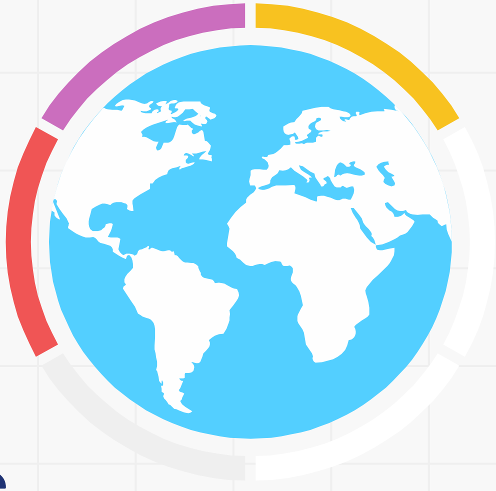
Ensemble Model Classification for Bank Customer Attrition Analysis Using Explainable AI (XAI)



By : Muhammad Haekal Akiyat

OBJECTIVE

- To extract meaningful insights from data and translate them into actionable strategies for informed decision-making.
- To make business recommendation base on classification model.

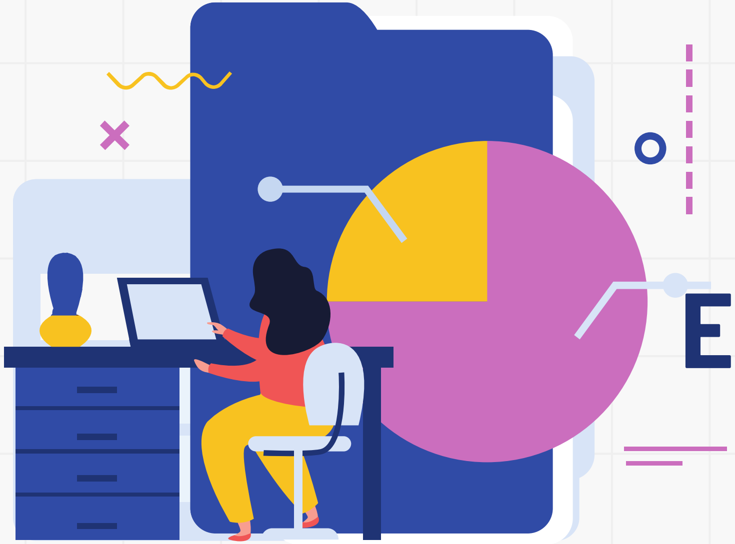
GOAL

- Create classification model using ensemble model (Extreme Gradient Boosting, Random Forest and Light Gradient Boosting)

Table of contents

Data Understanding

- **user_id:** customer account number.
- **attrition_flag:** customer status (Existing and Attrited).
- **customer_age:** age of the customer.
- **gender:** gender of customer (M for male and F for female).
- **dependent_count:** number of dependents of customers.
- **education_level:** customer education level (Uneducated, High School, Graduate, College, Post-Graduate, Doctorate, and Unknown).
- **marital_status:** customer's marital status (Single, Married, Divorced, and Unknown).
- **income_category:** customer income interval category (Less than $40K, $40K-$60k, $60K-$80K, $80K-$120K, $120K +, and Unknown).
- **card_category:** type of card used (Blue, Silver, Gold, and Platinum).
- **months_on_book:** period of being a customer (in months).
- **total_relationship_count:** the number of products used by customers in the bank.
- **months_inactive_12_mon:** period of inactivity for the last 12 months.
- **contacts_count_12_mon**: the number of interactions between the bank and the customer in the last 12 months.
- **credit_limit:** credit card transaction nominal limit in one period.
- **total_revolving_bal:** total funds used in one period.
- **avg_open_to_buy:** the difference between the credit limit set for the cardholder's account and the current balance.
- **total_amt_chng_q4_q1:** increase in customer transaction nominal between quarter 4 and quarter 1.
- **total_trans_amt:** total nominal transaction in the last 12 months.
- **total_trans_ct:** the number of transactions in the last 12 months.
- **total_ct_chng_q4_q1:** the number of customer transactions increased between quarter 4 and quarter 1.
- **avg_utilization_ratio:** percentage of credit card usage.



