Combatting fraudulent and predatory fintech apps with machine learning

Jonathan Fu[†] and Mrinal Mishra[§]

[†] *Department of Banking & Finance, University of Zurich and the Centre for Sustainability & Private Wealth*
[§] *Department of Banking & Finance, University of Zurich and the Swiss Finance Institute*

February 10, 2022

Contents

1 Executive summary 1

2 Background & motivation 2

3 Data & sample characteristics 3

3.1 Target apps and countries of interest 5

4 Classification methods & estimates on suspect apps 8

4.1 Manual classification 8

4.2 Market-based guidance 11

5 Exploratory analysis on demand and supply-side impacts 16

5.1 What are the impacts of “scam” lending apps on consumer outcomes? 16

5.2 What are the negative externalities of “scam apps” on legitimate providers? 17

6 Application of machine learning to predict suspect apps 18

6.1 Methodology for machine-learning phase 18

6.2 Results 20

6.3 Sensitivity tests 22

7 Conclusions and recommendations 28

7.1 Key conclusions 28

7.2 Recommendations for next steps 28

A Appendix: Country Sample 32

B Appendix: App classification examples 33

B.1 Example of “Legitimate” app class 33

B.2 Example of “Pure fraud” app class 34

B.3 Example of “Predatory, no service” app class 35

B.4 Example of “Predatory, problematic services” app class 36

C Appendix: Additional machine learning model figures 37

List of Figures

1 Number of newly released pure-play lending apps by week available in target countries 7

2 Overview of lending app typology and classification 9

3 Example of loan term disclosure 13

4 Examples of discrepancy between stated vs. actual received loan terms 16

5 Time-series trends for prevalence of suspect lending apps in Indian market 18

6 ROC curves—Baseline model using binary class outcomes 21

7 ROC curves—Baseline model using multi class outcomes 22

8 ROC curves—Ex-post model using binary class outcomes 23

9 ROC curves—Ex-post model using multi class outcomes 24

10 ROC curves—Restricted training time period model using binary class outcomes 25

11 ROC curves—Restricted training time period using multi class outcomes 26

12 ROC curves—SDK-based model using binary class outcomes 27

13 Decision tree for baseline model using binary class outcomes 37

14 Variable importance list for baseline model using binary class outcomes 37

15 Decision tree for baseline model using multi class outcomes 37

16 Variable importance list for baseline model using multi class outcomes 37

17 Example of variable importance list for SDK-based model using binary class outcomes 37

List of Tables

1 Categorization of finance apps in sample 4
2 Percentage of finance apps newly-released & unpublished during study period 5
3 Results from manual classification exercise on pure-play lending apps 10
4 Outcomes for pure-play lending apps 11
5 Market guidance on consumer credit among focus countries 13
6 Example of market-based classification 14
7 Relationship between exposure to "suspect" loan apps and prevalence of consumer protection issue 17
8 Relationship between prevalence of suspect apps in market and uptake of legitimate apps 18
9 Training and out-of-sample model accuracy–Baseline model 21
10 Training and out-of-sample model accuracy–Ex-post model 23
11 Training and out-of-sample model accuracy–Restricted training time period model 25
12 Training and out-of-sample model accuracy–SDK-based model 27
A.1 Overview of countries in data sample 32

1 Executive summary

Context & purpose of the study

Over the past decade, predatory and fraudulent practices in digital finance and financial technology have increased globally. The use of mobile applications for such purposes is a concern given the increased ubiquity of mobile devices, ongoing concerns about low financial and digital literacy in many population groups, and anecdotal evidence on a proliferation of methods and tools some finance app providers use to exploit vulnerable households and businesses. In addition to the direct harm caused to consumers, this can lead to mistrust of digital finance, which can delay financial inclusion efforts and undermine the benefits of financial technologies.

The objective of this project is to explore whether high-frequency app data and applied machine learning techniques can be leveraged to create a system for flagging and reporting highly suspect apps. As a "proof of concept", we draw on historical app meta and review data for 63 countries covering from January 2020 to April 2021 to document the prevalence of such problematic apps and test the efficiency and accuracy of such methods. To keep the project tractable, we focus on a targeted subset of personal loan apps. The pilot informs future experimentation with the approach and suggests possible real-world applications, which could help provide targeted shortlists of highly suspect individual apps, country- or global-level monitoring to understand prevalence of problematic apps at a given point in time, or be fed into buyer-beware labelling on the app stores.

Methodology

Focusing on personal loan apps, we test two approaches for labelling problematic apps as falling into several suspect and legitimate classes using: 1) manual classification and 2) classification based on market-specific guidance. The manual classification approach is done at a broad scale using heuristic techniques and the labels subsequently applied in supervised "gradient boosting" machine learning models. We train and validate the models on apps that were released prior to 2021 and test their out-of-sample performance in predicting class for the apps that are newly-released in 2021. The market-specific guidance approach is tested on three target countries—India, Nigeria, and the Philippines—by benchmarking personal loan apps' terms and conditions against local lending regulations and policies pertaining to usury rates. It is used as a case study to provide insights into feasibility of the approach and for purposes of comparison against the manual classification. For both approaches, we benchmark against actual app removals.

Main findings

1. We observe a large recent proliferation in personal loan apps of interest, along with signs that they exhibit considerably considerably more frequent entrance of new apps and exit of existing apps, relative to general finance apps.
2. The manual classification approach yields a high ratio of "likely suspect" (as opposed to "likely legitimate" apps). The current market-guidance approach yields a lower ratio of apps flagged as problematic, but this is partially explained by its narrower focus on usury pricing and by discrepancies between stated loan terms versus those actually received by users. For both approaches, apps flagged as problematic have significantly higher rates of removal from the app stores, which serves as one benchmark of validating "true positives".
3. The "suspect" apps are found to have significant negative impacts on both consumer outcomes and on other legitimate personal loan apps in the same markets.
4. The applied machine learning models are shown to have high (80-90%) out-of-sample accuracy in binary

class predictions and fair (70-80%) out-of-sample accuracy in multi-class predictions. This holds even for models using "static" input features—e.g, that would not be reliant on reviews or ratings to come in.

5. Differences in the set up of the machine learning models are found to have some influence on predictive accuracy, however. On the one hand, this could imply need for further fine-tuning of the classification methods and model specifics, with direct input from key stakeholders who may benefit from use of applied vetting or monitoring tools. On the other hand, this also suggests that using an ensemble of approaches may actually offer a promising way to provide a rank-ordering on apps of interest.
6. In the future, lessons learned from both the manual and market-guidance approaches can be integrated into a rule-based and automated classification system.

Conclusions

The study demonstrates that available high-frequency app data and applied machine learning techniques can efficiently and accurately flag apps exhibiting suspect characteristics or behavior. This can increase efficiency in vetting the tens of thousands of newly-released finance apps per year and improve efforts to monitor available apps already published on the app stores. Findings have potential use-cases for a wide range of stakeholders, including financial regulators, international organizations, civil society organizations, and the app stores themselves.

1. **Real-time or periodic flagging of target apps:** the models can be transferred and applied to real-time app data to predict whether apps currently available in the given country market are likely to fall under legitimate versus various suspect categories, allowing for targeted action;
2. **Information for demand-side interventions:** information gleaned from real-time monitoring could also feed into consumer protection interventions (e.g., just-in-time public awareness campaigns);
3. **Global monitoring:** the methods could be readily converted into aggregated-level reporting to flag countries that are current hotspots of activity based on prevalence of "suspect" apps in market;
4. **"Buyer beware" labeling:** there may be collaborative opportunities to develop alternative or complementary rating or labeling system on app store fronts for finance-related apps in order to help consumers better spot or take extra caution on apps exhibiting suspect characteristics.

2 Background & motivation

Over the past decade, predatory and fraudulent practices in digital finance and financial technology have increased. The use of mobile applications for such purposes is of particular concern for several reasons. First, an increasing share of the global population now has access to mobile devices and uses them to access financial services.[1] Second, there are ongoing concerns about low financial and digital literacy in many population groups—particularly among first-time users of formal financial services.[2] Finally, there is anecdotal evidence that the COVID-19 pandemic has led to a proliferation of the methods and tools such "scam" app providers use to exploit vulnerable households and businesses.[3] In addition to the direct harm caused to consumers, this can lead to mistrust of digital finance, which

[1] At the end of 2019, 90% of the world's population was covered by mobile networks, 67% of the global population (5.2 billion people) had a subscription to mobile services, and 49% of the global population (3.8 billion people) were mobile internet users (GSMA (2020)). Demirguc-Kunt et al. (2018) meanwhile document large improvements in financial access within the last decade, with advances in digital and mobile technology playing a large contributing role.

[2] See, for example: Heidhues & Kőszegi (2010); Lusardi & Mitchell (2011, 2014)) for overview of this literature.

[3] Examples of recent media coverage include: https://www.nytimes.com/2021/03/27/world/easy-to-get-hard-to-repay-predatory-loans-flourish-in-pandemic-gripped-india.html ; https://www.forbes.com/advisor/in/loans/how-to-avoid-digital-lending-scams/ ; https://www.economist.com/finance-and-economics/2020/01/30/how-digital-financial-services-can-prey-upon-the-poor ; https://qz.com/africa/1975202/kenya-prepares-to-crack-down-on-high-interest-loan-apps/

can delay financial inclusion efforts and undermine the benefits of financial technologies.

While the scale and scope of issues observed in the finance app markets are troubling, a silver lining is that they also highlight steps that could be taken to mitigate issues on both the supply and demand side. In particular, a promising method could be to leverage high-frequency app data and apply machine learning techniques to create a system for flagging and reporting highly suspect apps. This could feed into various tools or interventions to mitigate the risk and impact from such problematic apps. For example, such analysis could give regulators and supervisors real-time snapshots of the state of their current finance app markets and predict propensity of currently published apps to fall in suspect categories. This information could be used to improve efficiency for vetting the thousands of new finance apps being published each month and reducing the time to spot and take down problematic ones.

To explore this possibility and opportunity, IPA has set up a pilot project as a "proof of concept" to demonstrate the effectiveness and efficiency of such methods applied to available app data. To make the pilot tractable, the approach has been developed and tested on historical app data from January 2020 to April 2021 and focused on assessing "personal loan apps" in those markets. The decision to focus on personal loan apps is motivated by the amount of media coverage received on them and because they show signs of having particularly high adoption during the pandemic period.[4] Nevertheless, it would be important to increase scope in the future to also investigate other types of finance-related "scam" apps purportedly providing other products and services.[5] The machine learning models are trained and validated on earlier data from our historical time series and then tested on a withheld out-of-sample dataset covering the latter time period–i.e., we train and test on entirely independent samples. This allows us to simulate how such an approach might perform in a "real-time / real-world scenario" when the model has been trained on historical data and is then applied to make predictions on completely new data.

In particular, the objectives of the pilot project and involved analyses are several fold:

1. To document different forms of problematic lending apps found in the finance app markets
2. To systematize evidence on the prevalence and impacts of such apps
3. To assess whether machine-learning approaches trained on historical data can be used to efficiently and accurately predict problematic apps in unseen data
4. To assess how sensitive are results to classification methods and specifics of methodology
5. To inform additional work that would be required to transition this into a real-time monitoring tool.

3 Data & sample characteristics

This project utilizes mobile application meta data, review data, and download data from the Google Play Store. These data cover the universe of finance category apps that were available on the app store at any point during the period from January 1st, 2020 to April 30th, 2021 for 63 countries. In practice, this amounts to 134,744 unique app IDs. The meta data include the various information (both visible and hidden), text, and images seen on the app store for a given app package ID.[6] This is obtained in monthly data cuts across the study period and allow us to

[4]See, for example, Fu & Mishra (2021).

[5]For example, there have also been reports of fraudulent government-to-persons (G2P) social assistance apps, bank account balance checkers, etc.: https://diariodonordeste.verdesmares.com.br/negocios/aplicativos-falsos-do-auxilio-emergencial-sao-usados-em-golpes-saiba-como-se-proteger-1.3076913, https://www.businesstoday.in/personal-finance/banking/story/reserve-bank-rbi-warns-against-app-to-check-account-balance-50751-2015-04-11.

[6]Examples of the visible data include: Title, Short/long text description, Screenshots, Date last updated, Byte size, External version number, App developer, Developer website, Developer email, Developer physical address, Download / installs estimate (bucket), # of ratings, # of reviews, Promo video. Examples of hidden data include the initial release date, internal version code, Software Development Kit (SDK) lists,

Table 1: Categorization of finance apps in sample

	# of apps	% of apps	Est. # downloads (in M)	Avg. downloads per app
Panel A. Finance apps by EN vs. non-EN main language				
All finance	134,744	100.0%	6,970M	51,728
- All finance (EN main lang.)	87,426	64.9%	4,080M	46,668
- All finance (non-EN main lang.)	47,318	35.1%	2,890M	61,076
Panel B. Finance apps by product categories (EN only)				
All finance (EN main lang.)	87,426	–	4,080M	46,668
- General banking	18,015	20.6%	1,640M	91,035
- Payments	17,544	20.1%	1,830M	104,309
- Insurance	4,999	5.7%	397M	79,416
- Investment	4,325	4.9%	251M	58,035
- Credit, excl. personal loans	16,252	18.6%	1,290M	79,375
- Personal loans	5,106	5.8%	826M	161,770

track both newly released apps and apps that are removed ("unpublished") from the app store across time. The review data include disaggregated reviews including rating, date of review, and review text going back to the app's initial release date. Finally, the data include historical download estimates at the country and daily-level back to its initial release date. The sample countries combined represent roughly 75% of the world population. Appendix Table A.1 provides a full list of the country sample.

For this pilot, we focus on apps with English listed as main language to facilitate the methods, both for manual review of app information and for applied text and natural language processing techniques. In practice, two-thirds of the sample finance category apps have English listed as the main language, with the remaining one-third being non-English, as depicted in Table 1, Panel A.[7]

For the apps listed with English as the main language, we subcategorize them by product type by applying text-based content analysis on the apps' titles, short descriptions, and full descriptions. For example, we start off using broader keyword searches on "credit", "loan(s)", "financing", etc. to first identify credit-related apps. We do similar keyword searches on other product types to flag general banking, payment, insurance, and investment apps. From the subset of credit-related apps, we flag the more narrow subset of personal loan apps by using regular expressions to conduct keyword searches on more precise terms such as "personal loan(s)", "quick loan(s)", "instant loan(s)", "consumer loan(s)", "payday loan(s)" etc.) on app titles and their short and full text descriptions. Moreover, for the "personal loan" apps, we then focus only on *pure-play* apps by excluding apps that have been flagged for other product types.[8] In Table 1, Panel B, we observe that general banking apps, payment apps, insurance, and investment apps comprise roughly 21%, 20%, 5.7%, and 4.9% of the English main language apps, respectively. General credit apps (excluding personal loan apps) and pure-play personal loan apps meanwhile make up 18.6% and 5.8% of these apps. It is worth noting that while pure-play loan apps thus appear to comprise a relatively smaller percentage share of overall apps, they show signs of being in considerable demand. Their estimated downloads are around 826 million and their average number of downloads per app are roughly 4 times higher than the finance category average.

