# Vehicle Financing Cross Sell Insights and Modeling

By Partha Chowdhury

Email: partha2@gmail.com

Linkedin

# Table of Contents:

# Problem Statement and Solution Approach

The company's current cross-sell rate is 12%, which is lower than the industry average. The company wants to increase the cross-sell rate to at least 25% to boost its revenue and improve customer loyalty.

**Approach:**

Data Gathering & Synthetic data generation → Data cleaning, missing value handling etc. → EDA → Model building → Insights

- For the purpose of this case study,  I have found a relatable dataset on vehicle loan from Kaggle.
- I have created certain features synthetically using Faker library and other methods to make it suitable for cross sell analysis

<u>Assumptions:</u>

- As some data are synthetically generated and rest of data source is from open platform, this may not yield in real-life insights or model. But this analysis is to highlight the approach and solutioning.

# Executive summary: Customer Behavior, Preferences, Product Usage, Patterns and Correlations

- ❖ Customers aged 20-40 shows higher cross-sell rates, specially among self-employed individuals
- ❖ Customers with higher loan amounts (>50K) or insurance usage are more likely to cross-sell.
- ❖ High engagement via emails (click rates >50%) is positively correlated with cross-sell.
- ❖ Customers with ‘Asset to Loan value’ ration of 20-80% is showing better cross sell rate

❖ *The most important features driving cross-sell success are*

- ✓ Age of the customer
- ✓ loan to asset value ratio (ltv)
- ✓ Interaction intensity
- ✓ SMS count
- ✓ Email click rate
- ✓ Credit history length etc.

All the above features are showing in top10 important features across models

Pivot Table: Cross-Sell Rate by Age Bracket

| Employment_Type | Salaried | Self employed |
|---|---|---|
| age_bracket | | |
| 20-30 | 24.686235 | 25.433234 |
| 30-40 | 23.505966 | 23.340679 |
| 40-50 | 10.048891 | 9.902628 |
| 50-60 | 9.844811 | 9.731371 |

Pivot Table: Cross-Sell Rate by Loan Bracket

| insurance_flag | 0 | 1 |
|---|---|---|
| loan_bracket | | |
| 0-50K | 17.151943 | 20.767370 |
| 50K-100K | 17.370820 | 21.049192 |
| 1L-2L | 20.796460 | 23.333333 |

Pivot Table: Cross-Sell Rate by Email Engagement (%)

| | cross_sell_flag |
|---|---|
| email_click_rate_bracket | |
| 0-10% | 18.063856 |
| 10-20% | 17.805348 |
| 20-50% | 17.907100 |
| 50-70% | 17.676963 |
| >70% | 18.248473 |

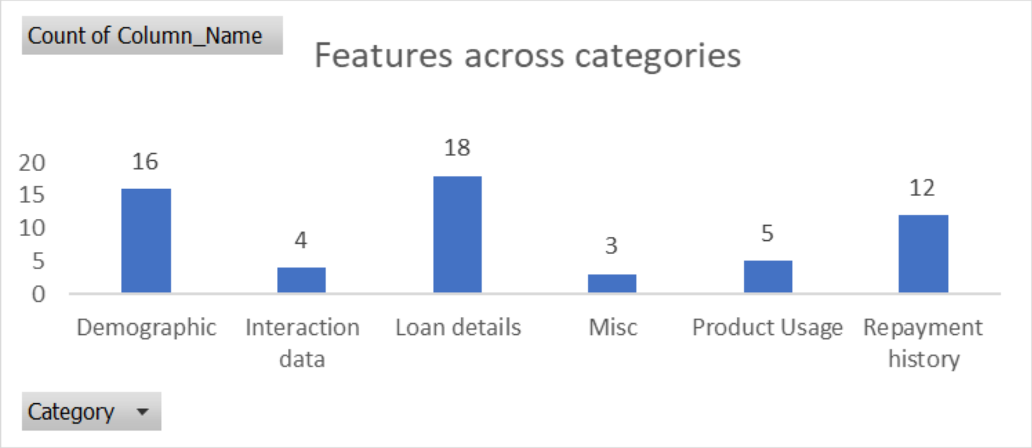
Cross-Sell Rate by Asset-value-to-Loan Ratio:

