edX | Microsoft DAT102x:
Microsoft Professional Capstone - Data Science

Project Report:
"Predicting Poverty Around the World"

Ricardo Lobato, July 2019

EXECUTIVE SUMMARY

The present document refers to the approach followed while participating on the competition "Predicting Poverty Around the World", which took place under the July 2019 edition of the Microsoft course reference DAT102x.

The **goal** of the above-mentioned competition was to **predict the probability of living under a specified poverty line** of a group of individuals from seven different countries. This was to be made based on set of socioeconomic factors associated with each individual and taking into account the known "poverty" probabilities of another group of individuals from the same countries and characterised by the same set of factors.

Approach wise, two phases were carried out:

1. **Data Exploration** – during which the data provided was analysed and internal relationships identified (by calculating summary and descriptive statistics and creating visualizations of the data), as well as some data processing was carried out.
2. **Prediction** – where the algorithm **Boosted Decision Trees Regression** was used with the data processed from the previous phase to obtain the aimed "poverty probabilities".

Both phases were performed making use of the **Azure Machine Learning Studio** tool, **Python scripts** (with Jupyterlab notebook) and **Microsoft Excel 365** software. Details regarding the data provided, the prediction model and the two phases can be found in the following sections of this report.

As a result of the above-mentioned approach, the following conclusions are presented:

- The maximum **Coefficient of Determination** obtained with the chosen prediction model was **0.4240**. This value sits above the +90% benchmark of the competition, revealing that such model performs well;
- Within the different socioeconomic factors characterising individuals, the top 5 ones having a more significant effect to the "poverty" probability are (by descending order):
  - *"education_level"*
  - *"country"*
  - *"phone_technology"*
  - *"is_urban"*
  - *"can_use_internet"*
- Notwithstanding the previous bullet point, during the several experiments made, it was observed that the model's maximum levels of accuracy were obtained when all of the socioeconomic factors were considered (meaning, no filtering).
- Also the score mentioned above was achieved when training the model with all the data available.

DATASET

The **dataset** under analysis was split into 3 files:

- One file containing **12600 observations** detailed through **58 variables**, with each observation referring to an individual and each variable to a socioeconomic factor related to that same individual;
- A second file containing the **"poverty probability"** for each individual on the previously mentioned file (thus also being made of 12600 observations);
- A third file, similar to the first one but related to 8400 individuals only (thus with 8400 observations and 58 variables). It was over this group of individuals that the poverty probability prediction was to be made.

Note: Within the Machine Learning space, it is common practice to refer to *variables* as ***features*** (in this case, the socioeconomic factors) and to the item to be predicted as ***label*** (in this case, the "poverty probability"). Going forward in this document, this will be the terminology used.

Within the dataset all possible **sensible information** had been appropriately masked. In particular:

- **Individuals** were only referred to by a sequential number, meaning there was no personal information being provided which could lead to an individual's identification;
- Similar to the previous point, the seven different **countries** were referred to only by letters, thus being impossible to related any of the data with a specific country;
- Again on the same token, **religion** was referred only by letters as well making null the possibility of identifying a certain religion.

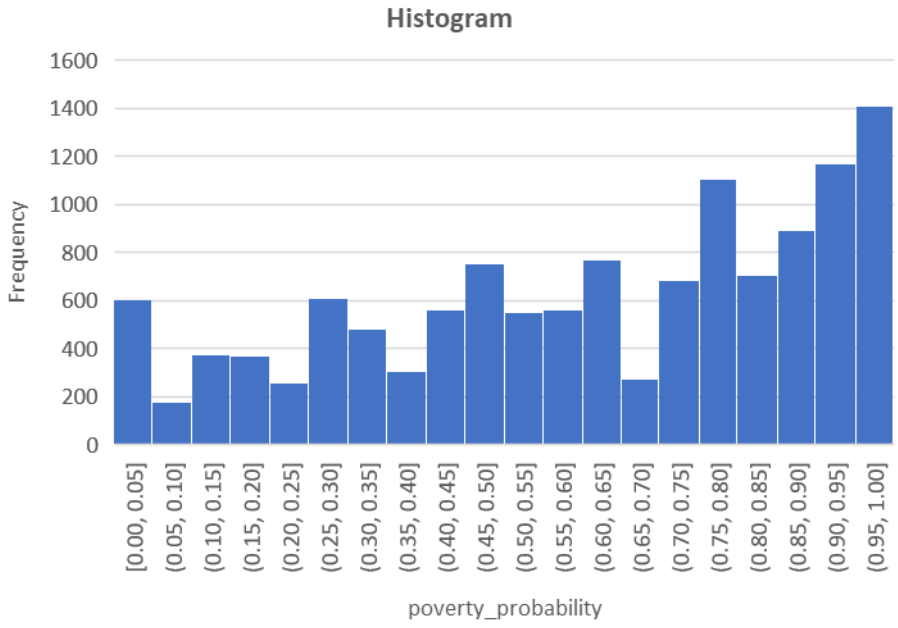
As to **gender,** female individuals were identified and although not evenly distributed against the non-female individuals, in this author's report opinion, there was no need to mask this data.

DATA EXPLORATION PHASE

LABEL ("poverty probability")

Understanding the **label** (the known "poverty probabilities") was one of the first steps of the exploration phase. To that effect and since this is a numerical variable, its Summary Statistics was calculated, and a Histogram plotted:

Summary Statistics *poverty_probability*	
Mean	0.6113
Median	0.6330
Std. Deviation	0.2915
Skewness	-0.4538
Minimum	0
Maximum	1
Count	12600



As it can be observed from the Summary Statistics, the mean and the median are not far from each other, however the Standard Deviation is considerable high which means there is a high variance amongst the poverty probabilities. In turn, the Histogram reveals a left-skewed and multi modal distribution for the poverty values.

FEATURES

TYPE IDENTIFICATION

Of the 58 features, only one, "age", was of **Numeric** type, the remaining were all **Categorical** ones. Within these, the number of categories ranged from 2 to 41 and were identified by boolean values (True-False), numbers or text strings. In summary:

Feature Type		**Count**
Numerical		1
Categorical	2 categories (True-False values)	37
	3 and more categories (Numeric or String Values)	20

NULL VALUES PROCESSING

For the 6 features containing **null values**, the adopted processing approach was as follows:

