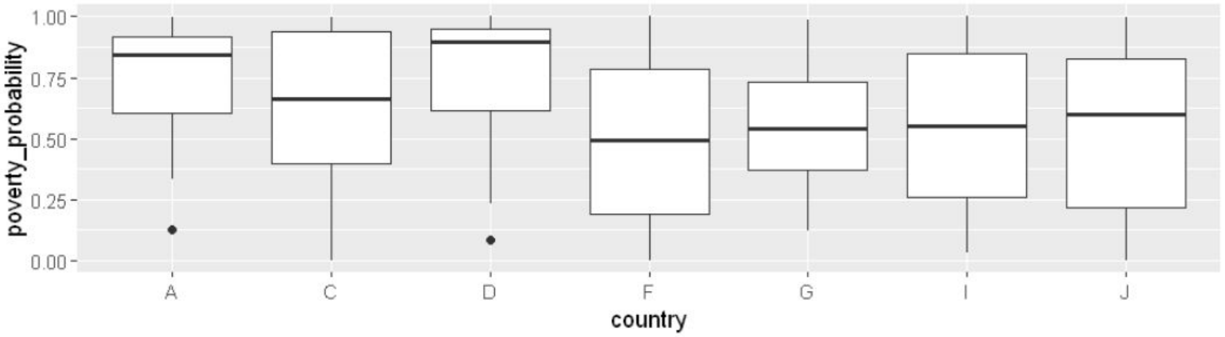
Categorical Relationships

Having explored the relationship between poverty probability and numeric features, an attempt was made to discern any apparent relationship between the extensive list of categorical features and poverty probability. The following box-plots show the categorical columns that seem to exhibit a relationship with poverty probability:

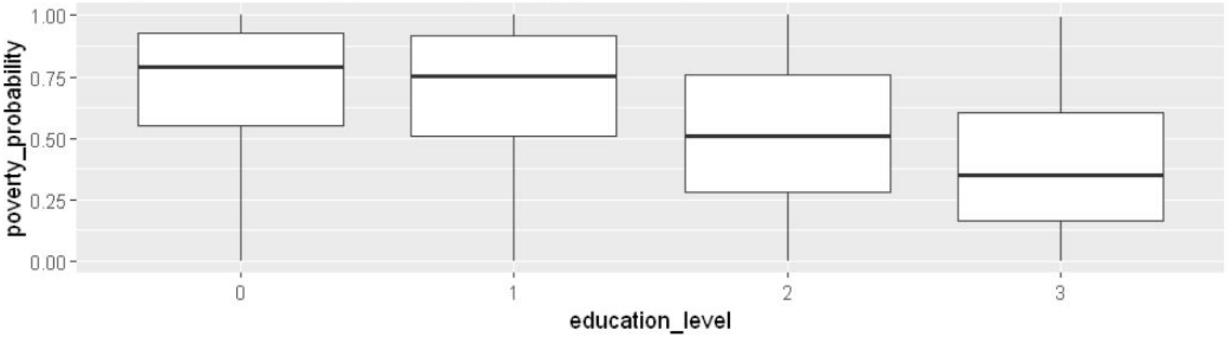
Box plot of is_urban vs. poverty_probability



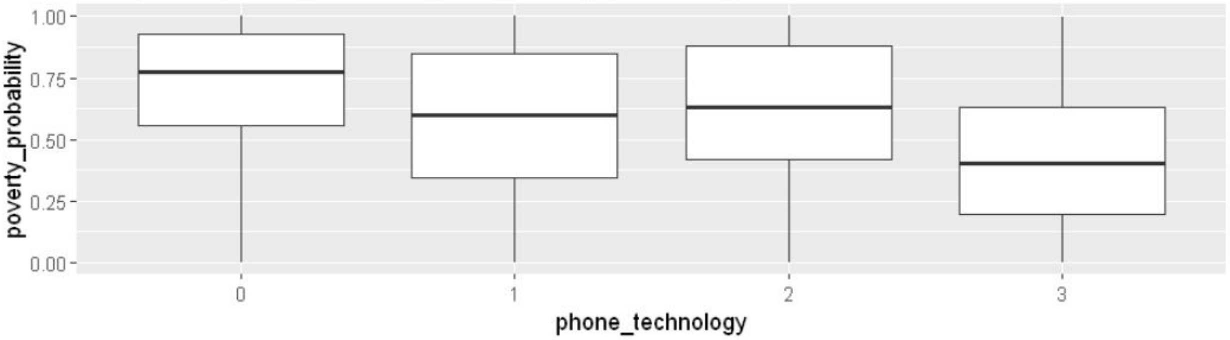
Box plot of country vs. poverty_probability



