Banking the un-banked

Team Members – Reet Kanjilal, Mridula Arora, Souravi Banerjee, Aditya P Dutta

IIMB
तेजस्वि नावधीतमस्तु

Background – PM JDY Scheme

Before PM JDY

Financial Exclusion

Leakages in subsidies, benefits not reaching the targeted population due to beneficiaries not having own authentic savings bank accounts.

- Launched in India on Aug 28th, 2014
- World's largest financial inclusion program
- Aims to provide access to banking services for all unbanked households in India. As part of this program, every unbanked adult is supplied a no-frills bank accounts at no cost.
- Sudden shock in the supply of banking of unprecedented scope. On the inauguration day, 1.5 Crore (15 million) bank accounts were opened under this scheme.
- There is lot of interest and need among policy makers to understand the impact of this scheme on the citizens.

Impact – PM JDY Scheme

Huge Volumes

PRADHAN MANTRI JAN-DHAN YOJANA

MERA KHATA BHAGYA VIDHATA

28.23 Crore Accounts

22.14 Crore RuPAY Cards

Rs 63971 Crore Deposits

27.41 lakh* **accounts availed ODs**

As on 05.04.2017

* As on 31.03.2017

Department of Financial Services, Ministry of Finance, Government of India

Project Objectives

- Evaluate the difference in the transaction activity between Pre JDY account and Post JDY accounts. Find whether PMJDY has increased the probability of a customer transacting on his/her account

- Expand PMJDY Scheme to non-inclusive population in a targeted manner. Identify a non-inclusive population who have high probability of using bank accounts to transact and hence can be prioritized target for inclusion in PMJDY scheme

MetaData

- Anonymized monthly transaction details of 1500 customers - pre and post JDY scheme
- 3 states Data – Odisha, Uttarakhand and Tamil Nadu
- 31 Months transactional data for Pre JDY phase accounts
- 10 months transactional data for Post JDY accounts - August 2014 and May 2015
- 33 Customer Profile attributes and 15 Transaction attributes available for each monthly transaction
- We identified that the most important attributes were Monthly Balance, Deposit, Withdrawal, In Remit and Out Remit amounts