Permissions lists, Interactive elements, Countries available in, Main language, Languages available in, Other app-stores available in.

[7]The apps' main language is identified through a specific variable in the meta data. App titles, text, and descriptions will be in the corresponding language. Future work could center around expanding the methodology to integrate non-English language apps.

[8]In practice, we setup the "credit-related" apps and "personal loan" apps to be mutually exclusive. "Pure play" denotes a company that focuses its efforts and resources on only one line of business. This is to distinguish from some other apps that are from financial intermediaries providing credit products along with a wider range of non-credit products and services.

Table 2: Percentage of finance apps newly-released & unpublished during study period

	App product category						
	(1) All	(2) Banking	(3) Payments	(4) Insurance	(5) Investment	(6) General credit	(7) Personal loan
Panel A. Entry and exit							
% New release	34%	22%	43%	30%	35%	39%	70%
% Unpublished	24%	13%	25%	20%	20%	26%	52%
Panel B. Time on app store							
Avg. app age (in mths.)	29.3	41.0	24.8	27.9	27.5	27.5	11.4
Observations	134,744	18,015	17,544	4,999	4,325	16,252	5,106

3.1 Target apps and countries of interest

Are there high-level signs of problems in personal loan apps?

To keep the pilot tractable, we focus on "pure-play" personal or consumer loan apps. This is motivated by the greater anecdotal evidence and considerable media scrutiny concerning them in the recent period. There are several high-level and suggestive signs of problematic activity in this subsample of pure play personal loan apps of interest, relative to other finance app categories. As depicted in Table 2:

1. The personal loan apps are much more likely to be "new releases" during the study period.
2. The personal loan apps are much more likely to be "unpublished" (i.e., removed from app store) during the study period
3. The personal loan apps' average time on the app store is 11.4 months, compared to 29.3 for the general finance category average.

What is the scale of proliferation in personal loan apps?

The time-series patterns in the release of these pure play personal lending apps gives a cursory sense of how prolific these apps are and why it's likely been difficult for either internal or external stakeholders to vet or monitor them all. Figure 1a depicts the number of newly-released pure-play lending apps by week across our study period, aggregated across the 63 countries in our sample. We observe that there were on average around 80 new (English main language) pure-play personal lending apps being released each week. This translates into several hundred new (English main language) personal loan apps per month and several thousand per year in our sample countries.[9] This is likely a lower bound estimate since it omits non-English apps, which may in practice constitute an important share of the personal lending market.

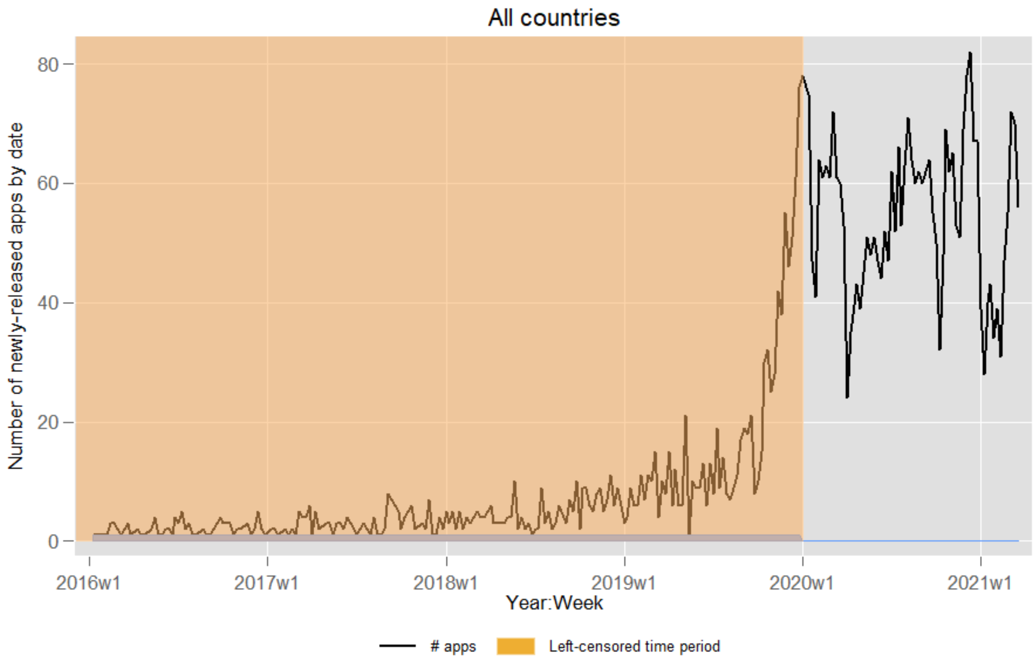
There may be some country markets that are more or less exposed to these apps. To investigate this further, we examine country-level differences in exposure across a few target countries—namely, the United States, India, Nigeria, and the Philippines. Our interest in these four countries are severalfold. First, in Fu & Mishra (2021), we observe that some of the most downloaded mobile apps in Nigeria, the Philippines, and India following the outbreak of COVID-19 include finance-related apps that are either predatory or entirely fraudulent. Note that our

[9]It is worth noting that there is left-censoring occurring since our meta-data does not include apps that were unpublished prior to 2020. As previously noted, about 30% of these target apps are from before that date, and the figure captures the distribution of their dates of release.

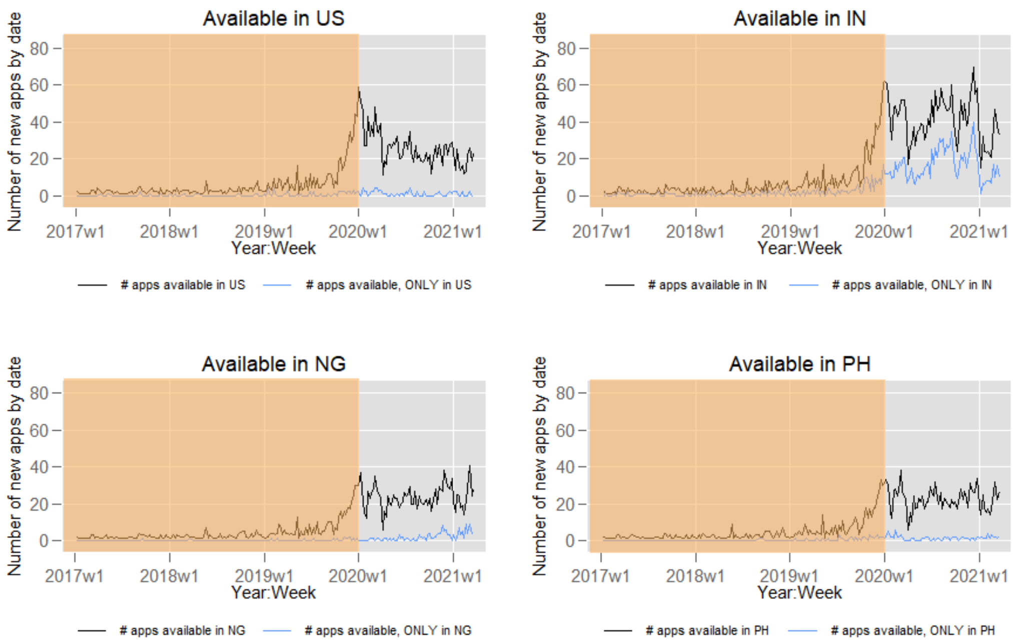
use of the term “predatory” goes beyond usurious practices and captures broader cases where lending occurs but consumers face abusive practices or are otherwise taken advantage of. The United States, meanwhile, is used as a general benchmark country. Second, focusing generally on countries with larger population size for this pilot study is logical since it is expected to yield more data points of all types (apps, review data, etc.) for the analyses. Finally, it would also be logical to focus on countries where English is the dominant language (or commonly used) to facilitate text-based analysis.

Figure 1b repeats the previous time series, but tracks the number of personal loan apps by week that are i) newly-available in each target country and ii) newly-availably *only* in each target country. We observe that India in particular seems to have the most proliferation of supply and, in particular, targeted apps that are only available within country. Nigeria and the Philippines both see a steady flow of new personal lending apps coming into their market per week. However, Nigeria appears to have a slight increase in targeted apps only available within country towards the end of 2020 and beginning of 2021, whereas the Philippines exhibits some signs of having targeted apps at the beginning of 2020, which mostly disappear as the pandemic prolongs. The US generally sees a marked decrease over time in terms of the number of new personal loan apps being released during the study period.

Figure 1: This figure depicts time series trends for the number of newly released apps available in the app store during our study period. Panel A depicts these trends in terms of apps available across the full country sample. Panel B depicts these trends in terms of apps available in particular countries of interest (the United States, India, Nigeria, and the Philippines). These are further broken out between apps that are available in the country (but may be simultaneously available in others) and apps that are *only* available in the given country. Orange shading denotes the time period where left censoring occurs. We are unable to unable to track apps that were both newly released and removed within the left-censored time period. Thus, the true number of new apps being released during that period may be obscured.



(a) Number of newly released pure-play lending apps by week available in full country sample



(b) Number of newly released pure-play lending apps by week available in targeted countries (US, IN, NG, PH)

4 Classification methods & estimates on suspect apps

Focusing on the pure-play personal loan apps, we next study the prevalence of problematic apps using two approaches.[10] Specifically, we apply: 1) a system of manual classification and 2) a classification based on market-specific guidance. The former approach is done across all personal loan apps in our sample through manual review and application of heuristic techniques on the apps' meta data, review data, and Google Play storefronts. The labelled data from the manual classification is retained and used in the subsequent section to test how well machine learning models perform in predicting them out-of-sample. The latter approach is tested on three target countries (India, Nigeria, and the Philippines) by benchmarking the personal loan apps terms and conditions against local lending regulations and policies. It is used as a case study to provide insights into feasibility and for purposes of comparison against the manual classification. In general, comparing these two approaches to classification help us to better understand the scale and scope of problematic lending apps in the app stores and test for the influence of different classification methods on results.

4.1 Manual classification

The method

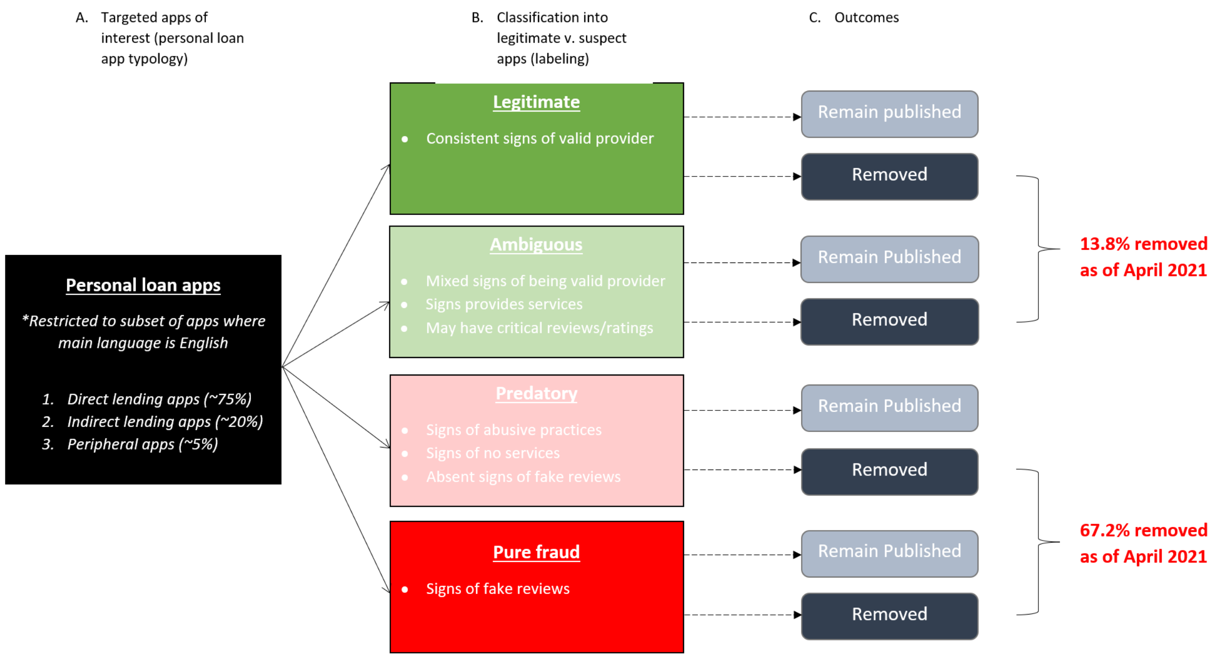
We start by conducting a systematic manual review of the personal loan apps' individual meta data, review data, and app web pages (i.e., going through the actual Google Play url). During this review, we classify the apps as falling into "legitimate", "ambiguous", "predatory", and "pure fraud" categories, based on heuristic signals revealed in the app meta data, review data, and Google Play storefront pages. Again, we remind that our use of the term "predatory" goes beyond its common application to usurious pricing or interest rates and captures broader cases where lending occurs but consumers face abusive practices or are otherwise taken advantage of. In particular, we categorize apps into several classes of suspect vs. legitimate apps, based on a few defining characteristics:

- "Pure fraud": characterized by existence of fake reviews. For example, we look for a number of signals where we examine the higher rated (4-5 star) reviews for existence of idiosyncratic review dates (i.e., densely clustered around the initial release date and then disappearing as time goes on), complete lack of reviewer profile pics, abnormal user names not fitting the app market, and idiosyncratic text. By contrast, we examine the lower rate (1-3 star) reviews for existence of regularly distributed review dates, complaints about app/reviews being fake, and having paid "registration fees" but receiving no service.
- High complaints of "Predatory" practices: based on signs of i) no real services provided or ii) problematic lending practices. For the former, for example, we examine the higher rated (4-5 star) reviews for signs that the users are real but there are no indications services provided (e.g., reviews seem positive but signal that the user is still awaiting approval or disbursement of loan).[11] Further manual verification of the app developer's information (i.e., website, email, company, address, online presence) does not indicate company exists. For

[10] In practice, we also observe three broad types of lending apps. First, around 74.9% are pure-play lending apps that claim to directly provide personal loans (including payday, consumer, MSME, etc.). We are agnostic to the provider type (e.g., commercial bank, neobank, fintech, MFI, NBFI, others). Second, around 19.8% are indirect lender apps that do not directly offer loans, but take a role as a "marketplace/guide" (or at least claim to) to access personal loans. In practice, some are very transparent about their intermediary role, whereas others seem to mask this. Finally, there is a minority (aroud 5%) of peripheral apps related to personal lending such as loan / EMI calculators and credit report / score apps. For purposes of this exercise, we remain agnostic to the specific type of loan app and keep them all in the sample.

[11] For example, in practice, we see suggestive evidence that some apps seem to "coerce" or promote prospective borrowers to give 5-star ratings to increase their chances of loan approval.

