Identify Hidden Affluent Customers

Hiring assessment for Maybank

Presented by:
Naitik Shukla
Senior Data Scientist
30-Mar-2024

Introduction

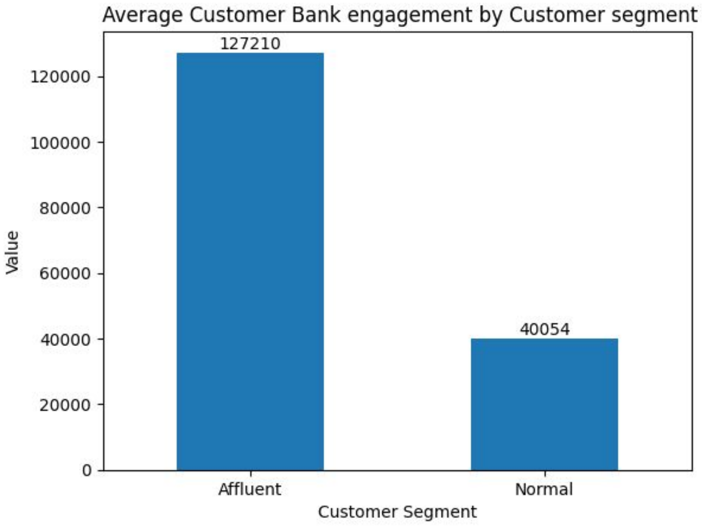
- The bank is always striving to enhance its customer base across various segments to maximize revenue opportunities.
- One crucial aspect of this endeavor is upgrading the segment of Existing To Bank (ETB) customers from Normal to Affluent.
- By identifying hidden affluent customers within the existing customer base, the bank can effectively target them for upselling relevant products and services.
- This presentation aims to outline our approach to identifying these hidden affluent customers and the potential impact on revenue growth.

Introduction

- **Business Problem**: The bank seeks to upgrade ETB customers to the affluent segment to enhance revenue opportunities.
- **Objective**: Identify hidden affluent customers within the existing customer base for targeted upselling.
- **Significance**: Upgrading customer segments can lead to increased revenue and improved customer satisfaction.
- **Presentation Overview**: We will discuss our methodology, data analysis, and data prep for approaching identified affluent customers, and ML based approach to identify hidden affluent customers.
- **Result**: Show Monetary benefit with new approach over original value.

