# A GENERALIZED AND AUTOMATED DATA MINING APPROACH FOR CUSTOMER ACQUISITION

MANEESH KUMAR MEENA

BANK OF BARODA

# PROBLEM STATEMENT

- Objective of this project is to build a digital KYC mechanism which is easy, secure, maintains history and can be shared consistently across entities within and outside the Bank.

# SUGGESTED IDEA

- We suggested a KYC solution based on data mining and AI(artificial intelligence).
- This idea works in 3 main steps.
  - Standardization of  data (from bank’s data to ML data)
  - Creating policies and predications (information to focus on)
  - Training and testing of multi-class model and PNL reports

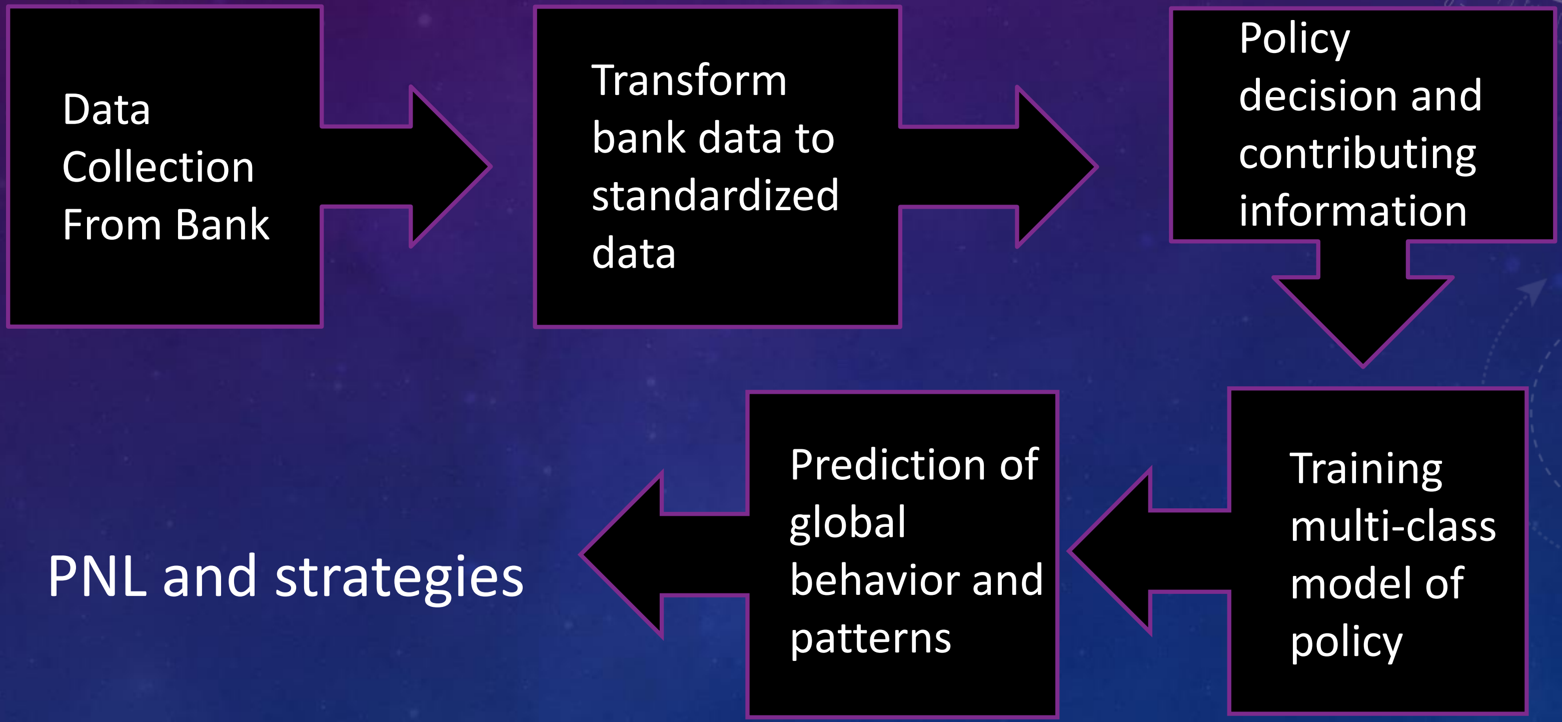
  Further we will create global policy and patterns (global: applicable to all customers in a specified group) and we will also apply these global policies to individuals.

