Global Poverty Analysis Report

JULY 2019

Authored by: Travis Klein

Investigating Global Poverty in Data

Executive Summary

Immense progress has been made in combating severe poverty over the last several decades. However, it remains a durable challenge in distressed regions around the globe—most notably in Sub-Saharan Africa. As poverty is a distinctly complex and multivariate issue, determining the true underlying causes contributing to it is paramount to addressing it effectively. This report details an inquiry into determinants of poverty worldwide and the development of two different machine-learning models used to predict probability of poverty based on these factors. Upon comparison of these models, it was found that variance in probability of poverty could be predicted with just over 40% accuracy according to variance in household survey data.

As poverty is a distinctly complex and multivariate issue, determining the true underlying causes contributing to it is paramount to addressing it effectively.

The data for this analysis comes primarily from Financial Inclusion Insights survey data collected by InterMedia across seven countries averaging below the poverty income threshold of $2.50/day (USD) per person. The survey data captures over 50 details for each respondent—key features upon analysis included, among others, education, access to phone technology, and number of recent financial activities. This survey dataset consists of 12,600 rows and serves as the ‘feature’ data for the ML model.

The ‘label’ values for the model—the characteristics we’re eventually trying to predict—consist of probability values (between 0 and 1) calculated using the Poverty Probability Index (PPI), which produces an estimated probability of an individual’s poverty status (i.e. whether or not they fall below the poverty threshold) using 10 characteristics of household characteristics and asset ownership. This data has been generated for each line of the feature dataset and therefore also consists of 12,600 rows.

For the purposes of this analysis, we will be investigating salient or otherwise interesting relationships between the InterMedia survey data and the PPI poverty probabilities, and then using pertinent insights therefrom to **predict** these probabilities in the absence of individually calculated values from the PPI. A notable **consideration** of this approach is that it assumes on the accuracy of the PPI—in other words, a 100% 'accurate' model built on this approach will still only be as accurate as the PPI in terms of predictive potency.

Methodology

Cleansing, transformation, and analysis of data was performed using both Python 3 and Microsoft Excel, as needed. Development of machine-learning model was performed in Microsoft Azure ML.

Preparation of Data

To begin investigating the relationships between survey data and PPI probability values, the feature and label datasets were first joined into a single, combined dataset that was also eventually used to train the ML model. This combined dataset consists of 12,600 rows and a total of 60 columns, including both numerical and categorical features. See below for a complete list of initial features:

