Project 1: Predicting Mobile Banking Usage Using Financial Inclusion Microdata

Econ 1680: Machine Learning, Text Analysis, and Economics

Nadya Tan

May 9, 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. There is a wealth of research in this field focusing 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)). When it comes to economic indicators, we know that, rather predictably, that consumers with higher income levels are more likely to use digital services, than those with lower income levels (Veríssimo (2016)). There has also been evidence that education, internet penetration and remittances are also correlated with the use of mobile banking in Sub-Saharan Africa (Asongu (2015)). Lastly, a study done by Fiserv on credit union members showed that mobile banking users tend to be younger, on payroll and have a higher debit/ credit usage (Fiserv (2016)).

This study aims to contribute to the existing body of research by leveraging machine learning techniques to identify the most powerful predictors of mobile banking usage, and to explore if introducing non-linear prediction techniques help improve the prediction of mobile banking usage.

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.

Tables 1-6 go over the main variables that will be used in this project. With the exception of age, all of the variables either were already on a 0-1 scale, or were one-hot encoded (education and income variables) or converted into a 0-1 scale. Definitions of each variable can be found in the Data Dictionary below.

The target variable is account_mob, which is a 0-1 variable on whether the individual has a mobile money account.

3 Method

I used two main methods in this paper. The first method is an OLS regression that was refined using Lasso and Ridge regressions, and the second method is a histogram gradient boost.

The first method estimates linear relationships between the variables using the following equation:

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

where, $y_c i \in \{0, 1\}$ 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. 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. I first shortlisted a list of variables using what we learnt from prior literature, and then ran Lasso and Ridge Regressions to pick out variables that are the most powerful predictors of 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):

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

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 regressions would push the coefficients of variables with less predictive power towards 0. I would then remove these less predictive variables, and then run another OLS regression with the refined list of variables, and tested its accuracy on a held-out test set. Since OLS predicts a continuous value between 0 and 1, I converted values ≥ 0.5 to be 1, and < 0.5 to be 0.

The other method is Histogram Gradient Boost, an ensemble machine learning method, where tree models are added to an ensemble sequentially, and each subsequent added tree model aims to correct the errors of the pre-existing models. The histogram-based gradient boost works by bucketing continuous feature values into discrete bins, and using those bins to construct histograms during training, making it faster and more efficient than traditional gradient boosts. Although this dataset was large, since most of the variables (apart from age) are discrete, the time saving would not be significant, but it was chosen since Sklearn's implementation of it was robust to NaN values in the dataset. The accuracy of this model was also tested on a held-out test set. To identify which variables were the most important in classification, the variables were sorted based on permutation feature error, which measures the increase in model error when the values of a feature are randomly shuffled. Important features should lead to a greater increase in error when their values are shuffled randomly compared to less important features.

4 Results

For the OLS regressions, I started off with a variable list that included age, gender, education, income, receiving wages, internet access, owning a mobile phone, paying utilities and debit/ credit usage. These variables were hand-selected based on the findings of prior literature. Table 7 shows the initial OLS results, and Table 8 shows the coefficients of the variables from the OLS, Lasso and Ridge regressions. Based on these results, although having an account is significantly correlated with having a mobile money account, Lasso and Ridge both deemed it to be an unimportant variable. I hence removed it from the list of variables, resulting in Table 9. Based on the coefficients, being a mobile owner is the strongest predictor of having a mobile banking account (22 percentage points more likely), followed by receiving wages (15 percentage points more likely) and then being in the highest income quintile (15 percentage points more likely).

The predictive accuracy of the refined OLS model on the held-out test set was 0.65, and Figure 1 shows a confusion matrix, which shows that the model is equally good at predicting the right value for both actual negative and actual positive samples.

Histogram gradient boost allowed us to include more variables, since its robustness against NaN values meant that we did not have to drop too many samples to make sure that the models could be run. The features are listed in order of importance in Table 10. Interestingly, account and account_fin ranks highly, and so does remittances, saving, having

internet access, receiving transfers and being of a high income quintile. The accuracy of this model was 0.87, which outperformed the OLS, Figure 2 shows the confusion matrix - which shows that this model does a better job at predicting actual negative samples than actual positive samples.