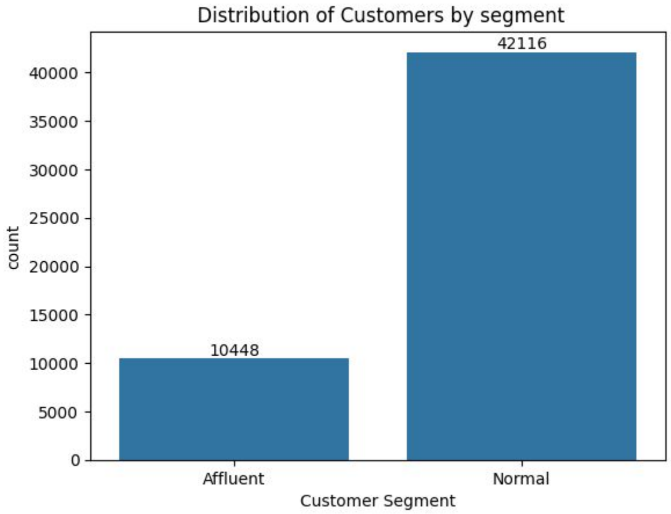
EDA Key finding

1. Average Bank Engagement of Customer by segment
    a. Affluent: 127,210
    b. Normal: 40,054



Count of Customers by segment

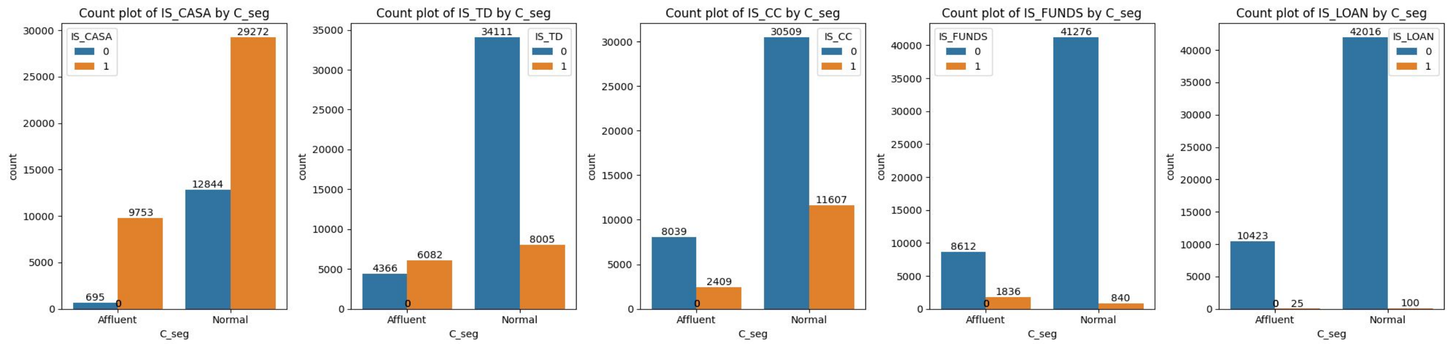
a. Affluent: 10,448
b. Normal: 42,116



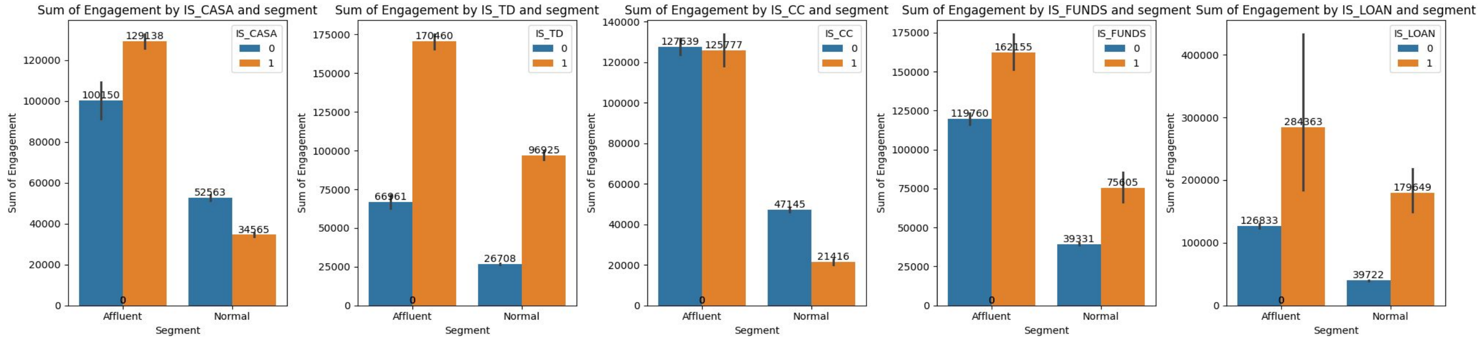
EDA Key finding contd.

5 Products identified in given data viz. CASA,TD,LOAN,FUNDS,CC(Credit Card)

- CASA, TD(inferred) and Credit Card is most sold products with high engagement.



Customer Count by Product

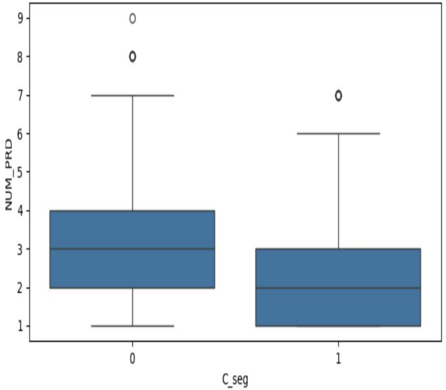


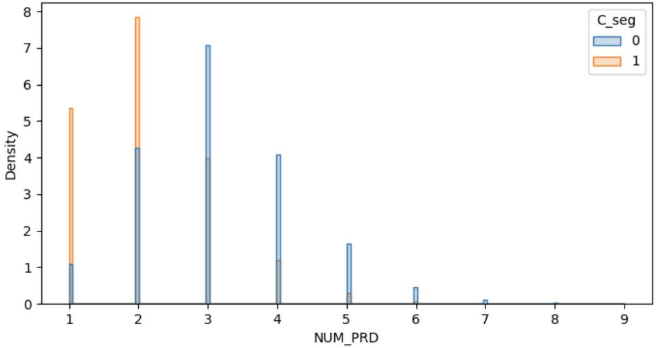
Customer Monetary Engagement by Product

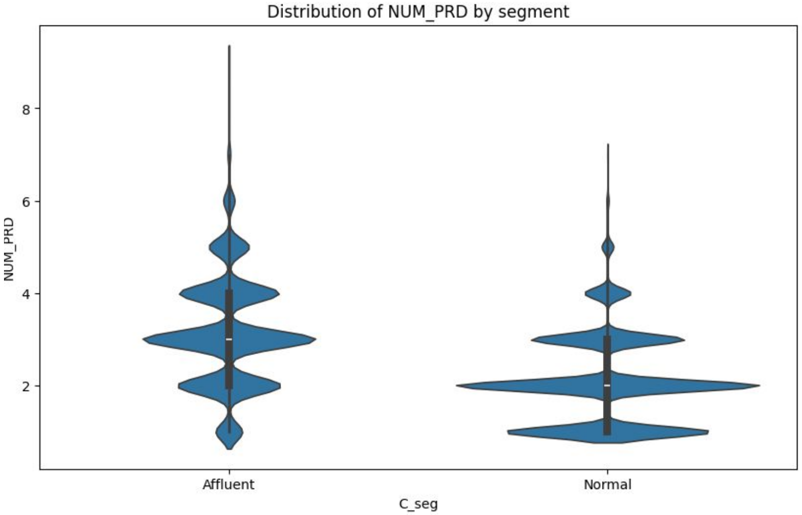
EDA Key finding contd. - NUM_PRD

- Few columns distribution with interesting variation between segments.