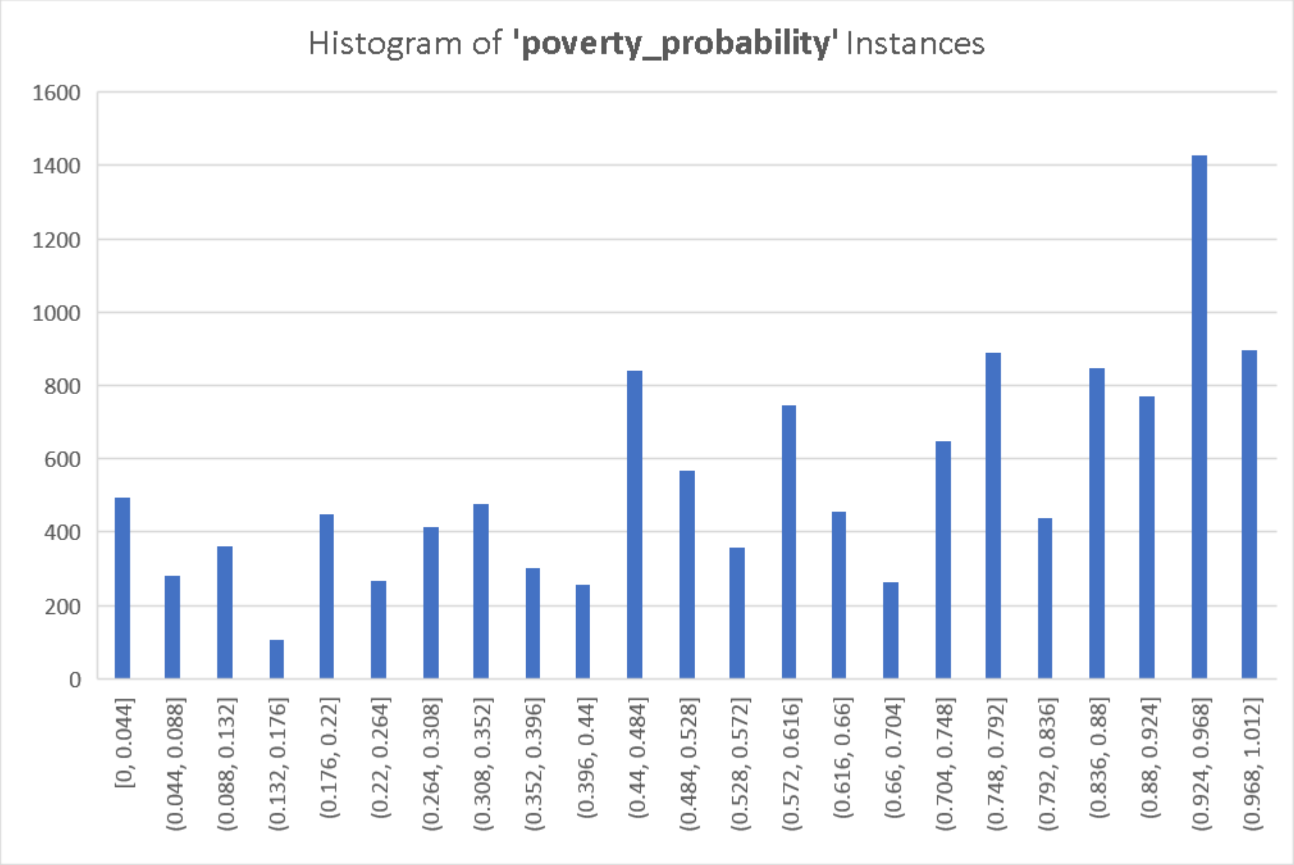
row_id	country
is_urban	age
female	married
religion	relationship_to_hh_head
education_level	literacy
can_add	can_divide
can_calc_percents	can_calc_compounding
employed_last_year	employment_category_last_year
employment_type_last_year	share_hh_income_provided
income_ag_livestock_last_year	income_friends_family_last_year
income_government_last_year	income_own_business_last_year
income_private_sector_last_year	income_public_sector_last_year

num_times_borrowed_last_year	borrowing_recency
formal_savings	informal_savings
cash_property_savings	has_insurance
has_investment	bank_interest_rate
mm_interest_rate	mfi_interest_rate
other_fsp_interest_rate	num_shocks_last_year
avg_shock_strength_last_year	borrowed_for_emergency_last_year
borrowed_for_daily_expenses_last_year	borrowed_for_home_or_biz_last_year
phone_technology	can_call
can_text	can_use_internet
can_make_transaction	phone_ownership
advanced_phone_use	reg_bank_acct
reg_mm_acct	reg_formal_nbfi_account
financially_included	active_bank_user
active_mm_user	active_formal_nbfi_user
active_informal_nbfi_user	nonreg_active_mm_user
num_formal_institutions_last_year	num_informal_institutions_last_year
num_financial_activities_last_year	poverty_probability

A combination of Z-score and min-max normalization was used to prepare numerical features, and missing values throughout the set were replaced using either probabilistic principle component analysis (PCA) or the column mean, as appropriate. Finally, the 'row_id' field and any features where less than 10% of values were present were excluded from eventual model training.