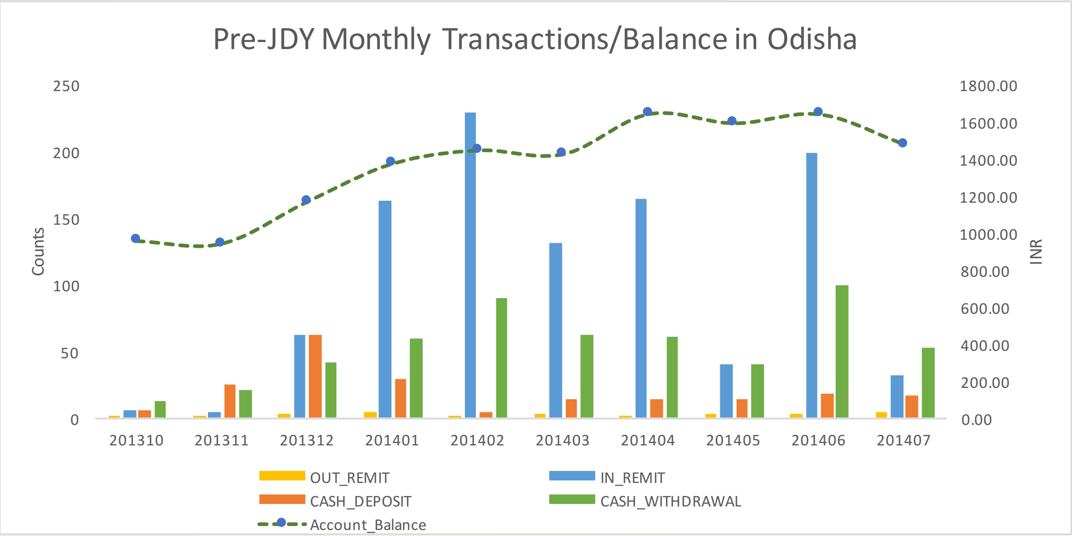
Data Exploration – Variable Summary

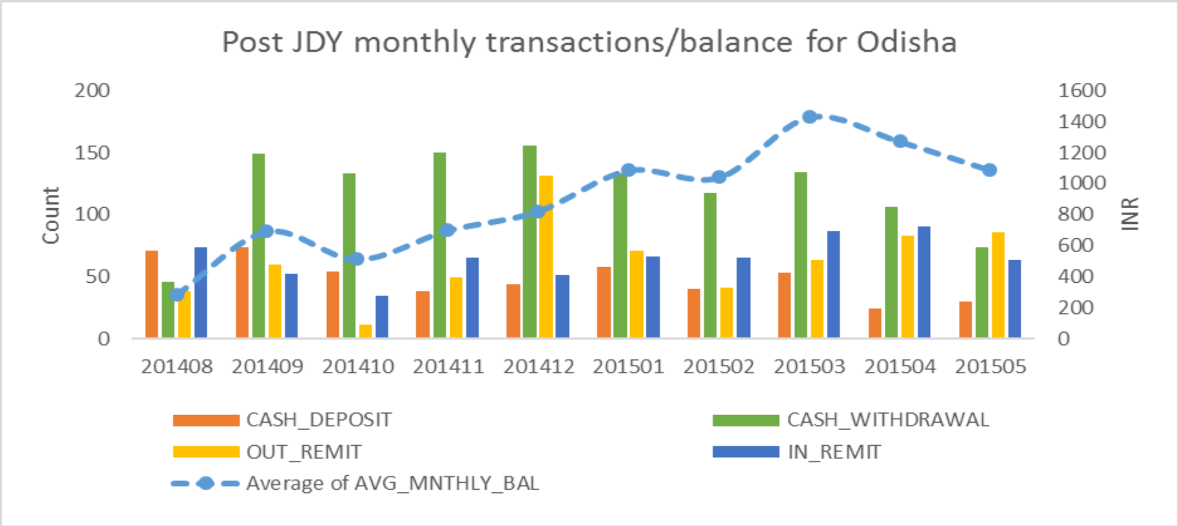
- Keeping our 1st objective in mind, we closely monitored the number of transactions in each phase, for each account and found **In Remit** to be a good indicator of account activity
- Out a total of 3010 transactions in the Pre-JDY phase and 3067 in Post-JDY phase, In Remit stood out contributing majorly to the total

Account Activity	Cash Deposit	Cash Withdrawal	Out Remit	In Remit
Pre-JDY	10%	28%	10%	52%
Post-JDY	25%	23%	18%	34%

Data Exploration –Summary

There has been a noticeable increase in average balance maintained in the post JDY phase, but have transactions really increased?





Data Exploration – Customer Profile

Are the customers transacting in the post JDY phase demographically different?

Transacting	Pre-JDY	Post JDY
Age Group	40-60	23-40
Marital Status	63% are married	62% are not married
Occupation	90% are in unaccounted professions	45% are students and housewives