Number of Product: Affluent customers tend to have more number of products than Normal.

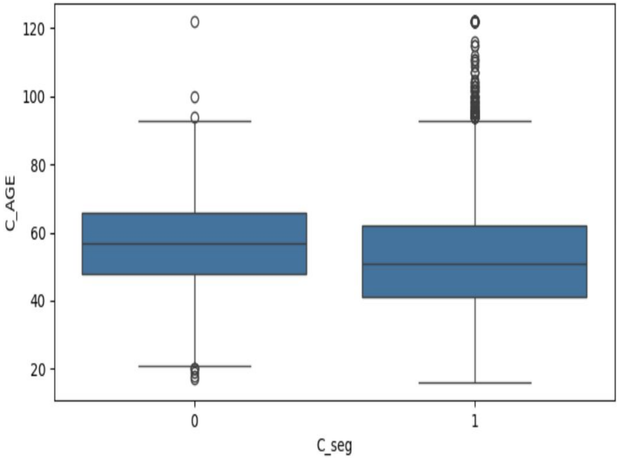


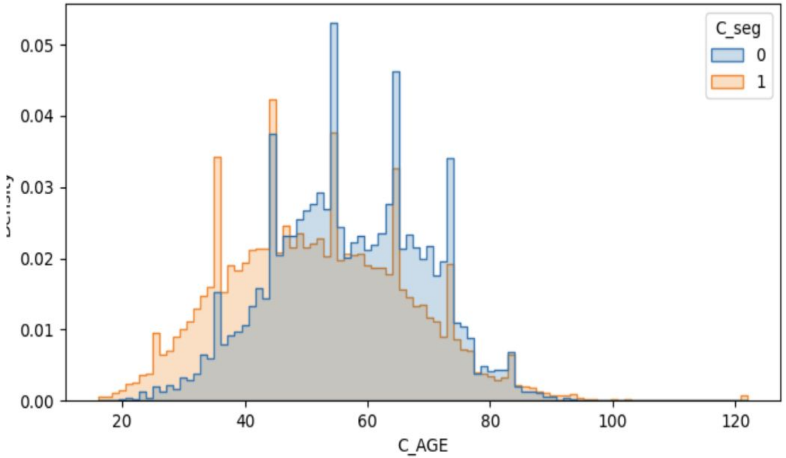




EDA Key finding contd. - C_Age

Age: Many Normal segment users lies in Outliers zone(Q3+1.5*IQR), but generally Q1-Median-Q3 is lower for Normal category than Affluent category.





Key Takeaways from RFM

R - Recency — No data to get this

F - Frequency — No data to get this

M - Monetary → Focus on this

MTHCASA	- average monthly balance in CASA
MTHTD	- average monthly balance in TD
Asset_value	- Total Asset value
UT_AVE	- Average Unit Trust(UT) value
AVG_TRN_AMT	- Average Credit card transaction amount in a month

Data Prep - NULL

While exploring data, it has been observed that data is not complete, there are lots of missing values exist in data, So before building any derived features first task is to cleanup what seems to be bad record.

For eg. For huge number of data, there exist no information about their product Monetary information like below, so we removed from our analysis.

	C_ID	NUM_PRD	CASATD_CNT	HL_tag	AL_tag	N_FUNDS	ANN_N_TRX	C_seg
65856	16777	2	NaN	NaN	NaN	NaN	NaN	NORMAL
65865	17211	2	NaN	NaN	NaN	NaN	NaN	NORMAL
65866	17293	2	NaN	NaN	NaN	NaN	NaN	NORMAL
65929	19536	2	NaN	NaN	NaN	NaN	NaN	NORMAL
66039	21940	2	NaN	NaN	NaN	NaN	NaN	NORMAL

Data Prep contd. - DUPLICATE

We identify there exists DUPLICATE entry within our dataset based on C_ID, which should able to help us identify repeating customer.

Looking into specific data from below where C_ID=11, we can see it's not possible for same customer to have both records.

Considering this fact, for further analysis we will consider each record as separate customer information, C_ID and PC are dummy variable with no value.

	C_ID	C_AGE	C_EDU	C_HSE	PC	INCM_TYP	gn_occ	NUM_PRD	CASATD_CNT	MTHCASA
44084	0	31	NaN	NaN	20184.0	2.0	PMEB	1	NaN	NaN
7920	0	61	NaN	NaN	29894.0	NaN	HOUSEWIFE	2	1.0	35373.02
34921	11	70	A-Levels	HDB 4-5 ROOM	22167.0	2.0	RETIREE	4	2.0	34867.23
31864	11	70	Masters	SEMI-DETACHED	9259.0	6.0	PMEB	2	1.0	47782.61

Data Prep contd. - BINNING

For features like Age, Edu, housing and occupation there exist many Null as well as multiple entry which have ordinal value.