Investigation of Data and Potential Relationships

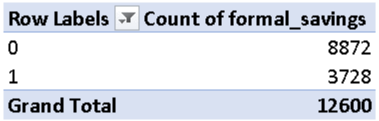
As the 'poverty_probability' field is what we'll eventually be trying to predict, it is of particular interest. See below for a histogram for our 12,600 instances of 'poverty_probability', bucketed from least to greatest:



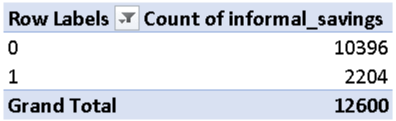
Note that we have a fairly opaque and non-normal distribution here, with no obvious patterns emerging from the distribution alone. However, the data does exhibit a modest left skew, indicating that our dataset is slightly weighted toward survey respondents on the higher end of the probability distribution.

Numerical features displayed differing levels of correlation with ‘poverty_probability’, but there are indeed circumstances where even the lack of correlation can still be informative. For example, ‘age’ only demonstrated a correlation coefficient of **0.007226** with our label set, but this yields the implied conclusion that **poverty doesn’t appear care how old you are**—at least for the countries in this dataset.

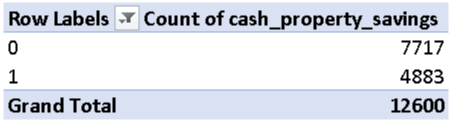
This seems to go against what might be the natural intuition—that younger people would be more highly correlated with poverty since they’ve had less time to accrue cumulative wealth—but this intuition is based implicitly on an assumed correlation of *savings* with age. Three types of savings data are captured in the survey as Boolean values. The frequency distributions of these values are below:



Row Labels	Count of formal_savings
0	8872
1	3728
Grand Total	**12600**



Row Labels	Count of informal_savings
0	10396
1	2204
Grand Total	**12600**



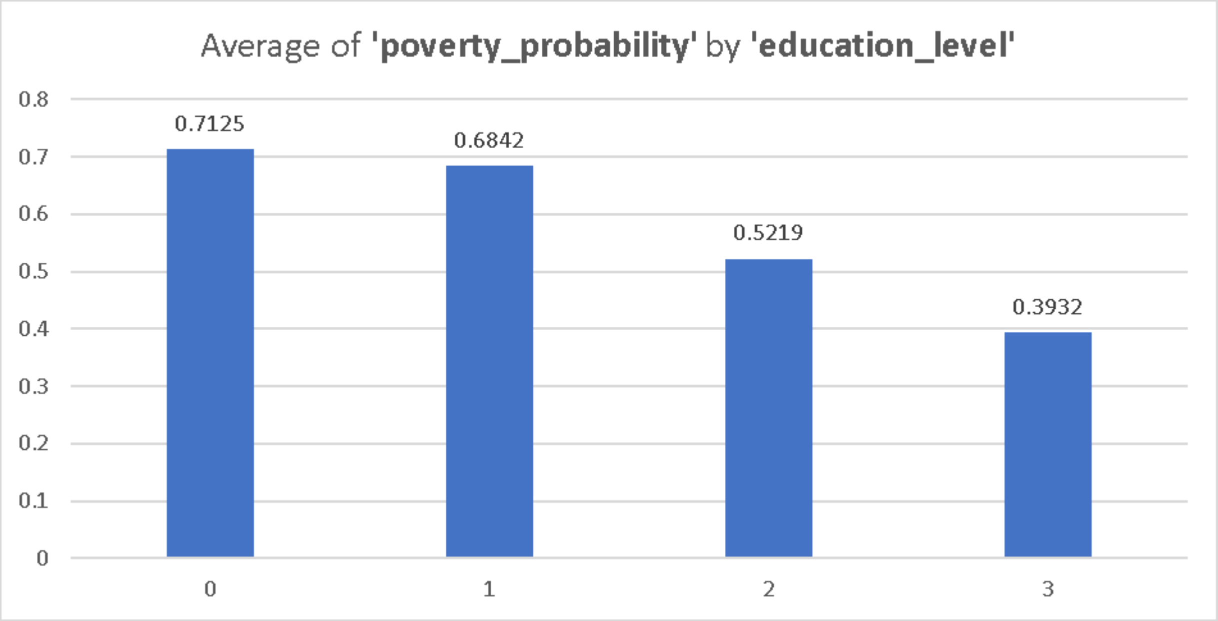
Row Labels	Count of cash_property_savings
0	7717
1	4883
Grand Total	**12600**