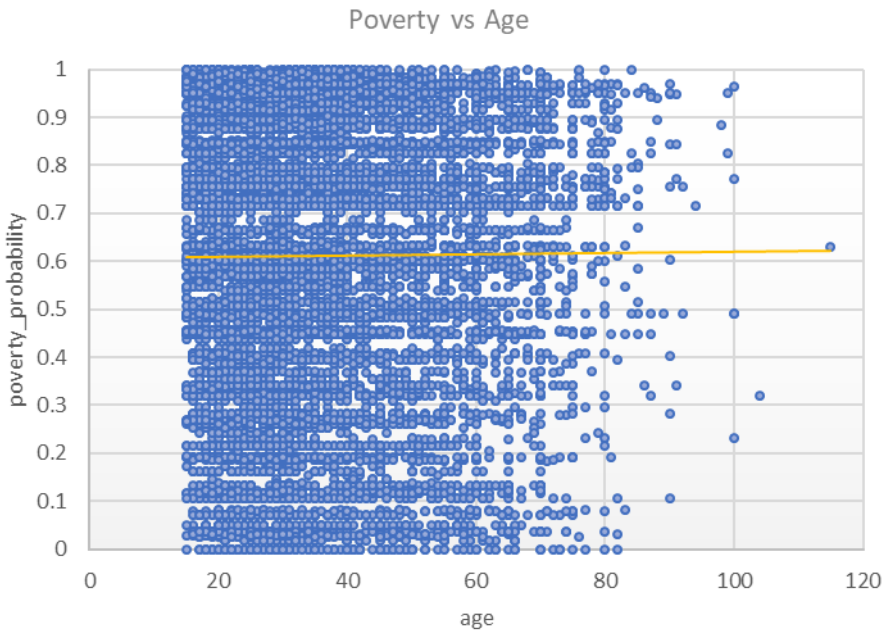
Feature	**Amount of null values**	**Processing**
"education_level"	Less than 5%	nulls replaced by "-1"
"share_hh_income_provided"	Less than 5%	nulls replaced by "-1"
"bank_interest_rate"	Almost in its entirety	Feature deleted
"mm_interest_rate"	Almost in its entirety	Feature deleted
"mfi_interest_rate"	Almost in its entirety	Feature deleted
"other_fsp_interest_rate"	Almost in its entirety	Feature deleted

ONE-HOT-ENCODING/TEXT TO INDICATOR VALUES

Being this the case of a regression problem and intending to use predictive models which require the use of numeric features only, the need to transform all the *boolean* and *string* features to *numeric* type was compulsory. Concurrently, the awareness that splitting features with a considerable number of categories into several other features (one per category) frequently results in better outcomes, made it an intended task from the early stages of the data exploration phase. The **Azure ML Studio** framework provides different ways of performing such tasks. After a few trials obtaining all the exact same result, it was decided to use the convert to "Categorical Features" option within the "Edit Metadata" module of that platform for the sake of simplicity.

Relationships – Label & Numerical Features

To understand the relationship between the label and the unique numeric feature a **scatter plot** was used:



This graphic representation shows that there is no specific relationship between the label and the "*age*" feature, thus implying the same feature was potentially of little use when trying to predict poverty levels.

Relationships – Label & Categorical Features

To study these relationships, **for each categorical feature 3 types of graphical representation** were created:

a. One **bar chart**, where each bar represented the total **count** of each category of the feature – the aim here was to understand how the different categories where distributed within the feature;

b. A second **bar chart** where each bar represented the average of poverty level of a category – useful to understand how poverty levels would vary between categories of the same feature;

c. A **"violin" chart** showing the approximate distribution of the poverty probabilities for each category – in order to obtain more detail on how the poverty levels were changing within each category.

The resulting output (53 x 3 = 159 charts) can be observed in the **Appendix 1** of this report.

Such representation allowed, not only to further **process the data**, but also to understand **which features would contribute better** for predicting poverty values.

Processing wise, based on the information provided by the count charts, for some features their categories were merged. In particular, feature "*avg_shock_strength_last_year*" was originally divided into 41 categories, and since the majority of these had very few counts, creating big imbalances within the feature itself, and also due to the fact that a few of them did not exist on the corresponding feature of Test file (the third file exposed on the "Dataset" section of this report), it was decided to merge the categories with few counts (basically all of the ones represented by a floating number) to their contiguous neighbours with higher counts. The result was a feature with 6 categories only (from 0 to 5).

Still on the processing side, some experiments of merging categories within the same feature were carried out while running the predictive model, in order to understand their effect on the prediction outcome. The charts In Appendix 1 are showing the features in their final version after the merging.

To identify which features would have a **better predicting power,** the following criteria was followed:

- The bigger the difference of average poverty between categories, the better;
- The bigger the difference in distribution shapes between categories("violin" charts), the better.

The reasoning behind such criteria was that features with categories "behaving" similarly, would be less useful to understand where are the differences which make the label vary.

Such criteria resulted in the choice of the following **initial set of 22 features** (identified by the order they are in the dataset):

- *country*
- *is_urban*
- *education_level*
- *literacy*
- *employment_type_last_year*
- *income_private_sector_last_year*
- *formal_savings*
- *has_investment*
- *num_shocks_last_year*
- *phone_technology*
- *can_call*
- *can_text*
- *can_use_internet*
- *can_make_transaction*
- *phone_ownership*
- *advanced_phone_use*
- *reg_bank_acct*
- *financially_included*
- *active_bank_user*
- *active_mm_user*
- *num_formal_institutions_last_year*
- *num_financial_activities_last_year*

For ease of reference, in Appendix 1 these features are identified with the symbol ★ .

In addition to graphic representation described above, while building the prediction model on **Microsoft Azure ML Studio**, it was used the **"*Filter Selection*"** module (on its *"Pearson Correlation"* option) which helped validate the choice of features explained here. The outcome can be found on Appendix 2 of this report.

As referred above, this was the **initial set** of features with which the "Prediction Phase" of this competition approach took off. However, as explained in the following section of this report, it ended up not being the final one used to obtain the best predicting score.

PREDICTION PHASE

The choice of the predicting model came out from initial trials with 2 different types of classical regression models:

- Linear Regression with L1 Normalization;
- Boosted Decision Trees Regression.

The trials were made using 70% of the data to train the models and the remaining 30% to test their results. Since in all trials the best results were obtained with the **Boosted Decision Trees Regression** model, it ended up being the clear winner to take forward.

Once the model had been chosen, a few runs to tune its hyperparameters were carried out, all with the initial set of 22 features obtained from the Data Exploration and using the data split above mentioned. In face of unsatisfactory outcomes and some difference against the results obtained from the competition submissions, a **10 folds Cross Validation** step was introduced. At this stage the data split was set to 80% for training and validating and the remaining 20% for pre-testing (or in this case for double checking) before submission.

With a Cross Validation step in place helping to spot potential "overfitting" situations, several experiments were made, not only by changing **the number of features** and **the amount of data** to train the model with, but also by **merging some categories** within features.

As a result, apart from merging the categories of the feature "*avg_shock_strength_last_year*" previously described in this report, none of the other cases tried obtained better scores. Regarding to the number of features, the best results were observed when the model was **trained with all the features** of the dataset. This somewhat unexpected situation might have justification on the fact that the chosen model already includes a feature selection tool, which according to the documentation from the Azure ML Studio is advisable to make full use by avoiding any previous manual selection. Finally, as to the amount of data used, the bigger the amount used, the better. In fact, making also use of the 20% of data initially left for pre-testing before submission, lead to significant higher scores.