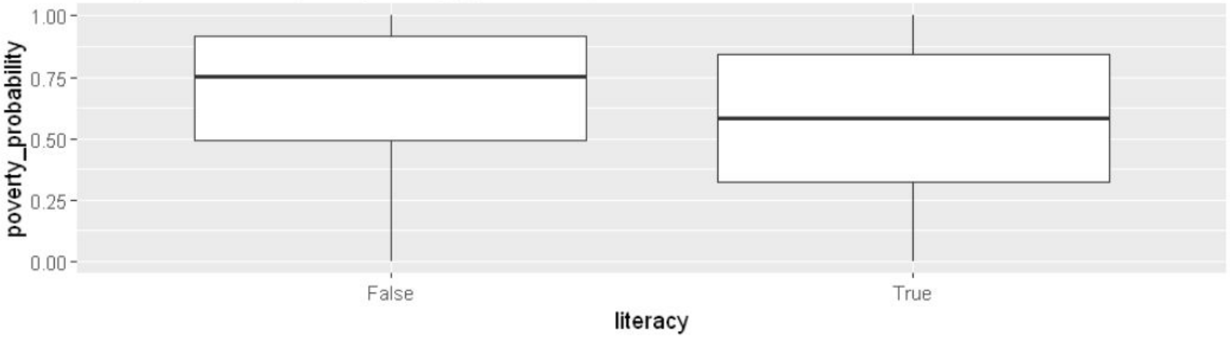
Box plot of education_level vs. poverty_probability



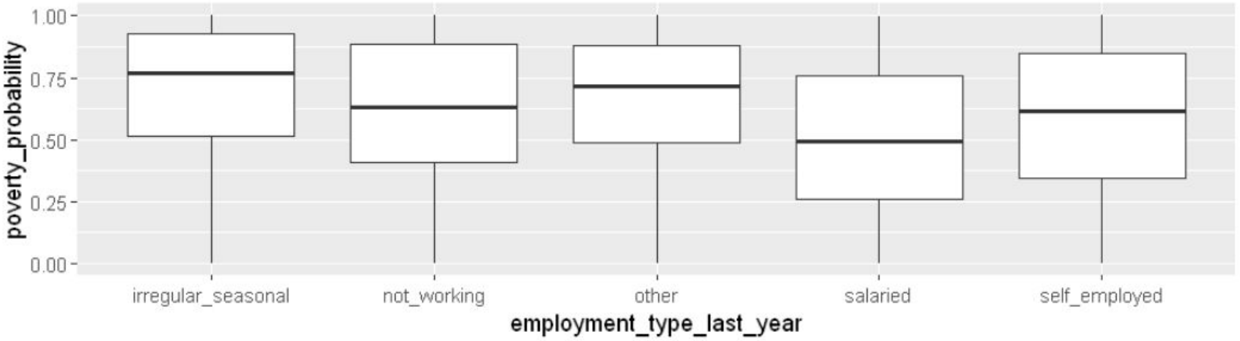
Box plot of phone_technology vs. poverty_probability



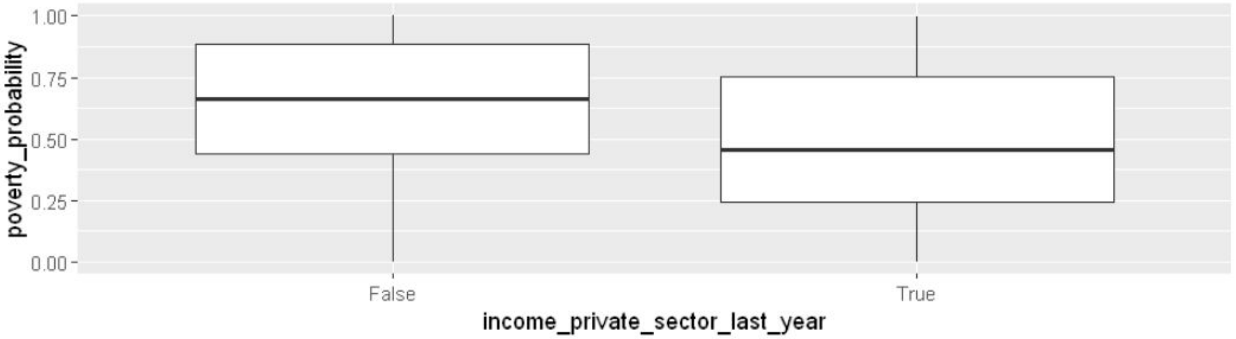
Box plot of literacy vs. poverty_probability