5 Conclusion

The histogram gradient boost model greatly outperformed refined OLS in terms of accuracy, and this could be for several reasons. 1) It was able to utilize a greater sample due to its robustness to NaN values, 2) it accounts for non-linearity and 3) It was able to utilize a greater volume of variables. In terms of the most powerful predictors of mobile banking usage, both the histogram gradient boost and refined OLS agreed that receiving income (whether through wages, remittances or transfers) and income quintiles, as well as age and internet access, are some of the best predictors of whether a household uses mobile banking, and both models agreed that variables such as gender and education are less important predictors.

References

Asongu, Simplice (2015) "Conditional Determinants of Mobile Phones Penetration and Mobile Banking in Sub-Saharan Africa," *African Governance and Development Institute (AGDI)*.

Fiserv (2016) "Mobile Banking Adoption: Where Is the Revenue for Financial Institutions?".

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*.

6 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

7 Tables and Figures

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

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
Count	142816	72811	88782	88807	88619	32569	88556	88534
Mean	0.51	0.74	0.59	0.65	0.37	0.83	0.79	0.80
Stdv	0.499	0.44	0.49	0.48	0.48	0.38	0.41	0.40
Min	0	0	0	0	0	0	0	0
Median	1	1	1	1	0	1	1	1
Max	1	1	1	1	1	1	1	1

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: 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

Table 7: OLS Summary Result (Initial)

Dep. Variable:	account_mob	**R-squared:**	-0.119
Model:	OLS	**Adj. R-squared:**	-0.121
Method:	Least Squares	**F-statistic:**	-41.96
Date:	Mon, 06 May 2024	**Prob (F-statistic):**	1.00
Time:	13:12:13	**Log-Likelihood:**	-3751.6
No. Observations:	5553	**AIC:**	7533.
Df Residuals:	5538	**BIC:**	7633.
Df Model:	14		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
age	-0.0062	0.000	-12.821	0.000	-0.007	-0.005
female	-0.0018	0.013	-0.139	0.890	-0.027	0.024
educ_2	-0.0196	0.025	-0.793	0.428	-0.068	0.029
educ_3	-0.0447	0.026	-1.718	0.086	-0.096	0.006
inc_q_2	-0.0158	0.030	-0.527	0.598	-0.075	0.043
inc_q_3	0.0352	0.028	1.236	0.217	-0.021	0.091
inc_q_4	0.0641	0.027	2.400	0.016	0.012	0.116
inc_q_5	0.1058	0.025	4.168	0.000	0.056	0.155
account_fin	0.3208	0.054	5.908	0.000	0.214	0.427
receive_wages	0.1486	0.014	10.870	0.000	0.122	0.175
mobileowner	0.0319	0.046	0.698	0.486	-0.058	0.122
internetaccess	0.0718	0.024	2.985	0.003	0.025	0.119
pay_utilities	0.0924	0.015	6.133	0.000	0.063	0.122
fin4	0.0876	0.016	5.618	0.000	0.057	0.118
fin8	0.0720	0.016	4.430	0.000	0.040	0.104

Omnibus:	26119.535	**Durbin-Watson:**	1.949
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	625.141
Skew:	-0.002	**Prob(JB):**	1.79e-136
Kurtosis:	1.356	**Cond. No.**	442.

Table 8: OLS, Lasso and Ridge Coefficients

	var	OLS	Lasso	Ridge
0	age	-0.006155	-0.084740	-0.084838
1	female	-0.001799	-0.000842	-0.000895
2	educ_2	-0.019575	-0.009144	-0.009741
3	educ_3	-0.044663	-0.021339	-0.021985
4	inc_q_2	-0.015814	-0.005121	-0.004944
5	inc_q_3	0.035185	0.011878	0.012219
6	inc_q_4	0.064104	0.026389	0.026790
7	inc_q_5	0.105754	0.052000	0.052464
8	account_fin	0.320836	0.000000	0.000000
9	receive_wages	0.148609	0.072887	0.072907
10	mobileowner	0.031931	0.004595	0.004680
11	internetaccess	0.071777	0.021005	0.021104
12	pay_utilities	0.092354	0.040788	0.040840
13	fin4	0.087599	0.037199	0.037271
14	fin8	0.072039	0.029127	0.029170

