Fintech and Credit Scoring for the Millennials: Evidence using Social and Mobile Footprints

Authors: Sumit Agarwal, Shashwat Alok, Pulak Ghosh, and Sudip Gupta

NSE-NYU Conference -2019

Research Questions

- Does social and mobile footprints available from individuals' mobile phones predict loan outcomes?

    1. Likelihood of loan approval
    2. Loan Default
    3. Impact on loan approval & outcome if a social credit score is used for denied borrowers without traditional credit score

        - Prediction (not causal) Counterfactuals : What would have happened if we offered loans using social and mobile footprints to borrowers for whom traditional credit scoring is not available and hence generally denied credit by traditional banks?

        a) How many more would have been approved?
        b) What would have been the new overall default rate if we replaced the high risk approved with a medium risk denied?
        c) Estimation challenges: application of machine learning (ML)

- Use data from a large fintech lending firm in India

Motivation

- 2 billion people around the world still lack bank accounts (IMF 2017)
  - Even those with bank account lack credit access
  - Primarily due to the lack of credit history/credit score
- Not just a developing country phenomena: 45 million Americans do not have a credit score (Consumer Financial Protection Bureau, 2015)
  - Could be 'good borrowers' if their `credit worthiness' could be evaluated by alternate means.
- In India, about 850 million individuals have never taken credit
  - Out of these, Transunion estimates - 220 million are credit eligible
  - But , 150 million such credit eligible customers lack a credit history and score

Motivation

- While credit history is absent, access to internet through mobile phones has been growing
  - **98.7 per cent mobile phone adoption in developing countries**
  - **Internet and Social media use has increased exponentially**
- An overlap between potentially creditworthy individuals who lack credit access but have an online presence
  - Large traces of data
- Quest for an alternate credit Score
  - Zhima/Seasame credit from Ant financial
    - Social media interactions and purchase transactions
  - Fintech lending platforms that rely on digital footprint data for loan decisions have been mushrooming

Big Picture Question

- Can digital presence data be used to come up with an alternate credit score for unbanked customers?
- Why should we care?
  - A huge untapped market!
  - Lenders can potentially expand credit to the traditionally underserved
  - Policy implications for governments as they seek to expand credit access
- What do we know?
  - Berg, Burg, Gombovic and Puri (2019)
  - Iyer, Khwaja, Luttmer and Shue (2015)
  - D'Acunto, Rauter, Scheuch, and Weber (2019)

Our setting: Data

- Data: proprietary anonymized data on about 400,000 loan applicants from a mobile-only Fintech lending platform operating in India since 2016.
  - loans size: Rs 10,000 ($142) to 200,000 ($2846)
  - loan duration: minimum of 15 days to a maximum of 180 days.
  - 180,000 active users with 75% repeat borrowers
- All loan applications between February 2016 to November 2018
- The loan process: a customer has to download the app, enter all the requisite details, documentation and submit.
- Key variables:
  - Loan level: Loan size, interest rate, duration and purpose
  - Customer level: Age, salary, Job designation, Education level, Location
  - Customer's social and mobile footprint:
    - Mode of login (Linkedin vs Facebook),
    - Types of apps installed: E-commerce (Ex Amazon, Flipkart etc) , travel (Airbnb, Tripadvisor etc) , dating (Tinder etc), social media (Whastapp, Facebook, Twitter etc), financial apps (Banking and Stock trading)
    - Type of smartphone: IOS vs Android
    - Social footprints: Number of calls/sms/contacts/social media connections, incoming/outgoing/missed calls, durations

Key findings

1. Digital footprint data has a significant correlation with the loan approval probability.

   Those with a significant digital mobile presence are significantly more likely to get approved

2. Digital and Social footprint variables have higher (relative to credit score) predictive power in predicting defaults

   - as compared to the credit score, the digital footprint variables taken together are able to explain 10 percentage points higher variation in defaults.
   - AUC of credit score is 58.6% and of social footprint variables alone is 69%

3. The discriminatory ability of the digital footprint variables varies based on the loan purpose

   Customers who log in via Facebook are 20%, 26%, and 32% more likely to default when they take loans for making a purchase, meeting the EMI on an existing loan, or repayment of an existing loan respectively but less likely to default when they take a loan for medical needs.

4. (Prediction Counterfactual) If we choose a relatively conservative threshold of 10% predicted default probability, about 13% borrowers without the CIBIL score and denied loans would have been approved

Summary Statistics: Loan and Financial Variables

- Out of the 417,578 loan applications in our sample, 272,931 were approved while 144,647 were denied.
- Default rate in our full sample is quite high at approximately 13.5%
  - Has come down over time – 3% for loans granted in 2018
  - 3% otherwise in India for retail loans
- Average credit score is 634 (sub-prime borrowers) and 20% of borrowers don't have a credit score
  - caters to unbanked (higher risk?) borrowers
- Average loan size is Rs 22,000 ($314)
- Average annualized interest rate is 25%
- Average age is 32 (Younger customers)
- Average salary is Rs 37,524 ($527) per month or $6324 per annum
  - roughly 3 times the median per capita income of $2,134 in 2018.
  - Relatively well-off customers
- Loan purpose: 8% travel, 9% for EMI, 13 % for purchasing a good, 8% for the purpose of repaying a loan principal, and 22% for medical expenditure.

Summary Statistics: Digital Footprint Variables

- Log in
  - 27% of customers logged in using Facebook
  - 2% used Linkedin.
  - Rest by other means
- 68% customers have a mobile banking, mutual fund apps or stock trading app
  - Suggests potentially fintech savvy customers
- 51% of customers have installed another mobile-loan application
- 12% of the customers own an apple phone

Multivariate Analysis: Simple Logistic Regression

$$\text{Loan Outcome}_{ilt} = \beta_0 + \sum_{j=1}^{M} \beta_j \text{Loan Characteristics}_{lt} + \sum_{j=1}^{N} \beta_j \text{Customer financials}_{it} + \sum_{j=1}^{O} \beta_j \text{Customer mobile/social footprint}_{it} + \varepsilon_{ilt}$$

i refers to an individual, l refers to loan and t refers to the time of loan application

Loan outcome:

1. Approved is a dummy variable which takes the value one for loans that were approved and zero otherwise
2. Default which identifies loans in default