JDY – A Positive Impact

Percentile Study & Transaction Behavior Analysis

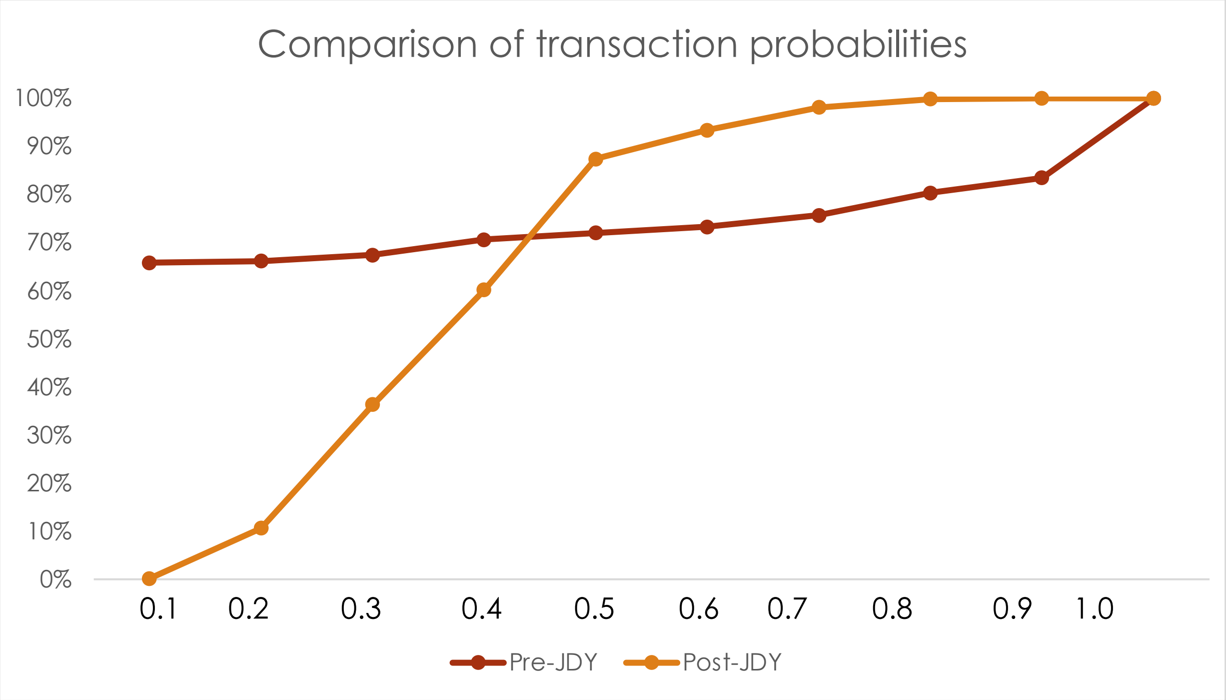
Data Exploration – Insights

To understand the significance of JDY we tried fitting a logistic regression model to understand:

- The probability of transaction for each percentile
- Significance of the variables in the two phases

Percentile Study

- Transaction behavior in the pre and post JDY phase



- There is considerable shift in the percentile measure for pre and post JDY phase, with a noticeable increase in transaction likelihood at each percentile

Transaction Behavior Analysis - Data Preparation

- Data rolled over to customer level
- Pre-JDY & Post-JDY Data analyzed in 4 timeframes
  - 1 month window – picked only 1 month data per customer
  - 3 month window – averaged 3 months transactional data per customer
  - 6 month window – averaged 6 months transactional data per customer
  - 10 month window – averaged 10 months transactional data per customer
- Tested on Sample data for 500 customers each of Uttarakhand, TN and Odisha and multiple observation window of their account activity

Data Analysis

- Logistic Regression was performed on Pre-JDY & Post-JDY combined data for the above time frames with Pre & Post JDY Indicator
  - Significance of the variables in the two phases

One Month Observation window for JDY and non JDY account				
Coefficients:				
	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	-2.7	0.2012	-13.423	2.00E-16
VISA_d	2.6497	0.6417	4.129	3.64E-05
PAN_d	2.201	0.7176	3.067	0.00216
JDY_IND	0.699	0.2542	2.75	0.00595

In spite of working with a small dataset and with limited number of accounts, JDY indicator is significant for both TN and Odisha

Efficient Unbanked to Banked conversion – Identifying right prospect

Modelling Approach

- Response variable is a IN Remit Amount Indicator - 1 if the customer has received money & 0 otherwise
- Algorithms which support distributed processing were preferred
- Following methods were chosen for modelling
  - SVM
  - k-NN
  - GBM
  - XGBoost
- Oversampling methods were used to balance the un-balanced data
- Model Validation was on an In time test sample as no out of time validation data is available
- To decide a suitable framework, we took Odisha data to train a machine learning model