Figure 2: Overview of lending app typology and classification



the latter, we examine the overall reviews for signs that the users are indeed receiving some services and loans, but they exhibit high prevalence of complaints on common consumer protection issues such as high interest rates, fees, or short tenure, abusive debt collection practices, and concerns over data privacy.

- Legitimate: in the absence of pure fraud and "predatory" signals and upon further manual verification of the app developer's information (i.e., valid website, email, verified company, address, online presence), we classify an app as "legitimate".
- "Ambiguous": cases where there may be some suspect signals, but they do not necessarily appear to be consistent; further manual verification of the app developer's information (i.e., website, email, company, address, online presence) reveals some validation of company's existence, but it may be only partial (e.g., they may appear to have valid email and physical location, but no website and/or additional online presence, etc.).

For a simpler binary classification, we aggregate the legitimate and ambiguous categories together as being "likely legitimate" and the predatory and "pure fraud categories as being "likely suspect". In Appendix B, we provide a number of illustrative examples of such apps and their review data, to demonstrate such signals. Figure 2 provides an overview of this manual review.

Manual classification results

Table 3 summarizes the results of this manual classification. In Panels A and B, we observe that of the 5,106 apps in our targeted sample, 16.8%, 9.9%, 68.8%, and 4.6% are ultimately classified as legitimate, ambiguous, predatory, or pure fraud, respectively, based on our review. Using the binary classification, we estimate 26.7% and 73.3% falling into the "likely legitimate" and "likely suspect" categories, respectively. There are some signs that India may have the highest overall exposure to problematic personal loan apps, but all of our target countries exhibit generally similar rates.[12]

Table 3: Results from manual classification exercise on pure-play lending apps

	Prevalence of app in category				
	All	Available in US	Available in IN	Available in NG	Available in PH
	(1)	(2)	(3)	(4)	(5)
Panel A. Number of personal lending apps					
# of apps	5,106	2,613	3,810	2,281	2,241
Panel B. Percent of apps falling in legitimate vs. suspect classification buckets					
% in 4-category classification					
- Legitimate	16.8%	16.6%	15.7%	16.4%	16.9%
- Ambiguous	9.9%	11.1%	9.2%	11.7%	12.2%
- Predatory	68.8%	69.1%	71.4%	68.1%	67.2%
- Pure fraud	4.6%	3.3%	3.7%	3.9%	3.8%
% in Binary classification					
- Likely legitimate	26.7%	27.7%	25.0%	28.1%	29.0%
- Likely suspect	73.3%	72.3%	75.0%	71.9%	70.9%
Panel C. Estimated downloads in millions (M)					
All personal lending apps	834M	418M	670M	394M	407M
4-category classification					
- Legitimate	616M	369M	522M	343M	344M
- Ambiguous	38M	17M	29M	16M	27M
- Predatory	149M	24M	106M	26M	29M
- Pure fraud	31M	8M	14M	9M	8M
Binary classification					
- Likely legitimate	654M	386M	551M	359M	371M
- Likely suspect	180M	32M	119M	35M	36M
Panel D. Average downloads per app in category					
All personal lending apps	163,337	159,969	175,853	172,731	181,615
4-category classification					
- Legitimate	716,279	850,230	871,452	914,667	910,053
- Ambiguous	76,342	59,170	82,102	59,023	97,802
- Predatory	42,246	13,172	38,757	16,582	18,857
- Pure fraud	144,651	116,020	109,756	124,801	113,034
Binary classification					
- Likely legitimate	479,824	533,887	579,390	560,062	569,892
- Likely suspect	48,103	16,781	41,638	21,293	22,782

Panels C and D meanwhile depict the amount of users exposed to these apps in terms of absolute downloads and average downloads per app across the different categories. We observe that the "likely legitimate" apps have the largest absolute uptake at around 654 million downloads and an average of 480,000 downloads per app. This meets *a priori* expectations since the associated providers are likely to have greater name recognition and trust. Moreover, it is worth noting that these download estimates (drawn from the meta data) are cumulative over the entire release of the apps. As the "likely legitimate" apps are more likely to remain on the app store for longer periods and include downloads from prior to the start of our main study period in January 2020, these comparisons may be particularly skewed in their favor. Nevertheless, the absolute number of users exposed to the "likely suspect"

[12]While this seems at first glance a particularly high percentage, we note further supporting evidence in Appendix Table #, which highlights that 70% of the predatory apps are ultimately removed as of April 2021.

apps is still quite considerable, with 149 million downloads of the predatory apps and 31 million downloads of the pure fraud apps. The average downloads per app of the "likely suspect" apps is still considerable at around 48,000 per app. This is particularly so for the apps classified as "pure fraud" apps, which see on average around 145,000 downloads per app.

In practice, we observe that 67.2% of the apps labeled as "likely suspect" through this manual classification system are removed during our study period (i.e., as of April 2021) compared to just 13.8% of the "likely legitimate" apps. The removal rates for the apps flagged as "predatory" and "pure fraud" classes also continue to increase as time goes on. This large gap in removals for the "likely suspect" apps relative to the "likely legitimate" apps provides initial suggestive evidence that the classification is indeed fairly successful at separating between apps with higher vs. lower propensity for problematic behavior. For those apps that are flagged as likely suspect but remain on the app store, future work could involve manually validating the apps and their actual services.

Table 4: Outcomes for pure-play lending apps

Prevalence to be "unpublished" (as of Apr. 2021)	
# of all personal loan apps	5, 106
4-category classification	
% Legitimate	16.4%
% Ambiguous	9.3%
% Predatory	70.1%
% Pure fraud	17.7%
Binary classification	
% Likely legitimate	13.8%
% Likely suspect	67.2%

4.2 Market-based guidance

The method

An alternative approach for classification could be to draw upon market-specific rules for separating between "legitimate" vs. "problematic" applications. A large number of countries have issued regulations or market guidance that governs various aspects of lending practices, such as interest rate caps, limits to fees associated with lending, permissible debt collection practices, treatment of client data, and so forth. For example, Ferrari et al. (2018), document that 76 countries around the world as of 2016 still set interest rate caps to protect consumers from usurious lending practices. These differ substantially regarding what types of providers they cover—e.g., whether they have rules specific to fintech lenders—and how specific restrictions are, such as whether they include only interest or all interest and fees (e.g. by capping APR not interest rate.). However, in cases where such guidance exists, attempts could be made to draw upon them to develop direct benchmarks for classifying digital lending apps as being legitimate vs. suspect.

An application via a case study example

While setting up a globally comprehensive system for market-based classification is beyond the scope of this pilot project and policy brief, we go through some of the requisite preliminary steps to see what insights it may yield and how results compare with our other classification approaches. We use our three target countries—India, Nigeria, and the Philippines—as examples and do a review of government and regulatory policies, laws, and guidelines on personal and consumer lending. Specifically, we focus on identifying pertinent usury laws on interest rate ceilings and fees, since they allow for a fairly transparent and direct method of identifying apps beyond permissible thresholds within a given market. In future work, such an approach could also be extended to try and benchmark apps based on other loan terms or lender practices.

Table 5 provides an overview. In practice, we observe that India and Nigeria have many consumer lending laws in place, but looser guidance on what specific thresholds would be considered usurious. For example, in the wake of the recent digital lending crisis, India has created new guidance on digital lending apps regarding mandatory disclosure on loan terms, usage of app permissions, and avoidance of strong-arm tactics for loan recovery following the crisis. However, these guidelines as of now do not currently set forth any specific interest rate ceilings, fee restrictions, or minimum loan durations.[13] Nigeria previously had regulation outlining specific interest rate caps and fees for mortgage loans and considered imposition of interest rate ceilings for broader lending markets. However, the mortgage loan laws were removed in 2019 and there are not any currently other rate caps imposed, to the best of our knowledge.[14]

By contrast, the Philippines has recently set concrete and specific usury guidance on various forms of consumer credit in light of the economic crisis induced by COVID. First, in September 2020, the Bangko Sentral ng Pilipinas (BSP) issued Memorandum "Circular No. 1098", which lays down the maximum interest and finance charges that banks and other non-bank financial institutions (NBFIs) can impose on credit card receivables. Specifically, permissible rates on credit card transactions are capped at an interest of 2% per month or 24% per annum and upfront processing fees should be limited to PHP 200 for each transaction.[15] Second, in late December 2021, the BSP released "Circular No. 1133" on the "Ceiling/s on Interest Rates and Other Fees Charged by Lending Companies (LCs), Financing Companies (FCs), and their Online Lending Platforms (OLPs)" which prescribed the caps on the interest rates and other fees of LCs and FCs to take effect on Jan. 3, 2022. Specifically, the circular places a 6% cap on monthly nominal interest rates (around 0.2% per day) for loans not exceeding P10,000 and are payable within four months.[16] Meanwhile, the effective interest rate, which takes into account processing and handling fees as well as other compounding effects, was set at 15% per month. For late payment or non-payment, the BSP set an an interest rate of 5% per month on the outstanding scheduled amount due. In relation to this, a total cost cap of 100% of the total amount borrowed was also set regardless of the length of time the loan is outstanding. These latter examples in the Philippine context allow us to test how such a market-specific benchmark approach could work.

[13]These guidelines have been published in a recent report by the Reserve Bank of India's "Working Group on Digital Lending including Lending through Online Platforms and Mobile Apps", as outlined here: https://www.rbi.org.in/Scripts/PublicationReportDetails.aspx?UrlPage=&ID=1189

[14]See: https://www.cbn.gov.ng/out/2019/fprd/removal%20of%20interest%20rate%20and%20cap.pdf

[15]See: https://www.bsp.gov.ph/Regulations/Issuances/2020/c1098.pdf

[16]See: https://www.bsp.gov.ph/SitePages/MediaAndResearch/MediaDisp.aspx?ItemId=6084

Table 5: Market guidance on consumer credit among focus countries

Country	Source	Description	Permissible interest rates	Permissible fees
India	RBI Working Group, January 2021	RBI set up a working group on digital lending (including mobile apps) to enhance customer protection and make digital lending ecosystem safe. Broad recommendations on aspects of data collection and security, transparency, etc. No specific guidance on rates or fees.	–	–
Nigeria	N/A	—	—	—
Philippines	Usury Law Act No. 2655; BSP Memorandum Circular No. 1098; BSP Memorandum Circular No. 1113	In the spirit of promoting responsible lending and considering prevailing economic conditions, BSP set a ceiling on cost of lending through credit card transactions to ease the financial burden on consumers, particularly micro-, small-, and medium-sized business enterprises during the COVID-19 pandemic.	24% annualized for credit card transactions; 6% monthly nominal rate and 15% APR for unsecured personal loans	200 PHP max. processing fees for credit card transactions; for unsecured personal loans, the fee prohibitions are incorporated into the maximum APR rates

Figure 3: Example of loan term disclosure

LOAN RATES AND FEES

• Loan Amounts: ₱1,000 to ₱15,000
• One-time service fee: 8.5-27.4%
• Monthly EIR: 14.86-18.32%
• APR: 172.9% -228.3%
• Late fee: one-time fee of 8% of outstanding amount due
• Applicable Tax (DST & GRT): 0.46% - 1.5%
• Minimum repayment period offered = 61 days
• Maximum repayment period offered = 61 days

• Example Fees: For a borrower who takes a ₱4,000 loan for our 61-day term, they would owe a 27.4% service fee of ₱1,096 and ₱60 in required tax for a total payment of ₱5,155.

₱4,000 (principal) + ₱1,096 (service fee) + ₱60 (required Tax) = ₱5,155 (total payment)

Market-specific guidance results: Philippines as a case study

We draw on the Philippines as a case study, as it has the most specific market guidance on usury rates among our target countries. We subset to the 122 personal loan apps that are *only available* in the country, scrape the apps' descriptions for key loan terms, and transform them into structured variables. In practice, we observe that the majority of the personal loan apps disclose their terms and conditions in a fairly standardized manner, which facilitates this task. Figure 3 provides an example of loan term disclosure that is typical of the more transparent apps.

In Table 6, Panel A, we observe that in practice, 108 out of the 122 relevant apps (about 89% of sub-sample) either directly provide APR or have enough information to calculate it. The remaining minority of apps are missing all or some key loan terms, which prohibit accurate calculation of pricing. It is worth noting, however, that some of the apps that report loan terms present the information in a way that appears to obfuscate assessment of true cost of

Table 6: Example of market-based classification
This table summarizes findings from comparisons of personal loan apps in relation to consumer lending market guidance in the Philippines. The sample is restricted to pure-play loan apps only available in the Philippines.

	# of apps	% of apps
Panel A. Number and percent of apps with disclosed interest rates and associated fees		
Disclosed loan terms	108	88.5%
Panel B. Number and percent of apps with disclosed loan terms exceeding permissible interest rates and fees		
Based on BSP Circular No. 1098	73 out of 108	67.6%
Based on BSP Circular No. 1113	13 out of 108	12.0%
Panel C. Number and percent of app removed from store		
All apps	75 out of 122	61.4%
Apps flagged based on BSP Circular No. 1098	43 out of 73	58.9%
Apps flagged based on BSP Circular No. 1113	11 out of 13	84.6%

the loans—e.g., reporting low interest rates, but taking various processing and service fees out of the principal in ways that translate into considerably higher *de facto* APR. Some illustrative examples are provided in Figure 4. As such, wherever possible, we recalculate the APR based on the full information on interest rates and fees presented to validate the apps' disclosed pricing.

We then compare the collated app loan terms against the two sets of country-specific benchmarks set forth by Circular No. 1098 and 1133—for credit card transactions and small, short term unsecured consumer loans, respectively. The apps at a cursory glance perform poorly against the former circular but fairly well against the latter and most recent BSP rules. Specifically, in Table 6 Panel B, we observe that for the 108 personal loan apps with fully disclosed loan terms, 73 (or 67.6%) would be classified as "usurious" based on the earlier BSP Circular on consumer credit card transactions, while only 13 (or 12%) would be flagged based on the more recent measures, which have been developed more specifically around unsecured personal loans and digital lending apps.

On the one hand, the apps thus perform fairly well when benchmarked against the most recent and more pertinent BSP guidelines on usurious pricing. In issuing the circular, the BSP acknowledged that the ongoing economic disruption caused by COVID has led to higher than usual rates of loan defaults, and they have thus tried to strike a balance in protecting both consumers and lenders serving riskier market segments. As such, it is interesting to find that the majority of apps fall under the pricing thresholds deemed reasonable by local regulators. On the other hand, it is also important to highlight that these apps still have a particularly high propensity for removal from the app store and are exhibiting other forms of suspect behavior. In Table 6 Panel C, we observe that 61% of the personal loan apps available only in the Philippines have been removed from the app store, including almost 85% of those exceeding the most recent BSP circular.

There are several reasons that may explain why the apps still see such high removals despite many of them having pricing that at first glance seemingly fall within permissible rates. First, when comparing the apps' meta data descriptions versus reviews, we observe that many are stating certain loan terms in their descriptions that differ considerably from those actually received by users, either being purposefully or unintentionally deceptive. (Again, Figure 4 provides several illustrative examples.) The current methodology of simply benchmarking against loan terms and conditions described by developers is insufficient to flag such cases. Second, there are also signs that many of the users' key concerns actually center more around abusive staff and loss of personally sensitive information, rather than high interest *per se*. We expand upon this subject in greater detail in Section 5.