Dependent Variable: Loan Approval

VARIABLES	Odds Ratio (1)	Odds Ratio (2)	Odds Ratio (3)
Log of cibil	1.190***	1.016	1.024*
	(0.000)	(0.229)	(0.083)
Log of Salary			1.171***
			(0.000)
Log Loan Amount			0.733***
			(0.000)
Log Age			1.491***
			(0.000)
High School Dummy			1.071
			(0.153)
College Dummy			1.065
			(0.252)
Supervisor			0.875***
			(0.001)
Manager			0.903***
			(0.005)
Travel.purpose cashe			0.990
			(0.873)
EMI.purpose cashe			0.910
			(0.115)
purchase.purpose cashe			0.982
			(0.726)
Loanrepayment.purpose cashe			0.994
			(0.920)
Other purpose.purpose cashe			0.952
			(0.219)
Log no of SMS			
Log No of Contacts			
Log no of Apps			
Log Callog			
Dating App			
Finsavy App			
Socialconnect App			
Travel App			
Mloan App			
Facebook status			
Linkedin status			
IOS Dummy			
Constant	4.531***	35.951***	35.708***
	(0.000)	(0.000)	(0.000)
State Fixed Effects	N	N	N
Observations	235,765	189,055	189,055
Pseudo R2	0.00881	3.21e-05	0.00509
AUC	0.581	0.508	0.567

Key Takeaways

- Credit score is not strongly related to approval
  - Especially for customers with digital footprint
  - suggesting that the Fintech lender relies primarily on other parameters for loan approval.

- Customers that earn more, are older, and need smaller loans have a higher chance of approval.

Dependent Variable: Digital Footprint and Loan Approval

VARIABLES	Odds Ratio (1)	Odds Ratio (2)	Odds Ratio (3)	Odds Ratio (4)	Odds Ratio (5)	Odds Ratio (6)
Log of cibil	1.190***	1.016	1.024*			1.014
	(0.000)	(0.229)	(0.083)			(0.297)
Log of Salary			1.171***			
			(0.000)			
Log Loan Amount			0.733***			
			(0.000)			
Log Age			1.491***			
			(0.000)			
High School Dummy			1.071			
			(0.153)			
College Dummy			1.065			
			(0.252)			
Supervisor			0.875***			
			(0.001)			
Manager			0.903***			
			(0.005)			
Travel.purpose cashe			0.990			
			(0.873)			
EMI.purpose cashe			0.910			
			(0.115)			
purchase.purpose cashe			0.982			
			(0.726)			
Loanrepayment.purpose cashe			0.994			
			(0.920)			
Other purpose.purpose cashe			0.952			
			(0.219)			
Log no of SMS				0.989	0.990	0.987
				(0.175)	(0.204)	(0.116)
Log No of Contacts				0.990	0.989	0.987
				(0.562)	(0.516)	(0.469)
Log no of Apps				1.172***	1.178***	1.183***
				(0.000)	(0.000)	(0.000)
Log Callog				1.034***	1.036***	1.034***
				(0.004)	(0.002)	(0.004)
Dating App				0.924	0.921	0.946
				(0.351)	(0.333)	(0.527)
Finsavy App				1.203***	1.205***	1.218***
				(0.002)	(0.002)	(0.001)
Socialconnect App				0.916	0.954	0.946
				(0.340)	(0.610)	(0.559)
Travel App				1.005	1.001	0.986
				(0.883)	(0.969)	(0.698)
Mloan App				1.076**	1.076**	1.080**
				(0.018)	(0.019)	(0.015)
Facebook status				1.020	1.018	1.017
				(0.555)	(0.584)	(0.624)
Linkedin status				0.963	0.968	0.951
				(0.689)	(0.735)	(0.603)
IOS Dummy					1.439***	1.427***
					(0.001)	(0.001)
Constant	4.531***	35.951***	35.708***	17.301***	16.055***	15.264***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
State Fixed Effects	N	N	N	N	N	N
Observations	235,765	189,055	189,055	194,093	194,093	189,055
Pseudo R2	0.00881	3.21e-05	0.00509	0.00225	0.00256	0.00260
AUC	0.581	0.508	0.567	0.541	0.544	0.544

Key Takeaways

- The number of contacts, the number of apps installed, Finsavvy dummy, and Mloan app dummy are positively associated with loan approval.
- Customers with an IOS device have a 50% higher likelihood of approval.
  - Prior studies highlight that owning an IOS device is a strong predictor of higher earnings (Bertrand and Kamenica (2018)).
  - and lower defaults (Berg, Burg, Gombovic and Puri (2019)

- Bottom line: digital footprint variables have significant explanatory power for the likelihood loan approval.
  - over and above the credit bureau score.

Dependent Variable: Loan defaults

- AUC (Area Under the Curve) is a measure of the goodness of the model
- Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. (True positives)

VARIABLES	Odds Ratio (1)	Odds Ratio (2)	Odds Ratio (3)
Log of cibil	0.877***	0.900***	0.872***
	(0.000)	(0.000)	(0.000)
Log of Salary			0.240***
			(0.000)
Log Loan Amount			4.136***
			(0.000)
Log Age			0.545***
			(0.000)
High School Dummy			0.822***
			(0.000)
College Dummy			0.724***
			(0.000)
Supervisor Dummy			0.994
			(0.749)
Manager Dummy			0.932***
			(0.000)
Travel.purpose cashe			0.759***
			(0.000)
EMI.purpose cashe			0.838***
			(0.000)
purchasc.purposc cashc			0.816***
			(0.000)
Loanrepayment.purpose cashe			0.806***
			(0.000)
Other purpose.purpose cashe			0.861***
			(0.000)
Log no of SMS			
Log No of Contacts			
Log no of Apps			
Log Callog			
Dating App			
Finsavy App			
Socialconnect App			
Travel App			
Mloan App			
Facebook status			
Linkedin status			
IOS Dummy			
Constant	0.321***	0.256***	8.816***
	(0.000)	(0.000)	(0.000)
State Fixed Effects	N	N	N
Observations	219,219	184,423	184,423
Pseudo R-squared	0.00417	0.00238	0.0903
AUC	0.601	0.586	0.723

Key Takeaways

- AUC of the model using only the credit score for predicting defaults is 59%
  - is lower than 62% reported by Iyer, Khwaja, Luttmer and Shue (2015) based on a sample of loans from a US based peer to peer lending platform, “Propser.com” and 68.3% reported by Berg, Burg, Gombovic and Puri (2019) based on a sample of purchases from a German e-retailer.
  - Comparable to 59% AUC reported for LendingCLub - Berg, Burg, Gombovic and Puri (2019)
  - Suggests that Discriminatory ability of credit score maybe lower in emerging markets
  - The marginal value of digital footprint variables (additional information) is likely to be higher in such environments
- Default likelihood is lower for all categories of loans (Travel, EMI, Purchase, Repayment, and other) relative to loans taken for medical needs.
  - Consistent with the idea that health shocks are correlated with financial distress (Kalda (2019)).
- Salary, education, and job designation are negatively related to defaults.