Although we have at least moderate representation of savings among the population as TRUE (1) values for all three types of savings, comparison with ‘age’ shows no compelling correlations—set against ‘age’, our features ‘formal_savings’, ‘informal_savings’, and ‘cash_property_savings’ yielded the following correlation coefficients:

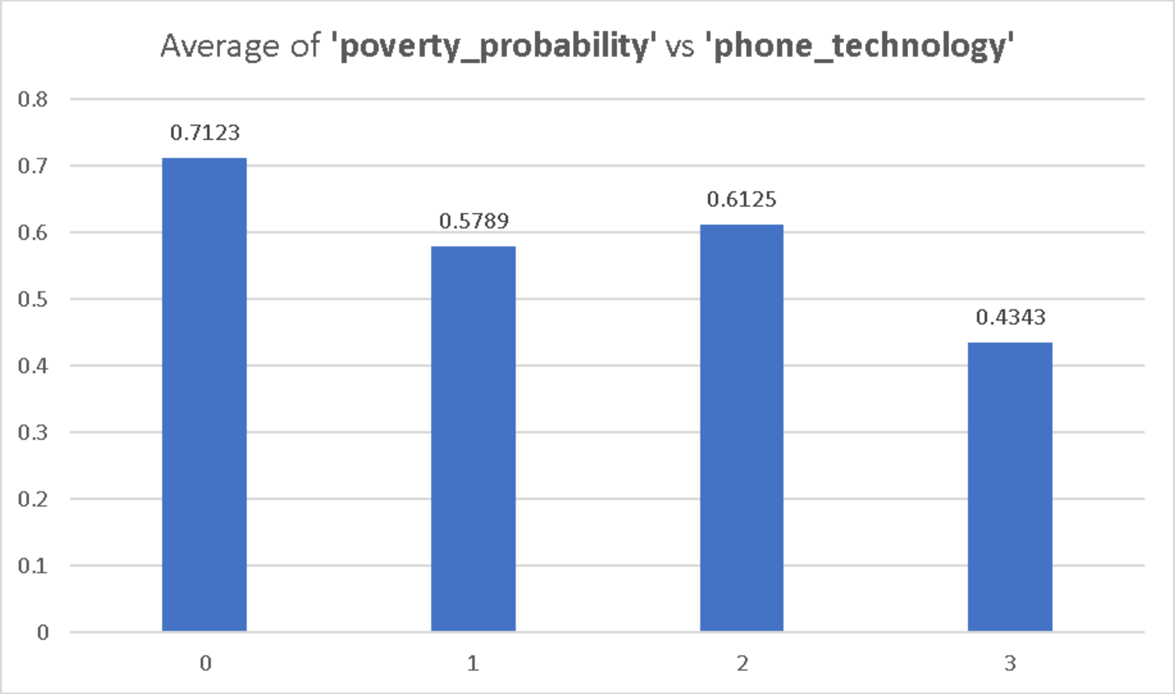
	‘formal-savings’	**‘informal_savings’**	**‘cash_property_savings’**
Corr. Coefficient compared to ‘age’	0.010131839	0.002928489	-0.006332393

This leads naturally to the further insight that **respondents are effectively no more likely to start saving more as they get older**.

However, other numerical features showed more compelling correlations with ‘poverty_probability’ – and perhaps ones more in line with expectations. For example, one might expect poverty to fall as education level rises, and the data indeed bear this out with a correlation of **-0.34549**, indicating a steady inverse relationship.