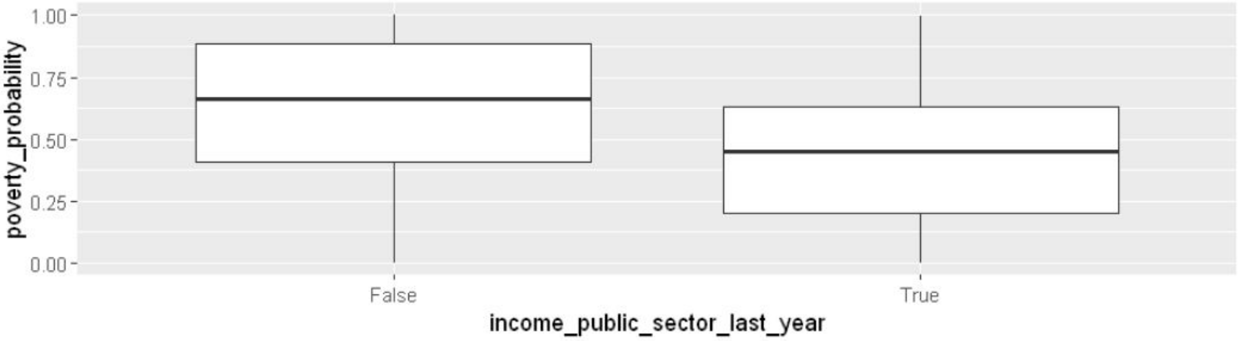
Box plot of employment_type_last_year vs. poverty_probability