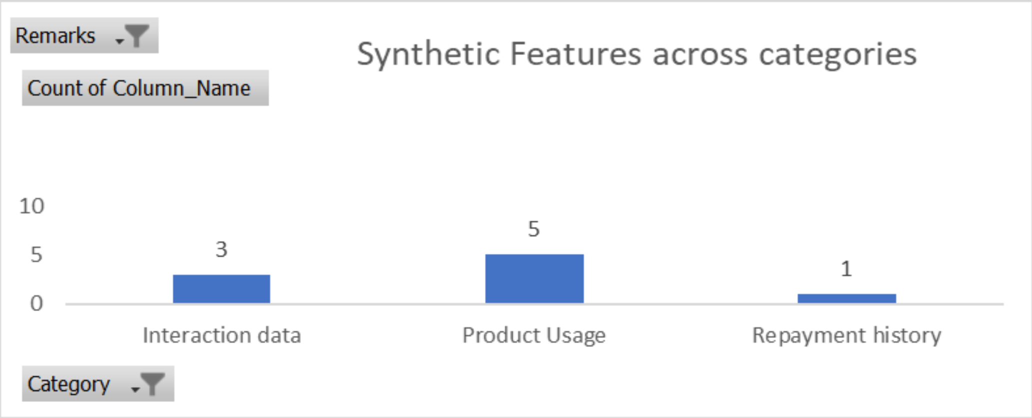
| | cross_sell_flag |
|---|---|
| Asset_value_to_loan_bracket | |
| Low (<20%) | 14.000000 |
| Moderate (20-50%) | 19.242166 |
| High (50-80%) | 18.326857 |
| Very High (>80%) | 16.908256 |

More insights from model and recommendations is presented at the end of this presentation

# About the Data:

- Dataset has 2.3 lakhs observations with 41 features.
- Some of the important features are:
  - Customer demographic details: DOB, Employment type, PAN_Flag, Adhaar_Flag etc.
  - Loan details: Disbursed amount, Asset cost, Installment amount, risk_level (Derived) etc.
  - Interaction data: Call Logs count, SMS count, email click rate
  - Product Usage data: personal loan flag, insurance flag, asset mgmt flag, cross sell flag




Out[16]: (233154, 41)

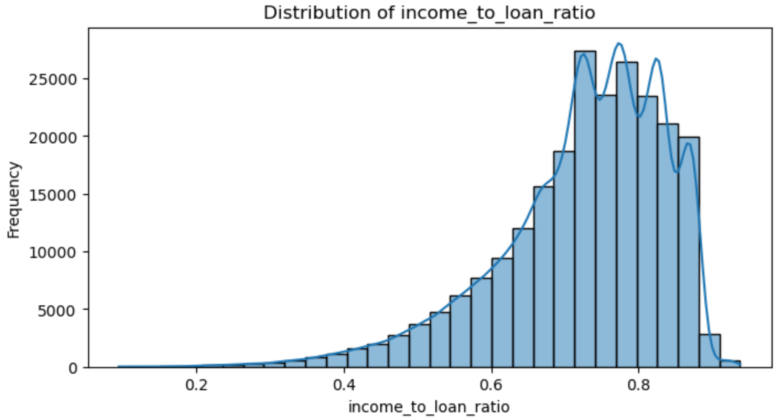
Out[16]:

