Project 1: Draft

Econ 1680: Machine Learning, Text Analysis, and Economics

Nadya Tan

March 8, 2024

1 Introduction

Mobile banking is an innovative solution for improving financial inclusion, and has been a key driver of financial inclusion (Purva Khera and Sahay (2021)); but the use of this technology is still very limited in developing countries (Hilal and Varela-Neira (2022)). To better inform what policy levers can be used to promote adoption of mobile banking and increase financial inclusion, it can be useful to study what features are the most predictive of active mobile banking usage to see where policymakers can target to increase financial inclusion. Predicting which households are likely to be using mobile banking could also help policymakers determine whether a given form of policy (e.g. mobile cash payouts) is likely to be accessible by the target audience. Existing research in this field tends to focus mainly on psychological attributes such as trust, perceived effort or perceived usefulness, and how those attributes affect the takeup of mobile banking (Yin and Lin (2022)). We also know, rather predictably, that consumers with higher income levels are more likely to use digital services, than those with lower income levels (Veríssimo (2016)). However, not much research has been done on more specific household economic indicators, such as spending or saving habits, are most predictive of active mobile banking usage, which is what this project seeks to address.

Research Question: Predicting active mobile banking usage at a household level based on data on household demographics and other financial activities

2 Data Sources and Descriptions

I used data from the World Bank's Global Financial Inclusion (Global Findex) database. A survey was conducted at a household level across 139 countries that collects data on the financial habits (banking, saving, payments) of respondents. Definitions of all variables that are used can be found in the data dictionary in the appendix.

The following tables go over the main variables that I have selected to be of interest. The data here has been processed, in the following steps:

- Education and income quartile variables were initially given as a scale (Table 1) - I applied one-hot encoding to create separate dummy variables (Table 7)
- One-hot encoding was also applied to generate dummy variables for each country
- Several variables (e.g. uses a debit/ credit card) were recorded as 1 if yes, 2 if no and 3/4 if the response was Na. I converted these to be a 0-1 scale, with a Na for 3 or 4.
- Other variables (e.g. receive wages) were recorded as different values depending on the method of transfer (cash, account, other) - I made it so that we now have two variables out of this. One was a scaled variable (see Table 6), where I coded a 2 if the transfer was done using a more high-tech method (account transfer), 1 if the transfer was done using cash, and 0 if there was no transfer. The initial variable was coded as a 0-1 variable, where the variable is 1 if any transfer was done, and 0 if no transfers were done.

Table 1: Demographic Statistics

	Female	Age	Educ	Inc_q	Emp_in	Urbanicity
Count	143887	143420	143887	143887	143887	143887
Mean	0.532	41	1.97	3.23	0.64	0.304
Stdv	0.50	17.3	0.73	1.42	0.48	0.46
Min	0	15	1	1	0	0
Median	1	38	2	3	1	0
Max	1	99	5	5	1	1

Table 2: Financial Habits Statistics

	Account	Account_fin	Account_Mob	Saved	Borrowed	Receive Wages
Count	143887	143887	82706	143887	143887	143114
Mean	0.71	0.66	0.26	0.54	0.53	0.38
Stdv	0.45	0.48	0.44	0.50	Fin4	Fin5
Min	0	0	0	0	0.50	0.49
Median	1	1	0	1	1	0
Max	1	1	1	1	1	1

3 Method

I will be comparing the following methods: OLS with country fixed effects, LASSO and ridge regression to see which has the most predictive power.

Since the relationships that I will be estimating are linear, they will be estimated using the following equation:

$$y_{ci} = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \beta_3 X3_i + ... + \theta_c + \varepsilon_{ci} \tag{1}$$

Table 3: Financial Habits Statistics Continued

	Receive Transfers	Receive Pension	Receive Agriculture	Pay Utilities	Remittances
Count	143067	143298	113897	143145	45438
Mean	0.194	0.121	0.134	0.579	0.917
Stdv	0.39	0.326	0.341	0.493	0.275
Min	0	0	0	0	0
Median	0	0	0	1	1
Max	1	1	1	1	1