We clubbed multiple categories into single bin based on Heuristic knowledge only.

House_Type	House_Value
HDB 1-3 ROOM	0
Others	0
NaN	0
HDB 4-5 ROOM	1
HDB EXECUTIVE APARTMENT/ MANSIONETTE	1
EXECUTIVE CONDOMINIUM	2
PRIVATE APARTMENT	2
PRIVATE CONDOMINIUM	2
SEMI-DETACHED	3
TERRACE	3
BUNGALOW	4
SHOPHOUSE	4
INDUSTRIAL BUILDING	5
COMMERICAL BUILDING	5
OFFICE	5
HOTEL/ SERVICE APARTMENT	5

Education	Education_Value
Below O-Levels	0
O-Levels	0
A-Levels	0
Others	0
NaN	0
Diploma	1
Technical/Vocational Qualifications	1
Degree	2
Professional Qualifications	2
Masters	3
PHD/Doctorate	3

Occupation	Occupation_Value
STUDENT	0
HOUSEWIFE	0
OTHERS	0
Others	0
NaN	0
BLUE COLLAR	1
RETIREE	1
PMEB	2
WHITE COLLAR	2

Data Prep contd. - New features

To show if any products any customer purchased, we will infer based on amount for each product identified columns, and create columns value 1 or 0.

For eg. IS_CASA can have values (0,1) which tells if that customer have that product or not.

Created 5 columns **IS_CASA, ID_TD, IS_FUNDS, IS_LOAN, IS_CC** for 5 products identified.

Data prep contd. - New Features

To focus on Monetary part, we will derive some new features which will give information about customer with less sparse data.

Some new features are:

- **CASA_DIFF** = Range of Min and Max specific customer balance over last year (This will signal activity and big range signify bigger transactions)
- **CC_MTH_TRN_AMT_DIFF** = provide Range of transactions using CC customer doing(signal for activity and how big spend)
- **LoanAsset_ratio** = Total Loan purchase price with Asset value ratio, signal buying strength
- **AssetvCValue** = Find ratio with total asset and total yearly assets accumulated, signal growth of customer.
- **Cus_engagement_val** = Total engagement of customer with Bank

Data prep contd. - Outliers

Age=2, occupation = RETIREE

	C_ID	C_AGE	C_EDU	C_HSE	PC	INCM_TYP	gn_occ
36911	57783	2	NaN	NaN	11212.0	NaN	RETIREE

Removing highest 5 records from Top 1 percentile customers based on

MTHCASA

DRvCR

Asset value

	C_ID	gn_occ	C_AGE	MTHCASA
10594	27951	HOUSEWIFE	42	6534839
48329	15555	PMEB	58	4206869
10981	3792	PMEB	85	4106541
29469	14133	PMEB	78	3154990
53573	6014	RETIREE	84	3150314

	C_ID	gn_occ	C_AGE	DRvCR
376	17189	PMEB	60	11635000
10627	31877	PMEB	59	4670670
15437	61121	HOUSEWIFE	95	3800000
3459	66104	PMEB	41	3256142
47715	418	RETIREE	77	3215349

	C_ID	gn_occ	C_AGE	Asset value
50418	56192	PMEB	55	7940605
10594	27951	HOUSEWIFE	42	7115850
9877	31752	HOUSEWIFE	85	4953129
376	17189	PMEB	60	4403973
48329	15555	PMEB	58	4223319

Model Training - Approach

Since we have to predict result for whole dataset, So classical train-test approach with 80:20 ratio won't work, Since predicted dataset have already been used in Train dataset.

To predict and train on complete dataset without data leak we use 2 concepts:

1. **`StratifiedKFold` -** each set contains approximately the same percentage of samples of each target class as the complete set.
2. **`Cross_val_predict` -** For each split, it fits the model on the training set and makes predictions on the test set.

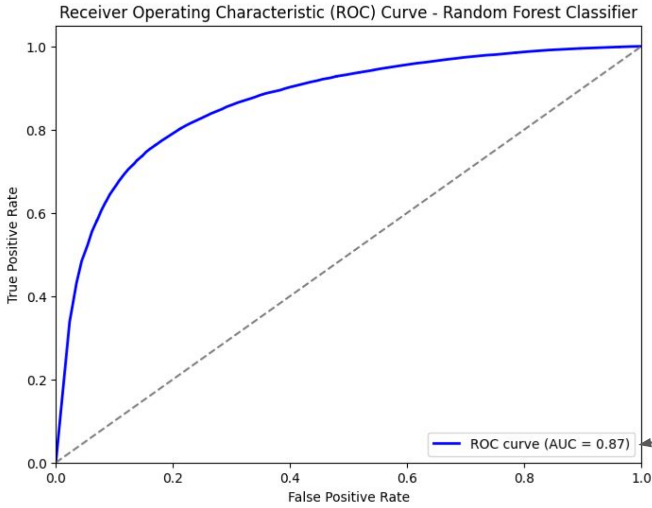
Model Training - Random Forest Classifier

```
##################################################
               Random Forest Classifier
##################################################
Cross-validated Precision for class 0 (affluent customers): 0.642007303569325
Cross-validated Recall for class 0 (affluent customers): 0.5216309341500766
Cross-validated F1-score for class 0 (affluent customers): 0.5755927549242225

Number of hidden affluent customers identified: 3039
Number of Existing affluent customers: 10448
```