# MOTIVATION

- It is cost efficient and efficient. While a lot of data mining is done manually, it can do it automatically with high accuracy.
- Use of multi-class classification reduces error thus low errors.
- Policy for group of customer which helps bank to regulates new services and campaign of its new strategies and from individual KYC bank can improve its relationship and trust with customers.
- We can have large scale policies(regulates in large region, large group) or can have small policies(targets specific group of customers) and we can also have policy for individual too.
- Use of PNL will help to keep track of these policies and their effectiveness.

# METHODOLOGY

- Basic idea



Data Collection From Bank

Transform bank data to standardized data

Policy decision and contributing information

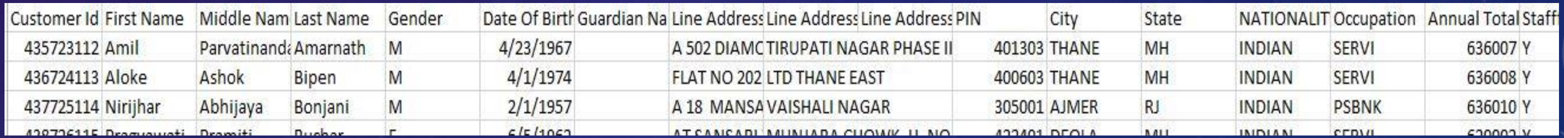
Training multi-class model of policy

Prediction of global behavior and patterns

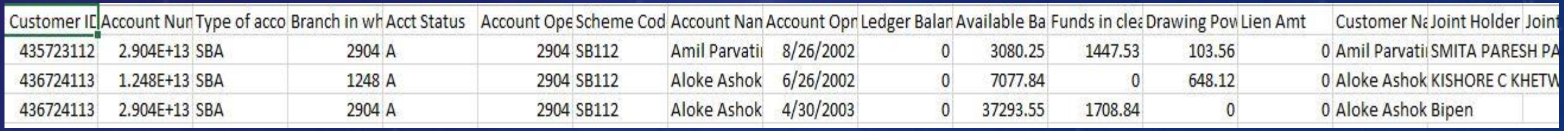
PNL and strategies

# DATA COLLECTION FROM BANK

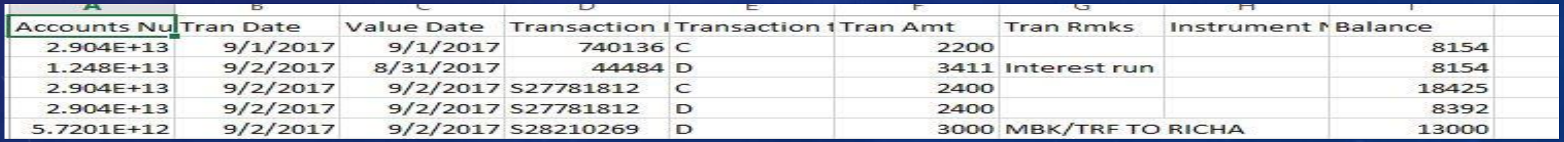
- Data collection of customer that are registered with bank or have business with bank. Ex- Account data, Customer Data, Transition data etc.
- This data is available to the bank and will be used in further data processing. Here are some snapshots of above mentioned available data:



| Customer Id | First Name | Middle Nam | Last Name | Gender | Date Of Birtl | Guardian Na | Line Address | Line Address | Line Address | PIN | City | State | NATIONALIT | Occupation | Annual Total | Staff |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 435723112 | Amil | Parvatinand | Amarnath | M | 4/23/1967 | | A 502 DIAMC | TIRUPATI NAGAR PHASE II | | 401303 | THANE | MH | INDIAN | SERVI | 636007 | Y |
| 436724113 | Aloke | Ashok | Bipen | M | 4/1/1974 | | FLAT NO 202 | LTD THANE EAST | | 400603 | THANE | MH | INDIAN | SERVI | 636008 | Y |
| 437725114 | Nirijhar | Abhijaya | Bonjani | M | 2/1/1957 | | A 18 MANSA | VAISHALI NAGAR | | 305001 | AJMER | RJ | INDIAN | PSBNK | 636010 | Y |
| 438726115 | Pragyawati | Pramiti | Puchar | F | 6/5/1962 | | AT SANSARI | MUNJABA CHOWK H NO | | 422401 | DEOLA | MH | INDIAN | SERVI | 630002 | Y |



| Customer ID | Account Nun | Type of acco | Branch in wh | Acct Status | Account Ope | Scheme Cod | Account Nan | Account Opn | Ledger Balan | Available Ba | Funds in clea | Drawing Pow | Lien Amt | Customer Na | Joint Holder | Joint |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 435723112 | 2.904E+13 | SBA | 2904 | A | 2904 | SB112 | Amil Parvatii | 8/26/2002 | 0 | 3080.25 | 1447.53 | 103.56 | 0 | Amil Parvatii | SMITA PARESH PA | |
| 436724113 | 1.248E+13 | SBA | 1248 | A | 2904 | SB112 | Aloke Ashok | 6/26/2002 | 0 | 7077.84 | 0 | 648.12 | 0 | Aloke Ashok | KISHORE C KHETW | |
| 436724113 | 2.904E+13 | SBA | 2904 | A | 2904 | SB112 | Aloke Ashok | 4/30/2003 | 0 | 37293.55 | 1708.84 | 0 | 0 | Aloke Ashok | Bipen | |

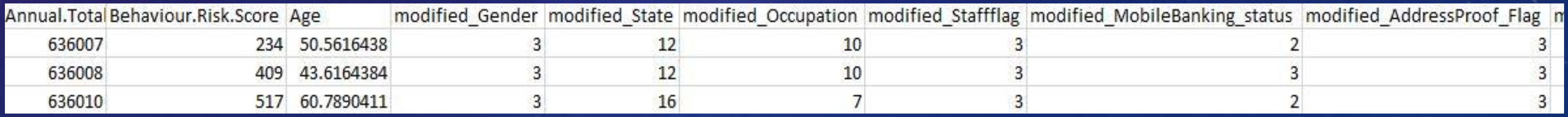


| A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|
| Accounts Nu | Tran Date | Value Date | Transaction I | Transaction 1 | Tran Amt | Tran Rmks | Instrument N | Balance |
| 2.904E+13 | 9/1/2017 | 9/1/2017 | 740136 | C | 2200 | | | 8154 |
| 1.248E+13 | 9/2/2017 | 8/31/2017 | 44484 | D | 3411 | Interest run | | 8154 |
| 2.904E+13 | 9/2/2017 | 9/2/2017 | S27781812 | C | 2400 | | | 18425 |
| 2.904E+13 | 9/2/2017 | 9/2/2017 | S27781812 | D | 2400 | | | 8392 |
| 5.7201E+12 | 9/2/2017 | 9/2/2017 | S28210269 | D | 3000 | MBK/TRF TO RICHA | | 13000 |

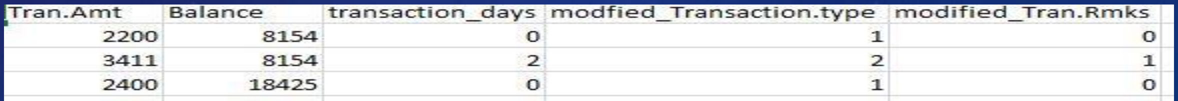
# STANDARDIZATION OF DATA

- Standardization based on levels of data
  - Lets say if we have customers belongs to 3 different states then we can assign levels to them as 1,2,3 and so on.
- Transformation from string to numeric data
  - If we have data such as joint holder where names are given. Here we will replace that with 1/0, 1: if joint account else 0.
  - R-scripts are used for data-processing