Dependent Variable: Loan defaults

VARIABLES	Odds Ratio (1)	Odds Ratio (2)	Odds Ratio (3)	Odds Ratio (4)	Odds Ratio (5)	Odds Ratio (6)
Log of cibil	0.877***	0.900***	0.872***			0.906***
	(0.000)	(0.000)	(0.000)			(0.000)
Log of Salary			0.240***			
			(0.000)			
Log Loan Amount			4.136***			
			(0.000)			
Log Age			0.545***			
			(0.000)			
High School Dummy			0.822***			
			(0.000)			
College Dummy			0.724***			
			(0.000)			
Supervisor Dummy			0.994			
			(0.749)			
Manager Dummy			0.932***			
			(0.000)			
Travel.purpose cashe			0.759***			
			(0.000)			
EMI.purpose cashe			0.838***			
			(0.000)			
purchasc.purposc cashc			0.816***			
			(0.000)			
Loanrepayment.purpose cashe			0.806***			
			(0.000)			
Other purpose.purpose cashe			0.861***			
			(0.000)			
Log no of SMS				0.969***	0.968***	0.972***
				(0.000)	(0.000)	(0.000)
Log No of Contacts				0.964***	0.966***	0.976***
				(0.000)	(0.000)	(0.007)
Log no of Apps				0.659***	0.653***	0.656***
				(0.000)	(0.000)	(0.000)
Log Callog				0.917***	0.913***	0.915***
				(0.000)	(0.000)	(0.000)
Dating App				1.246***	1.252***	1.218***
				(0.000)	(0.000)	(0.000)
Finsavy App				0.709***	0.706***	0.745***
				(0.000)	(0.000)	(0.000)
Socialconnect App				1.331***	1.233***	1.288***
				(0.000)	(0.000)	(0.000)
Travel App				1.034*	1.041**	1.038**
				(0.052)	(0.019)	(0.031)
Mloan App				1.002	1.002	1.003
				(0.908)	(0.892)	(0.822)
Facebook status				1.091***	1.095***	1.099***
				(0.000)	(0.000)	(0.000)
Linkedin status				1.270***	1.256***	1.251***
				(0.000)	(0.000)	(0.000)
IOS Dummy					0.495***	0.511***
					(0.000)	(0.000)
Constant	0.321***	0.256***	8.816***	1.760***	2.031***	2.979***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
State Fixed Effects	N	N	N	N	N	N
Observations	219,219	184,423	184,423	189,295	189,295	184,423
Pseudo R-squared	0.00417	0.00238	0.0903	0.0207	0.0222	0.0225
AUC	0.601	0.586	0.723	0.604	0.607	0.608

Key Takeaways

- AUC of this specification is 61% and 2% greater than the AUC estimate using just the credit bureau score.
  - Even a 1% increase is AUC is quite significant
- Explains about 2% additional variation in loan defaults as compared to just the credit bureau score.
- Individuals without a financial app are about one and a half times more likely to default relative to those that have such an app installed.
  - may be correlated with the financial sophistication of a customer.
- Customers with some other mobile loan application (Mloan dummy) are about 9% less likely to default.
- Those with a dating app (any other social network app) are 24% (33%) more likely to default.
- Customers with a travel app are about 4% more likely to default.
- those with an Android phone are twice as likely to default as those with an Apple phone.

- Caveat: We can't pin down the underlying causal mechanism
  - However, these results indicate that the nature of apps installed on the phone have significant discriminatory power in default prediction.

Credit Score vs Digital Footprint

- Focus on a model that includes all loan characteristics, customer characteristics, and digital footprint but excludes credit bureau score.
- Our objective here is two folds.
  1. Whether our results on digital footprint continue to hold once we control for other loan level and customer level characteristics.
  2. We want to examine if observable loan, customer, and digital foot print characteristics can predict loan defaults -- as compared to a model that includes just the CIBIL score along with customer and loan characteristics.

  - Bottom Line: Can digital footprint substitute for the credit score?

Credit Score vs Digital Footprint: Loan defaults

VARIABLES	Odds Ratio (1)	Odds Ratio (2)	Odds Ratio (3)	Odds Ratio (4)	Odds Ratio (5)	Odds Ratio (6)	Odds Ratio (7)	Odds Ratio (8)	Odds Ratio (9)
Log of cibil	0.877***	0.900***	0.872***			0.906***		0.882***	0.885***
	(0.000)	(0.000)	(0.000)			(0.000)		(0.000)	(0.000)
Log of Salary			0.240***				0.265***	0.265***	0.260***
			(0.000)				(0.000)	(0.000)	(0.000)
Log Loan Amount			4.136***				4.193***	4.272***	4.302***
			(0.000)				(0.000)	(0.000)	(0.000)
Log Age			0.545***				0.333***	0.365***	0.363***
			(0.000)				(0.000)	(0.000)	(0.000)
High School Dummy			0.822***				0.855***	0.852***	0.848***
			(0.000)				(0.000)	(0.000)	(0.000)
College Dummy			0.724***				0.745***	0.742***	0.740***
			(0.000)				(0.000)	(0.000)	(0.000)
Supervisor Dummy			0.994				0.979	1.003	1.020
			(0.749)				(0.281)	(0.874)	(0.321)
Manager Dummy			0.932***				0.951***	0.964**	0.992
			(0.000)				(0.007)	(0.048)	(0.682)
Travel.purpose cashe			0.759***				0.808***	0.807***	0.802***
			(0.000)				(0.000)	(0.000)	(0.000)
EMI.purpose cashe			0.838***				0.842***	0.863***	0.869***
			(0.000)				(0.000)	(0.000)	(0.000)
purchasc.purposc cashc			0.816***				0.845***	0.843***	0.846***
			(0.000)				(0.000)	(0.000)	(0.000)
Loanrepayment.purpose cashe			0.806***				0.817***	0.836***	0.840***
			(0.000)				(0.000)	(0.000)	(0.000)
Other purpose.purpose cashe			0.861***				0.869***	0.864***	0.851***
			(0.000)				(0.000)	(0.000)	(0.000)
Log no of SMS				0.969***	0.968***	0.972***	0.953***	0.957***	0.958***
				(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Log No of Contacts				0.964***	0.966***	0.976***	0.965***	0.971***	0.968***
				(0.000)	(0.000)	(0.007)	(0.000)	(0.002)	(0.001)
Log no of Apps				0.659***	0.653***	0.656***	0.632***	0.635***	0.636***
				(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Log Callog				0.917***	0.913***	0.915***	0.921***	0.922***	0.925***
				(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Dating App				1.246***	1.252***	1.218***	1.209***	1.187***	1.161***
				(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Finsavy App				0.709***	0.706***	0.745***	0.753***	0.802***	0.816***
				(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Socialconnect App				1.331***	1.233***	1.288***	1.358***	1.457***	1.661***
				(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Travel App				1.034*	1.041**	1.038**	0.915***	0.911***	0.907***
				(0.052)	(0.019)	(0.031)	(0.000)	(0.000)	(0.000)
Mloan App				1.002	1.002	1.003	0.976	0.978	0.977
				(0.908)	(0.892)	(0.822)	(0.115)	(0.162)	(0.153)
Facebook status				1.091***	1.095***	1.099***	1.094***	1.096***	1.086***
				(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Linkedin status				1.270***	1.256***	1.251***	1.196***	1.169***	1.172***
				(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)
IOS Dummy					0.495***	0.511***	0.427***	0.444***	0.458***
					(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Constant	0.321***	0.256***	8.816***	1.760***	2.031***	2.979***	109.509***	112.540***	142.081***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
State Fixed Effects	N	N	N	N	N	N	N	N	Y
Observations	219,219	184,423	184,423	189,295	189,295	184,423	189,295	184,423	180,701
Pseudo R-squared	0.00417	0.00238	0.0903	0.0207	0.0222	0.0225	0.111	0.113	0.115
AUC	0.601	0.586	0.723	0.604	0.607	0.608	0.742	0.744	0.746