Model performance

Additional customers identified



Higher AUC signifies learning from data

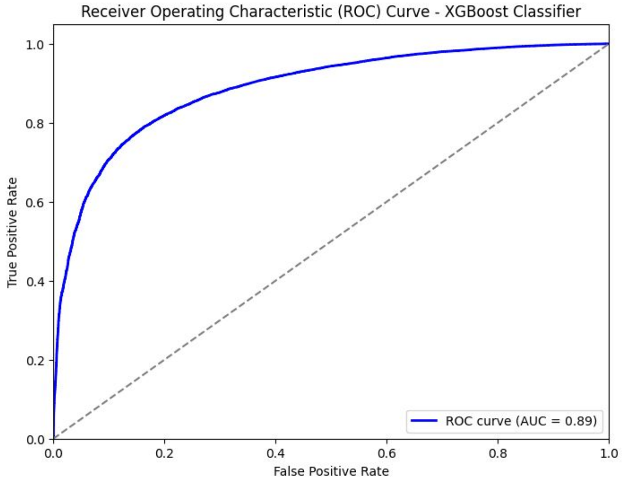
Model Training - XGBoost Classifier

```
##################################################
                XGBoost Classifier
##################################################
Cross-validated Precision for class 0 (affluent customers): 0.6824906728418886
Cross-validated Recall for class 0 (affluent customers): 0.5077526799387443
Cross-validated F1-score for class 0 (affluent customers): 0.5822951539432523

Number of hidden affluent customers identified: 2468
Number of Existing affluent customers: 10448
```

Model performance

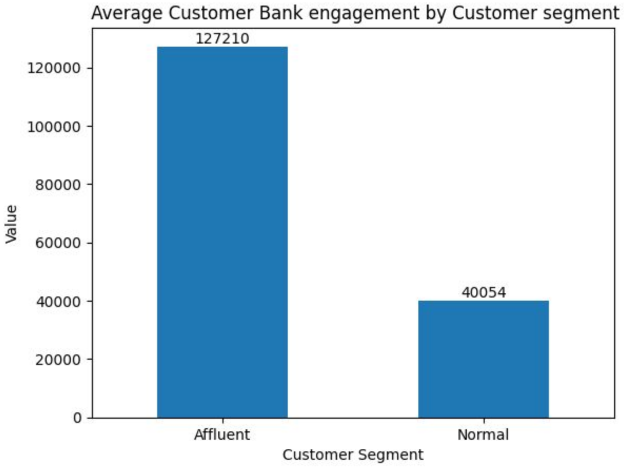
Additional customers identified



Higher AUC signifies learning from data

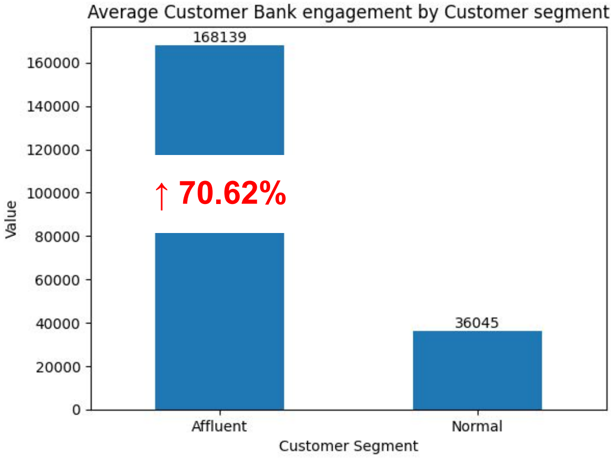
Model Training - Business value

Original from data



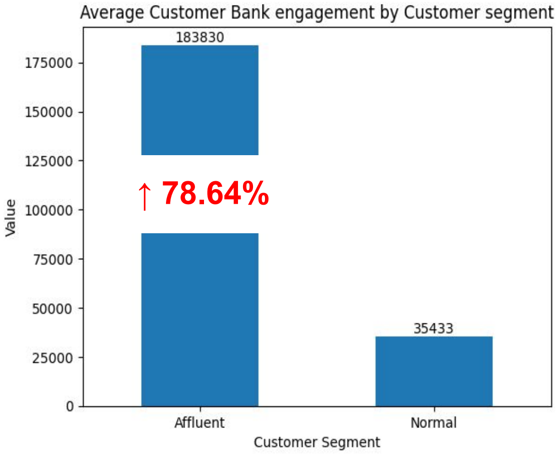
Total Monetary value of Affluent customers= **1,329,093,823**