Lessons learned and limitations of the approach

Based on the case study, we find that a market-specific classification approach could be implemented more broadly and, despite certain limitations, could be considered as one among an ensemble of approaches to further develop. Such an approach can be made fully transparent and readily tailored to the specific requirements and needs of a given regulator. For example, it can be used to immediately flag outlier apps *ex-ante* that would not pass existing usury laws in a given market. We imagine it may be possible to set up independent models using different classification approaches, and then triangulate on the top ranked or flagged suspect apps across the multiple models.

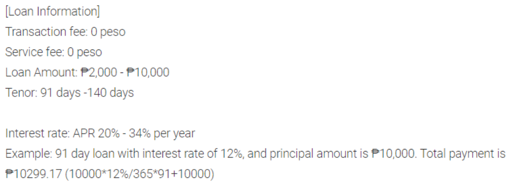
The approach does face certain limitations and challenges, however. First, not all countries have market-specific guidance in place on key loan terms, which may limit its applicability. Second, when market-specific guidance does exist, the rules are not always fully transparent, may not cover digital lending app markets, and/or may evolve frequently. As such, the approach would work best through close partnership with country regulators to aid in interpretation of existing rules or to develop new ones when they do not exist—i.e., rather than a top-down approach trying to gather data globally.[17] Finally, reliance on stated loan terms described by app developers appears prone to gaming and not fully reliable as developers appear to sometimes mask actual loan terms or find workarounds to existing rules. Alternative approaches could try and automate collection of real-world information on actual loan conditions from the disaggregated review data for a given app or also supplement with transaction auditing.[18] This may provide more accurate data for benchmarking, but would add considerably to complexity of the process.

[17] As regulation and guidance is market specific, such an approach and subsequent results would be by their nature country dependent. For apps that are available in only a single country, it would be relatively straightforward to benchmark them against existing market-specific lending rules. However, for apps that are simultaneously available in multiple countries, it is plausible that a given app could be considered “usurious” with respect to the market guidance in a given country but permissible in another country. As such, this would likely require some adjustments to a machine learning approach intending to use apps labelled in such a manner. (For example, a key change is that we would likely have to rerun models labeled at the app-country level, rather than the app-level.)

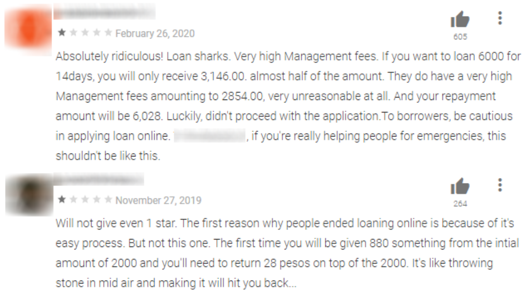
[18] For example, IPA and the Inclusion for All Initiative conducted an audit study measuring the costs of using digital financial services in Nigeria and compared those with both providers’ officially stated prices and the regulatory cap on fees set by the Central Bank of Nigeria. https://www.poverty-action.org/blog/how-can-we-deepen-transparency-nigerias-digital-financial-services.

Figure 4: Examples of mismatch between stated vs. actual received loan terms

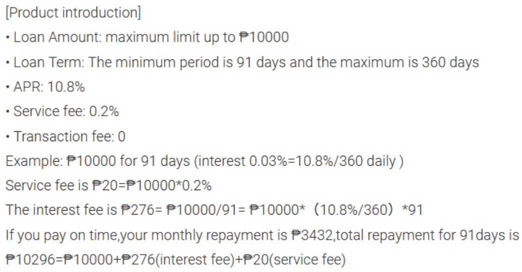
(a) Description of loan terms for sample app 1



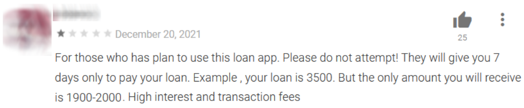
(b) Actual loan terms for sample app 1



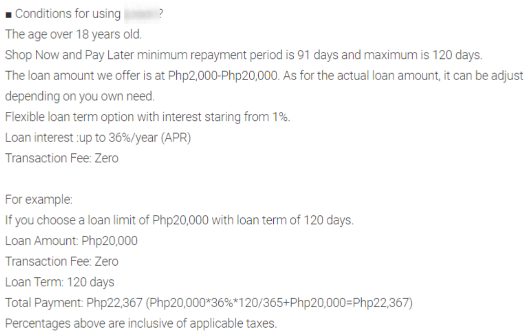
(c) Description of loan terms for sample app 2



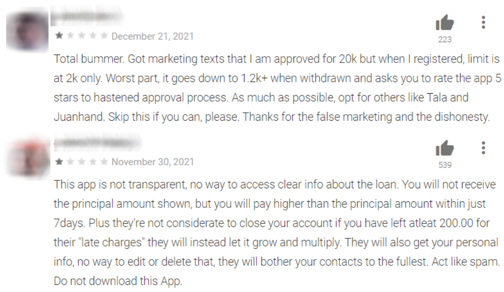
(d) Actual loan terms for sample app 2



(e) Description of loan terms for sample app 3



(f) Actual loan terms for sample app 3



5 Exploratory analysis on demand and supply-side impacts

5.1 What are the impacts of "scam" lending apps on consumer outcomes?

Having labelled the "suspect" personal loan apps, we assess some of their adverse impacts on consumer outcomes. To do so given our available data, we analyze users' complaints regarding a number of common consumer protection themes from the literature. Specifically, we use text-based content analysis on the disaggregated review data to tag individual reviews for the presence of complaints on: (i) fake/scam apps, ii) ex-ante fees, iii) high interest, iv) short tenure, v) abusive staff, and vi) loss of sensitive or personally identifiable data. We allow that the same individual review may trigger multiple tags. After tagging the individual reviews, we then calculate the prevalence of these varied issues at the app- and time-level (quarterly). An important qualification is that we ultimately subset to lower-rated reviews (i.e., with a rating from 1 to 3) to circumvent and address the "fake" review issue.[19] We then regress the *Prevalence of consumer protection issues* on the use of *Suspect loan apps*, holding constant the country

[19] As noted previously, the subset of "fake" lending apps appear to use systematic review farming to drive uptake from unsuspecting real users. As this leads these apps to have an unusually high number of reviews with positive sentiment, this ends up distorting the prevalence measures.

of app availability and time fixed effects (at the quarter level). On the one hand, it is important to acknowledge that our method of classifying "suspect" apps makes it somewhat mechanical that they may be related to higher likelihood of consumer protection issues. On the other hand, it is worth highlighting that even "legitimate" apps—lending or otherwise—also see frequent criticisms and complaints from users. Thus, for purposes of this analysis, pertinent question are about "additionality" in a given market and time – i.e., exposure to consumer protection issues above and beyond market norms.

In Table 7, we observe that relative to similar loan apps in the same country market and time period, users of the suspect apps are 45% more likely to have any type of complaint. In terms of thematic area of complaint, they are 48% more likely to be a complaint about having fallen victim to a fake or scam apps, 41% more likely to complain about having paid an ex-ante processing or registration fee, 109% more likely to have dealt with abusive staff, and 61% more likely to signal concern over loss of sensitive data. There are some signs of higher complaints about higher interest and short tenure, but these are economically and statistically less significant. We conjecture that this may partially be explained by the fact that pure-play personal loan apps in general (i.e., even those from legitimate providers) often have fairly aggressive loan terms. As a whole, they depict a large economic magnitude of "additionality" in consumer protection issues surrounding these suspect apps and that their users appear to have been subject to considerable adverse effects.

Table 7: Relationship between exposure to "suspect" loan apps and prevalence of consumer protection issue
Standard errors clustered at the country-quarter level in parentheses. ***p<0.01, ** p<0.05, * p<0.1.

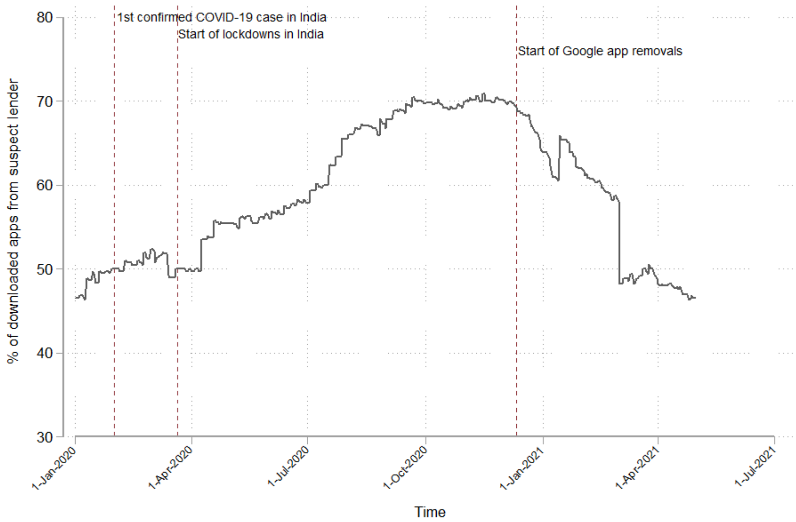
				Type of complaint			
DV= Ln(prevalence of CP issue)	Any	"Fake/scam" app	Process fee	High interest	Short term	Abusive staff	Sensitive data
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
"Suspect" loan app=1	0.451***	0.479***	0.411***	0.225	0.263	1.093***	0.613***
	(0.000)	(0.005)	(0.001)	(0.577)	(0.409)	(0.000)	(0.001)
Observations	3196	3196	3196	3196	3196	3196	3196
R^2	0.166	0.300	0.217	0.239	0.266	0.281	0.269
Additional controls							
Country (app availability in) FE	Y	Y	Y	Y	Y	Y	Y
Time (quarter) FE	Y	Y	Y	Y	Y	Y	Y

5.2 What are the negative externalities of "scam apps" on legitimate providers?

We also explore how the rise in "scam" finance apps imposes negative externalities on legitimate providers in a given market, focusing on the Indian market. We subset to the pure-play lending apps that are *only available* in the Indian market and draw on our prior categorization to separate the apps into *likely legitimate* versus *likely suspect* lending apps. For these two broad groups, we aggregate their downloads at the daily level for use as main dependent variables. We construct an explanatory variable, *Prevalence of suspect apps*, that captures the share of downloaded personal loan apps within a given market and point in time (i.e., at the daily level) that are likely suspect vs. likely legitimate. Then, we set up a basic regression model where we estimate the relationship between *Downloads of "likely legitimate" loan apps* on the *Prevalence of suspect loan apps*, using month fixed effects to account for seasonality.

We find evidence of a strong negative relationship in both absolute and relative terms. In Table 8, we observe that a one percent increase in the prevalence of suspect lending apps is associated with a reduction of $\tilde{1}$8,500 daily downloads of legitimate apps. This translates into a daily reduction in downloads of roughly 7.4% in relative terms. The results highlight that financial regulators and supervisors will need to increasingly address consumer protection issues in these app markets, which have until now been mostly lightly monitored. Left unchecked, such widespread instances could potentially erode trust in legitimate financial providers and further hinder efforts to promote financial inclusion in vulnerable populations, who are likely to be the primary users of such apps.

Figure 5: This figure depicts time series trends for the percentage of pure-play lending apps in India during our study period that are "likely suspect" (as opposed to "likely legitimate" based upon our manual classification). Note: this is a subset restricted to apps available only in the Indian market.



6 Application of machine learning to predict suspect apps

6.1 Methodology for machine-learning phase

Finally, having labelled the personal loan apps into various classes, we set up machine learning models as a proof-of-concept of the proposed solution. Our main task is to develop a supervised learning model that can predict the "class" of app as an output based on available app data as inputs. Generally, we use a machine learning technique called "gradient boosting" (specifically *xgboost*), which is considered state-of-art for many predictive classification tasks. It builds an ensemble of shallow and weak successive trees with each tree learning and improving on the previous. When combined, these many weak successive trees produce a powerful "committee" that are often hard

Table 8: Relationship between prevalence of suspect apps in market and uptake of legitimate apps
Standard errors clustered at the country-quarter level in parentheses. ***p<0.01, ** p<0.05, * p<0.1.

DV=Downloads of legitimate apps	Target study period (1) Absolute # DLs	(2) Relative (Ln.) Dls	Uncensored study period (3) Absolute # DLs.	(4) Relative (Ln.)
Prevalence of suspect apps (%)	-18,458*** (781.409)	-0.074*** (0.003)	-3,850*** (231.962)	-0.021*** (0.001)
Additional controls				
Month-of-year FE	Y	Y	Y	Y
Observations	486	486	684	684

Standard errors in parentheses; * $p < 0.10$ ** $p < 0.05$ *** $p < 0.01$

to beat with other algorithms. (Depictions of the model decision trees generated through our analysis are provided in Appendix Section C.) The key advantages of the gradient boosting approach are that it is generally more accurate compared to other modes, has high flexibility[20], and it handles missing data - imputation is not required. (Some of the disadvantages are that it may be more likely to overfit the data and training time scales with data size.) Key model features and steps are as follows:

1. The main output variables used for the model are the labeled data from the manual classification exercise (for the 5,106 personal loan apps). We use the labelled data to separate and test application of the machine learning model using a **binary classification** and **multi category classification**.
    - The **binary classification model** aggregates the legitimate and ambiguous categories together as being “likely legitimate” and the predatory and pure fraud categories as being “likely suspect”.
    - The **3-category classification model** also aggregates the legitimate and ambiguous categories together as being “likely legitimate”. However, it differs by further separating predictions between the “predatory” vs. “pure fraud” classes.
2. For the model input variables, we draw on the app meta and review data. Throughout the process of the manual classification task, we identify a subset of variables that are posited to be correlated with different classes. We filter to these variables and conduct transformations, when necessary. To provide a few examples: we use the date of initial app release and the date of last update to construct variables on number of days since release on the market and number of days since last update. We use the address variable to extract and construct dummy variables for whether any address is listed at all and if any specific street address is listed, etc. With the constructed variables, we test:
    - Models that are focused primarily on *ex-ante* indicators–i.e., those that are static and not reliant on review information. Example of such inputs include whether or not the app has a website listed, size of the app, whether the app lists a generic email (e.g., gmail, outlook), etc. We prioritize these inputs for establishing a baseline model, as one of our objectives is to see whether we can predict class even prior to high volumes of (critical) reviews coming in.
    - Models that incorporate both static and dynamic information from the review data. Examples of the latter inputs include the number of reviews, number and distribution of ratings, etc. These additional inputs are used to test how much inclusion of more dynamic data (particularly, those relying on more users and reviews to accumulate) influences the predictive accuracy of our approach through additional sensitivity tests.
3. In all of the above cases, we split the sample into a training-validation dataset and test dataset for the main analysis. Specifically, we train and validate the model on a dataset of apps released until January 15th, 2021. This produces an inferred function, which is used for mapping new examples. We then apply this to the test dataset, which includes apps released after January 15th, 2021. This is used to assess out-of-sample performance of the model on completely unseen data.
4. We evaluate model performance of the learned functions through several means. We look at their i) raw accuracy in predicting class, ii) depict standard receiver operating characteristic curves (or ROC curve), and

[20]For example, one can optimize on different loss functions and it has several hyperparameter tuning options that make function fit very flexible.

iii) measure the area under the ROC curve (or AUC). ROC curves are commonly used graphical plots that illustrate the diagnostic ability of a classifier system as its discrimination threshold is varied. Meanwhile, the AUC provides an alternative aggregate measure of performance across all possible classification thresholds.