Modelling - SVM

- SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible
- Sampling - Divided the data into Train & test in 7:3 ratio. As data is un-balanced thus we did over sampling on the data with a probability of rare as 0.4
- Tuning - model tuning was done using kernel as radial to arrive at best performing Cost & gamma values

	Observation Window	Kappa	Sensitivity	Specificity	Pos Pred Value	Accuracy
SVM	One Month	-0.01	0.94	0.05	0.88	0.84
	Three Months	0.23	0.85	0.39	0.88	0.78
	Six Months	0.06	0.67	0.44	0.91	0.65
	Ten Months	0.26	0.66	0.60	0.57	0.63

SVM Results

Modelling – k-NN

- k-NN classification output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors
- Sampling & Tuning
  - SMOTE sampling technique has been used to synthetically increase data points
  - Hyper parameter tuning has been done by using K fold cross validation method

	Observation Window	Kappa	Sensitivity	Specificity	Pos pred value	Accuracy
KNN Results	One Month	0.12	0.33	0.85	0.14	0.82
	Three Months	0.04	0.11	0.92	0.25	0.77
	Six Months	0.10	0.32	0.77	0.42	0.62
	Ten Months	0.12	0.58	0.54	0.54	0.56

KNN Results

Modelling – GBM

- GBM as a method is a slow learner, training itself takes a lot time and final outcome is a probability score from which optimal cut off point has been decided for each observation window from the training data and used that cut off to classify test data
- Sampling & Tuning
  - K fold validation techniques have been used while training the models
  - Data set has been divided into training and test data on 7:3 ratio.
  - For few observation window, Over sampling and SMOTE techniques have been used to balance the data

	Observation Window	Test AUC	Kappa	Sensitivity	Specificity	Pos pred value	Accuracy
GBM Results	One Month	0.78	0.15	1	0.62	0.13	0.64
	Three Months	0.7	0.26	0.64	0.71	0.33	0.7
	Six Months	0.69	0.26	0.87	0.45	0.45	0.6
	Ten Months	0.69	0.33	0.62	0.71	0.67	0.67

GBM Results

Modelling – XGBoost

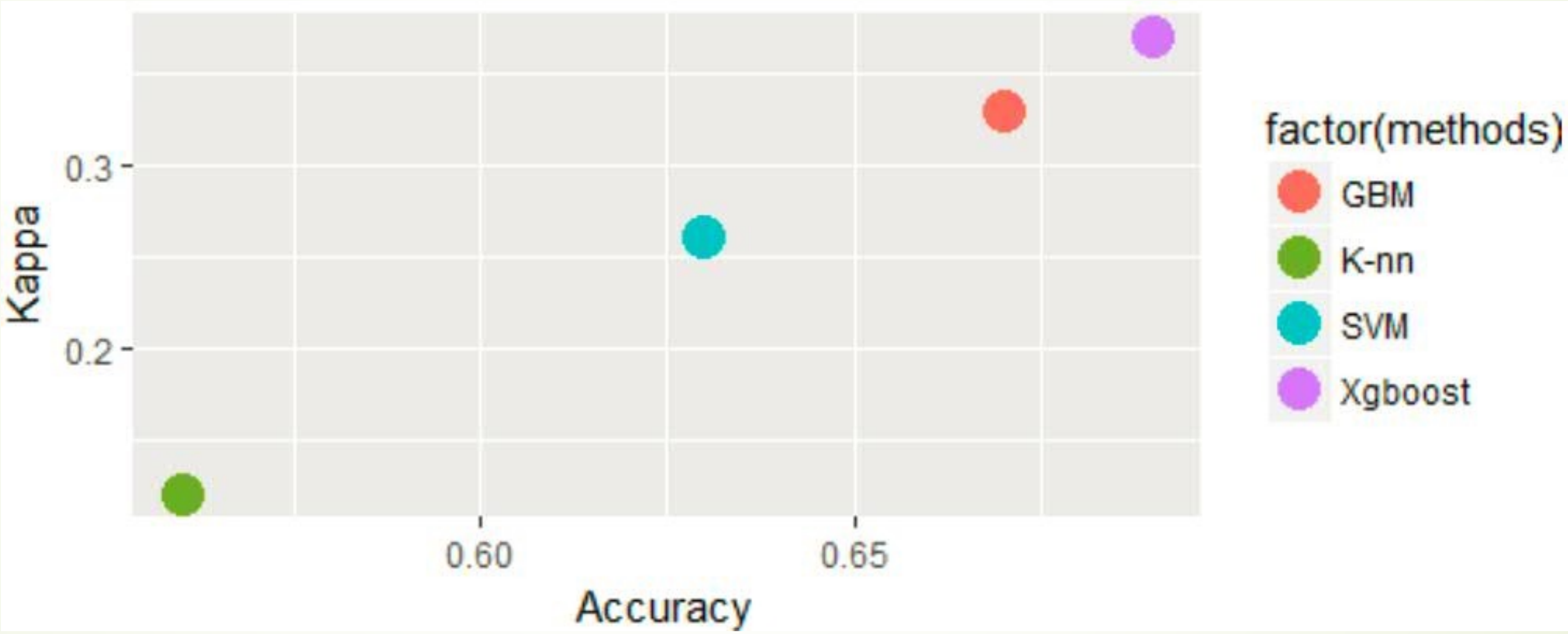
- XGBoost tries to reduce misclassification rate by assigning higher weights to misclassified points in successive iteration
- Output is a probability score which further divided into two class based upon a probability cut off obtained from training dataset.
- Sampling & Tuning
  - K fold validation techniques have been used while training the models
  - SMOTE technique has been used to balance the data wherever needed
  - Dataset has been divided into 7:3 ratio for training and test sample