Random Forest out



New Total Monetary value of Affluent customers= = **2,267,690,693**

XGBoost output



Total Monetary value of Affluent customers= = **2,374,348,280**

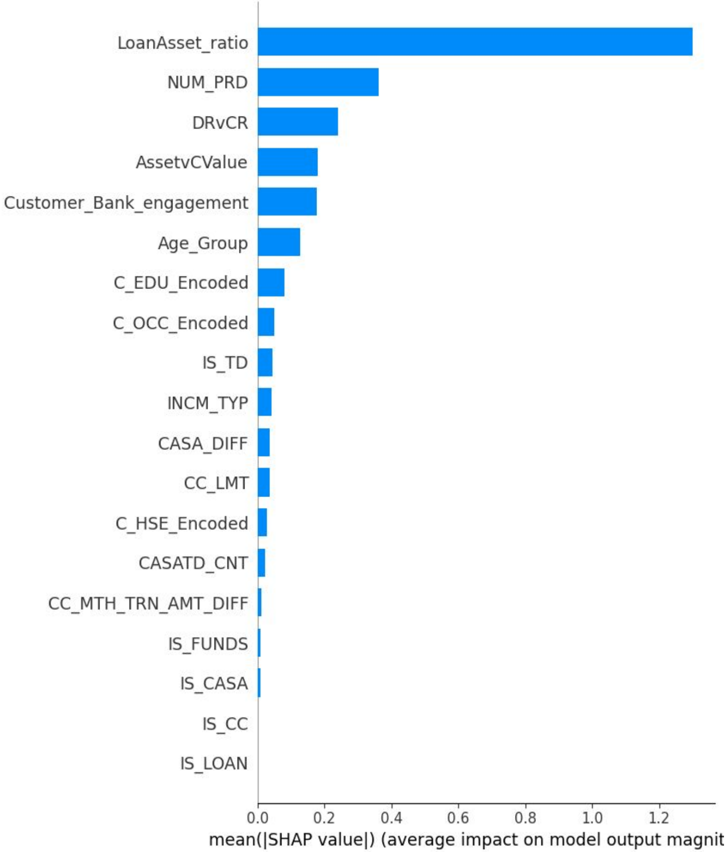
For more breakdown into each product details refer here: **[ProductWise Customer Engagement]**

Post Training Analysis - SHAP(**SH**apley **A**dditive ex**P**lanations)

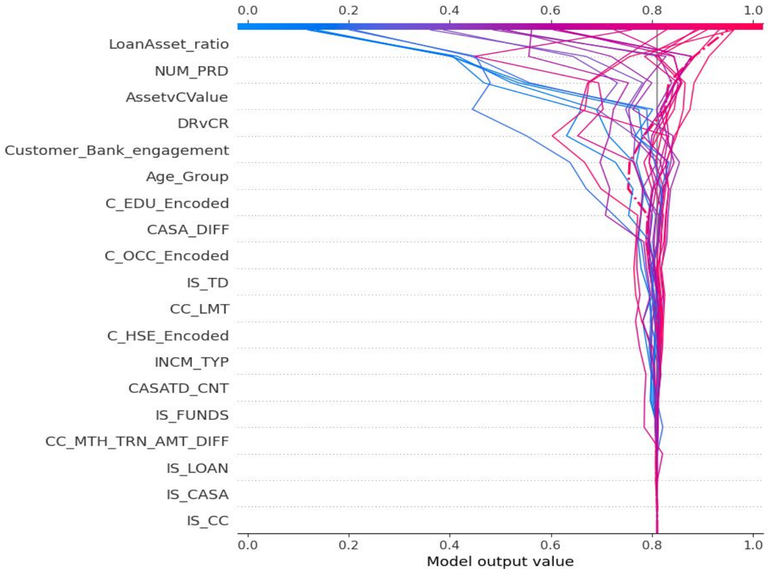
Post model train SHAP helps in model explainability by quantifies the contribution of each feature to a model's output, offering granular insights into prediction behavior. By revealing feature importance, it enhances model interpretability, aiding in decision-making and model refinement.

SHAP originated from cooperative game theory's ‘Shapley values’ and has been adapted for machine learning interpretability. This technique is model agnostic and really helps to understand our model prediction.