It was also observed that changing the "seed feeds" controlling the random choice of data and or model parameters, would impact the model scores.

Once the learning from all these experiments was incorporated, the model Hyperparameters were finally tuned.

Upon submission, the final tuned model obtained a **Coefficient of Determination** of **0.4240**. Internally (before submission) on that metric the very same model scored a value of **0.4417** (0.0177 of difference).

CONCLUSIONS

In summary, from the approach taken the following has been concluded:

- Considering the competition metric's maximum value of 1 (Coefficient of Determination), the final model obtained allows some level of label prediction;
- Based on the benchmarks of the competition, the same predictive model is performing well;
- It was clearly identified which features contribute best for predicting the label;
- Notwithstanding the previous point, the best results occurred when the model was trained with all the features of the dataset, as well as by making use of the entire data available.

edX | Microsoft DAT102x:
Microsoft Professional Capstone - Data Science

APPENDIX I

Project Report:
"Predicting Poverty Around the World"

Ricardo Lobato, July 2019

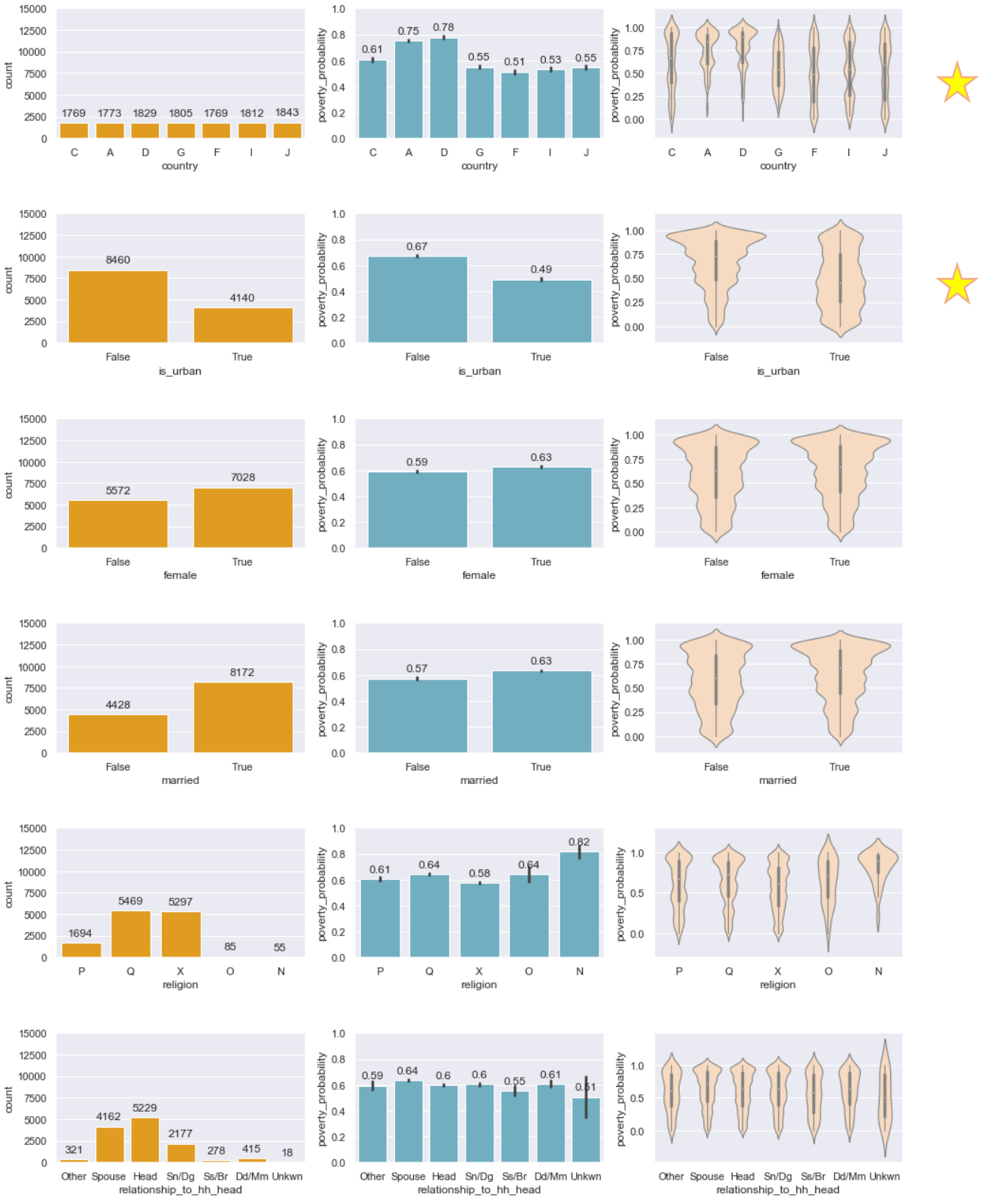
APPENDIX 1

Project Report: "Predicting Poverty Around the World"