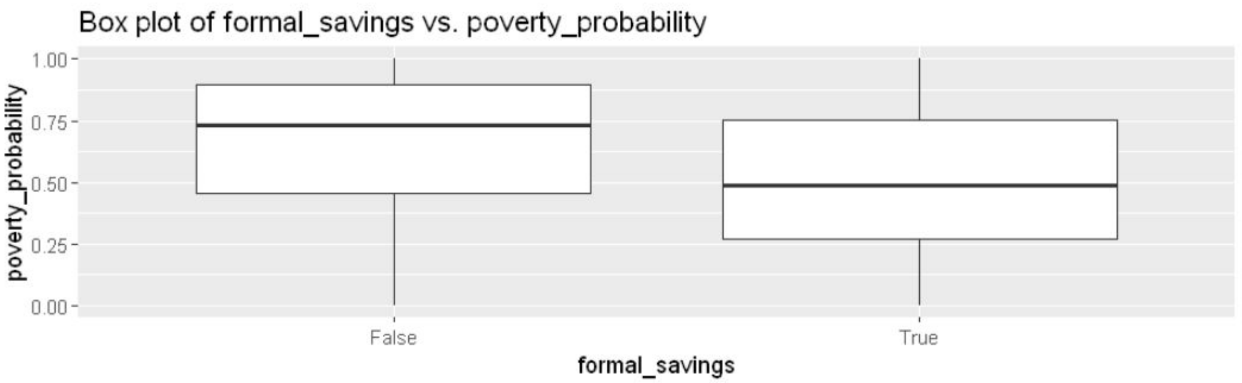


Box plot of income_private_sector_last_year vs. poverty_probability

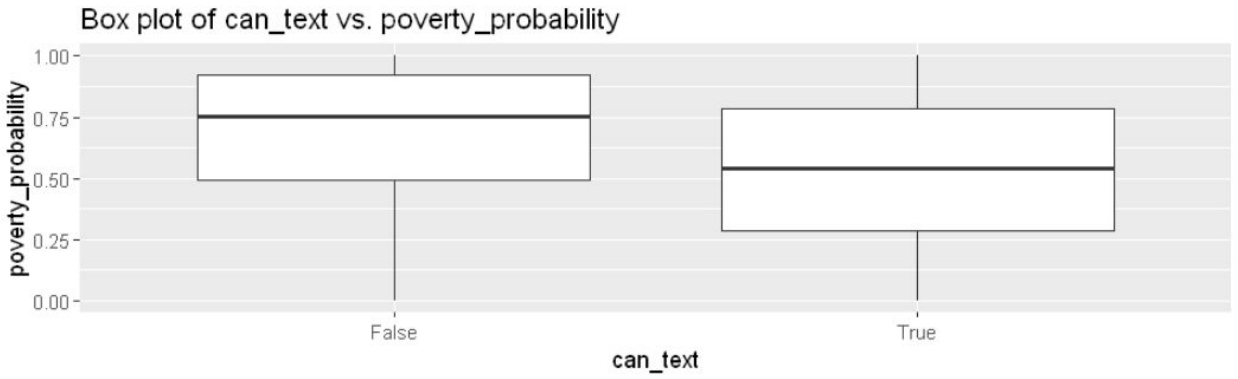


Box plot of income_public_sector_last_year vs. poverty_probability

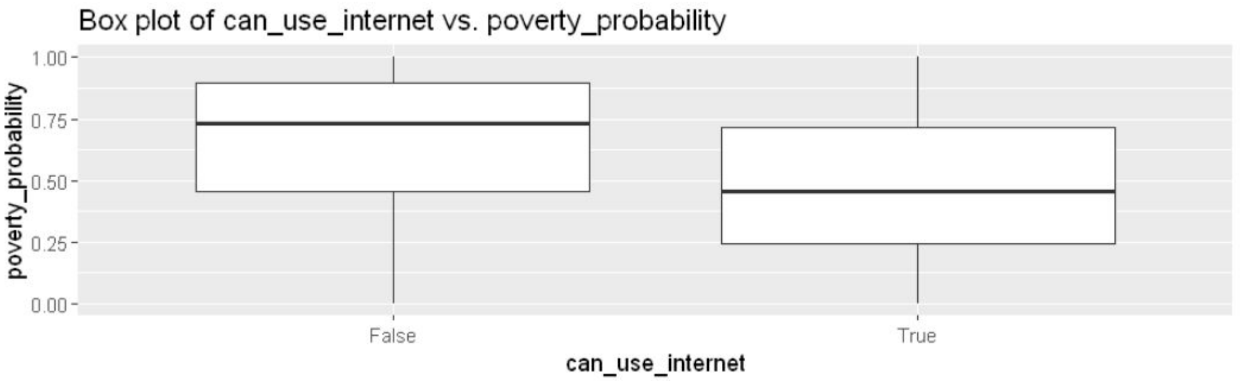




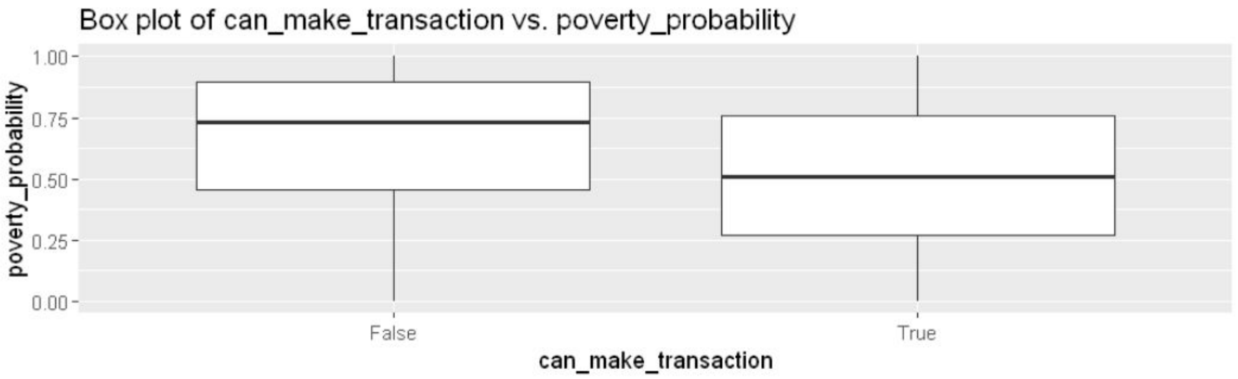
Box plot of formal_savings vs. poverty_probability