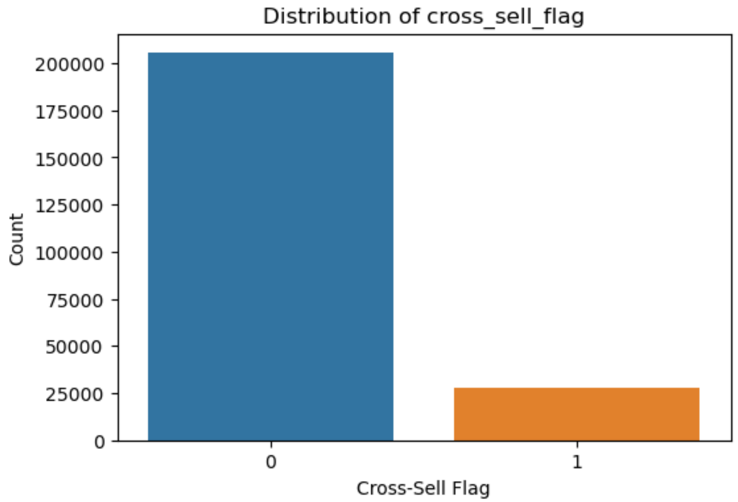
| | UniqueID | disbursed_amount | asset_cost | ltv | branch_id | supplier_id | manufacturer_id | Current_pincode_ID | Date.of.Birth | Employment.Type | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 420825 | 50578 | 58400 | 89.55 | 67 | 22807 | 45 | 1441 | 01-01-84 | Salaried | ... |
| 1 | 537409 | 47145 | 65550 | 73.23 | 67 | 22807 | 45 | 1502 | 31-07-85 | Self employed | ... |
| 2 | 417566 | 53278 | 61360 | 89.63 | 67 | 22807 | 45 | 1497 | 24-08-85 | Self employed | ... |
| 3 | 624493 | 57513 | 66113 | 88.48 | 67 | 22807 | 45 | 1501 | 30-12-93 | Self employed | ... |
| 4 | 539055 | 52378 | 60300 | 88.39 | 67 | 22807 | 45 | 1495 | 09-12-77 | Self employed | ... |

| Remarks | Sythetic data |
|---|---|
| **Row Labels** | **Count of Column_Name** |
| **Interaction data** | **3** |
| call_logs_count | 1 |
| email_click_rate | 1 |
| sms_count | 1 |
| **Product Usage** | **5** |
| asset_mgmt_flag | 1 |
| cross_sell_flag | 1 |
| insurance_flag | 1 |
| personal_loan_flag | 1 |
| total_products | 1 |
| **Repayment history** | **1** |
| payment_consistency_score | 1 |
| **Grand Total** | **9** |

# EDA: Exploration

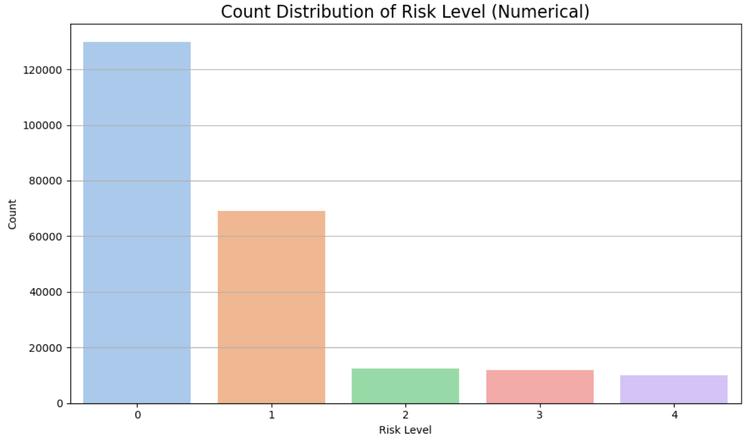
- Synthetic data shows around 12% of cross flag distribution in the dataset
- In the dataset, income to loan ratio is showing negative skew.
- Feature derived: Interaction intensity – aggregate of call, SMS and email intensity