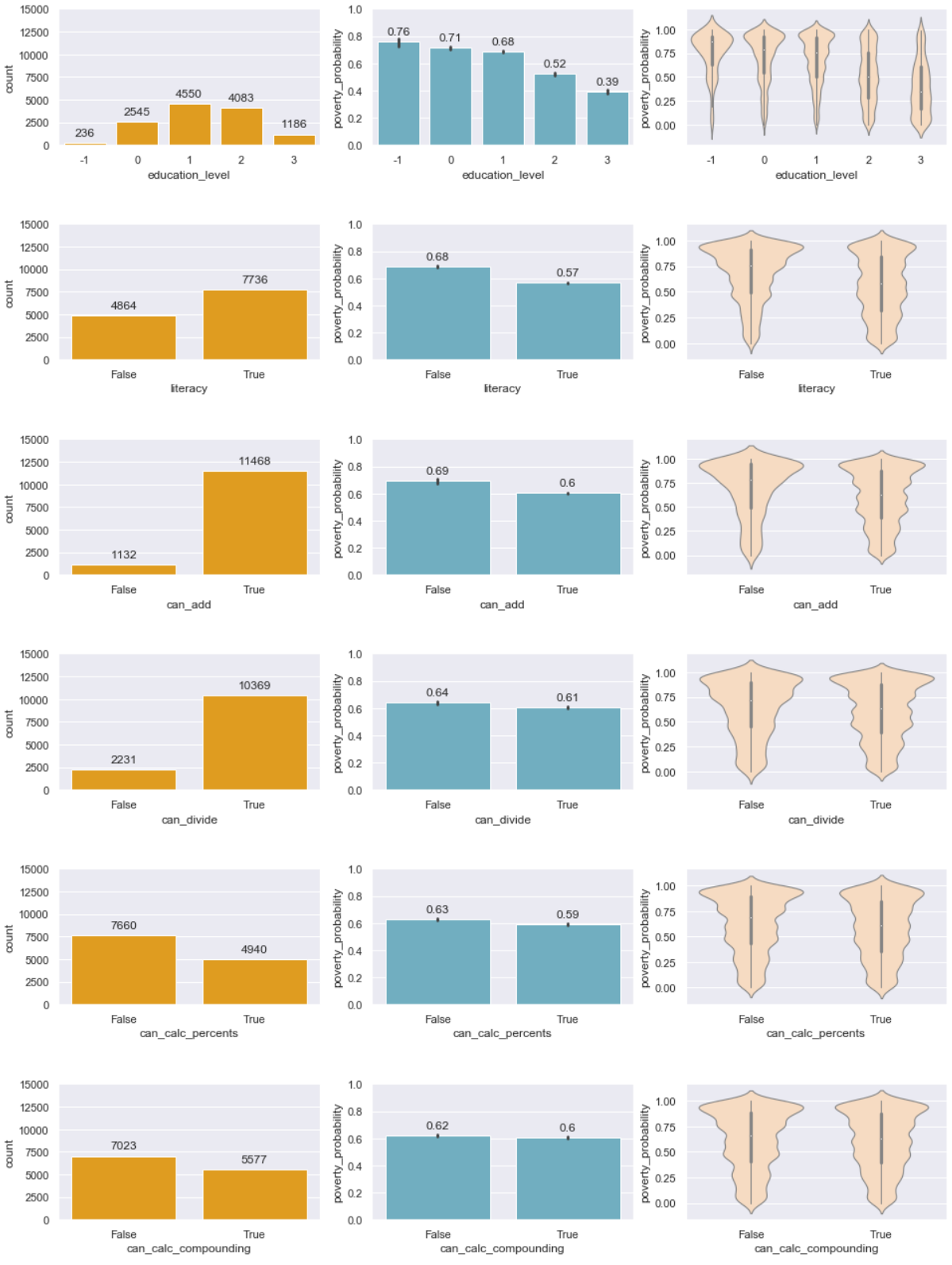
edX | Microsoft DAT102x: Microsoft Professional Capstone - Data Science

Ricardo Lobato, July 2019

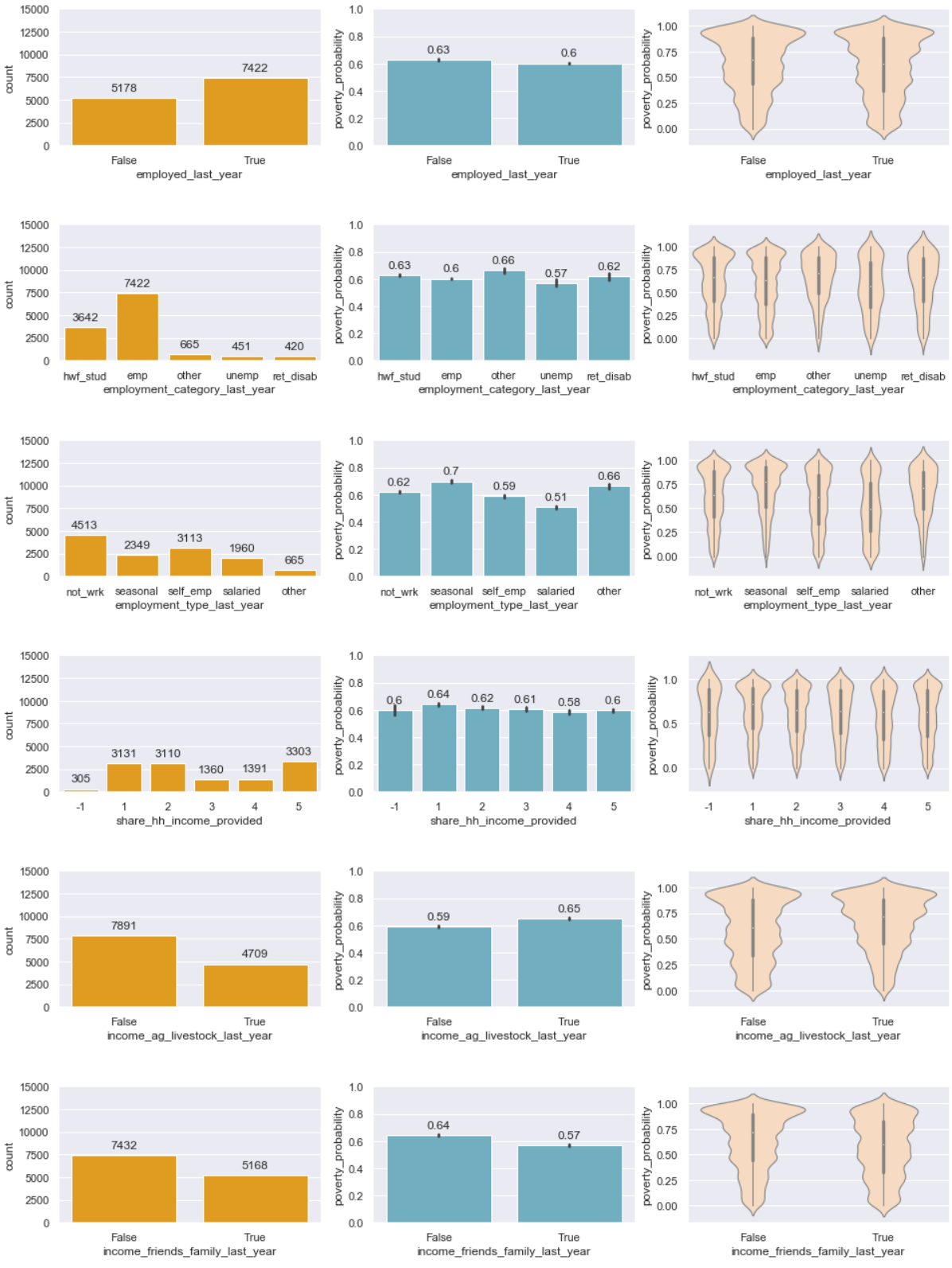
Count Plot | **Poverty Average Plot** | **Violin Plot**