Key Takeaways

- Mobile Digital footprint variables retain their discriminatory ability
  - captures unobservable aspects of borrower behaviour not captured by loan and customer characteristics.
  - For example, the coefficient estimate of IOS dummy remains statistically significant even after controlling for the customer's monthly salary.
    - implies that owing an Apple device captures an unobservable aspect of individuals which is not fully absorbed by earnings.
- The AUC of this specification is 74%
  - 15 percentage points higher than the AUC of the model using only the credit bureau score
  - 2 percentage points higher than the model which includes CIBIL score combined with customer and loan characteristics.
- Overall, digital footprint variables can be used to predict the likelihood of default and can perform at least as well as the credit score.

Customers Without a credit score

- So far we have focused on customers for which we have both credit score and digital variables
- But can digital variables predict default even for customers without a credit score?
  - Customers with a credit score may be very different from customers without a score
- So, focus on the sample of customers without a credit Score

Customers Without a credit score

VARIABLES	Odds Ratio (1)	Odds Ratio (2)	Odds Ratio (3)
Log of Salary	0.067***		0.068***
	(0.000)		(0.000)
Log Loan Amount	10.639***		12.061***
	(0.000)		(0.000)
Log Age	0.504***		0.410***
	(0.000)		(0.000)
High School Dummy	0.799***		0.846***
	(0.000)		(0.004)
College Dummy	0.655***		0.690***
	(0.000)		(0.000)
Supervisor	0.816***		0.870***
	(0.000)		(0.004)
Manager	0.883***		0.977
	(0.003)		(0.610)
Travel.purpose cashe	0.883		0.713***
	(0.113)		(0.007)
EMI.purpose cashe	0.812***		0.897
	(0.006)		(0.160)
purchase.purpose cashe	0.875**		0.907
	(0.032)		(0.133)
Loanrepayment.purpose cashe	0.608***		0.641***
	(0.000)		(0.000)
Other purpose.purpose cashe	0.963		1.015
	(0.402)		(0.756)
Log no of SMS		0.973***	0.953***
		(0.001)	(0.000)
Log No of Contacts		0.979	0.968
		(0.298)	(0.156)
Log no of Apps		0.875***	0.766***
		(0.000)	(0.000)
Log Callog		0.934***	0.942***
		(0.000)	(0.000)
Finsavy App		0.260***	0.265***
		(0.000)	(0.000)
Socialconnect App		8.625***	12.019***
		(0.000)	(0.000)
Travel App		0.927	1.211*
		(0.177)	(0.076)
Mloan App		0.957	0.852*
		(0.582)	(0.092)
Facebook status		0.919**	1.019
		(0.039)	(0.679)
Linkedin status		1.103	1.165
		(0.413)	(0.257)
IOS Dummy		0.846	0.616***
		(0.203)	(0.001)
Constant	181.810***	0.336***	482.318***
	(0.000)	(0.000)	(0.000)
Observations	47,152	45,473	45,425
Pseudo R2	0.0205	0.0237	0.0242
AUC	0.785	0.578	0.809

Key Takeaways

- The AUC of digital footprint model is 58% and comparable to AUC of credit score in the main sample

- Incremental predictive power over loan and customer characteristics
- Bottomline: Can use digital mobile footprints to score customers without a credit history

Social Footprint: Call Logs Data

- Can we infer aspects of borrower behavior from call logs
  - Likely to capture social capital of an individual – Singh and Ghosh (2017)
  - Social capital is an important determinant of default – Microfinance literature

  - We try to proxy for two aspects of social connections:
    - Breadth – Total Calls, Total Duration of Calls, number of persons called, diversity in calls as measured by the Herfindahl-Hirschman (concentration) Index of calls made to people in the contact list. etc
    - Depth – Average daily duration per person, average calls per person etc

Key Takeaways

- Both breadth and depth are important predictors of default
  - The AUC of the model with call logs only is 64.4% and better than a model with other digital footprint variables
    - An AUC of above 60% is considered quite good
    - Remarkably easy to collect information has better predictive power for defaults
  - The AUC of the model with call logs and other digital footprint variables is 66%
    - 8% higher than AUC of the model with credit score only
    - Better than the 5.7 percentage points AUC improvement reported in Iyer, Khwaja, Luttmer, and Shue (2016) who compare the AUC using the Experian credit score to the AUC in a setting where, in addition to the credit score, lenders have access to a large set of borrower financial information as well
    - Comparable to the improvement in the AUC by +8.8 percentage points in a consumer loan sample of a large German bank (Berg, Puri, and Rocholl, 2017) in a setting where, in addition to the credit score, lenders have access to account data, as well as socio-demographic data and income information.
  - Bottom line: Simple metrics using call logs does remarkably well in predicting defaults as compared to the traditional credit bureau scores

Counterfactual: Prediction Policy Problem

- Causal Problem vs. Prediction Problem
- Athey (2017), Mullainathan et al (2017, 2018)
- Example:
- Rain dance vs. Umbrella
- Invest in a rain dance to increase the chance of rain : Causal problem
- Is the chance of rain high enough to merit an umbrella? Prediction problem
- Our case: The borrower's credit worthiness (default risk) did not change by fintech borrowing
- Prediction policy question: Could we predict the default risk of the borrower better, so people have more and efficient access to credit?