	Observation Window	Test AUC	Kappa	Sensitivity	Specificity	Pos pred value	Accuracy
XGBoost Results	One Month	0.77	0.12	0.88	0.6	0.12	0.61
	Three Months	0.77	0.27	0.37	0.89	0.43	0.79
	Six Months	0.69	0.26	0.87	0.45	0.45	0.6
	Ten Months	0.7	0.37	0.7	0.68	0.67	0.69

XGBoost Results

Model Analysis Criteria - Kappa

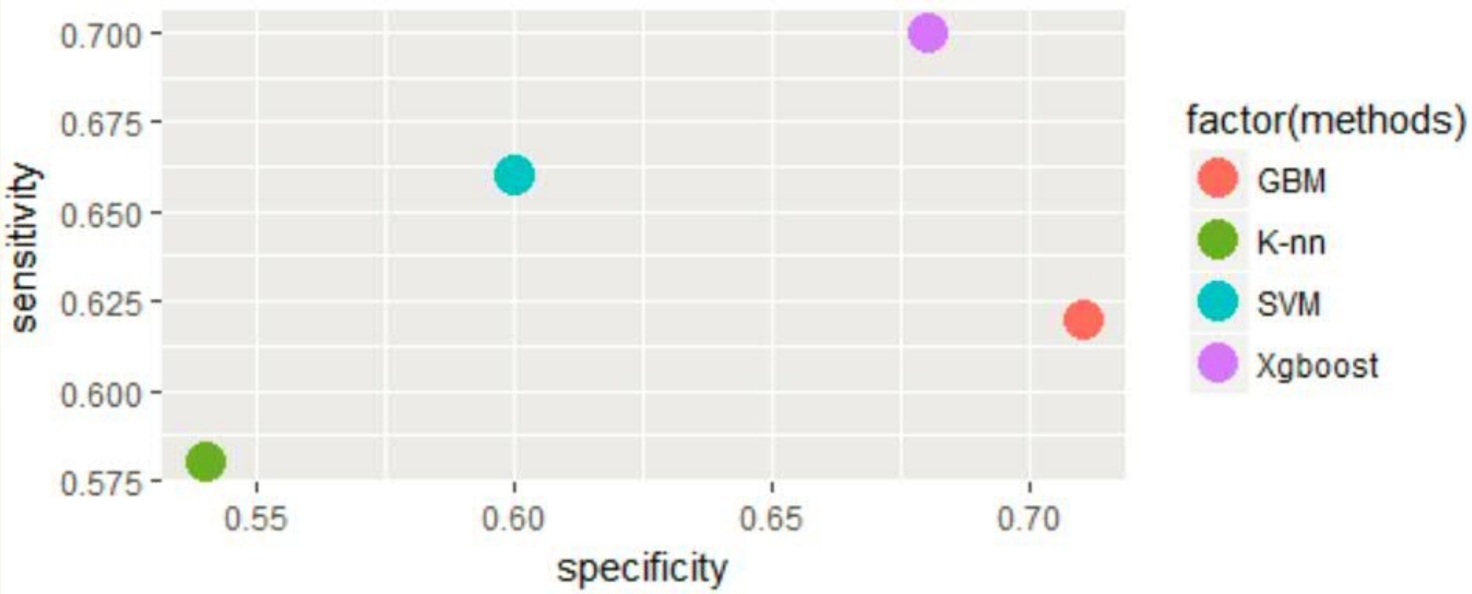
- Kappa value indicates the possibility of occurrence of an event without attributing it to mere chance