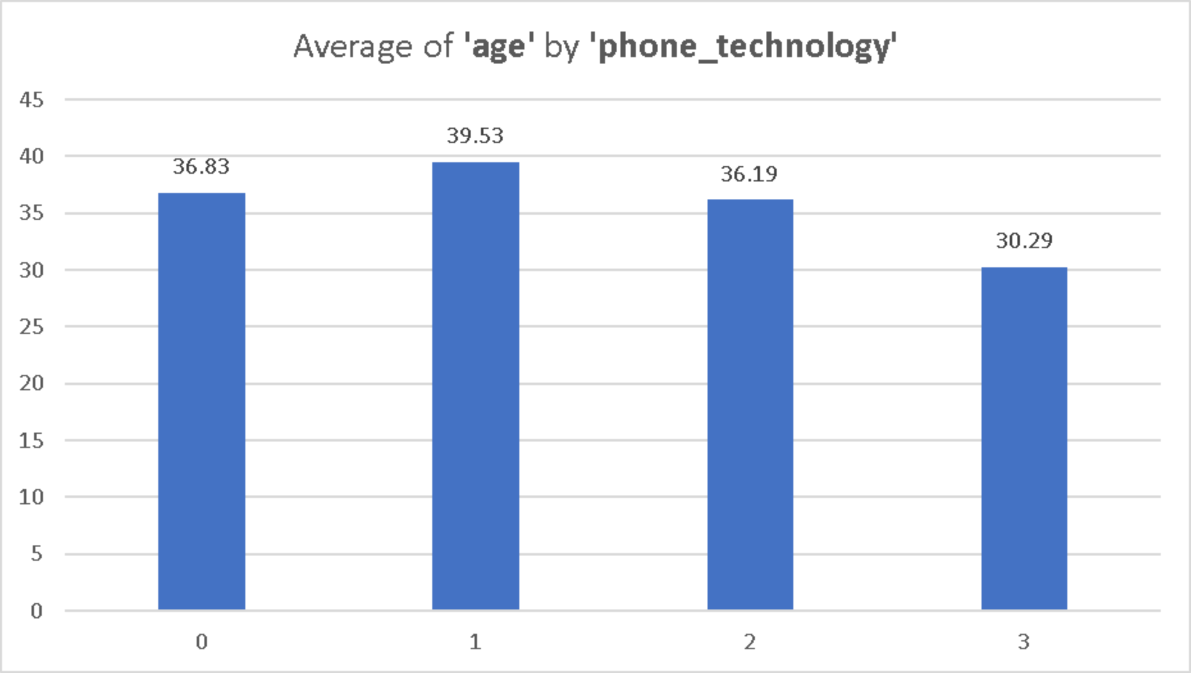


Finally, some comparisons yielded findings that were more of a mixed bag—partially expected and partially puzzling. For example, in a highly interconnected world where access to technology can afford vast arrays of potential opportunity, one might also expect access to a cell phone to be a strong predictor of financial success at this margin. Indeed, as ‘phone_technology’ increases we observe a **-0.28928** correlation with ‘poverty_probability’. However, upon visualization a strange phenomenon emerges:



Average ‘poverty_probability’ falls by a precipitous **18.72%** when a respondent goes from not having access to a phone to having access to a single phone, but likelihood of poverty actually **increases** when the number of accessible phones rises from one to two before plummeting again when the number rises to three.