Post Training Analysis - SHAP - Feature Importance

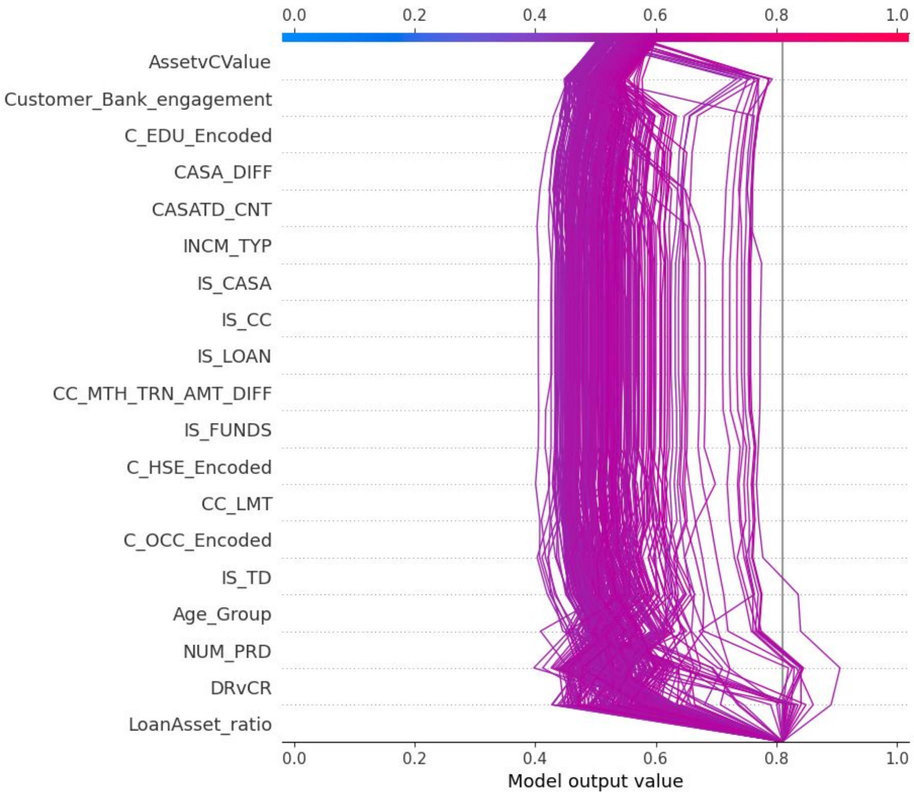


- Few columns have enough information to have similar model - we can truncate some columns from our model building.



Decision plot for sample records

Post Training Analysis - Decision Plot for prob(.5-.6)

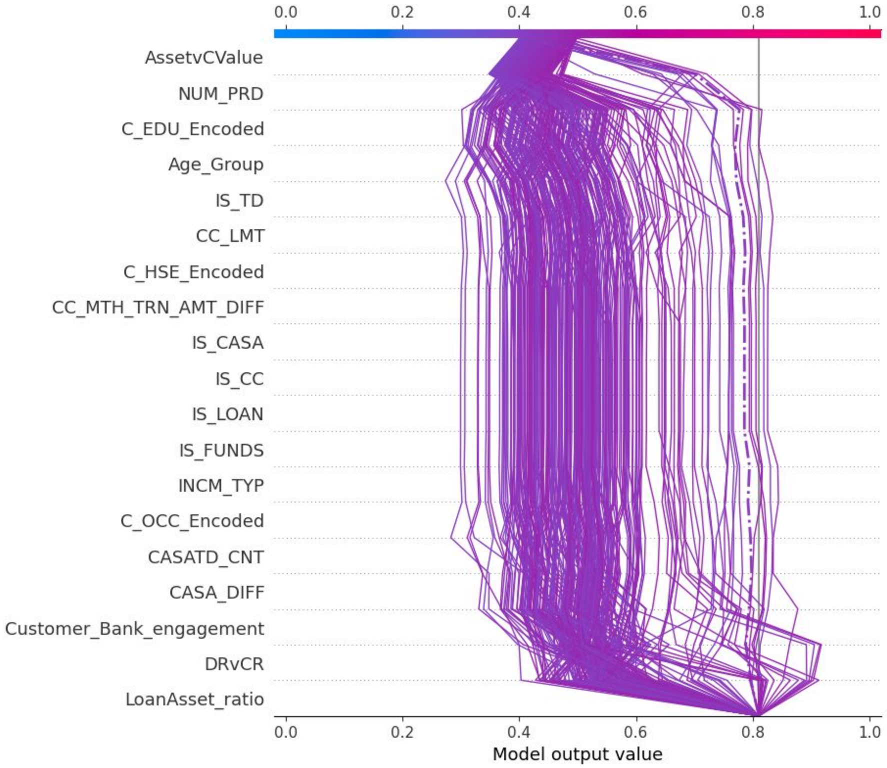


Our threshold for model prediction taken is .50, So Plotting inferences with a probability threshold between 0.5 and 0.6 allows us to focus on predictions near the decision boundary.

It aides in understanding model uncertainty and misclassifications within this range.