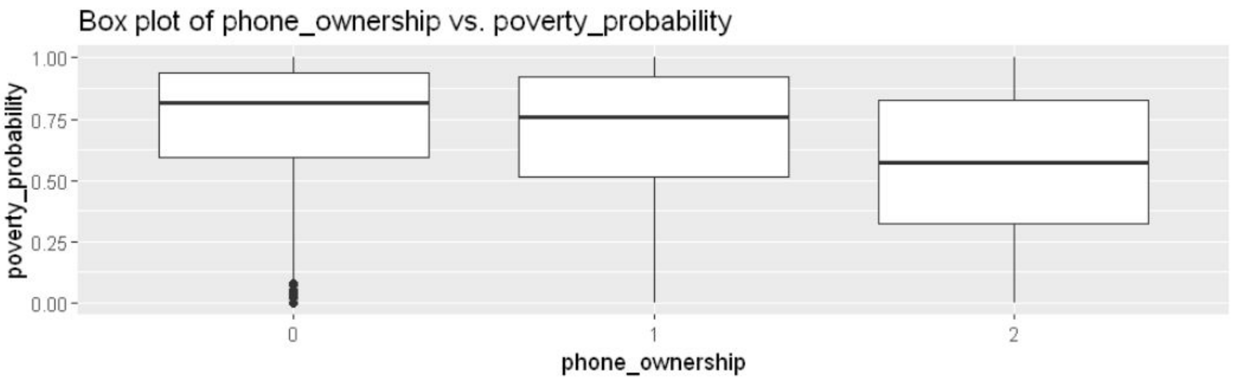
Box plot of can_text vs. poverty_probability

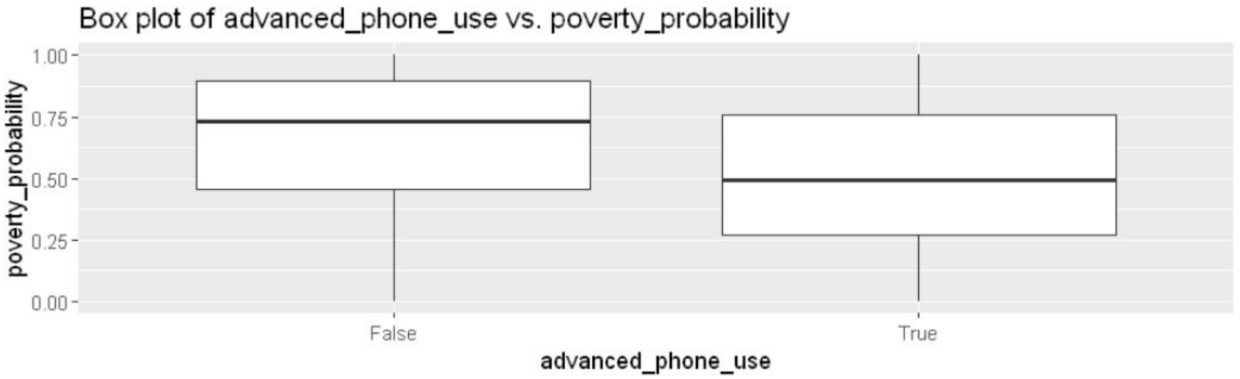


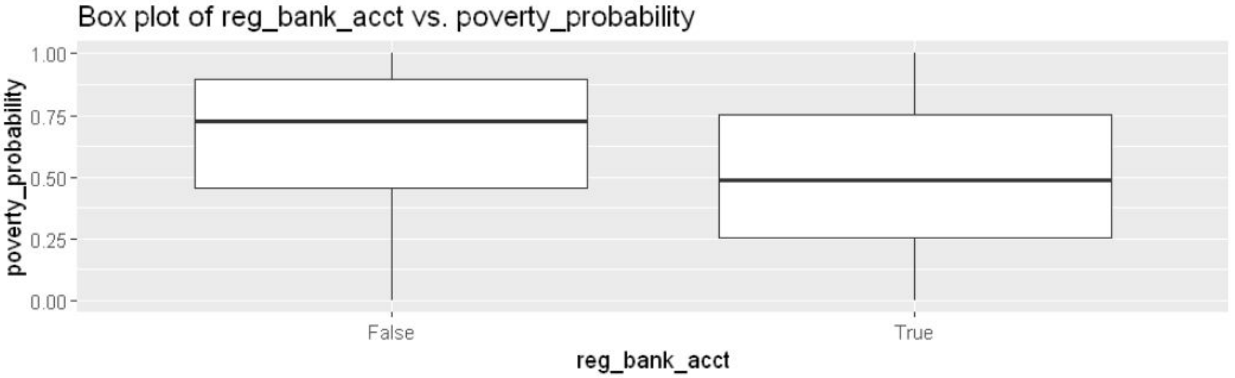
Box plot of can_use_internet vs. poverty_probability