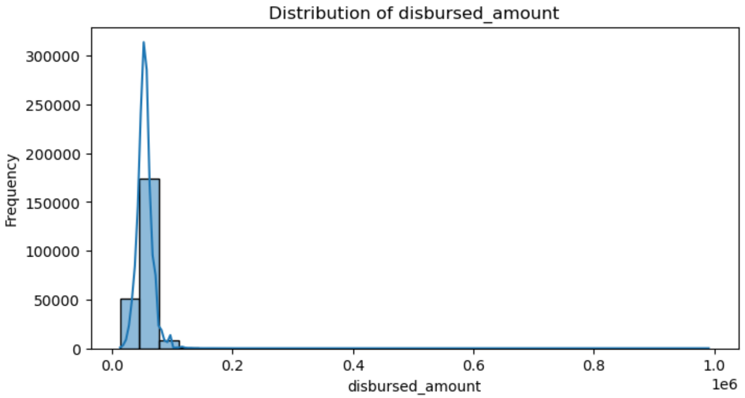


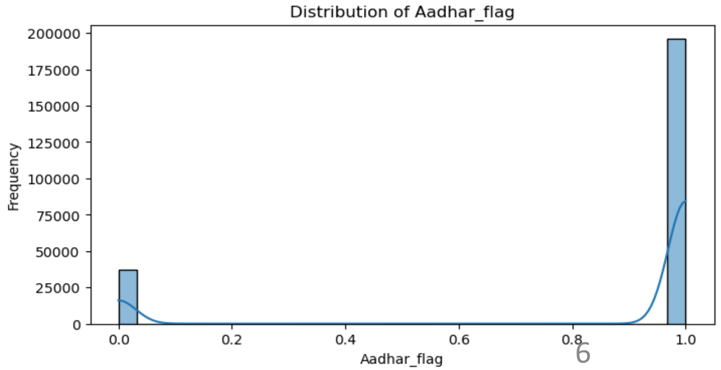

## Cleaning, Missing value handling and transformation

- Data format of date, ageing etc. features corrected.
- A risk level has been derived from risk score ranging from 0 to 4. ‘0’ indicating score not available, ‘4’ highest risk
- Some feature needed log transformation before usage as they were right skewed



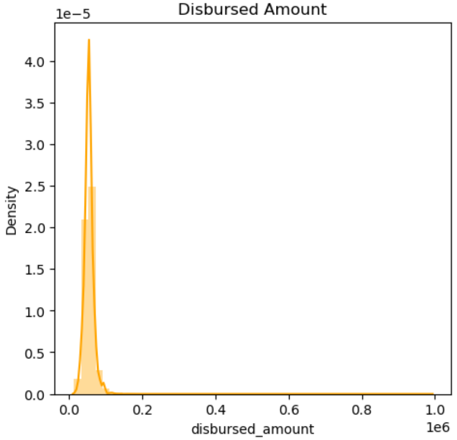



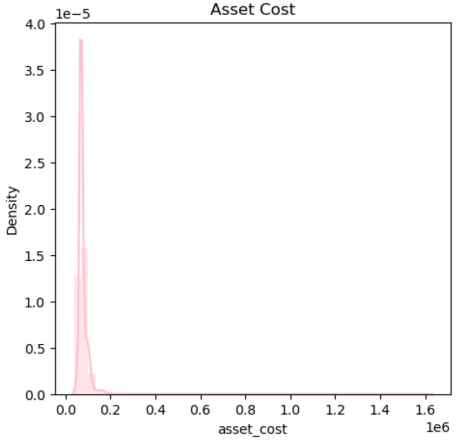

```python
risk_mapping = {
    "A-Very Low Risk": 1, "B-Very Low Risk": 1, "C-Very Low Risk": 1,
    "D-Very Low Risk": 1, "E-Low Risk": 1, "F-Low Risk": 1,
    "G-Low Risk": 1, "H-Medium Risk": 2, "I-Medium Risk": 2,
    "J-High Risk": 3, "K-High Risk": 3, "L-Very High Risk": 4,
    "M-Very High Risk": 4, "Not Scored: No Updates available in last 36 months": 0,
    "Not Scored: Sufficient History Not Available": 0,
    "Not Scored: More than 50 active Accounts found": 0,
    "Not Scored: No Activity seen on the customer (Inactive)": 0,
    "Not Scored: Only a Guarantor": 0, "No Bureau History Available": 0,
    "Not Scored: Not Enough Info available on the customer": 0
}
df['risk_level'] = df['PERFORM_CNS_SCORE_DESCRIPTION'].map(risk_mapping)
```