5. In practice, we also conduct a number of further sensitivity tests to compare performance and gain some insights on the decisions made on various aspects of the method influence results and predictive accuracy of the model. Specifically, we explore how inclusion of time dynamic review input variables, use of hidden software development kit (SDK) data, and using a narrower time frame for training data influence results.

6.2 Results

In this section, we summarize results from the machine learning models. For sake of exposition, we designate a "baseline" model in which: 1) we use 31 "static" input features drawn only from the meta data and 2) we do not impose any date restrictions for the apps included in the training dataset. We depict results from both the binary and multi-class model focusing on highlighting model accuracy and evaluation using ROC curves and the AUC metrics. In practice, the approach uses a bootstrap method which takes a very large number of randomly selected 80-20 sample splits from the training and validation dataset to build the initial functional form. The accuracy for the validation set thus reflects the average predictive accuracy across these bootstrapped models. The functional form learned from the training and validation dataset is then applied to the test dataset as to test accuracy of classification on completely unseen data.

In Table 9, we observe that the baseline model exhibits a 89.6% accuracy in predicting between "likely suspect" vs. "likely legitimate" apps among the pure-play personal loan apps in the training and validation dataset. More importantly, it still exhibits high accuracy (roughly 83.1%) in predicting these binary classes in the withheld out-of-sample test dataset, which reflects unseen and more recently released loan apps. When we move to a multi-class model, there is some reduction in accuracy, but it still performs fairly well. Specifically, the baseline model has 85.6% accuracy in predicting among 3 classes ("likely suspect" v. "likely predatory" v. "likely fraud") among the pure-play personal loan apps in the training and validation dataset and around 72.2% accuracy in the withheld out-of-sample test dataset.

The ROC curves for the binary and multi-class validation and out-of-sample test sets can be found in Figures 6 and 7, respectively. The area under the ROC curve (or AUC) for the binary classifier is outstanding for the training and validation dataset and still strong for the out-of-sample test dataset.[21] For the multi-class classifier, the AUC is particularly strong in separating out the "likely legitimate" class (denoted as "class 0" in the figure). However, we do observe some drop off in the AUC for the "likely predatory" and "likely fraud" (denoted a classes 1 and 2, respectively in the figures). We interpret this to imply that the current baseline model may be sometimes mixing identification between the latter two classes. In other words, it performs well in identifying that there is some suspect behavior, but does sometimes have greater difficulty in separating between "pure fraud" vs. "predatory" cases.

Nevertheless, the overall results demonstrate that available high-frequency app data and applied machine learning techniques can perform quite well in efficiently and accurately flagging apps that exhibit suspect characteristics or behavior. On the one hand, we believe even the current methods can increase efficiency in vetting the

[21] Hosmer Jr et al. (2013) note that there is no "magic number" for what area under the ROC curve describes good discrimination, but provide a general rule of thumb that very high AUCs (.95 or higher) are sought, but AUC values of .70 and higher would be considered strong effects.

tens of thousands of newly-released finance apps per year and improve efforts to monitor available apps already published on the app stores. On the other hand, there is understandably still room to improve both the specific classification systems and the predictive accuracy of models. This is currently the case particularly in separating between suspected "predatory" vs. "fraud" classes. As such, in cases where suspect applications are flagged, subsequent manual review would understandably be recommended as a further check against false positives, before taking further action (e.g., to request direct take-downs from the app stores). However, the approach would still be expected to considerably reduce burden of manually verifying the thousands of new finance apps being newly released on a given country's app market per year and speeding up the rate at which they can be caught.

Table 9: Training and out-of-sample model accuracy–Baseline model

	Validation Set	**Out of Sample**
% Accuracy, Binary class model (Baseline)	89.57%	83.14%
% Accuracy, Multi-class model (Baseline)	85.61%	72.16%

Figure 6: ROC curves—Baseline model using binary class outcomes

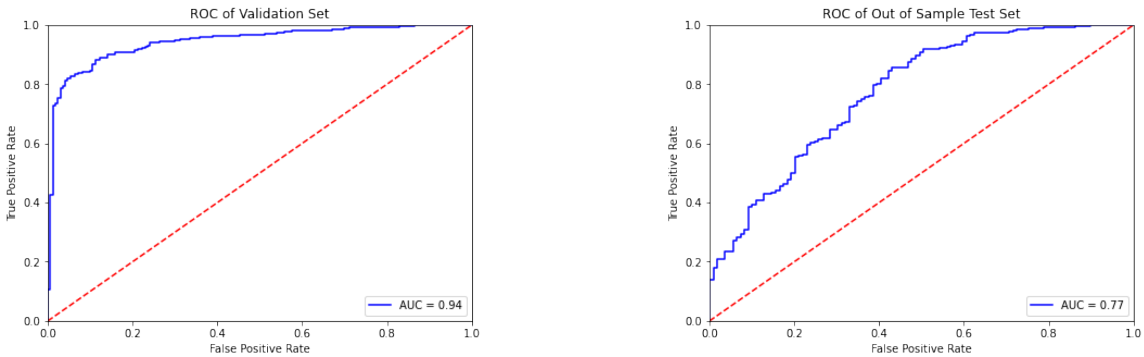
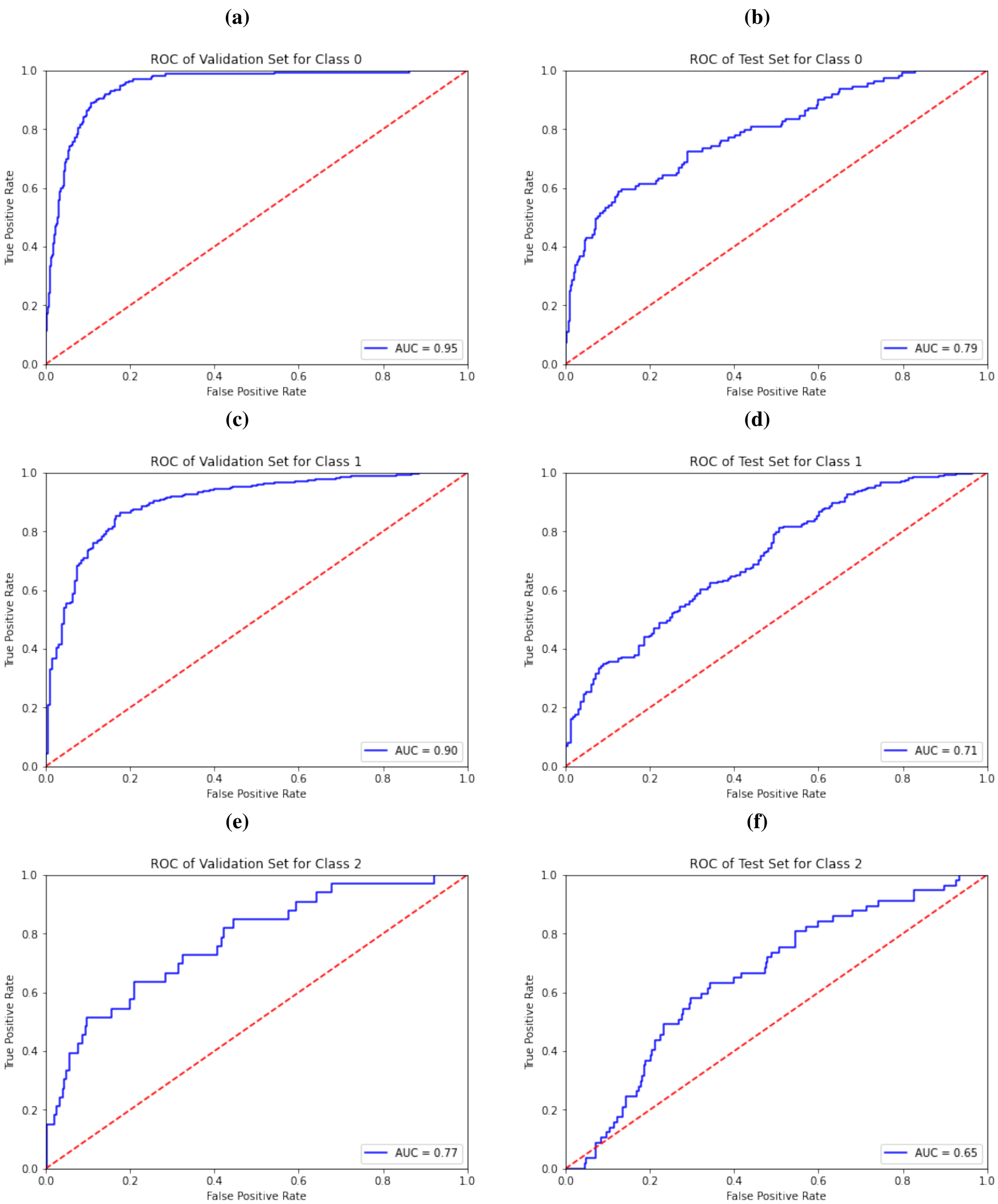


Figure 7: ROC curves—Baseline model using multi class outcomes



6.3 Sensitivity tests

Does inclusion of time dynamic review inputs improve model performance?

In the baseline model, we focus primarily on *ex-ante* indicators–i.e., those that are static and not reliant on review information. The motivation was to see whether we can predict class without having to be dependent on high volumes of reviews—likely critical ones—coming in. Nevertheless, it is important to also assess whether and how much inclusion of time dynamic meta and review data may influence the predictive accuracy of the approach. Consequently, we set up additional tests that include a number of such dynamic "ex-post" variables as input features, to see how their inclusion influences model performance. For example, this includes the number of downloads, number of combined ratings, number of 1-5 star ratings, mean ratings, and so forth.

In practice, we do observe that this time dynamic "Ex-post" model sees some improvement in accuracy when applied to classifying both the training and validation dataset and the out of sample dataset. However, the size of this increase is not particularly large. As depicted in Table, 10, the improved accuracy is typically less than 1 percentage point, in both the binary and multi-class model. On the one hand, this may offer some suggestion that the predictive models can function effectively absent use of review data and that dynamic features do not necessarily lead to drastic improvements to performance. On the other hand, it is also possible that to yield greater improvements to model accuracy, we may need to draw on more detailed further use of the disaggregated reviews– e.g., application of more sophisticated natural language processing techniques on the disaggregated reviews, to more accurately tag varied positive and negative issues associated with a given app. This latter approach would be an important further step for the future.

Table 10: Training and out-of-sample model accuracy–Ex-post model

	Validation Set	**Out of Sample**
% Accuracy, Binary class model (Baseline)	89.57%	83.14%
% Accuracy, Multi-class model (Baseline)	85.61%	72.16%
% Accuracy, Binary class model	90.17%	84.45%
% Accuracy, Multi-class model	85.26%	72.69%

Figure 8: ROC curves—Ex-post model using binary class outcomes

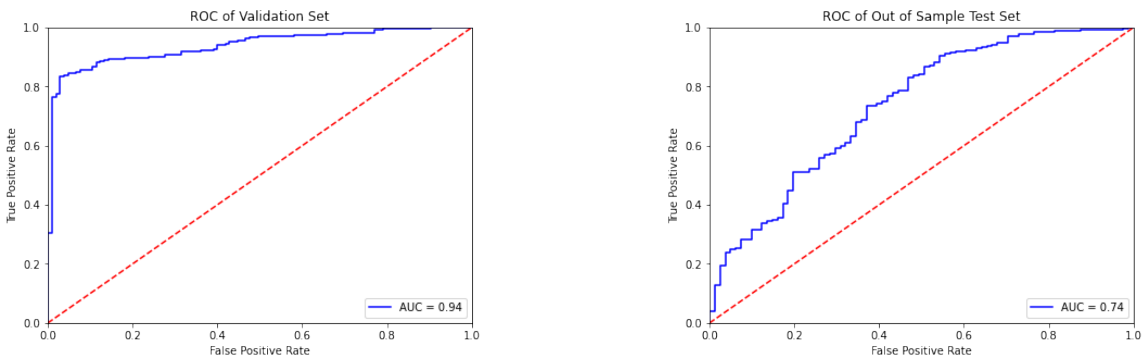
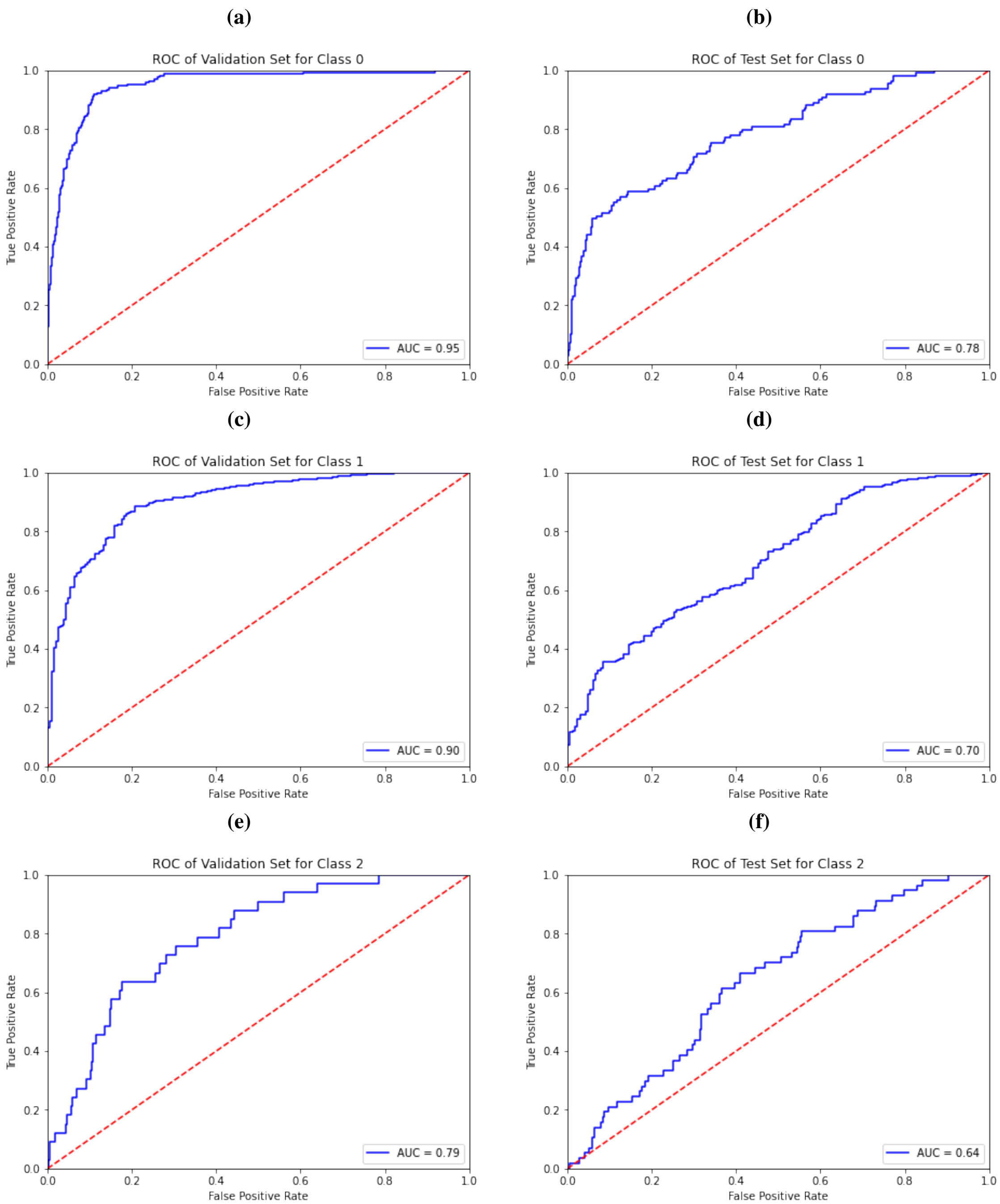


Figure 9: ROC curves—Ex-post model using multi class outcomes



Does a narrower timeframe for training data improve model performance?