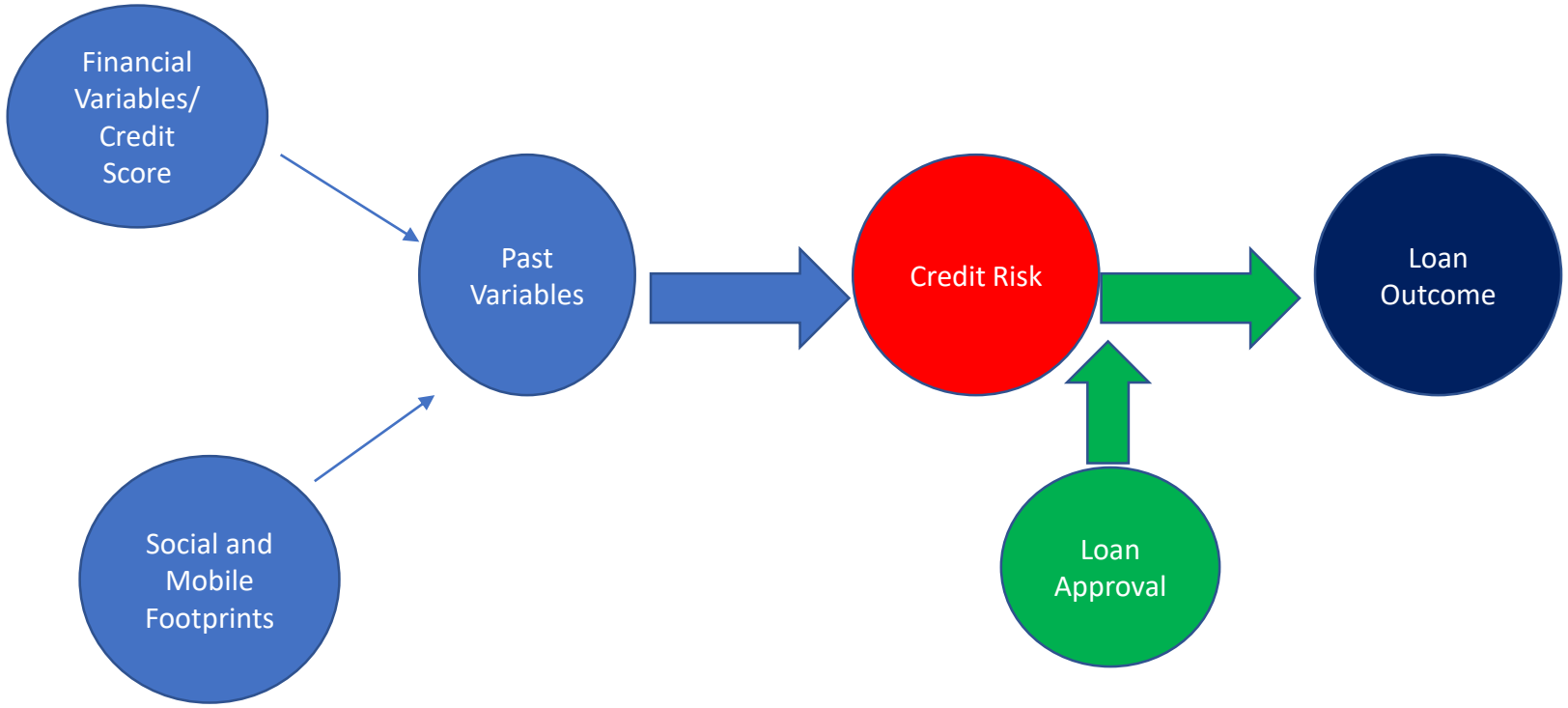
Prediction and Decision-Making: Predicting a State Variable

Kleinberg, Ludwig, Mullainathan, and Obermeyer (2018)

- Motivating examples:
  - Will it rain? (Should I take an umbrella?)
  - Which teacher is best? (Hiring, promotion)
  - Unemployment spell length? (Savings)
  - Risk of violation of regulation (Health inspections)
  - Riskiest youth (Targeting interventions)
  - Creditworthiness (Granting loans)
- Empirical applications:
  - Will defendant show up for court? (Should we grant bail?) Kleinberg et al (QJE 2018)
  - Will patient die within the year? (Should we replace joints?) Kleinberg et al (AER 2017)
  - Prediction of creditworthiness using alternative credit score (This paper)

Prediction and Decision-Making: Predicting a State Variable

- Prediction of default risk



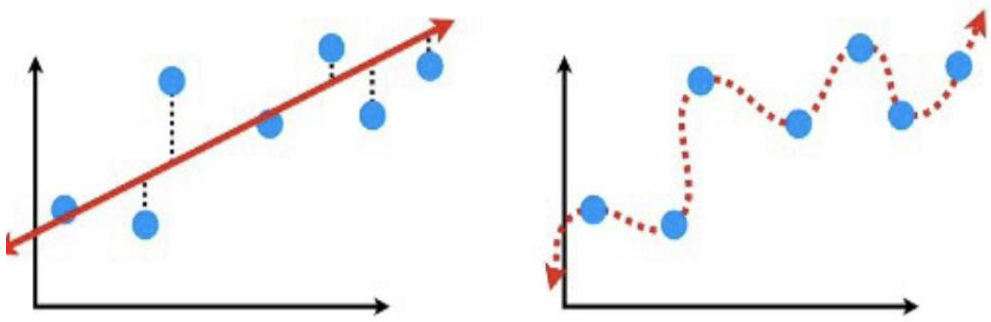
Prediction Policy Problem: Estimation Challenge

- OLS (in-sample) estimation is not good for prediction problems
- We need good prediction for out of sample
  - Machine Learning (ML) methods
- OLS =>poor predictions
  - consider a two variable example where OLS estimation produced $\beta_1 = 1 \pm 0.001$ and $\beta_2 = 4 \pm 10$,
  - a predictor of $x + 4x$.
  - But given the noise in β_2 , for prediction purposes one would be tempted to place a smaller (possibly 0) coefficient on x_2 . Introducing this bias could improve prediction by removing noise.
  - Bias-variance trade-off
  - OLS first minimizes bias and not treat it as a joint problem

Prediction Policy Problem

- Mean Square Error in out-of-sample
- MSE(x)= $E_D[(\hat{f}(x) - y)^2]$

$$\underbrace{E_D[(\hat{f}(x) - E_D[\hat{y}_0])^2]}_{\text{Variance}} + \underbrace{(E_D[\hat{y}_0] - y)^2}_{\text{Bias}^2}$$



When the model tries to reduce bias it tends to overfit the data. Left fig have high bias & right fig have low bias.

- Because the f varies from sample to sample, it produces variance (the first term). This must be traded off against bias (the second term).
- By ensuring zero bias, OLS allows no trade-off.
- ML: $\hat{f}_{ML} = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda R(f)$

- Here R(f) is a regularizer that penalizes functions that create variance.