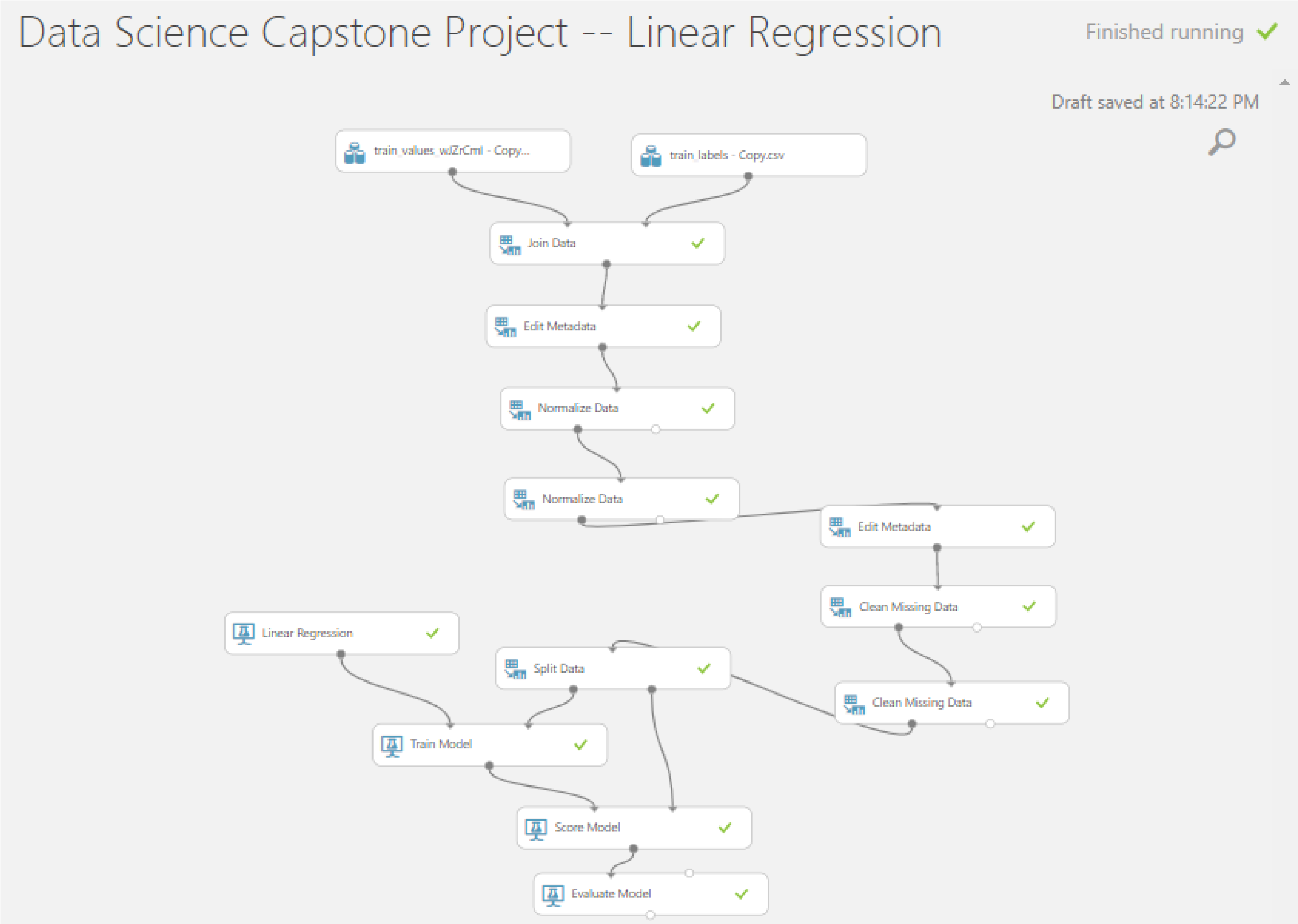
This is an interesting aberration from the expected pattern. The data seems to suggest either that something about having access to **two phones specifically** is detrimental—which seems unlikely—or that **access to phone technology is a proxy for other variables that may better explain the pattern**. One possibility is age:



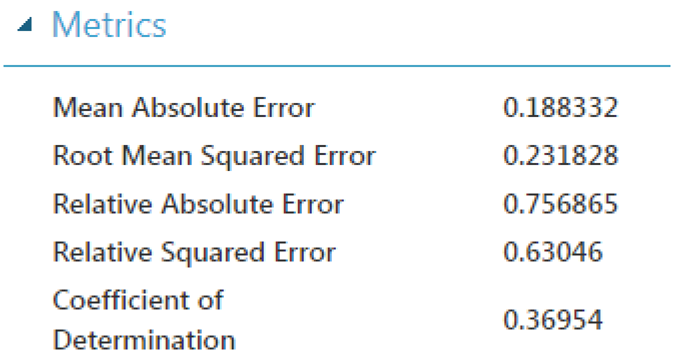
Note here how the average age of someone with no phone access is nearly identical to the age of someone with access to two—even though judging by the rest of the trend one might expect those with zero phone access to be the oldest group. Could something about that age of 36 tell at least some of the earlier story between ‘phone_technology’ and ‘poverty_probability’? The hypothesis may seem unlikely given the weakness of correlation between age and poverty outlined above, but indeed upon further inspection it turns out that **starting after age 15, respondents are more likely to be in poverty at age 36 than at any other age until they’re 71.** In other words, overrepresentation of 36-year-olds in the group with access to two phones could in theory at least partially explain the otherwise puzzling probability of poverty among that group. A peculiar finding—and broadly the correlation between age and poverty is still quite weak—but something about this unlucky age seems at least like interesting grounds for further inquiry.