01

Exploratory Data Analysis

Descriptive Analysis, Univariate Analysis,
Multi Variate Analysis

Descriptive Analysis

Data Types

12 Categorical
8 numerical

Missing Value

0 missing value

Duplicated Rows

0 duplicated rows

Numerical Columns statistics

	count	mean	std	min	25%	50%	75%	max
customer_age	10127.0	46.325960	8.016814	26.0	41.000	46.000	52.000	73.000
months_on_book	10127.0	35.928409	7.986416	13.0	31.000	36.000	40.000	56.000
credit_limit	10127.0	8631.953698	9088.776650	1438.3	2555.000	4549.000	11067.500	34516.000
total_revolving_bal	10127.0	1162.814061	814.987335	0.0	359.000	1276.000	1784.000	2517.000
avg_open_to_buy	10127.0	7469.139637	9090.685324	3.0	1324.500	3474.000	9859.000	34516.000
total_amt_chng_q4_q1	10127.0	0.759941	0.219207	0.0	0.631	0.736	0.859	3.397
total_trans_amt	10127.0	4404.086304	3397.129254	510.0	2155.500	3899.000	4741.000	18484.000
total_trans_ct	10127.0	64.858695	23.472570	10.0	45.000	67.000	81.000	139.000
total_ct_chng_q4_q1	10127.0	0.712222	0.238086	0.0	0.582	0.702	0.818	3.714
avg_utilization_ratio	10127.0	0.274894	0.275691	0.0	0.023	0.176	0.503	0.999

Outliers potential :

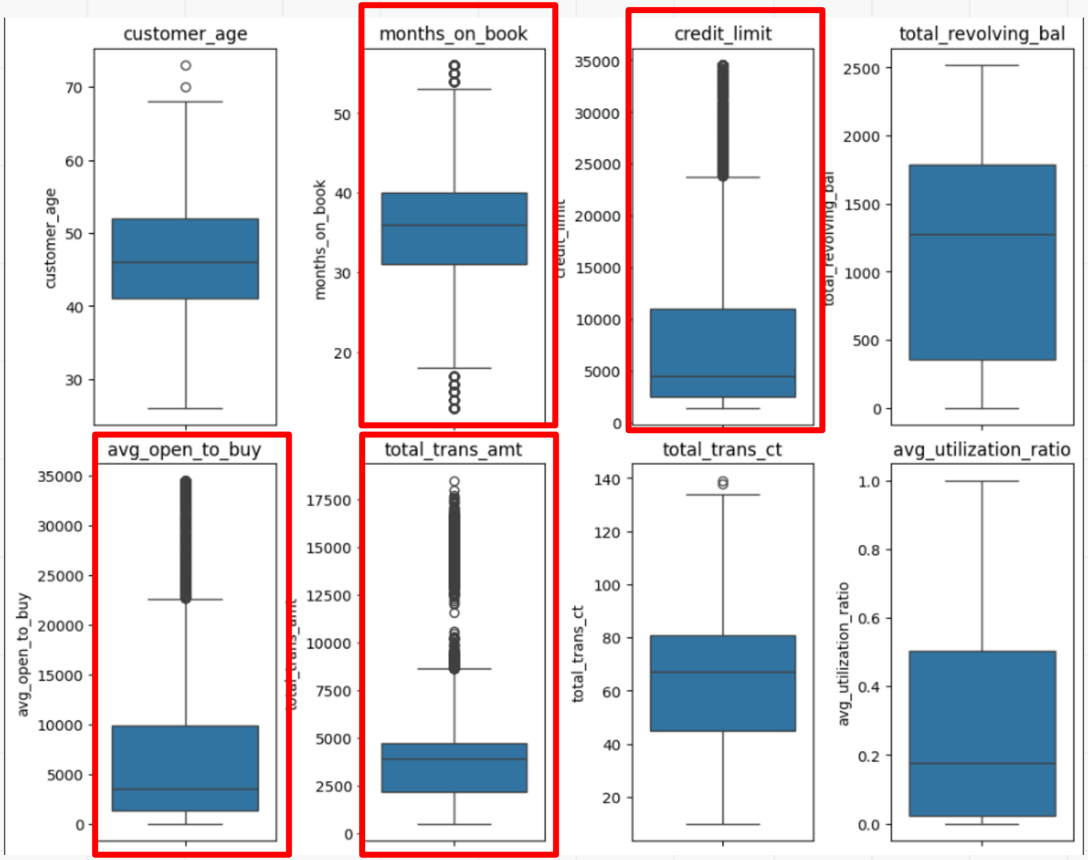
avg_utilization_ratio, avg_open_to_buy, total_trans_amt, credit_limit, months_on_book. Just by looking at the data distribution