Prediction Policy Problem: Machine Learning

Because the f varies from sample to sample, it produces variance (the first term). This must be traded off against bias (the second term). By ensuring zero bias, OLS allows no trade-off.

- ML: $\hat{f}_{ML} = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda R(f)$

- Here R(f) is a regularizer that penalizes functions that create variance.
- A key insight of machine learning is that this price λ can be chosen using the data itself. Imagine we split the data into f subsets (often called “folds”). For a set of λ , we estimate the algorithm on f − 1 of the folds and then see which value of λ produces the best predictionin the f th fold. This cross-validation procedure effectively simulates the bias-variance trade-off by creating a way to see which λ does best “out of sample.”

Prediction Policy Questions

1. Which set of variables best predict the default risk?
   - (credit score vs. social and mobile footprints)

2. How does the default risk predictability of social and mobile footprints vary across different levels of traditional credit scores?

3. What would have happened if we offered loans using social and mobile footprints to borrowers for whom traditional credit scoring is not available and hence generally denied credit by traditional banks?
   - Prediction (not causal) Counterfactuals

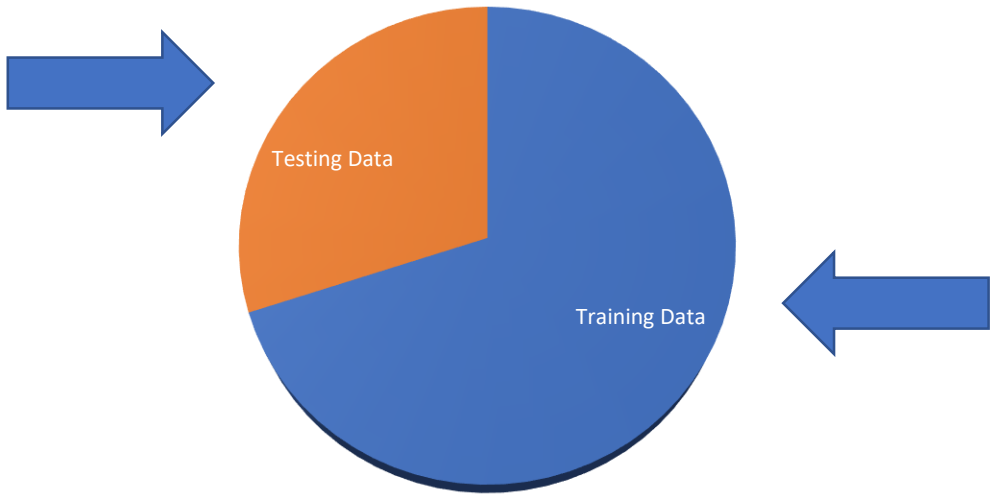
   a) How many more would have been approved?

   b) What would have been the new overall default rate if we replaced the high risk approved with a medium risk denied?

Prediction Policy Problem: Machine Learning

- ML:

Use this part for prediction counterfactual



Use this part for estimation

- We use 70:30 split

- ML methods: Random Forest, XGBoost, logistic

Prediction Policy Problem

- Which set of variables best predict the default risk?

Feature Groups	Logistic Regression		Random Forest		XGBoost	
	Recall	AUC	Recall	AUC	Recall	AUC
Only CIBIL	0.142	0.581	0.536	0.586	0.551	0.59
Only Mobile Footprint	0.142	0.581	0.542	0.587	0.551	0.59
Only Social Variables(CallLogs)	0.528	0.645	0.541	0.69	0.579	0.703
Non Banking Features	0.547	0.656	0.542	0.69	0.579	0.708

- mobile and social footprint variables alone has a much higher AUC score in predicting the probability of default relative to the borrower's credit score (CIBIL) across all three methods of machine learning algorithms.

Prediction Policy Problem: Borrower Heterogeneity

- Low CIBIL (bottom 10%)

Feature Groups	Logistic Regression		Random Forest		XGBoost	
	Recall	AUC	Recall	AUC	Recall	AUC
Only CIBIL	1	0.541	0.93	0.558	0.971	0.565
Only Mobile Footprint	0.965	0.589	0.724	0.658	0.919	0.61
Only Social Variables(CallLogs)	0.872	0.665	0.749	0.716	0.87	0.72
Non Banking Features	0.854	0.673	0.741	0.718	0.864	0.727

- High CIBIL (>750)

Feature Groups	Logistic Regression		Random Forest		XGBoost	
	Recall	AUC	Recall	AUC	Recall	AUC
Only CIBIL	0.105	0.583	0.501	0.585	0.515	0.59
Only Mobile Footprint	0.541	0.6	0.556	0.596	0.585	0.603
Only Social Variables(CallLogs)	0.541	0.652	0.545	0.688	0.592	0.703
Non Banking Features	0.565	0.665	0.559	0.693	0.589	0.707

- Social variables have higher predictive power relative to the credit score (CIBIL)
- Also true for the borrowers who belong to the higher end of the spectrum of the CIBIL score (> 750)

Prediction Policy Problem: Predicting default risk for Borrowers without CIBIL Score

Feature Groups	Training Sample (Borrower with CIBIL Score)		Test Sample (Borrower without CIBIL Score)	
	Recall	AUC	Recall	AUC
Logistic Regression	0.701	0.76	0.72	0.8
Random Forest	0.625	0.762	0.664	0.764
XGBoost	0.746	0.802	0.82	0.83

- Use the set of approved borrowers as a training sample and treat the borrowers without the CIBIL score as testing sample.
- Use our training sample data to train our model and select the optimal features using Logistic, random forest and XGBoost.
  - Then use the predicted features to predict the default probability of the testing sample; the set of borrowers who were approved without the CIBIL score.