Table 4: Financial Habits Statistics Continued

	Fin2	Fin4	Fin5	Fin6	Fin7	Fin8	Fin9	Fin10	Fin13a
Count	142816	72811	88782	88807	88619	32569	88556	88534	17634
Mean	0.51	0.74	0.59	0.65	0.37	0.83	0.79	0.80	0.74
Stdv	0.499	0.44	0.49	0.48	0.48	0.38	0.41	0.40	0.44
Min	0	0	0	0	0	0	0	0	0
Median	1	1	1	1	0	1	1	1	1
Max	1	1	1	1	1	1	1	1	1

where, $y_ci \in \{0, 1\}$ is the outcome we are interested in (active mobile banking usage measured at the household level) and $X1_i, X2_i...$ etc are various independent variables that were described above. θ_c represents the country fixed effects. An assumption that we make in OLS is that in order to interpret the coefficients as being unbiased estimates of the marginal effects on the dependent variable, conditional independence must hold (i.e. no omitted variable bias).

$$E[\varepsilon_i|X_i] = 0 \tag{A1}$$

As evident from the section above, there are a lot of variables that I will be using to try to predict active mobile banking usage. To handle this, I will be running Lasso and Ridge Regressions to pick out variables that are the most powerful predictors of active mobile banking usage. These two regressions work by adding a penalty term that forces a tradeoff between having more regressors and smaller coefficients vs. fewer regressors and larger coefficients. In comparison to OLS, these regressions aim to minimize the squared error, as well as the penalty term, which can be seen in the following equations:

Ordinary Least Squares (OLS):

$$\hat{\beta}^{\text{OLS}} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

Lasso (L1 Penalty):

$$\hat{\beta}^{\text{Lasso}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Table 5: Financial Habits Statistics Continued

	Mobile Owner	Internet Access	Any Digital Payment	Merchant Pay Digital
Count	143750	143296	143887	114281
Mean	0.88	0.70	0.65	0.325
Stdv	0.323	0.456	0.477	0.47
Min	0	0	0	0
Median	1	1	1	0
Max	1	1	1	1

Table 6: Financial Habits Statistics (Intensity Scaled)

	Wages	Transfers	Pensions	Agriculture	Utilities	Remittances
Count	140726	138840	140944	112613	136885	37588
Mean	0.67	0.316	0.20	0.155	0.90	1.63
Stdv	0.90	0.714	0.59	0.44	0.88	0.66
Min	0	0	0	0	0	0
Median	0	0	0	0	1	2
Max	2	2	2	2	2	2

Ridge (L2 Penalty):

$$\hat{\beta}^{\text{Ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

Due to the inclusion of the penalty term, Lasso and Ridge Regression coefficients will be biased. It is also important to note that all 3 regressions assume a linear relationship between the independent variables and active mobile banking usage. I will test out all 3 regressions using both just the one-hot encoded data, and using the scaled data for the relevant variables to see which ones perform better. Once the models have been fine-tuned to a satisfactory level, I will evaluate their performances on the held-out test set.

4 Results or Expected Results

Table 8 shows an initial comparison of coefficients from an initial run of OLS, Lasso and Ridge Regressions. Unfortunately, when dropping observations for which any one of the variables had a null value, the size of the dataset collapsed from 143887 to 1600. Out of 143887 observations, only 17634 of them have an observation for the dependent variable (Fin13a). These regressions were also only run using the one-hot encoded versions of the variables (and not the scaled version). Any of these factors could be contributing to how the Lasso regression coefficients are all 0. Moving forward, I will try running the regressions using the scaled versions of the variables, dropping variables where there are fewer observations (e.g. remittances) or looking at filtering the dataset (e.g. only looking at people who have a mobile account and internet access, instead of tossing those in as control variables).

Table 7: Education and Income Variables (One-hot Encoded)

	educ1	educ2	educ3	incq1	incq1	incq3	incq4	incq5
Count	143132	143132	143132	143887	143887	143887	143887	143887
Mean	0.269	0.505	0.22	0.17	0.17	0.19	0.21	0.26
Stdv	0.4438	0.499	0.42	0.37	0.38	0.39	0.41	0.44
Min	0	0	0	0	0	0	0	0
Median	0	1	0	0	0	0	0	0
Max	1	1	1	1	1	1	1	1