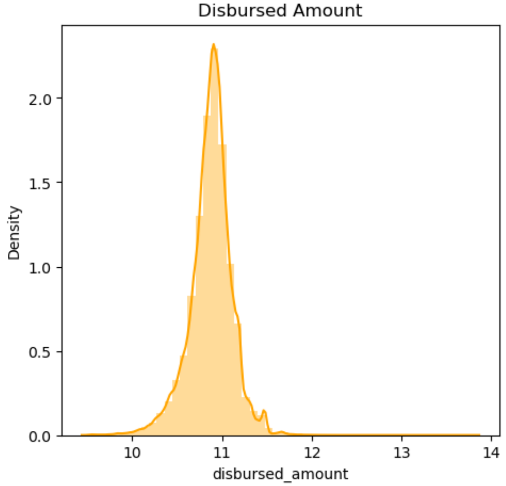
# EDA: Exploration contd..

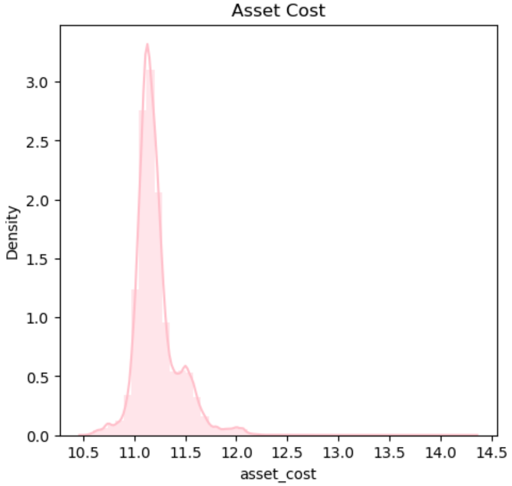
## Before Log transform

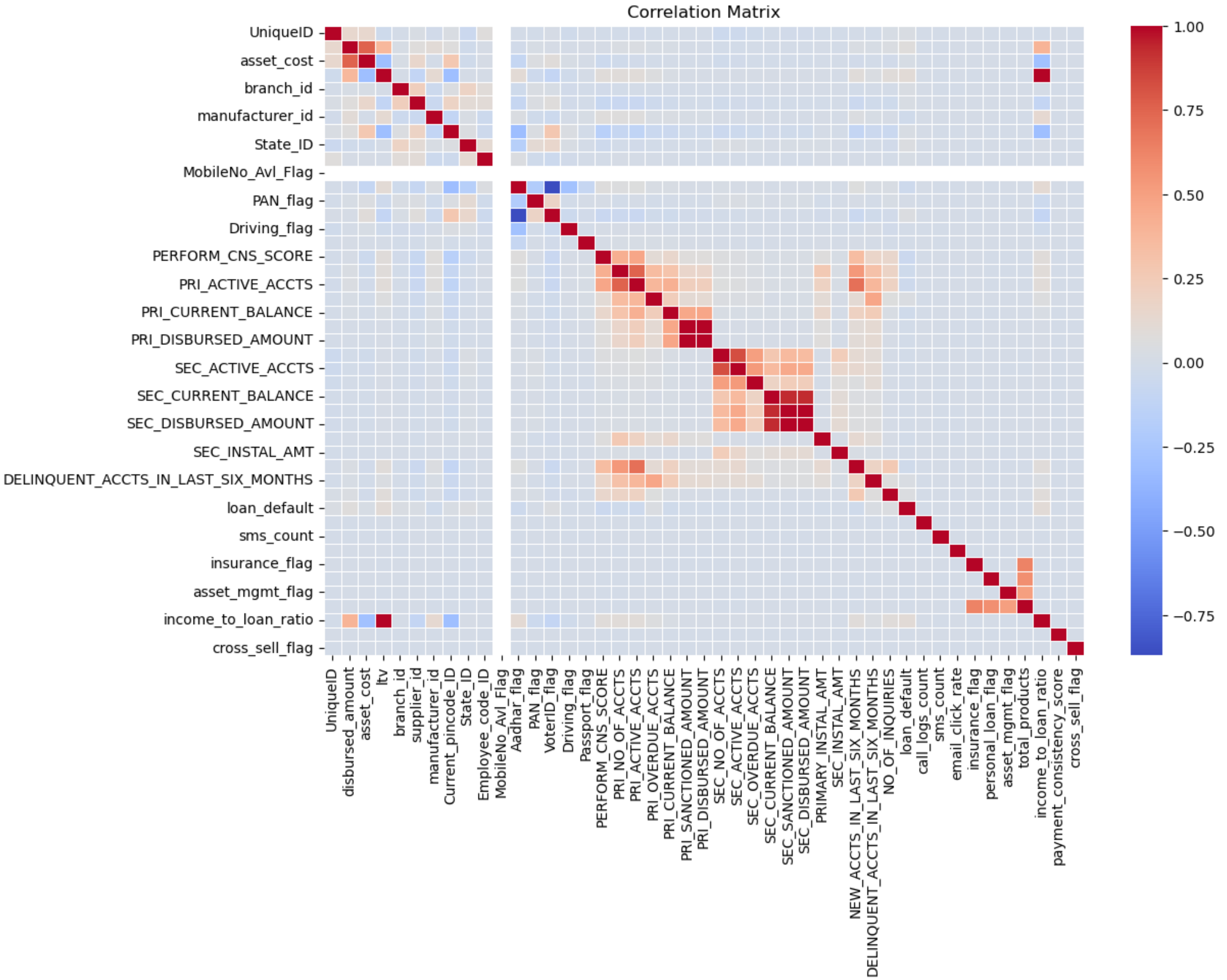



## After Log transform




❖ Some features are found to be correlated and hence dropped

# Feature importance and Model Performance

- As we have around 12% classes in the data with our target variable (cross sell flag), I have used SMOTE to balance the dataset for LR
- 3 models have been testes- Logistic Regression, Random Forests,  XGBoost model
- XGBoost is showing the best performance with highest balanced accuracy score
- Random Forest with Class Weights- ‘Balanced’ and XGBoost with ‘scale_pos_weight’ has been used to take care of class imbalance
- *The most important features  with high mean importance across models*
  - ✓ Age of the customer
  - ✓ loan to asset
  - ✓ Disbursed amount etc.

**Model Performance Comparison:**

| | Model | Accuracy | Balanced Accuracy |
|---|---|---|---|
| 0 | Logistic Regression | 0.522571 | 0.526317 |
| 1 | Random Forest | 0.820270 | 0.500067 |
| 2 | XGBoost | 0.566190 | 0.613952 |

```
Mean Feature Importances (Top 10):
age                          0.164204
asset_cost                   0.053434
ltv                          0.053114
disbursed_amount             0.051617
interaction_intensity        0.049940
sms_count                    0.048502
email_click_rate             0.048171