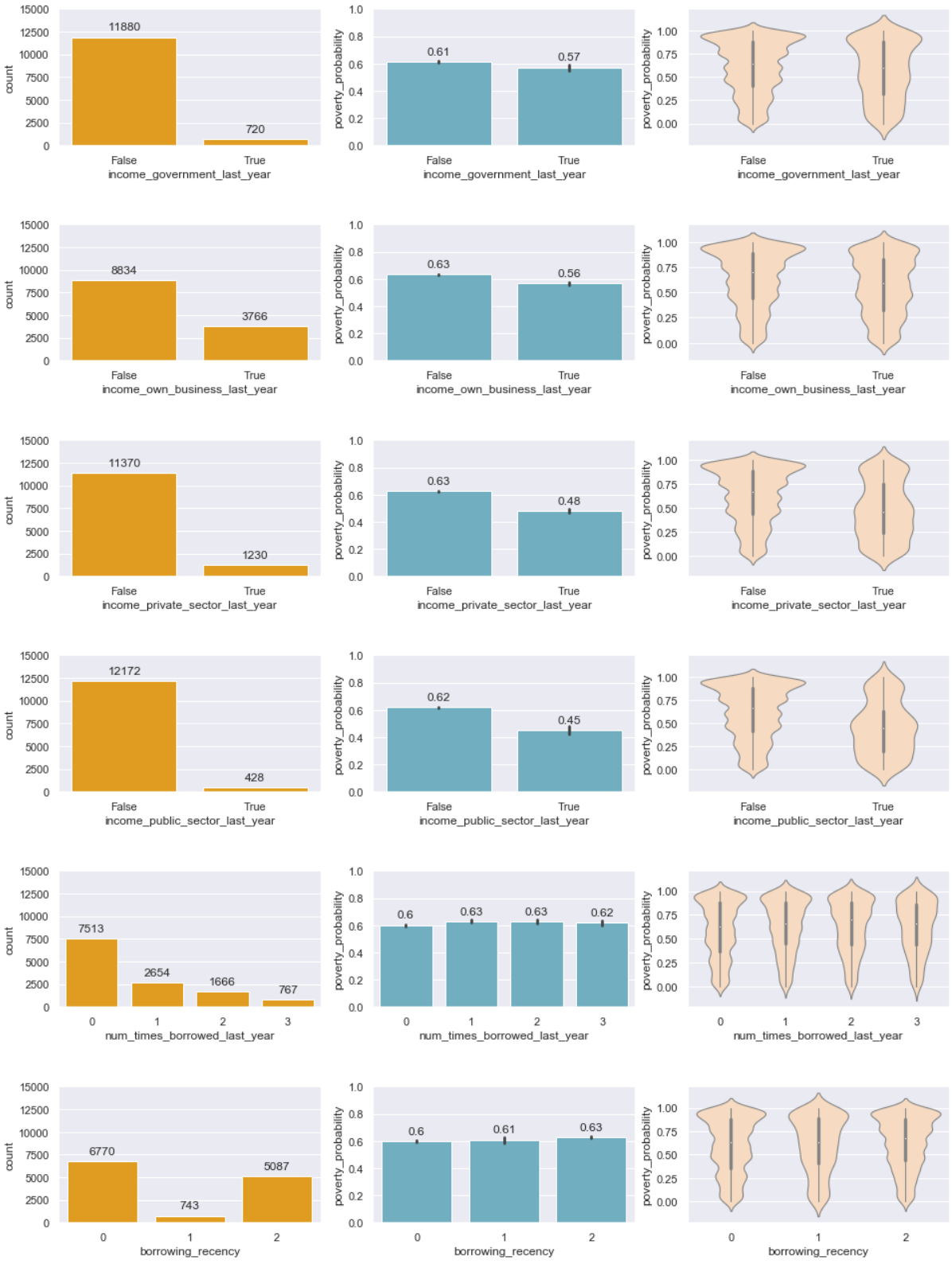
Count Plot | **Poverty Average Plot** | **Violin Plot**



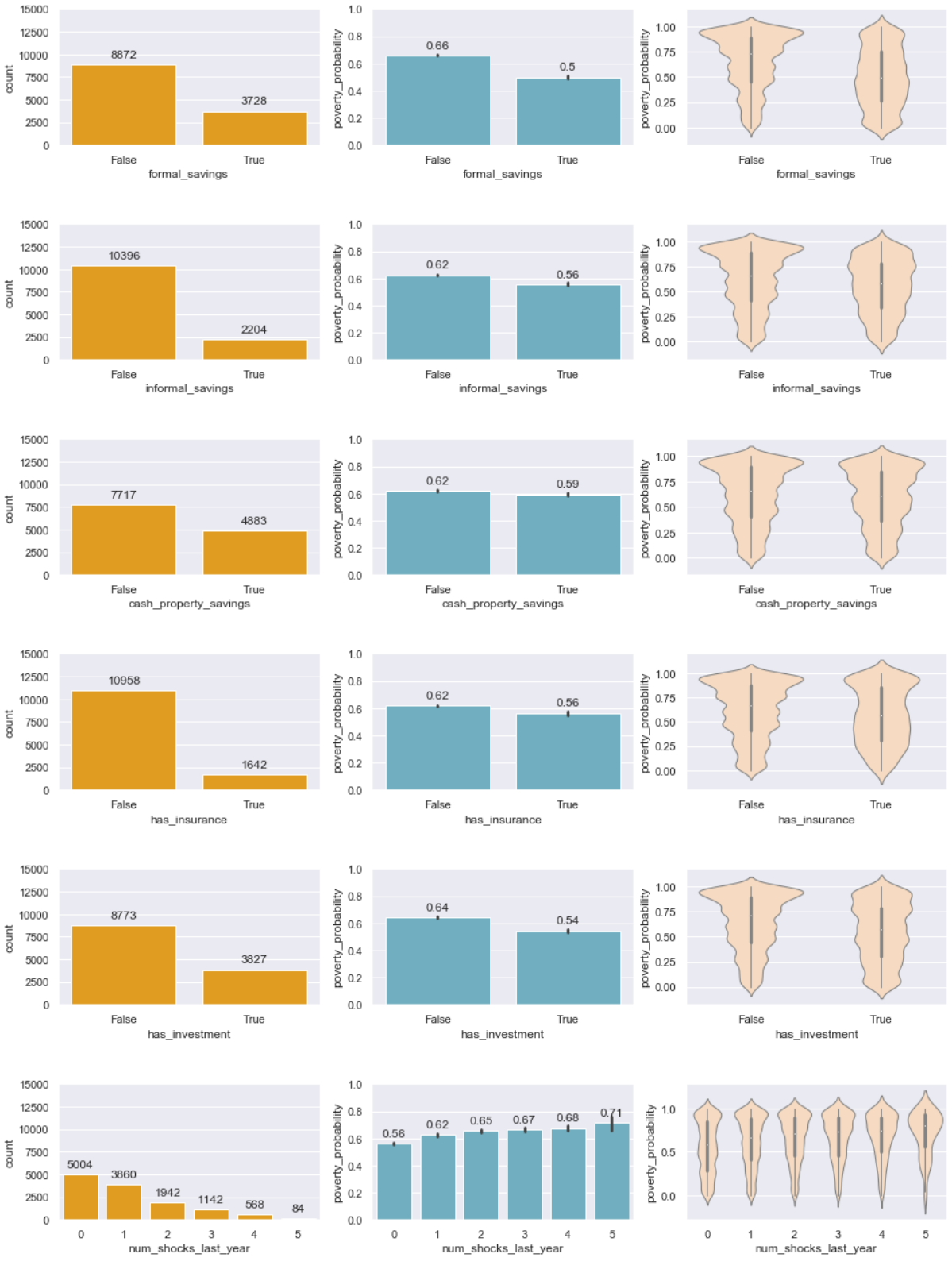
Count Plot | **Poverty Average Plot** | **Violin Plot**