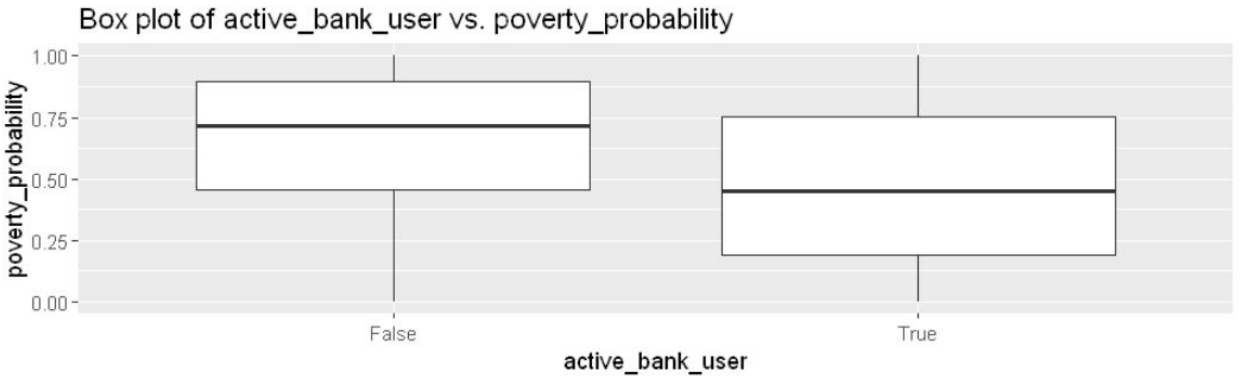


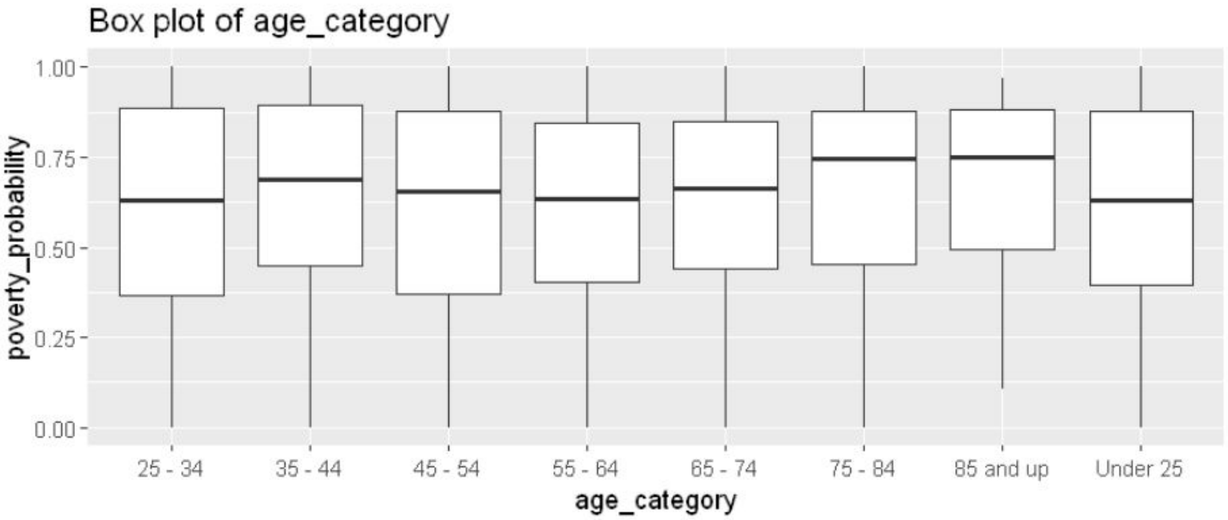
Box plot of can_make_transaction vs. poverty_probability