call_logs_count              0.046080
PRI_DISBURSED_AMOUNT         0.024139
CREDIT_HISTORY_LENGTH_MONTHS 0.023625
```

# Recommend Strategies for Effective Cross-Selling

| | Insights | Strategy | |
|---|---|---|---|
| **1** Target High Engagement Customers | Features like **interaction_intensity, email_click_rate, and call_logs_count** are among the top predictors of cross-sell likelihood | **Leverage interaction data to prioritize customers** with higher email click rates and call engagement. | Deploy **personalized marketing campaigns** for these segments. |
| **2** Focus on Young Customers (20-40) | Age has the highest importance in predicting cross-sell success. **Customers aged 20-40 show higher engagement** and responsiveness. | Design offerings tailored to younger demographics, such as **bundled insurance or personal loan discounts**. | Use **digital channels (e.g., targeted social media campaigns- Instagram, FB)** to connect with this segment. |
| **3** Incentivize Customers with Higher Loan-to-Value Ratios | Features like **ltv, asset_cost, and disbursed_amount** are strong predictors. Customers with higher loan-to-value ratios are likely to require additional products. | Offer **bundled packages (e.g., asset management services)** for customers with high-value loans. | Provide **tiered benefits** based on disbursed loan amounts. *(explained in next slide) |