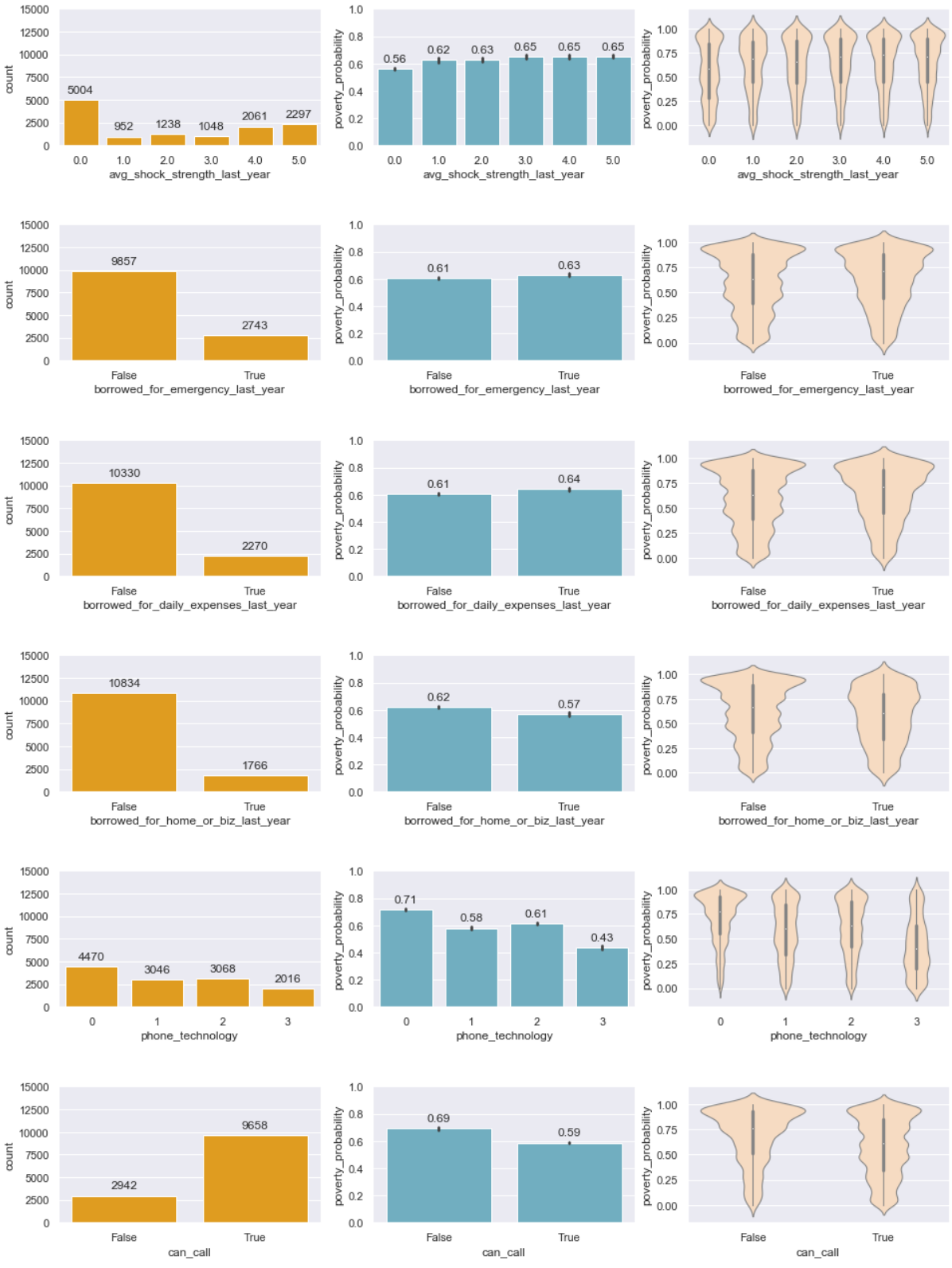
Count Plot | **Poverty Average Plot** | **Violin Plot**



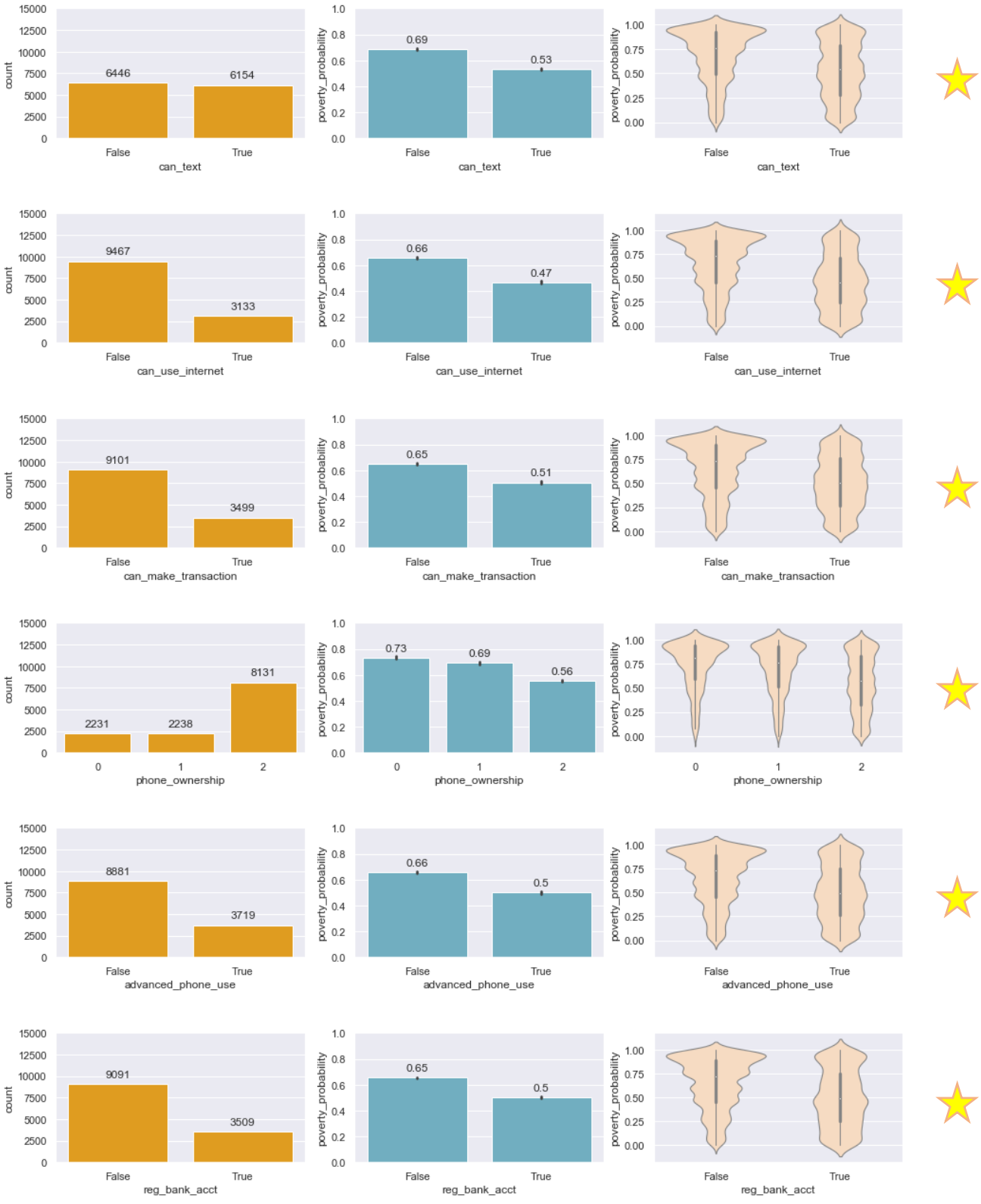
Count Plot | **Poverty Average Plot** | **Violin Plot**