Here are some screenshots of standardized data:



| Annual.Total | Behaviour.Risk.Score | Age | modified_Gender | modified_State | modified_Occupation | modified_Staffflag | modified_MobileBanking_status | modified_AddressProof_Flag | m |
|---|---|---|---|---|---|---|---|---|---|
| 636007 | 234 | 50.5616438 | 3 | 12 | 10 | 3 | 2 | 3 | |
| 636008 | 409 | 43.6164384 | 3 | 12 | 10 | 3 | 3 | 3 | |
| 636010 | 517 | 60.7890411 | 3 | 16 | 7 | 3 | 2 | 3 | |



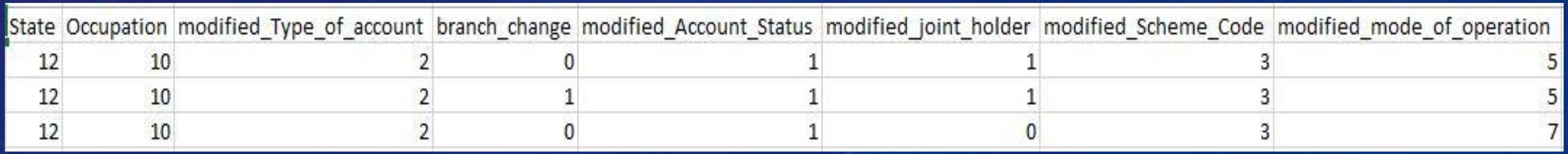
| Tran.Amt | Balance | transaction_days | modfied_Transaction.type | modified_Tran.Rmks |
|---|---|---|---|---|
| 2200 | 8154 | 0 | 1 | 0 |
| 3411 | 8154 | 2 | 2 | 1 |
| 2400 | 18425 | 0 | 1 | 0 |

# POLICY MAKING

- Prediction for the attributes of a standardized data.
- Selection of relevant information for prediction(this is taken care of during standardization).
- Forecasting there global behavior and pattern.
- Look for strategic point and information and further usages in PNL and individual KYC.

for example prediction for account status based on other attributes present in data.

status = predict(occupation, state, branch_change……….)

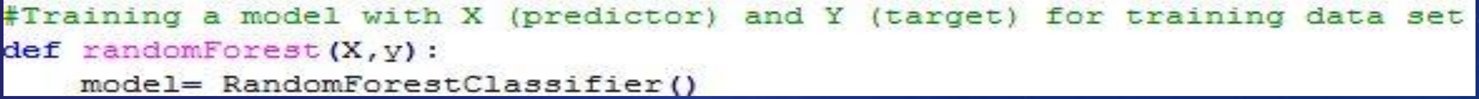


| State | Occupation | modified_Type_of_account | branch_change | modified_Account_Status | modified_joint_holder | modified_Scheme_Code | modified_mode_of_operation |
|---|---|---|---|---|---|---|---|
| 12 | 10 | 2 | 0 | 1 | 1 | 3 | 5 |
| 12 | 10 | 2 | 1 | 1 | 1 | 3 | 5 |
| 12 | 10 | 2 | 0 | 1 | 0 | 3 | 7 |

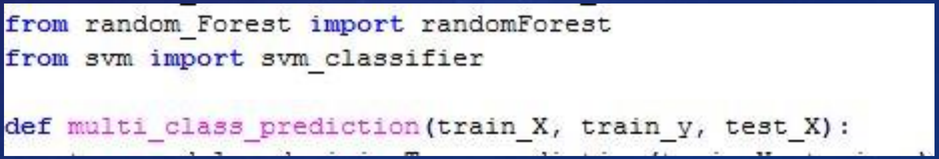
# TRAINING OF MODULES

- Training of multi-class classification model and hence training of all ML models. Ex- Regressions, Baysians etc.
- Training of all models on standardized data(These models will be trained on each policy separately).
- Data for training can be controlled, and of course same can be done for test data. These modules usually work on large data, so, here we will be using same data for training and testing.

here are some screen shots of training of random forest and multiclass model training:



```
#Training a model with X (predictor) and Y (target) for training data set
def randomForest(X,y):
    model= RandomForestClassifier()
```