Another concern may be that "scam" app developers are evolving various aspects of their apps over time. If such adaptions occur over time, then training and labeled data can become obsolete and older apps less predictive as comparison groups. As one way of testing this hypothesis, we run another sensitivity test where we restrict the training and validation dataset to apps that are initially released after January 1st, 2020. In practice, this removes roughly 1,500 of the pure-play personal loan apps from the training and validation dataset. The test dataset is left as is. That is, after training the model on the time restricted validation dataset, we apply it to the same test dataset as before.

Table 11 depicts the accuracy for these additional models. We observe that this "restricted time period" model indeed outperforms the baseline model in the training and validation and out-of-sample dataset in terms of accuracy for the binary classifier. For the multi-class classifier, it also outperforms the baseline model in predicting the validation dataset, but sees slightly worse performance in the out-of-sample dataset. Closer examination of the ROC curves and their AUCs in Figures 10 and 11 also show that these restricted training time models perform quite well in predicting some classes. This is despite being run on a sample size that is roughly one-third smaller. The slightly worse performance for the multi-class classifier seems to derive from prediction on the "pure fraud" cases (Class 2). This may signal that some of the useful app examples of fraud from which the model learned useful patterns may be from an earlier period.

Ultimately, however, our assessment is that the overall accuracy does not appear to dramatically change. On the one hand, this may suggest that changes to sample time frame has limits in terms of whether and how much it can improve model accuracy. On the other hand—and as a silver lining, it also provides some initial evidence that mitigates concerns that older apps and labeled data may become too quickly rendered obsolete for predicting the classes of more newly-released apps.

Table 11: Training and out-of-sample model accuracy–Restricted training time period model

	Validation Set	**Out of Sample**
% Accuracy, Binary class model (Baseline)	89.57%	83.14%
% Accuracy, Multi-class model (Baseline)	85.61%	72.16%
% Accuracy, Binary class model (Restricted training time)	92.66%	84.74%
% Accuracy, Multi-class model (Restricted training time)	86.71%	71.96%

Figure 10: ROC curves—Restricted training time period model using binary class outcomes

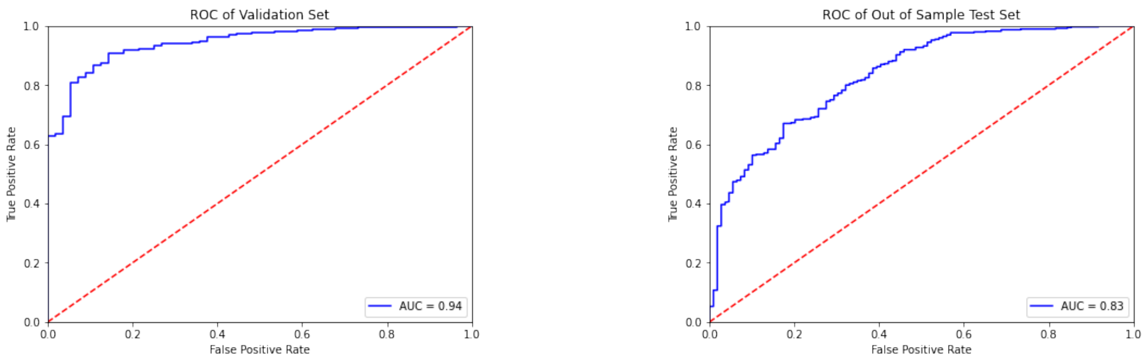
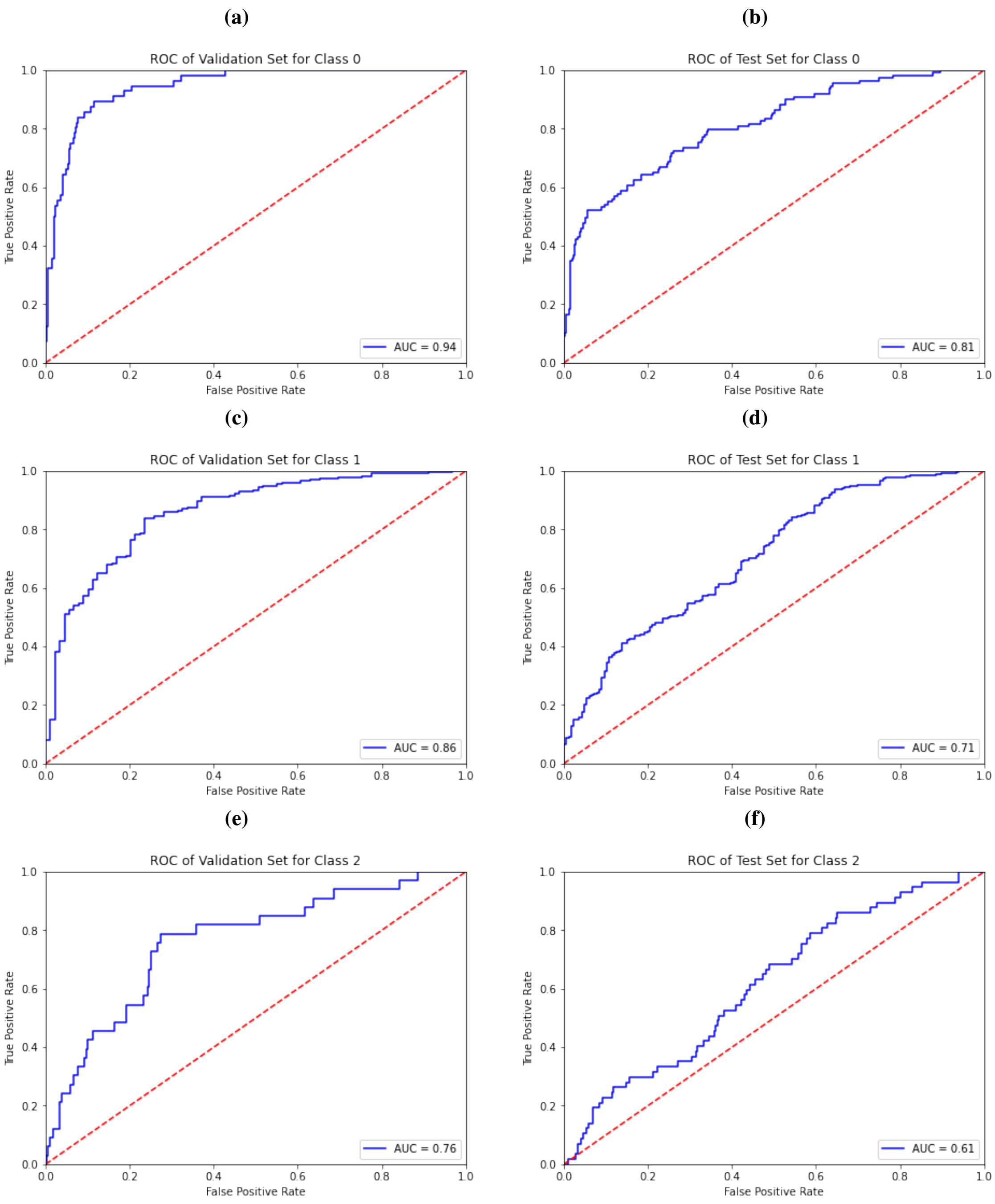


Figure 11: ROC curves—Restricted training time period using multi class outcomes



Can use of further hidden app data improve model performance?

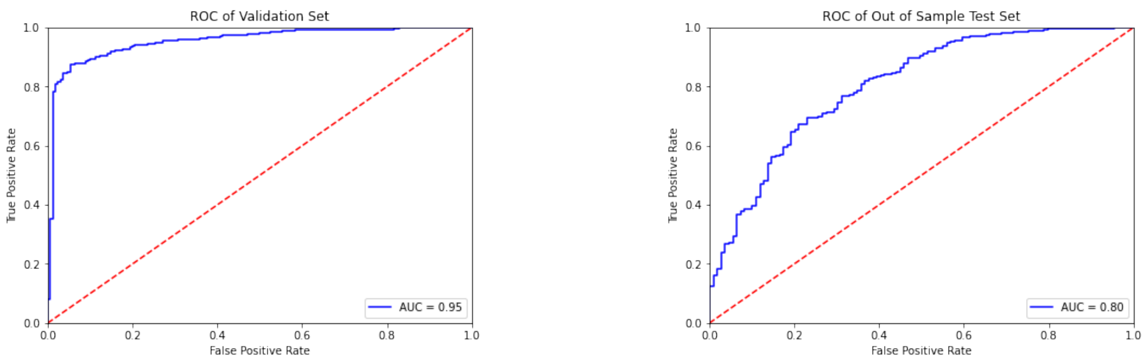
The app meta data also includes various hidden data on the underlying software development kit (SDKs) that a given app uses. In practice, there are over 5,000 SDKs and the combinations of SDKs used may be idiosyncratic and offer ways to identify related apps. We thus also test some preliminary models drawing on the SDK data as input variables. As an initial exploration, we identify the top 100 SDKs in terms of usage across the loan apps. For now, we avoid using SDKs that are unique to single or only a handful of apps to avoid potential overfitting at this stage. However, in the future, one could imagine using the more specific SDK signatures as a way to track down related apps. This seems pertinent since there are signs that many of the underlying apps are permutations of the same underlying shell. We construct dummies for this range of individual SDKs and then add them to our prior baseline model. Appendix Figure 17 depicts the variable importance list and show that the approach then considerably expands the input feature list relative to our baseline model.

In terms of performance, we observe that this “SDK-based” model slightly outperforms the baseline model in the training and validation dataset in terms of accuracy, but sees slightly worse performance in the out of sample data. This likely signals that the current exploratory approach is still only retaining SDKs that are too ubiquitous and not adding much further information value. As such, we posit that a more productive use of the SDKs may be for targeted cases, when we have a concrete lead on a specific suspect app and want to use its unique SDK signatures to identify parallel apps (e.g.., permutations by the same developer). We do not attempt for the pilot given time constraints, but believe this to be a promising method that warrants further work.

Table 12: Training and out-of-sample model accuracy–SDK-based model

	Validation Set	**Out of Sample**
% Accuracy, Binary class model (Baseline)	89.57%	83.14%
% Accuracy, Binary class model (With SDKs)	90.65%	82.75%

Figure 12: ROC curves—SDK-based model using binary class outcomes



7 Conclusions and recommendations

7.1 Key conclusions

This pilot study offers a “proof of concept” for the use of high-frequency app data and applied machine learning techniques to create a system for flagging and reporting on problematic finance applications. It demonstrates that such an approach can be used to efficiently and accurately flag apps that are exhibiting suspect characteristics or behavior. In particular, model input features can consist solely of static indicators that circumvent the need for relying upon review or rating data—e.g., would not necessarily require waiting for negative consumer complaints to come in to identify likely problematic apps. Given the documented scale of the finance app markets and signs of high “entry and exit” rates amongst apps of interest, this can improve efforts both to vet newly released apps and to monitor available apps already published on the app stores.

In cases where suspect applications are flagged, subsequent manual review would understandably be recommended as a further check against false positives, before taking further action (e.g., to request direct takedown from the app stores). However, such an approach could considerably reduce the burden of manually verifying the thousands of new finance apps being newly released on a given country’s app market per year and speeding up the rate at which they can be caught.

Different machine learning model set-ups are found to have some influence on predictive accuracy, however. On the one hand, this will require further fine-tuning of the classification methods and model specifics, with direct input from key stakeholders who may benefit from use of applied vetting or monitoring tools. On the other hand, this may generally suggest that using an ensemble of approaches may actually offer a promising way to triangulate upon suspects apps of interest.

7.2 Recommendations for next steps

Application to real-time data and real-world scenarios

The findings from the pilot have potential relevance and use-cases for a wide range of stakeholders, including country financial regulators, global standard setting bodies and international organizations, civil society organizations, and the app stores themselves. Real-world applications of the methods and findings include the following:

1. **Country-level financial regulators and supervisors:** the models can be transferred and applied to real-time app data to predict whether apps currently available in the given country market are likely to fall under legitimate versus various suspect categories. Real-time dashboards or periodic (e.g., monthly or quarterly) reporting could then be set up both to provide i) broad overviews of a given country’s app market and ii) targeted shortlists of highly suspect apps for further review. Following review of individual problematic apps or spikes in cases in the broader market, regulators could ask the app stores for targeted app take-downs or set up demand-side interventions to (e.g., just-in-time public awareness campaigns) warning consumers and profiling details around the problematic apps.
2. **Global standard setting bodies and international development organizations:** we observe that the “scam” finance app providers are highly fluid across markets. They are seen to rapidly proliferate in a given country market and then shift to new country locations once they are exposed and no longer seeing uptake in the former market. As such, some amount of global-level coordination and monitoring is required to track

these broader shifts in targeted countries. Global standard setting bodies and international development organizations would be well-placed to serve such a role. The methods used in this pilot could be readily converted into global-level reporting that could flag countries that appear to be current "hotspots" of suspect activity. To give two concrete examples, we could construct *aggregated* high-frequency indices measuring: i) the current levels of app "churn" across each countries' app markets or ii) the prevalence or predicted "scam" app. This could help more efficiently channel attention or funding to the most critically exposed locations at a given point in time.