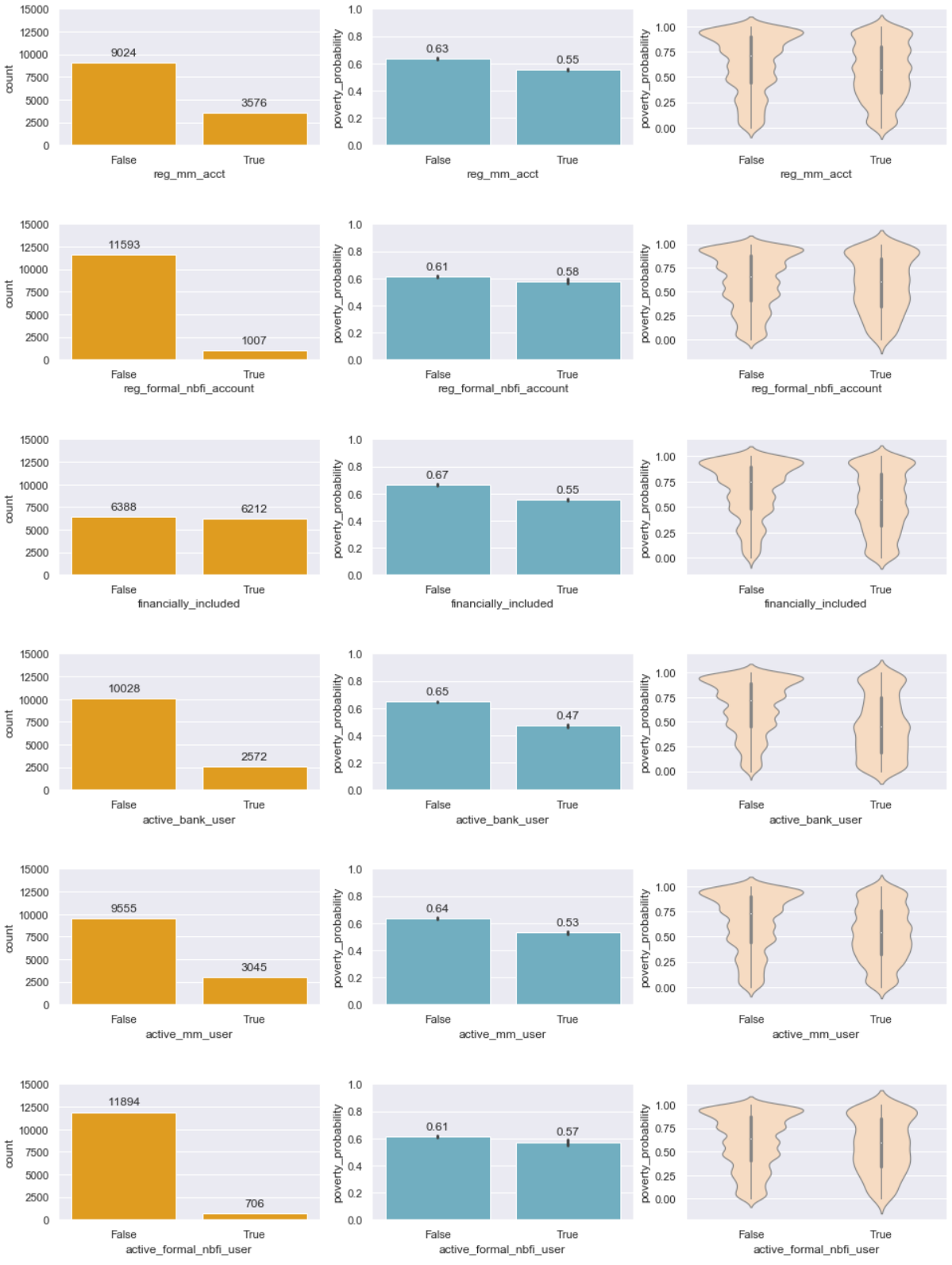
Count Plot | **Poverty Average Plot** | **Violin Plot**