GBM and XGBoost methods provide better results on this aspect compare to the other classification methods

Model Analysis Criteria – Specificity & Sensitivity

- Sensitivity gives us the true positive rate & Specificity is the true negative rate



XGBoost performs much better than GBM, SVM & KNN in these two aspects as well.

Model Analysis - Conclusion

- Preferred Modelling Algorithm - XGBoost
- XGBoost gives the best results for Kappa, Accuracy, Specificity & Sensitivity
- XGBoost is much faster and can be parallelized very easily.

Limitations and Recommendations

- Data Challenge – To scale up, more data on customers and greater granularity would help. The categorical variables had limited information.
- Income based analysis for targeting new customers - This could be used as a prototype to understand how an income bracket behaves before and after JDY scheme launch.
- Scaling up the solution - Best result found using XGBoost algorithm which can be parallelized very easily leading to flexibility to train or use the model in large dataset. The proposed framework is highly scalable and flexible in nature i.e. feasible to deploy on a large distributed ecosystem.
- Implementation strategies - The model helps us to predict – using available customer profile – that if the customer opens an account whether the customer is likely to transact or not. This can be used to find the most optimal target population for JDY rollouts and campaigns in specific states or areas.

Check

- Benefits being offered under the Jan Dhan accounts include interest on deposits, accidental insurance cover (Rs. 1 lakh) and life cover (Rs. 30,000) subject to fulfilment of eligibility. They also provide access to pension and insurance products

  Interest rate applicable for Saving Bank Accounts (presently at 4 per cent in most of the banks) will be admissible to accounts opened under PMJDY scheme, as per the **State Bank of India** website.

  Besides, a Jan Dhan account can also be used to receive Direct Benefit Transfer - under which the government transfers subsidies directly to the people.

  An overdraft facility will be permitted after satisfactory operation for six months. An overdraft up to Rs. 5000 is offered by the government in only one account per household ".
- "Today, we have been able to connect 240 million people additionally to the banking system. About 80 per cent of them are today functional with cash balances," Mr Jaitley said at a panel discussion on 'Financial Inclusion not Exclusion: Managing De-Risking' here on Friday.

  "Eighty per cent of these accounts didn't have money, in the first instance. So all government programmes -- from central to regional to local bodies which give assistance to the weaker sections -- now transfer payments in these accounts. The payments come from schemes like rural employment guarantee scheme and subsidies on fuel, food and ferlilisers,"
- The biometric identity (Aadhar) helps the government to identify people who are entitled to state support and exclude those who are not, he added.

  "Collectively we have been able to bring a lot of people under financial inclusion, but it is still a work in progress," Mr Jaitley said.