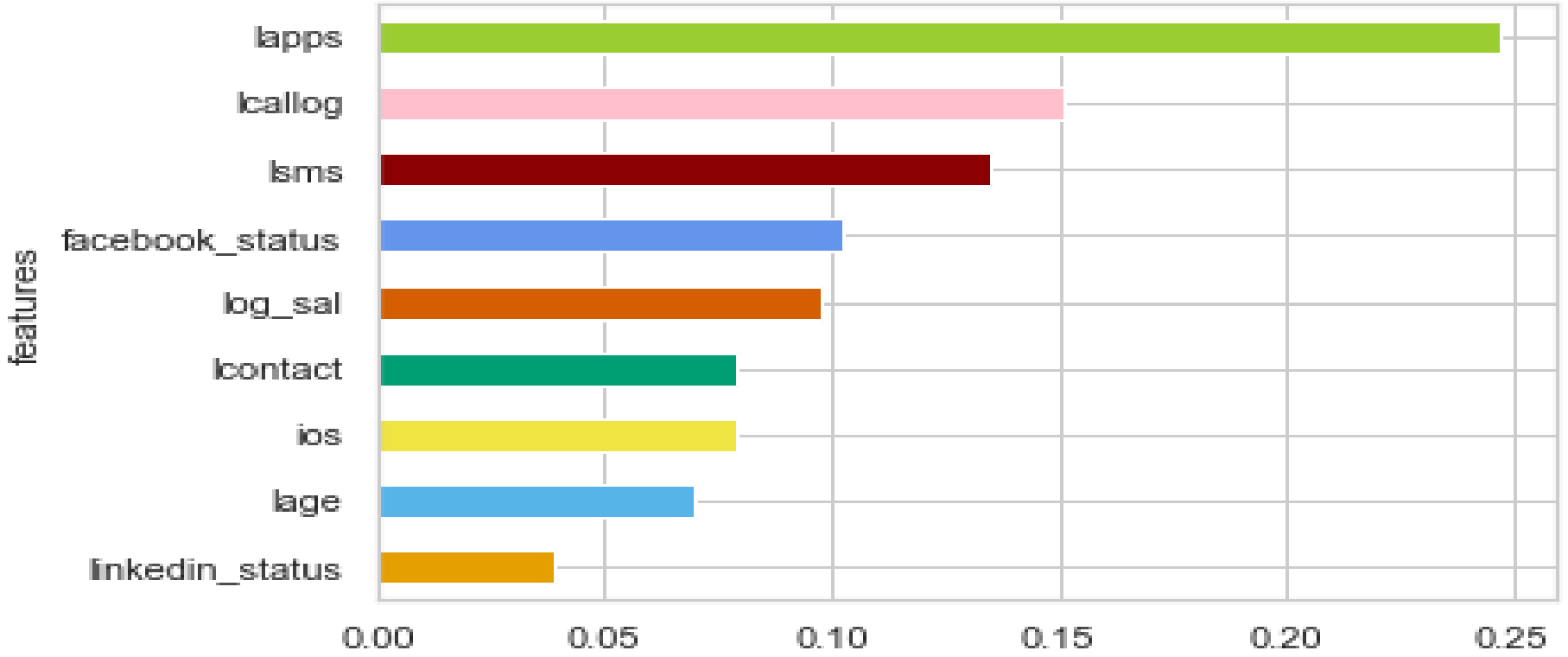
Prediction Policy Problem: Predicting for Borrowers without CIBIL Score

- Our algorithm (following Mullainathan (2018)) proceeds in the following steps:

1. Use the sample of all borrowers who were approved
   - Use ML algorithms (Logistic Regression, Random Forest, XGBoost) to estimate the model parameters based on non-CIBIL variables only
2. Use the predicted model from step 1 and apply it to the borrowers without credit score who were not approved for a loan to predict their probability of default.
   - Use different thresholds for the predicted defaults to find out how many borrowers who were not approved would have been approved based on the social footprint features

Prediction Policy Problem: Predicting for Borrowers without CIBIL Score

- Step 1: Important variables (features) for predicting default risk without CIBIL



- Digital footprints (# of apps, calls, sms) dominates financial variables like salary

Prediction Policy Problem: Predicting for Borrowers without CIBIL Score

Predicted Default Threshold	What proportion would have been approved
0.95	0.87
0.9	0.81
0.8	0.72
0.7	0.65
0.6	0.57
0.5	0.51
0.4	0.44
0.3	0.35
0.2	0.24
0.1	0.13
0.5	0.07

- Step 2: If we choose a relatively conservative threshold of 10% predicted default probability, about 13% borrowers without the CIBIL score and denied loans would have been approved. In our sample about 3500 borrowers did not have CIBIL score and denied a loan. Therefore about 450 of those borrowers would have been approved if we had used the social and mobile footprints.

Conclusion

- We document statistically and economically significant role of individuals' digital footprint variables in the loan approval process.
  - In absence of sufficient credit history and credit scores for millennial customers to judge their credit worthiness, the fintech lender uses individuals digital footprint as an alternative credit screening process.
- Individual's digital mobile footprint have significant predictive power in predicting default.
  - Importantly, these variables have incremental predictive power over and above the CIBIL credit score.
- The discriminatory power of digital footprint variables varies conditional on the loan purpose
- About 13% borrowers without the CIBIL score and denied loans would have been approved.
- Overall, our findings has implications for expanding access to credit to those who don't have a credit history but who leave a large trace of unstructured information on their mobile phones that can be used to predict loan outcomes.

Thank you!

Defaults, Digital Footprint, and Loan Purpose

- Interact loan purpose with digital footprint variables
  - examine whether digital footprint variables have greater discriminatory power in predicting defaults depending on the purpose of the loan.
  - Facebook login may capture the propensity of a consumer to engage in conspicuous consumption (Immorlica et al. (2017))
  - Default rates - higher for loans taken for the purpose of purchase by such customers?
- Base loan category in these tests is Medical loans
  - default rates are measured relative to the default rates for medical loans.

Key Takeaways

- as compared to customers who do not have financial apps installed on their phones, those who do are 34%, and 56% more likely to default when they take EMI loans and Repayment loans, respectively. Along similar lines,
- Customers who have installed another loan application app, are also more likely to default when they undertake a loan for EMI or loan repayment
- Are these customers are on average of low creditworthiness?
  - NO difference in any observable characteristic including credit score

Marginal Contribution

- Our work complements and builds on Berg, Burg, Gombovic and Puri (RFS forthcoming)
  - data covering approximately 250,000 purchases from an E-Commerce company located in Germany
  - digital footprint complements rather than substitutes for credit bureau information
  - informative even for customers who do not have credit bureau scores
- Our paper is similar in spirit to their work
  - While related, there are important differences
  - our paper further builds on and complements their findings

Marginal Contribution

1. Our data is from a stereotypical fintech lender operating in a developing country and covers all kinds of loans and not just those for e-commerce purchases.
   - allows us to extrapolate the importance of digital footprints in measuring creditworthiness for loans taken for different purposes and not just an e-commerce purchase.
2. Our data capture very different aspects of the digital footprint from the mobile phones of customers.
   - the large majority of customers in their sample access the digital world through desktop
   - important given that globally, about 50% of the users access the Internet through mobile phones, and 5% through tablets.
   - particularly true in a developing country setting. For instance, 80% of the Internet access time in India is through mobiles.
   - even in developed countries like the UK, USA, and Germany, the fraction of users that access the Internet primarily through mobile phones is increasing.
   - findings are potentially generalizable to other developing countries and the millennial generation.