Figure 1: Confusion Matrix - Refined OLS Model

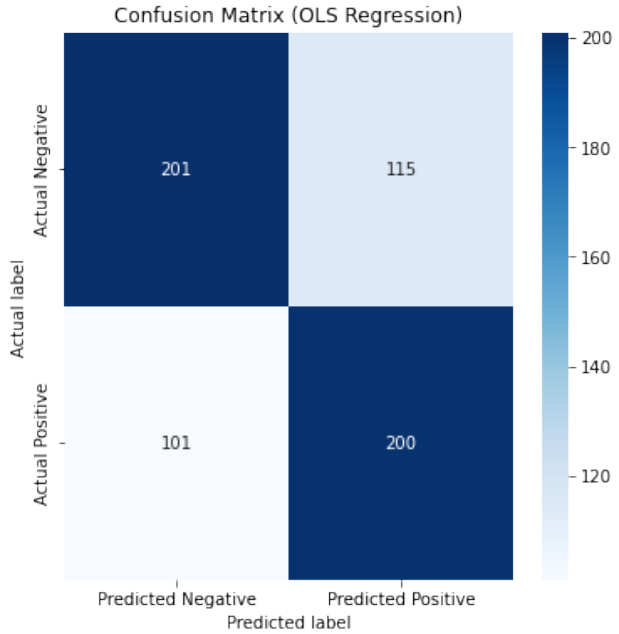

Table 9: OLS Summary Result (Refined)

Dep. Variable:	account_mob	**R-squared (uncentered):**	0.534
Model:	OLS	**Adj. R-squared (uncentered):**	0.533
Method:	Least Squares	**F-statistic:**	453.2
Date:	Mon, 06 May 2024	**Prob (F-statistic):**	0.00
Time:	13:13:15	**Log-Likelihood:**	-3769.0
No. Observations:	5553	**AIC:**	7566.
Df Residuals:	5539	**BIC:**	7659.
Df Model:	14		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
age	-0.0052	0.000	-11.444	0.000	-0.006	-0.004
female	0.0090	0.013	0.702	0.483	-0.016	0.034
educ_2	0.0111	0.024	0.457	0.647	-0.036	0.059
educ_3	-0.0218	0.026	-0.845	0.398	-0.072	0.029
inc_q_2	0.0321	0.029	1.108	0.268	-0.025	0.089
inc_q_3	0.0828	0.027	3.024	0.003	0.029	0.137
inc_q_4	0.1078	0.026	4.187	0.000	0.057	0.158
inc_q_5	0.1461	0.025	5.959	0.000	0.098	0.194
receive_wages	0.1537	0.014	11.228	0.000	0.127	0.180
mobileowner	0.2164	0.034	6.442	0.000	0.151	0.282
internetaccess	0.0790	0.024	3.280	0.001	0.032	0.126
pay_utilities	0.0949	0.015	6.286	0.000	0.065	0.125
fin4	0.0953	0.016	6.117	0.000	0.065	0.126
fin8	0.0848	0.016	5.243	0.000	0.053	0.116

Omnibus:	27363.700	**Durbin-Watson:**	1.951
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	600.966
Skew:	-0.007	**Prob(JB):**	3.18e-131
Kurtosis:	1.388	**Cond. No.**	318.

Table 10: Feature Importance (Hist Gradient Boost)

Feature	Importance
account	0.296306
account_fin	0.128050
remittances	0.017815
age	0.007291
saved	0.005284
receive_transfers	0.002763
receive_agriculture	0.002394
inc_q_5	0.002194
internetaccess	0.001487
fin4	0.001384
receive_wages	0.001354
fin10	0.001203
emp_in	0.001191
female	0.001167
fin9	0.000895
inc_q_3	-0.000858
educ_3	0.000768
pay_utilities	0.000550
inc_q_4	-0.000369
borrowed	0.000320
urbanicity_f2f	0.000254
mobileowner	0.000218
fin8	0.000206
receive_pension	-0.000157
fin2	0.000036
educ_2	-0.000030
fin7	-0.000018
inc_q_2	-0.000006

Figure 2: Confusion Matrix - Histogram Gradient Boost Model