Box plot of phone_ownership vs. poverty_probability


Box plot of advanced_phone_use vs. poverty_probability


Box plot of reg_bank_acct vs. poverty_probability


Box plot of active_bank_user vs. poverty_probability



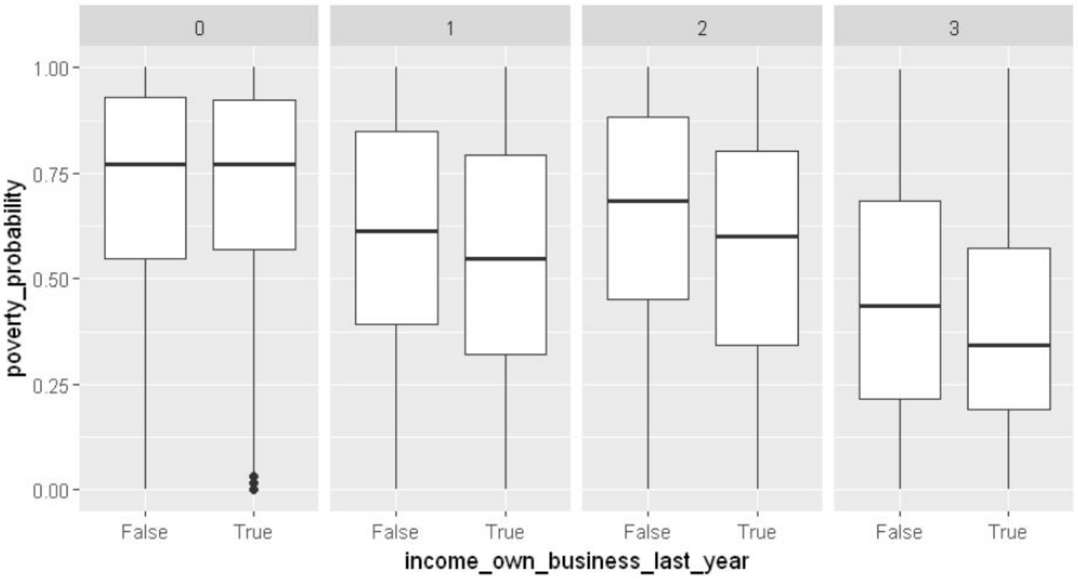
The box plots show some clear differences in terms of the median and range of poverty probability values for different categorical features. For example:

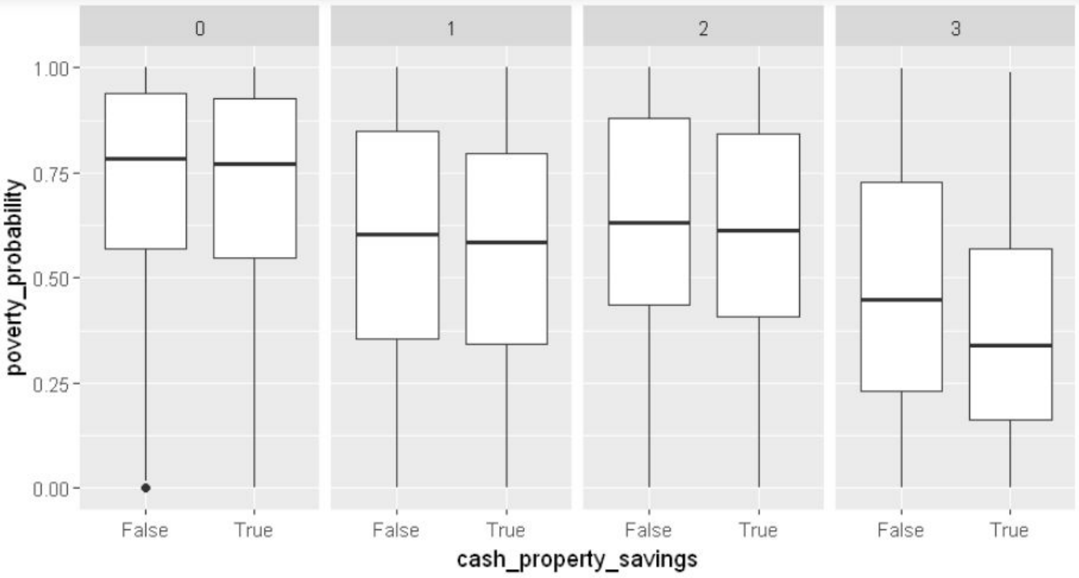
- Non-urban-dwelling individuals tend to have higher poverty probability levels but there is a wider range of probabilities for urban-dwellers.
- There are some countries that tend to have individuals with higher poverty probability levels.
- Individuals that have completed secondary or higher education typically have lower poverty probability levels.
- Smartphone owners typically have lower poverty probability levels than non-smartphone users.
- Non-literate individuals typically have higher poverty probability levels.
- Salaried individuals typically have lower poverty probability levels.
- Individuals receiving income from public or private sectors typically have lower poverty probability levels.
- Individuals that have a formal savings typically have lower poverty probability levels.
- Individuals that can send a text message typically have lower poverty probability levels.
- Individuals that can use the internet on their phone typically have lower poverty probability levels.
- Individuals that can make a transaction on their phone typically have lower poverty probability levels.
- Individuals that own their own phone typically have lower poverty probability levels.
- Individuals that can do advanced tasks on their phone typically have lower poverty probability levels.
- Individuals that have a bank account in their own name typically have lower poverty probability levels.
- Individuals that have used their bank account recently typically have lower poverty probability levels.