Marginal Contribution

3. Because we have data on the salary, education, and job of the customers we can disentangle whether digital footprint simply proxies for these characteristics or provides incremental information.
   - For, instance we find that owning an IOS device has predictive power even after controlling for earnings.
   - Collecting additional data on overall savings and investments portfolios of customers

4. We find that the default prediction can be improved significantly by using proxies that capture deeper aspects (“deep digital footprint”) of an individual’s digital presence.

Marginal Contribution

- Finally we, document that digital footprints can allow lenders to estimate the likelihood of default based on the end use of the loans.
- So, two customers with otherwise same credit scores and earnings may have a differing propensity to default for different kinds of loans.
- Implication: The same customer can have different creditworthiness (and consequently credit score) conditional on the purpose of the loan and digital footprints

Variables

Financial Transactions	Debit to Credit ratio	CIBIL	Log of cibil
	No. of Transactions	Customer Characteristics	Log of Salary
	Log Expenditure to Income ratio		Log Age
	Avg 2 Month Appreciation of Account Balance		High School Dummy
			College Dummy
			Supervisor Dummy
			Manager Dummy

Variables

Mobile and Social Footprint Variables	Log no of SMS	Log of Total No. of SMS.
	Log no of Contacts	Log of No. of people in contact list.
	Log no of Apps	Log of no. of applications in phone.
	Log Callog	Log of Total No. of calls.
	Dating App	Dummy takes 1 if customer has a dating app.
	Finsavy App	Dummy takes 1 if customer has a financial services app (stocks, banking, payment and wallet).
	Socialconnect App	Dummy takes 1 if customer has a social connect app (messaging app, video streaming app, music streaming app, social network app, dating app, video call app).
	Travel App	Dummy takes 1 if customer has a Travel app.
	Mloan App	Dummy takes 1 if customer has another loan app.
	Facebook Status	Dummy takes 1 if customer logged into Cashe app using Facebook.
	Linkedin Status	Dummy takes 1 if customer logged into Cashe app using Linkedin.
	IOS Dummy	Dummy takes 1 if customer has an Apple phone.

Variables

Per day Per person Avg No. of Outgoing calls
Per day Per person Avg No. of Missed calls
Per day Per person Avg Duration of Incoming calls
Per day Per person Avg Duration of Outgoing calls
Per day No. of persons called
Log of Per day Total Duration of Incoming calls
Per day Total No. of Incoming calls
Per day Total No. of Outgoing calls
Per day Total Duration of Outgoing calls
Per day Total No. of Missed calls
HHI of No. of Incoming calls

Variables

HHI of Total Duration of Incoming calls	Herfindahl-Hirschman index of duration of incoming calls. To compute this measure, we first calculate the duration of calls received from a person for every day (for a customer). We then take average across all days to get duration of calls per day. We then assign share of durations to every person and compute HHI for the customer.
HHI of Total Duration of Outgoing calls	Herfindahl-Hirschman index of duration of outgoing calls. To compute this measure, we first calculate the duration of calls made to a person for every day (for a customer). We then take average across all days to get duration of calls per day. We then assign share of durations to every person and compute HHI for the customer.
HHI of No. of Missed calls	Herfindahl-Hirschman index of missed calls. To compute this measure, we first calculate the no. of missed calls received from a person for every day (for a customer). We then take average across all days to get the no. of missed calls received from the

Univariate Analysis: Loan approval

	Approved (1)	Not Approved (2)	Difference (3)
Loan Amount	22174.26	17182.04	-4992.22***
Log Interest Rate	1.445	0.892	-0.552***
Loanpurpose Medical	0.214	0.095	-0.118***
Loanpurpose Travel	0.082	0.024	-0.057***
Loanpurpose EMI	0.087	0.081	-0.005***
Loanpurpose purchase	0.133	0.068	-0.065***
Loanpurpose Loanrepayment	0.081	0.047	-0.034***
Loanpurpose Other	0.405	0.232	-0.172***
Age	31.89	29.45	-2.44***
Salary	37524.53	30346.39	-7178.13***
CIBIL (>0, N=219k & 16k)	634.40	470.82	-163.58***

Univariate Analysis: Loan approval

	Approved (1)	Not Approved (2)	Difference (3)
Facebook Status	0.267	0.296	0.029***
Linkedin Status	0.021	0.015	-0.006***
Googleplus_status	1.712	1.690	-0.021***
Referral	0.116	0.039	-0.077***
Sales App	0.195	0.198	0.003
Dating App	0.029	0.028	-0.001
Finsavy app	0.679	0.034	-0.645***
Socialconnect app	0.714	0.036	-0.677***
Travel app	0.567	0.048	-0.518***
Mloan app	0.423	0.020	-0.403***
Referrer	0.234	0.034	-0.200***
# of SMS	2481.71	1109.00	-1372.71***
# of Apps	54.53	41.26	-13.27***
# of Contacts	844.84	683.81	-161.02***
# of Connections	525.89	452.39	-73.50
# of Calls	3136.50	2071.97	-1064.53***
IOS	0.119	0.066	-0.053***

Univariate Analysis: Loan Defaults

	Default (4)	Not Default (5)	Difference (6)
Loan Amount	35228.33	20509.83	-14718.49***
Log Interest Rate	1.857	1.393	-0.463***
Loanpurpose Medical	0.247	0.209	-0.037***
Loanpurpose Travel	0.075	0.082	0.007***
Loanpurpose EMI	0.073	0.088	0.014***
Loanpurpose purchase	0.129	0.133	0.004***
Loanpurpose Loanrepayment	0.082	0.081	-0.0006
Loanpurpose Other	0.395	0.405	0.010***
Age	32.00	31.88	-0.117***
Salary	39262.32	37342.43	-1919.89***
CIBIL (>0, N=219k & 16k)	602.04	639.10	37.06***

Univariate Analysis: Loan Defaults

	Default (4)	Not Default (5)	Difference (6)
Facebook Status	0.274	0.263	-0.011***
Linkedin Status	0.023	0.021	-0.002**
Googleplus_status	1.700	1.714	0.013***
Referral	0.115	0.118	0.002*
Sales App	0.188	0.196	0.007***
Dating App	0.029	0.029	0.0006
Finsavy app	0.677	0.862	-0.019***
Socialconnect app	0.760	0.708	-0.051
Travel app	0.576	0.566	-0.010***
Mloan app	0.423	0.423	-0.0002
Referrer	0.167	0.243	0.075
# of SMS	1949.25	2548.19	598.94***
# of Apps	47.07	55.47	8.40***
# of Contacts	827.64	847.03	19.38***
# of Connections	413.23	539.15	125.92***
# of Calls	2394.96	3229.05	834.08***
IOS	0.112	0.120	0.007***