3. **Consumer protection organizations and fintech associations:** similar to the use case for financial regulators, real-time dashboards or periodic (e.g., monthly or quarterly) reporting could be set up to provide such organizations overviews of the prevalence of "scam" finance apps in a given country's app market. Upon signs of spikes in problem cases in the market. This information could feed into targeted consumer protection interventions (e.g., public awareness campaigns) profiling details surrounding such issues;
4. **App stores/platform companies:** there may be collaborative opportunities for the app stores to draw on some of the labelled test datasets and methods to improve in-house detection. The app stores understandably have dedicated resources to general fraud detection and mitigation. However, we observe that a certain subset of these "scam" finance apps persists in reappearing and remaining published for some period. This suggests there may be important differences in some of the specific methods "scam" or "predatory" finance app developers are using that pass the app stores' current vetting methods. As such, some of these data and model learnings may be relevant to helping close off these gaps. There may also be opportunities to collaborate with other stakeholder groups to develop an alternative rating system (or "buyer beware" labeling) on app store fronts for finance-related apps to help consumers better spot and avoid risky apps.
5. **Financial crime / law enforcement agencies:** Finally, it is worth mentioning that there are signs that many of the fraud apps in a given market are coming from the same suppliers or developers. By creating a global catalogue of high-risk apps, it may improve ability to try and track down the most prolific of these app providers and target them with legal action (e.g., block their payment, bring criminal or legal cases, enact fines, etc.) in order to dissuade future replication from others.

Further work on the methodology

The pilot also informs further work required to improve the efficiency, utility and accuracy of the methods and subsequent tools and interventions. A few high priority next steps include the following:

- The current primary classification method used heuristic techniques that were acquired upon close review of the app's data. The resulting classification yielded "suspect" classes that indeed seem much more likely to be removed from the app store. To increase efficiency moving forward, these signals could now be converted into a rule-based classification method to automate and ensure standardization of the approach.
- The pilot workflow should be adapted to allow for intake of real-time data feeds. This could leverage API feeds of existing data providers and allow for increasing the frequency of model updating and reporting.
- The targeted app coverage should increasingly be expanded to include the full range of finance product types (beyond personal loan apps). This will increase overall sample size and likely contribute to further improvements to the accuracy of the models. However, it will be important to similarly scrutinize a (randomly drawn) sample of these non-personal loan apps to similarly develop heuristic signals and then convert them into rule-based classification systems.

- In a similar vein, the non-English apps should incorporated into the process. This is understandably an important gap of the current pilot since a large number of potentially exposed emerging and developing economies are non-English speaking countries. This effort will require careful attention, however, to find ways to integrate methods using text-based analysis or natural language processing across multiple languages. It would be advised to similarly pilot such a process with other languages relevant for large combined market segments (e.g., Spanish, Portuguese, Indonesian, etc.) in a stepwise manner.
- More sophisticated natural language processing techniques—such as Bidirectional Encoder Representations from Transformers (BERT)—can be integrated into various stages of the process to improve: 1) the initial categorization of apps by product types, 2) the tagging of disaggregated review data by key consumer protection theme, and 3) helping with other classification tasks where analysis of app descriptions play a role.

References

Demirguc-Kunt, A., Klapper, L., Singer, D., Ansar, S., & Hess, J. (2018). *The Global Findex Database 2017: Measuring financial inclusion and the fintech revolution*. The World Bank.

Ferrari, A., Masetti, O., & Ren, J. (2018). Interest rate caps: The theory and the practice. *World Bank Policy Research Working Paper*(8398).

Fu, J., & Mishra, M. (2021). Fintech in the time of COVID-19: Technological adoption during crises. *Journal of Financial Intermediation*, 100945.

GSMA. (2020). The mobile economy 2020. *London: GSM Association*.

Heidhues, P., & Kőszegi, B. (2010). Exploiting naivete about self-control in the credit market. *American Economic Review*, *100*(5), 2279–2303.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Lusardi, A., & Mitchell, O. S. (2011). Financial literacy around the world: An overview. *Journal of pension economics & finance*, *10*(4), 497–508.

Lusardi, A., & Mitchell, O. S. (2014). The economic importance of financial literacy: Theory and evidence. *Journal of economic literature*, *52*(1), 5–44.

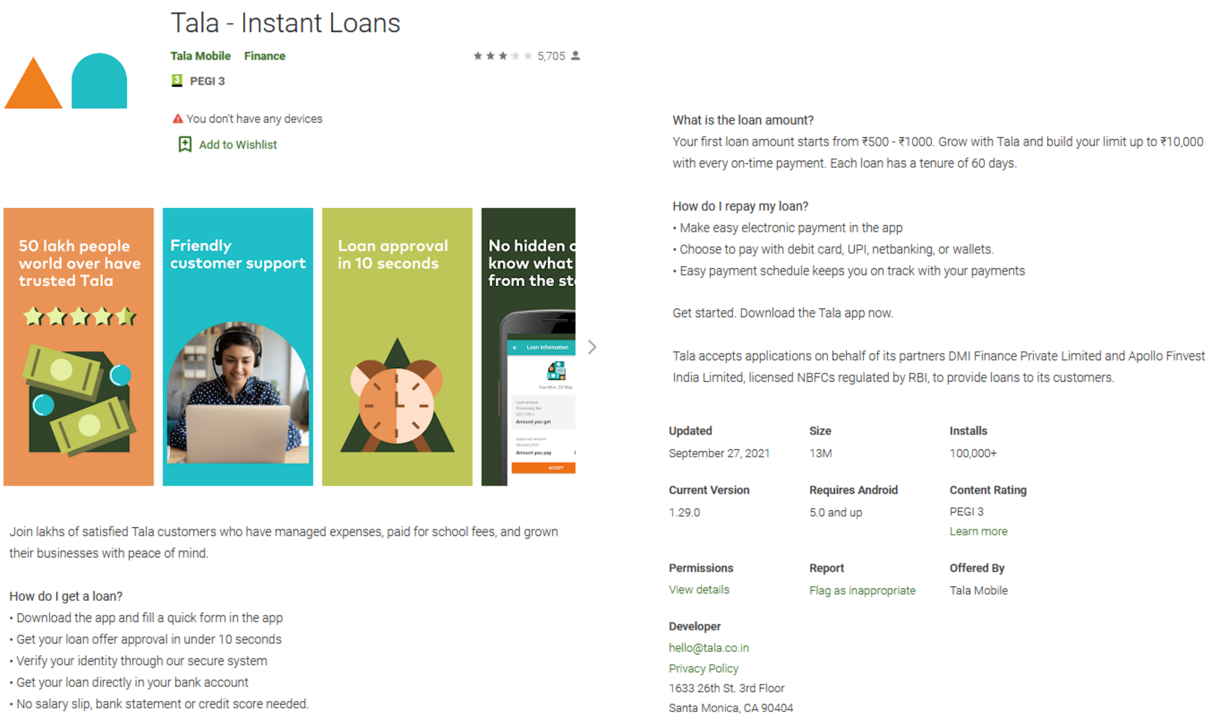
A Country Sample

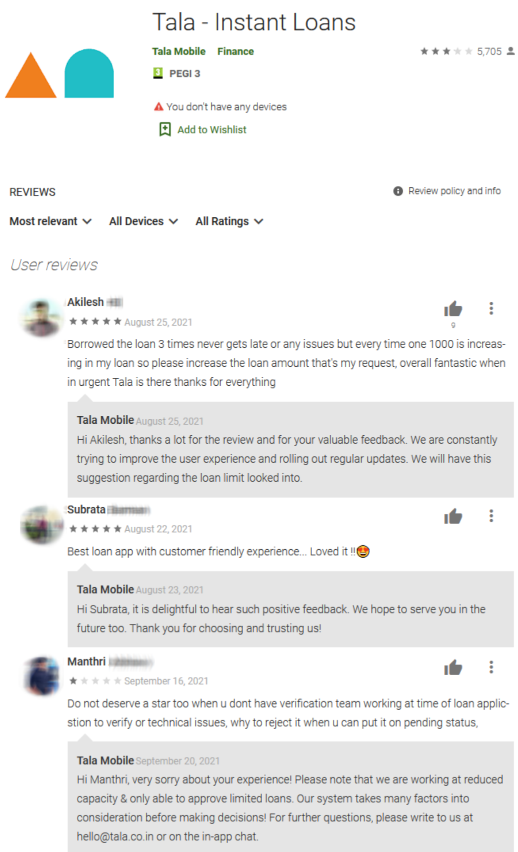
Table A.1: Overview of countries in data sample
This table lists the 63 countries in our data sample.

Country Code	Country	Country Code	Country
AE	United Arab Emirates	IR	Iran
AM	Armenia	IT	Italy
AR	Argentina	JP	Japan
AT	Austria	KR	South Korea
AU	Australia	LB	Lebanon
BE	Belgium	LT	Lithuania
BG	Bulgaria	LV	Latvia
BR	Brazil	MX	Mexico
BY	Belarus	MY	Malaysia
CA	Canada	NG	Nigeria
CH	Switzerland	NL	Netherlands
CL	Chile	NO	Norway
CN	China	NZ	New Zealand
CO	Colombia	PE	Peru
CY	Cyprus	PH	Philippines
CZ	Czech Republic	PL	Poland
DE	Germany	PT	Portugal
DK	Denmark	RO	Romania
EE	Estonia	RS	Serbia
EG	Egypt	RU	Russian Federation
ES	Spain	SA	Saudi Arabia
FI	Finland	SE	Sweden
FR	France	SG	Singapore
GB	United Kingdom	SK	Slovakia
GE	Georgia	TH	Thailand
GR	Greece	TR	Turkey
HK	Hong Kong	TW	Taiwan
HR	Croatia	UA	Ukraine
HU	Hungary	US	United States
ID	Indonesia	VN	Vietnam
IL	Israel	ZA	South Africa
IN	India		