Multi-faceted Relationships

Apparent relationships between poverty probability and individual features are helpful in determining predictive heuristics. However, relationships are often more complex, and may only become apparent when multiple features are considered in combination with one another. To help identify these more complex relationships, some faceted plots were created.

The following plots show some interesting aspects of the education_level column. It can be seen from these plots that education_level can be predictive of poverty probability levels in combination with such factors as income_own_business_last_year and cash_property_savings.

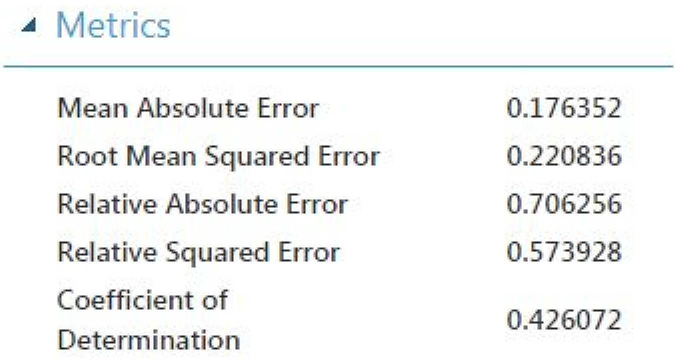




Regression

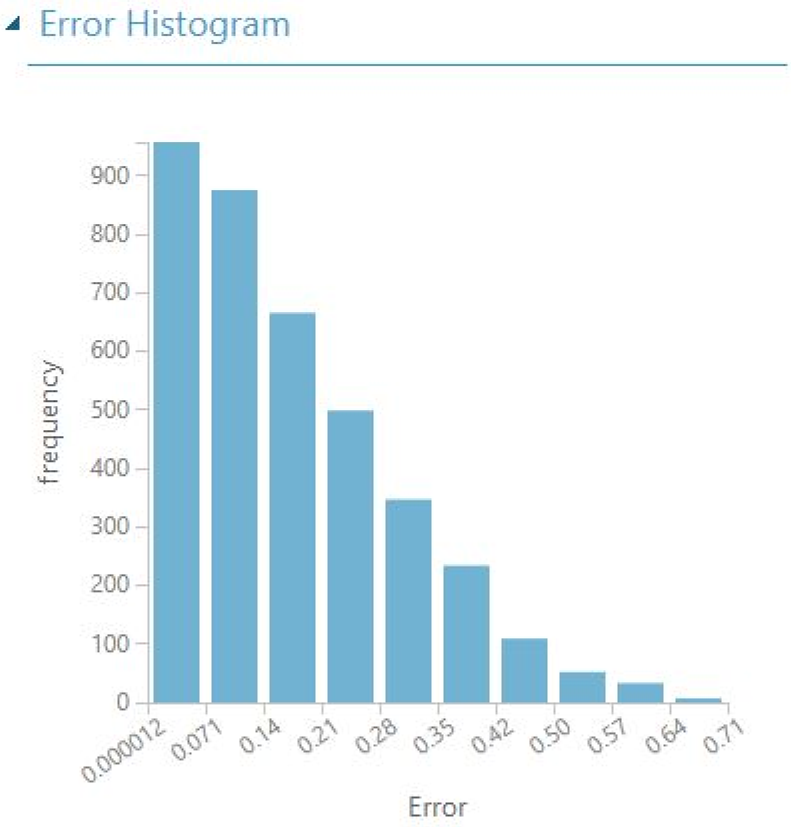
A regression model to predict the actual poverty probabilities of individuals was created. Based on the apparent relationships identified when analyzing the data, a Boosted Decision Tree regression model using one-hot encoding with 5 folds and parameter sweeping was created to predict the poverty probability.

The model was trained with 70% of the data, and tested with the remaining 30%. The statistical results of the regression model are shown below:



Metrics

Mean Absolute Error	0.176352
Root Mean Squared Error	0.220836
Relative Absolute Error	0.706256
Relative Squared Error	0.573928
Coefficient of Determination	0.426072



Error Histogram

Conclusion

This analysis has shown that the poverty probability of an individual can be confidently predicted from their socioeconomic and financial indicators. In particular, urban/non-urban dwelling, phone technology, ability to use the internet, ability to send a text message, and having a formal savings account have a significant effect on the poverty probability of an individual. Secondary features, such as education level can help further classify individuals and determine the poverty probability groupings to which they belong.