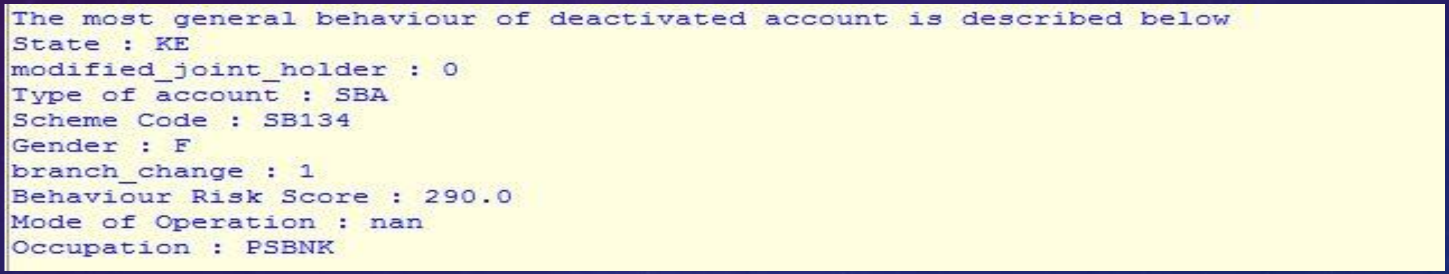


```
from random_Forest import randomForest
from svm import svm_classifier

def multi_class_prediction(train_X, train_y, test_X):
```

# PREDICTIONS

- Prediction of test data based on trained model
- Prediction mostly includes global behavior of particular policy(for that we are predicting). Ex: for account status policy following global is predicted

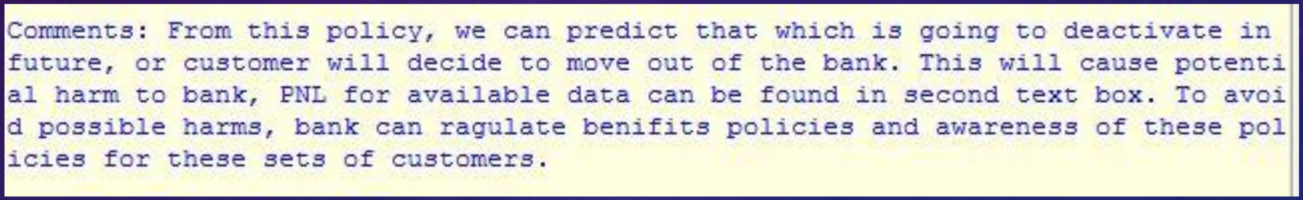


```
The most general behaviour of deactivated account is described below
State : KE
modified_joint_holder : 0
Type of account : SBA
Scheme Code : SB134
Gender : F
branch_change : 1
Behaviour Risk Score : 290.0
Mode of Operation : nan
Occupation : PSBNK
```

- These results shows that what kind of accounts are getting deactivated. As we can see customer with scheme code SB134 and occupation PSBNK ……… information are deactivating there accounts.

# PREDICTING STRATEGIES

- Prediction strategies includes the  conclusion of global behaviors, these are the strategies to prevent losses and promoting market campaign and rolling out new services.
- It can be seen in the comments of individual strategy.

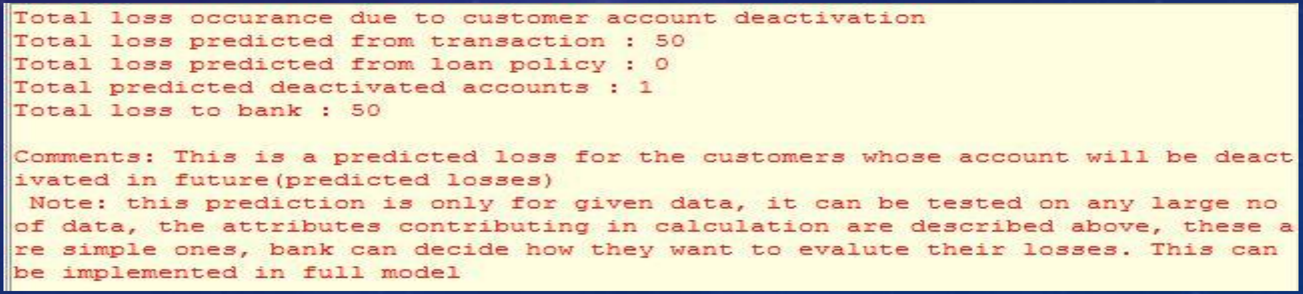


```
Comments: From this policy, we can predict that which is going to deactivate in
future, or customer will decide to move out of the bank. This will cause potenti
al harm to bank, PNL for available data can be found in second text box. To avoi
d possible harms, bank can ragulate benifits policies and awareness of these pol
icies for these sets of customers.
```