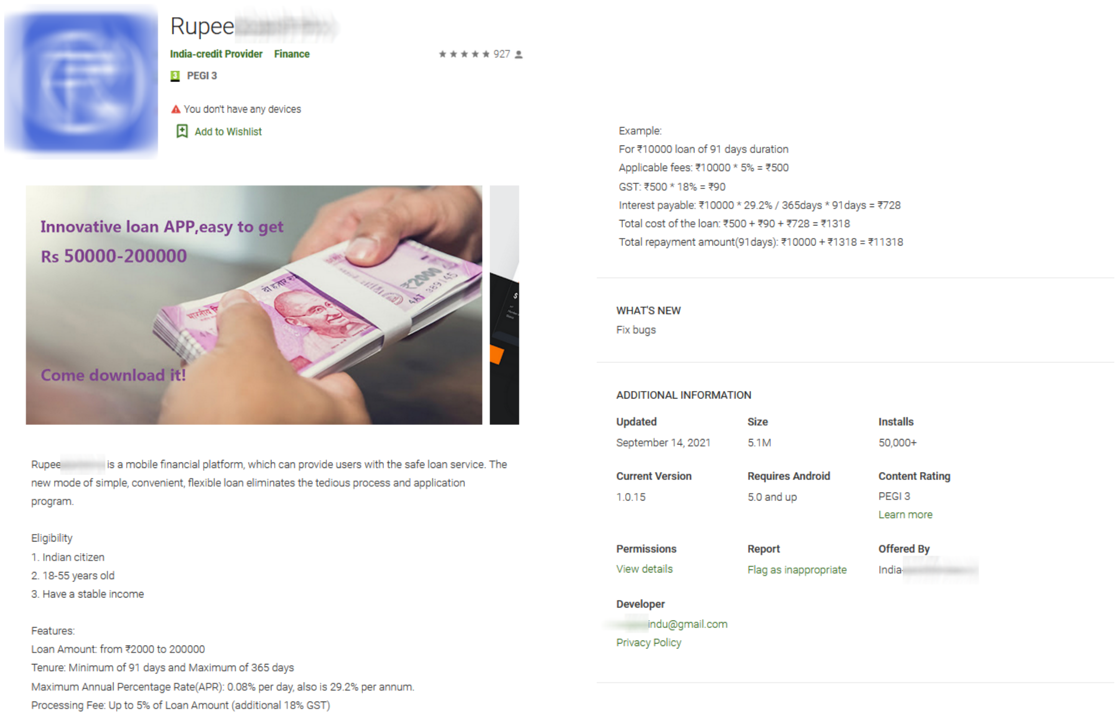
B App classification examples

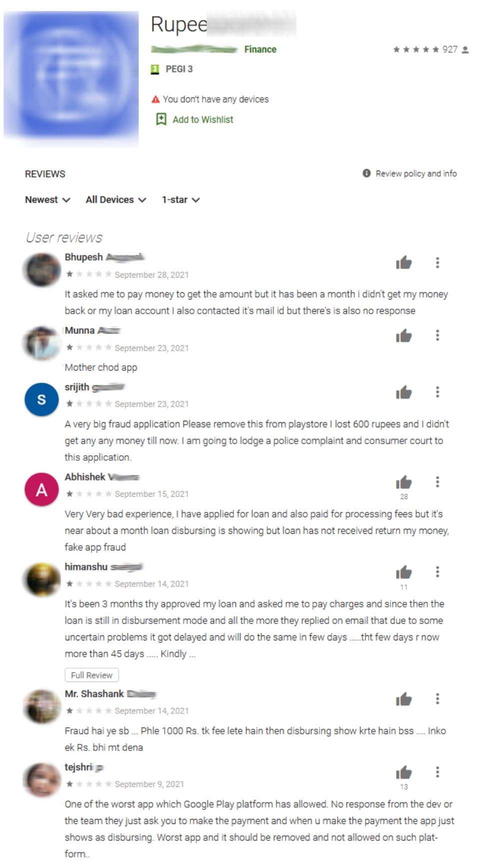
B.1 Example of "Legitimate" app class

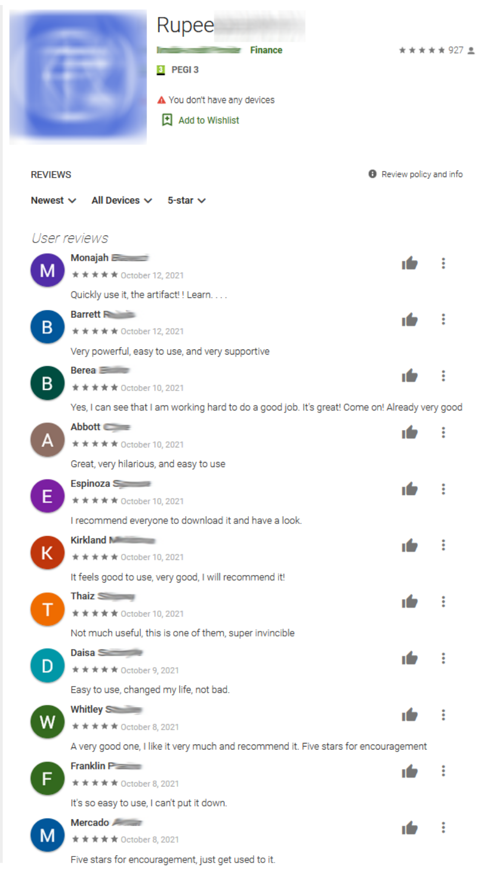




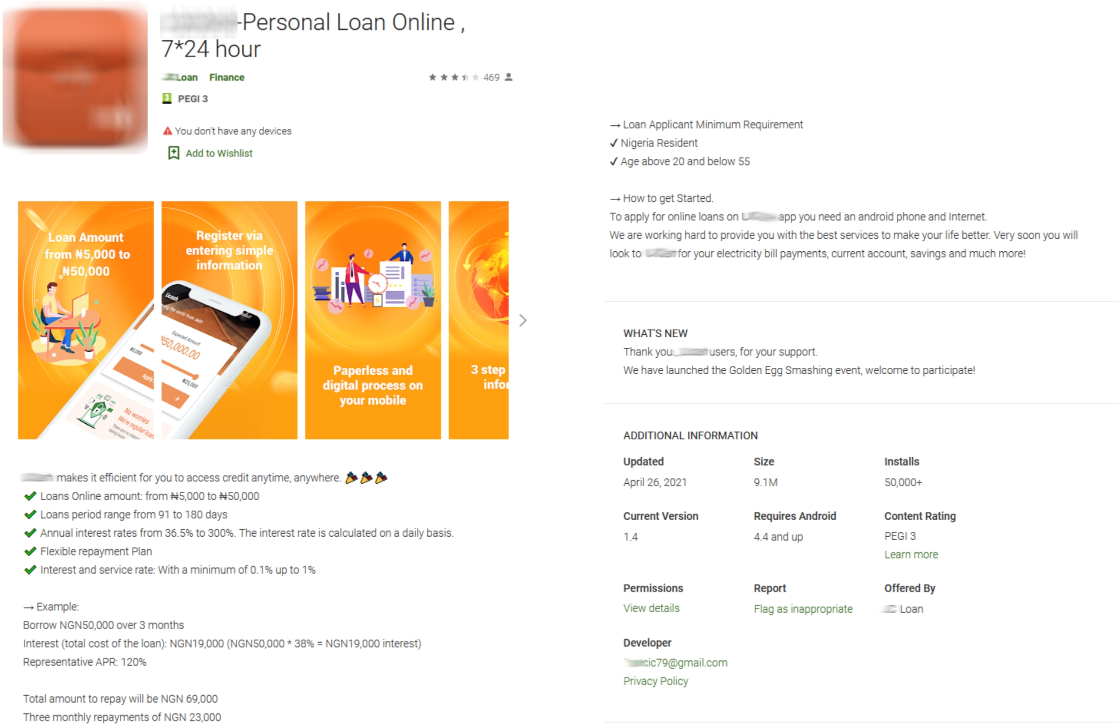
B.2 Example of “Pure fraud” app class

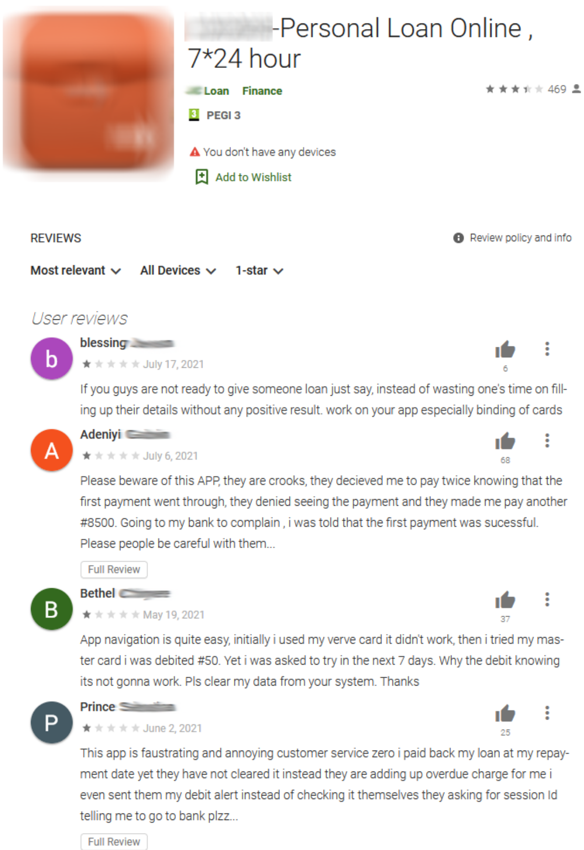


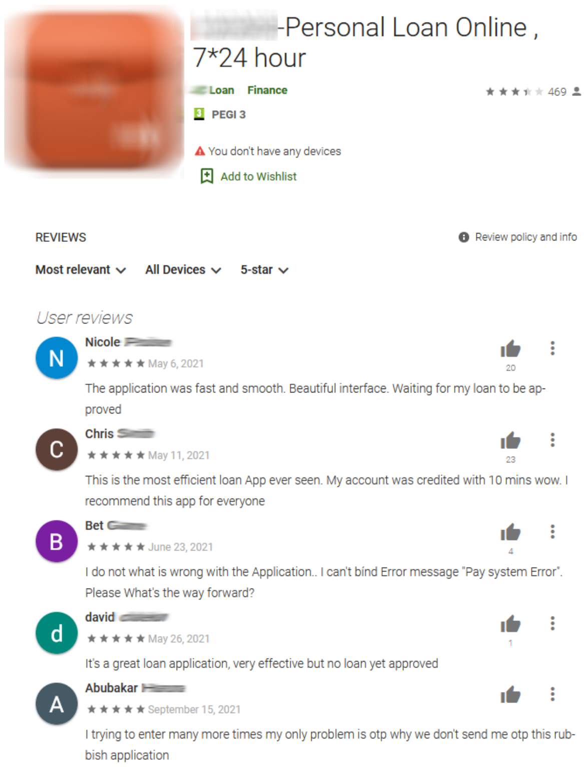




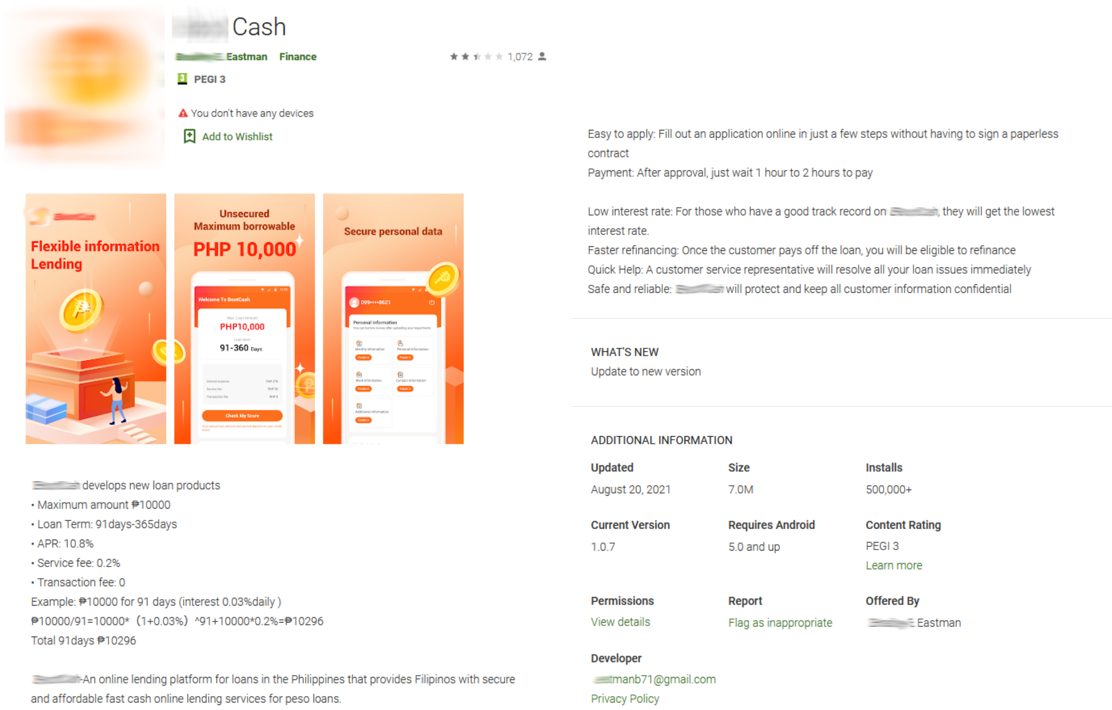
B.3 Example of “Predatory, no service” app class

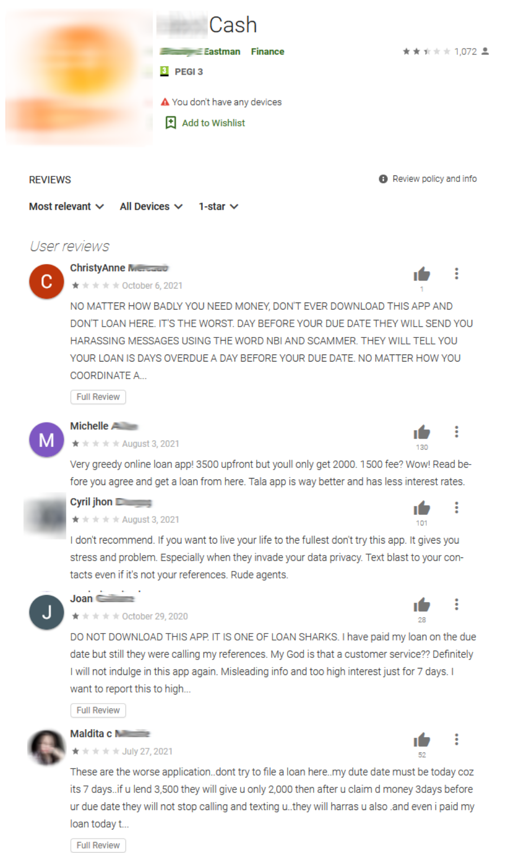


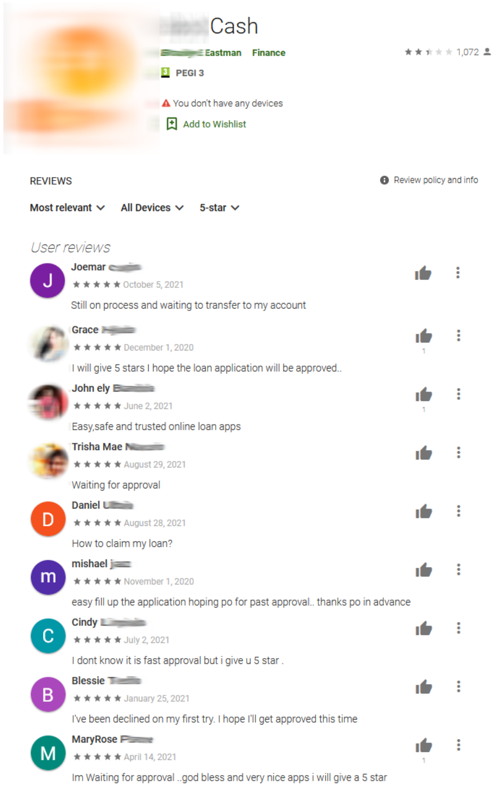




B.4 Example of “Predatory, problematic services” app class







C Additional ML Model Figures

Figure 13: Decision tree for baseline model using binary class outcomes

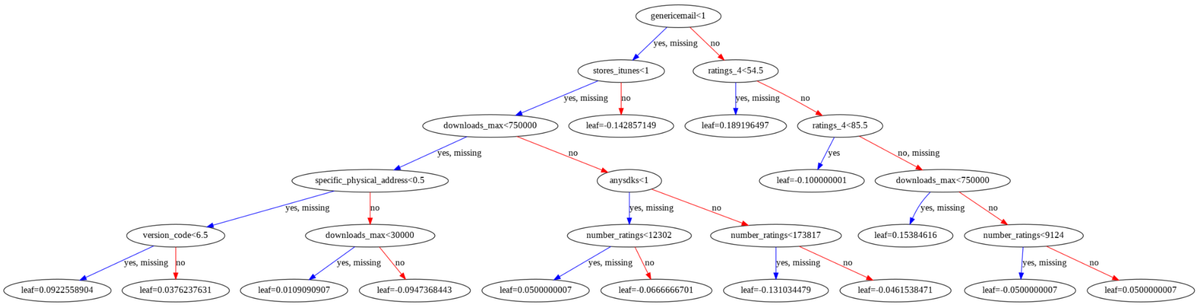


Figure 14: Variable importance list for baseline model using binary class outcomes

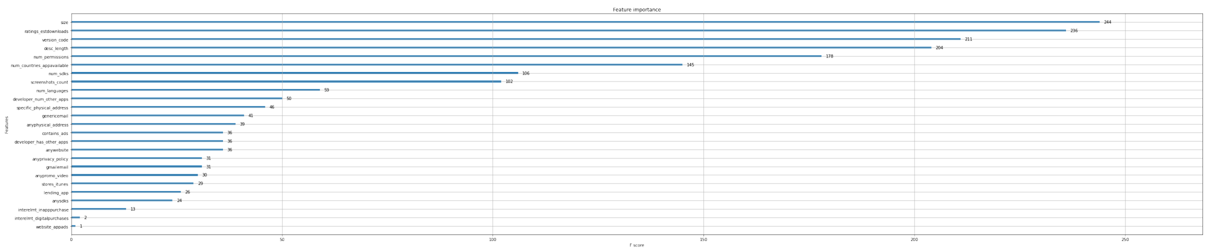


Figure 15: Decision tree for baseline model using multi class outcomes

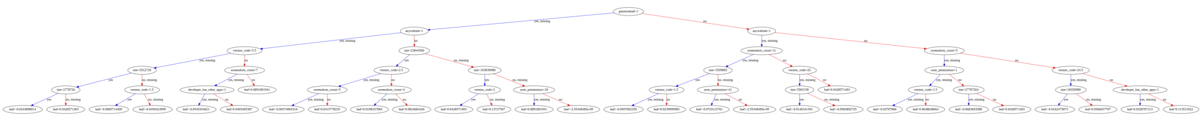


Figure 16: Variable importance list for baseline model using multi class outcomes

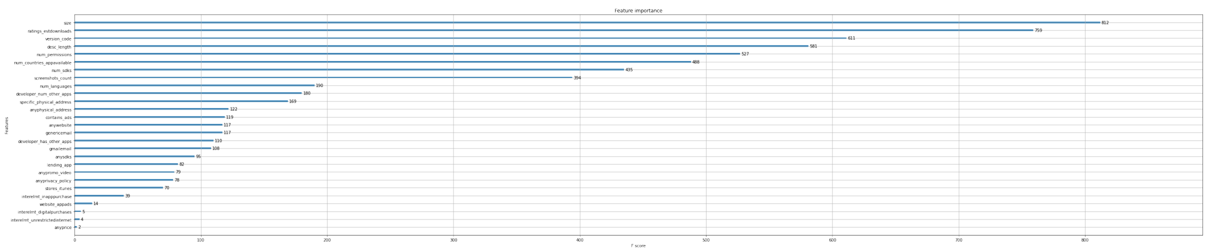


Figure 17: Example of variable importance list for SDK-based model using binary class outcomes