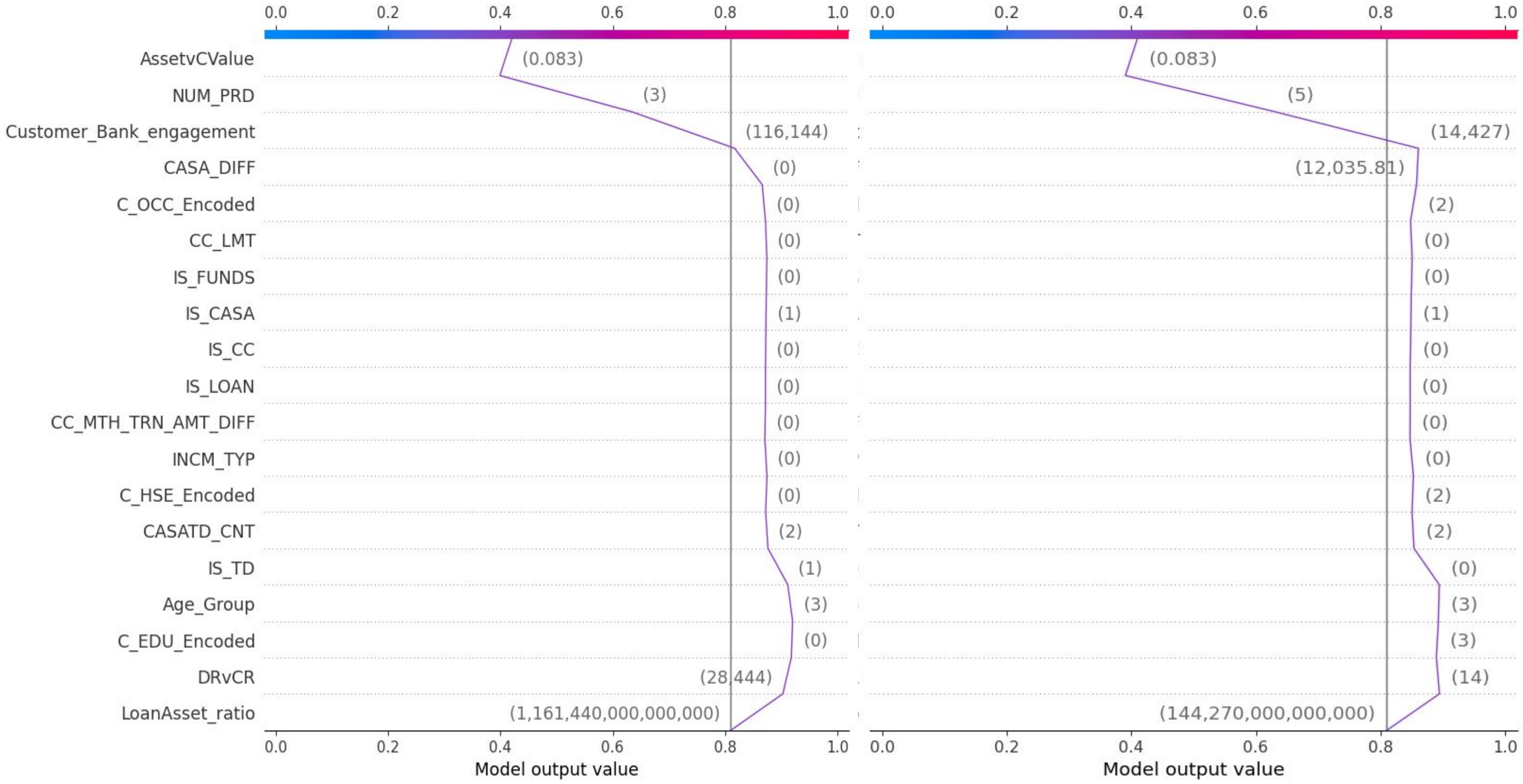
There are not many outliers in the output, Effect of **LoanAsset_ratio**, **DRvCR**, **NUM_PRD**, **AssetvCValue**,**Customer_engagement** immediately stand out

Post Training Analysis - Decision Plot for prob(.4-.5)



There are few outliers in the output, Effect of **LoanAsset_ratio**, **DRvCR**, **NUM_PRD**, **AssetvCValue**,**C_EDU and Customer_engagement** immediately stand out

Visualize 2 outliers from range(.4-.5) prob



Thankyou


Random Forest Classifier
ProductWise Customer Engagement
XGBoost Classifier
Sum of Engagement by IS_CASA and Segment
IS_CASA
0
1
170841
119251
53221
28723
Sum of Engagement by IS_TD and Segment
IS_TD
195891
106548
80803
25811
Sum of Engagement by IS_CC and Segment
IS_CC
167186
171152
43187
16902
Sum of Engagement by IS_FUNDS and Segment
IS_FUNDS
168747
166233
36109
31588
Sum of Engagement by IS_LOAN and Segment
IS_LOAN
388539
167775
176887
35689
Customer Bank engagement
Segment
Affluent
Normal
Sum of Engagement by IS_CASA and Segment
184337
166010
53215
27904
Sum of Engagement by IS_TD and Segment
212113
117312
75959
25757
Sum of Engagement by IS_CC and Segment
188316
171984
42623
15519
Sum of Engagement by IS_FUNDS and Segment
189844
167261
35580
24745
Sum of Engagement by IS_LOAN and Segment
425983
183237
160192
35137