- Rolling out new strategies for each policy based on it’s behavior, from global we can move to specific strategies, providing new services to group of specific customers(This is available  in fully developed model)

# PREDICTING PNL

- Predicting PNL based of the global behavior.
- Includes, loss due to current policies, investment in policies that bank will be rolling out, expenses in new market campaigns, and prediciting benefits from these market campaigns, and also future trends.
- In demo version we have calculated losses part only, since standardization of investment in new policies and strategies can differ from bank to bank (full model will have all these features). Ex:- loss report from deactivated accounts.



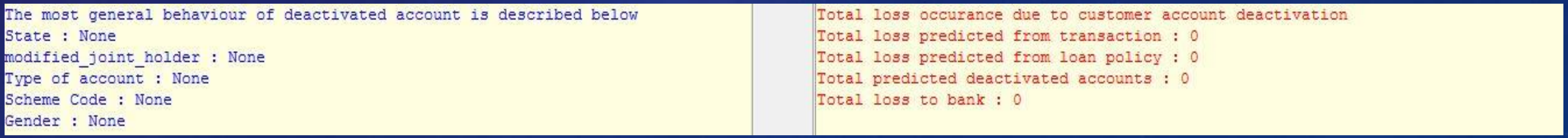
```
Total loss occurance due to customer account deactivation
Total loss predicted from transaction : 50
Total loss predicted from loan policy : 0
Total predicted deactivated accounts : 1
Total loss to bank : 50

Comments: This is a predicted loss for the customers whose account will be deact
ivated in future(predicted losses)
 Note: this prediction is only for given data, it can be tested on any large no
of data, the attributes contributing in calculation are described above, these a
re simple ones, bank can decide how they want to evalute their losses. This can
be implemented in full model
```

# GLOBAL BEHAVIOR FOR INDIVIDUAL

- Once done predicting policies/strategies/campaigns etc, we will look for individual now.
- Predict where individual  stand in that policies, effect of policies and new benefits from new services.
- Individual can be a small group with almost same behavior or can be a HNI.

in this demo, we developed for global policy for individual , fully model will have all these functionalities. Here is an example: account is not in deactivated group so no loss-



```
The most general behaviour of deactivated account is described below
State : None
modified_joint_holder : None
Type of account : None
Scheme Code : None
Gender : None
```

```
Total loss occurance due to customer account deactivation
Total loss predicted from transaction : 0
Total loss predicted from loan policy : 0
Total predicted deactivated accounts : 0
Total loss to bank : 0
```

# RESULTS

- Results includes new policies, services, and their cost and benefits and much more, since it is a demo model, fully developed model have more elaborate report of all results with individual strategies.

# CONCLUSIONS

- Data standardization is done on basis of random forest, bank can manually decide what part they want to focus on.
- Multiple ML models are used in classification, we can add any best performing model to multiclass classification, since it will further reduce the errors.
- Bank can decide what policy they want based on that final PNL can be implemented.
- Large amount of data is required to work this model effectively. More the data better the performance.
- Easy to implement model, can be implemented in bank's regular system.
- Further upgrades and modification can be done like automation of changes and auto-evaluated customer rating and much more.

# REFERENCES

- ML references from my own project.

  Maneesh, Varun and others, “Data Mining and Knowledge Discovery”, IME 672 (course project).

- A little bit of IOT.

# QUESTIONS

# THANK YOU