# Recommend Strategies for Effective Cross-Selling contd..

| | Insights | Strategy | |
|---|---|---|---|
| **4 Utilize Credit History and Payment Data** | CREDIT_HISTORY_LENGTH_MONTHS and PRI_DISBURSED_AMOUNT **indicate financial reliability and cross-sell potential.** | Offer **additional products (e.g., insurance or personal loans) to customers with long credit histories** or higher prior disbursed amounts. | Use payment consistency metrics to identify low-risk customers for upselling. |

*Explanation on #3- Provide **tiered benefits** based on disbursed loan amounts.

**1. Low Loan Segment (₹0 - ₹50,000)**
**Challenge**: Customers with small loans may have less disposable income.
**Benefits**:
- Discounts on basic insurance plans (e.g., vehicle insurance at certain discount).
- Easy access to low-interest personal loans for small amounts.
- Cashback rewards for timely repayments.

**Example**: "Get 5% cashback on the first 3 personal loan installments when you take a loan above ₹20,000."

**2. Medium Loan Segment (₹50,000 - ₹200,000)**
**Challenge**: Customers in this range may be exploring additional financial products but need strong incentives.
**Benefits**:
- Bundled products (e.g., combine a personal loan with an investment or savings plan).
- Loyalty rewards for purchasing multiple products.
- Lower processing fees or preferential rates for personal loans or asset management services.

**Example**: "Apply for a personal loan above ₹100,000 and get vehicle insurance at a flat 20% discount."

**3. High Loan Segment (₹200,000 and Above)**
**Challenge**: High-value customers are likely to cross-sell but expect premium service and high-value benefits.
**Benefits**:
- Priority customer service and dedicated relationship managers.
- Deeply discounted financial products like health or life insurance.
- Exclusive offers, such as investment advisory or asset management services.
- Waivers on prepayment penalties or interest rate reductions for additional products.

**Example**: "Customers with loans above ₹300,000 receive free life insurance coverage for the first year“

# Assumptions

## Key Assumptions

**Synthetic Data Creation**:

- Features such as cross_sell_flag and additional behavioral metrics were synthetically generated to simulate real-world scenarios.
- Assumed the synthetic features closely represent actual customer behaviors and preferences.

**Feature Independence**:

- Synthetic data generation did not enforce interdependence between some features, potentially oversimplifying relationships.

**Model Generalization**:

- Models are built assuming the training data distribution matches unseen real-world data.

# What Could Have Been Done Better

1. **Access to Real Data**:
    1. Using real customer behavioral and transactional data could reveal authentic patterns and improve model robustness.
    2. Real data would enable accurate dependency modeling between features like loan tenure, income, and engagement.

2. **Advanced Feature Engineering**:
    1. Derive features from available data (e.g., ratios, trends, and lags in customer interactions).
    2. Incorporate domain-specific metrics like financial risk scores or customer loyalty indices.

3. **Enrich Data Sources**:
    1. Use external data such as credit bureau reports, demographic databases, or customer surveys for additional insights.

4. **Modeling Enhancements**:
    1. Experiment with advanced techniques like:
        1. Gradient boosting algorithms (e.g., LightGBM, CatBoost).
        2. Ensemble methods combining predictions from multiple models.
    2. An alternative approach to SMOTE could have been under-sampling the majority class to create a balanced dataset. While under-sampling avoids synthetic data, it may lead to information loss, making it suitable for large datasets with highly imbalanced classes.

5. **Customer Segmentation**:
    1. Perform clustering or segmentation analysis to group customers into actionable cohorts before cross-sell targeting.

6. **Long-Term Tracking**:
    1. Develop a framework for monitoring model predictions and validating outcomes over time.

# Thank you


- Partha Chowdhury
- Email: partha2v@gmail.com
- Linkedin: www.linkedin.com/in/parthach