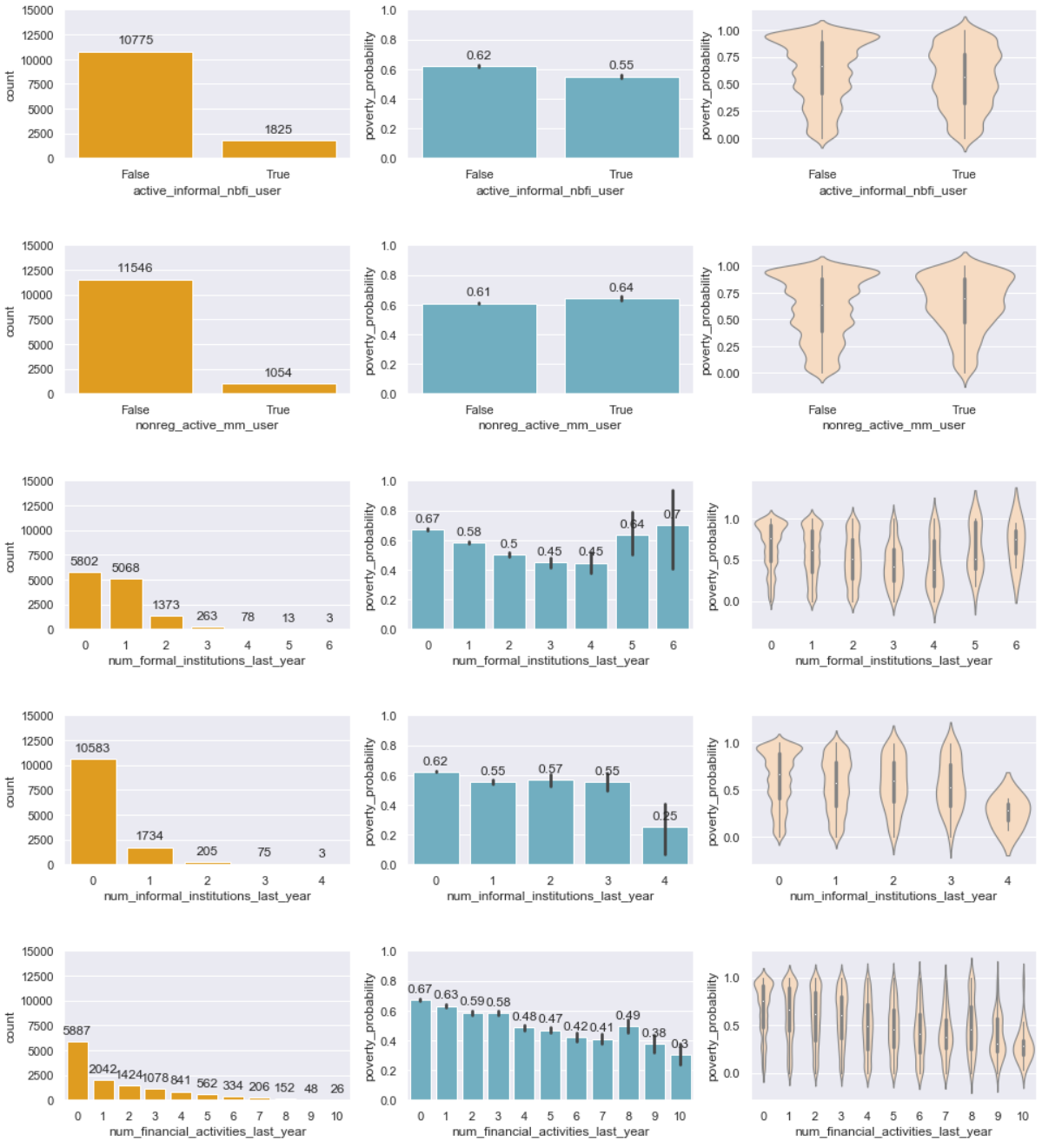
Count Plot | **Poverty Average Plot** | **Violin Plot**



Count Plot | **Poverty Average Plot** | **Violin Plot**



Count Plot | **Poverty Average Plot** | **Violin Plot**



edX | Microsoft DAT102x:
Microsoft Professional Capstone - Data Science

APPENDIX II

Project Report:
"Predicting Poverty Around the World"

Ricardo Lobato, July 2019

APPENDIX 2

Project Report: "Predicting Poverty Around the World"

edX | Microsoft DAT102x: Microsoft Professional Capstone - Data Science

Ricardo Lobato, July 2019

FEATURES RANKED BY "PEARSON CORRELATION" VALUE
(FEATURE SELECTION MODULE – AZURE ML STUDIO)

Id (on the dataset)	Group	Feature Name	FeatureType	Feat. Select. - bin to Num - avg_shock binned - hot-one
8	Education	education_level	Num_Cat_4	0.367
1	Demographics	country	String_Cat_7	0.348
40	Phone	phone_technology	Num_Cat_4	0.323
2	Demographics	is_urban	Cat_2	0.290
43	Phone	can_use_internet	Cat_2	0.284
58	Financial Inclusion	num_financial_activities_last_year	Num_Cat_11	0.269
42	Phone	can_text	Cat_2	0.262
45	Phone	phone_ownership	Num_Cat_3	0.260
26	Economic	formal_savings	Cat_2	0.253
46	Phone	advanced_phone_use	Cat_2	0.248
51	Financial Inclusion	active_bank_user	Cat_2	0.245
47	Financial Inclusion	reg_bank_acct	Cat_2	0.235
44	Phone	can_make_transaction	Cat_2	0.224
56	Financial Inclusion	num_formal_institutions_last_year	Num_Cat_7	0.219
16	Employment	employment_type_last_year	String_Cat_5	0.199
9	Education	literacy	Cat_2	0.199
50	Financial Inclusion	financially_included	Cat_2	0.192
52	Financial Inclusion	active_mm_user	Cat_2	0.156
30	Economic	has_investment	Cat_2	0.155
41	Phone	can_call	Cat_2	0.152
36	Economic	avg_shock_strength_last_year	Num_Cat_41	0.151
22	Employment	income_private_sector_last_year	Cat_2	0.147
35	Economic	num_shocks_last_year	Num_Cat_6	0.145
19	Employment	income_friends_family_last_year	Cat_2	0.124
48	Financial Inclusion	reg_mm_acct	Cat_2	0.122
6	Demographics	religion	String_Cat_5	0.113
23	Employment	income_public_sector_last_year	Cat_2	0.104
21	Employment	income_own_business_last_year	Cat_2	0.104
18	Employment	income_ag_livestock_last_year	Cat_2	0.103
5	Demographics	married	Cat_2	0.098
57	Financial Inclusion	num_informal_institutions_last_year	Num_Cat_5	0.090
27	Economic	informal_savings	Cat_2	0.087
54	Financial Inclusion	active_informal_nbfi_user	Cat_2	0.086
10	Education	can_add	Cat_2	0.086
17	Employment	share_hh_income_provided	Num_Cat_5	0.066
7	Demographics	relationship_to_hh_head	String_Cat_7	0.065
29	Economic	has_insurance	Cat_2	0.062
12	Education	can_calc_percents	Cat_2	0.062
15	Employment	employment_category_last_year	String_Cat_5	0.062
39	Economic	borrowed_for_home_or_biz_last_year	Cat_2	0.060
4	Demographics	female	Cat_2	0.058
28	Economic	cash_property_savings	Cat_2	0.047
38	Economic	borrowed_for_daily_expenses_last_year	Cat_2	0.046
24	Economic	num_times_borrowed_last_year	Num_Cat_4	0.045
25	Economic	borrowing_recency	Num_Cat_3	0.045
11	Education	can_divide	Cat_2	0.044
14	Employment	employed_last_year	Cat_2	0.042
49	Financial Inclusion	reg_formal_nbfi_account	Cat_2	0.034
55	Financial Inclusion	nonreg_active_mm_user	Cat_2	0.033
20	Employment	income_government_last_year	Cat_2	0.033
53	Financial Inclusion	active_formal_nbfi_user	Cat_2	0.033
37	Economic	borrowed_for_emergency_last_year	Cat_2	0.033
13	Education	can_calc_compounding	Cat_2	0.026
3	Demographics	age	Num	0.007

Legend: *Feature selected during Data Exploration Phase*