Predictive Model Development

After initial exploration and investigations into intriguing relationships within the dataset, two different machine-learning models were created in Azure ML Studio in order to predict 'poverty_probability' based on characteristic features in the training dataset. The evaluated accuracy of each model was then compared to determine which method was superior.



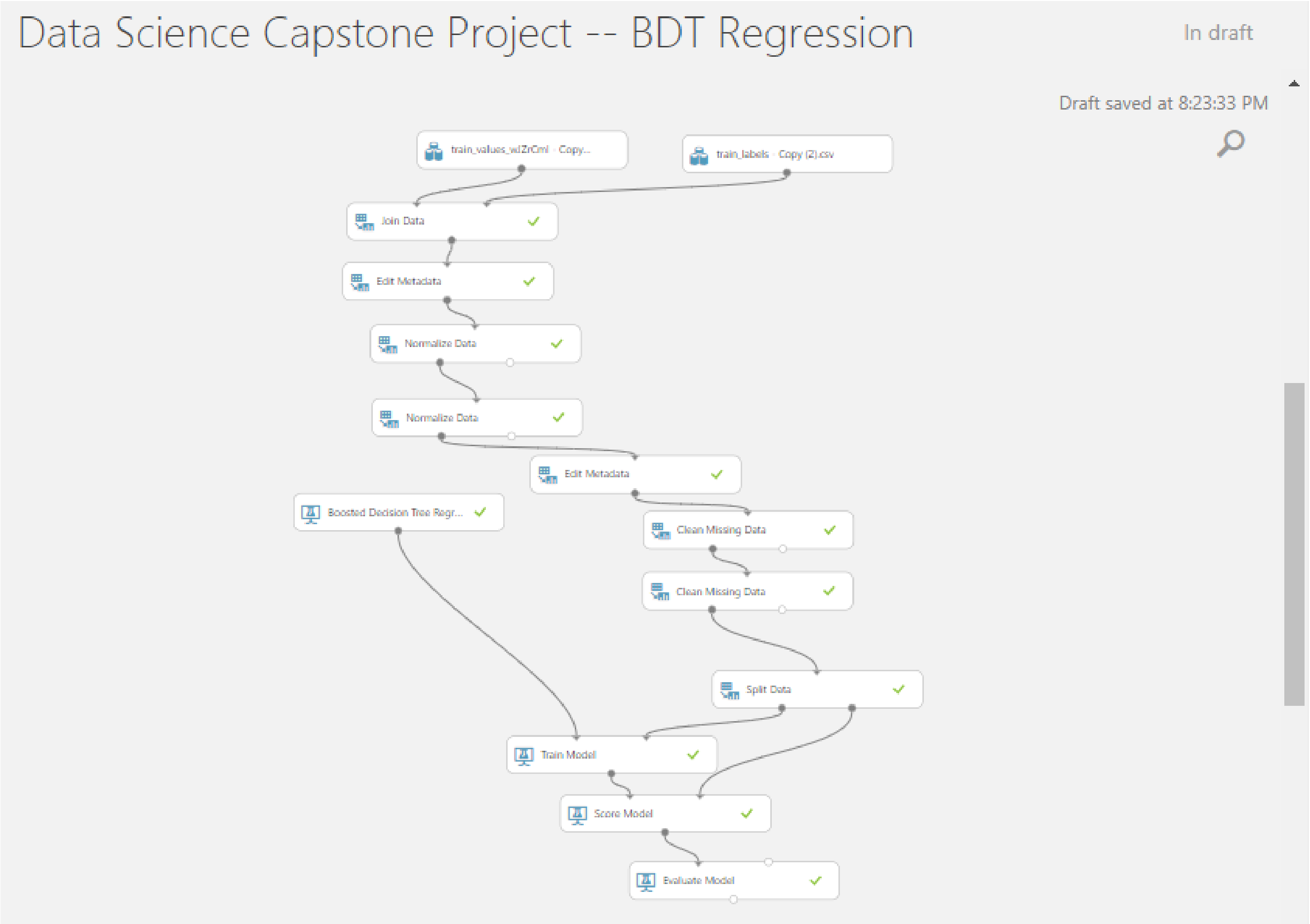
The first model created used linear regression. As illustrated above, the training features and labels were first joined into a single table, then various data preparation, normalization, and cleaning steps were performed before splitting, training, scoring, and finally evaluating the model.