5 Conclusion

Pending

6 Appendix

Data Dictionary

- **Female**: Gender of the individual
- **Age**: Age of the individual
- **Educ**: Education level, where 1 (Primary), 2 (Secondary), 3 (Tertiary)
- **Inc_q**: Within-economy household income quintile
- **Emp_in**: Employment status, indicating whether the individual is in the workforce
- **Urbanicity_f2f**: Urbanicity status, indicating whether the individual resides in a rural area
- **Account**: Whether the individual has an account
- **Account_fin**: Whether the individual has an account at a financial institution
- **Account_Mob**: Whether the individual has a mobile money account
- **Saved**: Whether the individual saved in the past year
- **Borrowed**: Whether the individual borrowed in the past year
- **Receive Wages**: Whether the individual received a wage payment
- **Receive Transfers**: Whether the individual received a government transfer payment
- **Receive Pension**: Whether the individual received a government pension payment
- **Receive Agriculture**: Whether the individual received a payment for the sale of agricultural goods

- **Pay Utilities**: Whether the individual paid a utility bill
- **Remittances**: Whether the individual made or received a domestic remittance payment
- **Fin2**: Whether the individual has a debit card
- **Fin4**: Whether the individual used a debit card
- **Fin5**: Whether the individual used a mobile phone/internet to access an account
- **Fin6**: Whether the individual used a mobile phone/internet to check account balance
- **Fin7**: Whether the individual has a credit card
- **Fin8**: Whether the individual used a credit card
- **Fin9**: Whether the individual made any deposit into an account
- **Fin10**: Whether the individual withdrew from an account
- **Fin13a**: Whether the individual used a mobile money account two or more times a month
- **Mobile Owner**: Whether the individual owns a mobile phone
- **Internet Access**: Whether the individual has internet access
- **Any Digital Payment**: Whether the individual made or received a digital payment
- **Merchant Payment Digital**: Whether the individual made a digital merchant payment

References

Hilal, Ashraf and Concepción Varela-Neira (2022) “Understanding Consumer Adoption of Mobile Banking: Extending the UTAUT2 Model with Proactive Personality,” *Sustainability.*

Purva Khera, Sumiko Ogawa, Stephanie Ng and Ratna Sahay (2021) “Measuring Digital Financial Inclusion in Emerging Market and Developing Economies: A New Index,” *IMF Working Paper.*

Veríssimo, José Manuel Cristóvão (2016) “Enablers and restrictors of mobile banking app use: A fuzzy set qualitative comparative analysis (fsQCA),” *Journal of Business Research.*

Yin, Lan-Xiang and Hsien-Cheng Lin (2022) “Predictors of customer’s continuance intention of mobile banking from the perspective of the interactive theory,” *Economic Research-Ekonomska Istraživanja.*

Table 8: Comparison of Coefficients Across OLS, Lasso and Ridge

	var	OLS	Lasso	Ridge
0	female	-0.017341	-0.000000	-0.008361
1	age	-0.000803	-0.000000	-0.009266
2	emp_in	-0.021976	-0.000000	-0.006511
3	urbanicity_f2f	-0.016858	0.000000	-0.005941
4	account	0.117257	0.000000	0.000000
5	account_fin	0.117257	0.000000	0.000000
6	account_mob	0.117257	0.000000	0.000000
7	fin2	0.117257	0.000000	0.000000
8	fin4	-0.000162	-0.000000	-0.000030
9	fin5	0.047076	0.000000	0.012692
10	fin6	0.000741	0.000000	0.000373
11	fin7	0.117257	0.000000	0.000000
12	fin8	-0.022816	-0.000000	-0.007857
13	fin9	-0.036945	-0.000000	-0.008876
14	fin10	0.096816	0.000000	0.027222
15	saved	0.010266	0.000000	0.004385
16	borrowed	-0.019680	-0.000000	-0.004203
17	receive_wages	0.000816	0.000000	0.000591
18	receive_transfers	-0.022529	-0.000000	-0.009528
19	receive_pension	0.049998	0.000000	0.014627
20	receive_agriculture	0.013327	0.000000	0.005242
21	pay_utilities	0.043454	0.000000	0.017188
22	remittances	-0.010861	0.000000	-0.001879
23	mobileowner	0.019491	-0.000000	0.002350
24	internetaccess	0.040449	-0.000000	0.009558
25	anydigpayment	0.117257	0.000000	0.000000
26	merchantpay_dig	0.033532	-0.000000	0.008292
27	educ2	0.025811	0.000000	0.010307
28	educ3	0.020307	-0.000000	0.006969
29	incq2	-0.065579	-0.000000	-0.016255
30	incq3	0.008089	0.000000	0.003377
31	incq4	-0.017804	-0.000000	-0.005997
32	incq5	-0.003270	0.000000	0.000172