◢ Metrics

Mean Absolute Error	0.188332
Root Mean Squared Error	0.231828
Relative Absolute Error	0.756865
Relative Squared Error	0.63046
Coefficient of Determination	0.36954

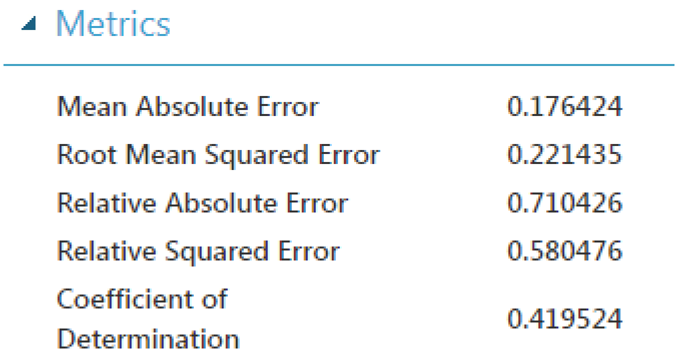
Trained on a 70/30 data split, the model's predictions had a coefficient of determination (R^2) value of **0.36954** – which was below the targeted accuracy for the model.



Data Science Capstone Project -- BDT Regression
In draft
Draft saved at 8:23:33 PM
train_values_vd2rCml - Copy...
train_labels - Copy (2).csv
Join Data
Edit Metadata
Normalize Data
Normalize Data
Edit Metadata
Boosted Decision Tree Regr...
Clean Missing Data
Clean Missing Data
Split Data
Train Model
Score Model
Evaluate Model

The second model used a boosted decision tree regression and all the same data preparation steps as for the linear model. The model's characteristics were as follows:

- 20 maximum leaves per tree
- 10 minimum samples per leaf node
- 0.2 learning rate
- 100 total trees created



Metrics

Mean Absolute Error	0.176424
Root Mean Squared Error	0.221435
Relative Absolute Error	0.710426
Relative Squared Error	0.580476
Coefficient of Determination	0.419524

Trained on a 70/30 split, the model produced a coefficient of determination (R^2) value of **0.419524** – which was above the targeted accuracy threshold for the model. As such, the BDT regression model was chosen for deployment over the linear model.

Finally, the selected model was modified to be trained on the *entire* training set and then output predicted values of ‘poverty_probability’ for the test dataset. The model performed suitably in deployment as well, producing a coefficient of determination of **0.4058**.

Conclusion

The analysis of the InterMedia survey data and PPI poverty probabilities yielded many interesting findings, and in the end the development of a predictive model determined that just over 40% of the variance in ‘poverty_probability’ can be determined by characteristic variations within the feature set. While a further investigation into the underlying accuracy of the PPI values themselves may also be worthwhile, the BDT regression model developed in this analysis process is now ready for web-service deployment and could potentially be of use to researchers and activists combatting poverty in these regions.