Categorical Columns statistics

	count	unique	top	freq
attrition_flag	10127	2	Existing Customer	8500
gender	10127	2	F	5358
dependent_count	10127	6	3	2732
education_level	10127	7	Graduate	3128
marital_status	10127	4	Married	4687
income_category	10127	6	Less than $40K	3561
card_category	10127	4	Blue	9436
total_relationship_count	10127	6	3	2305
months_inactive_12_mon	10127	7	3	3846
contacts_count_12_mon	10127	7	3	3380

Unique counts for categorical feature seems fair with less than 10 number of unique value

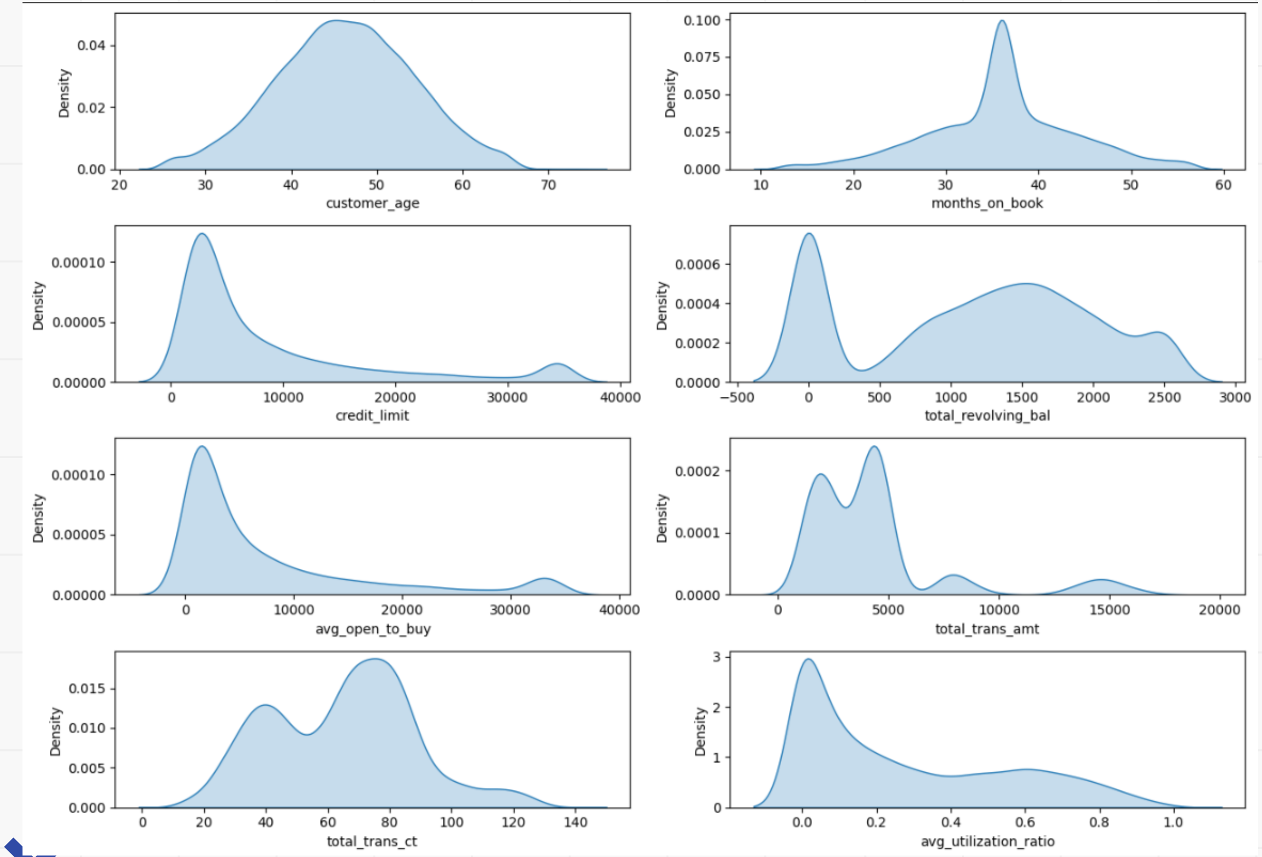
Univariate Analysis (Box plot)



Key takes:

- **avg_utilization_ratio, avg_open_to_buy, total_trans_amt, credit_limit, months_on_book** features have outliers
- Columns with outliers have a right-skewed distribution because they have values that are much larger than the mean

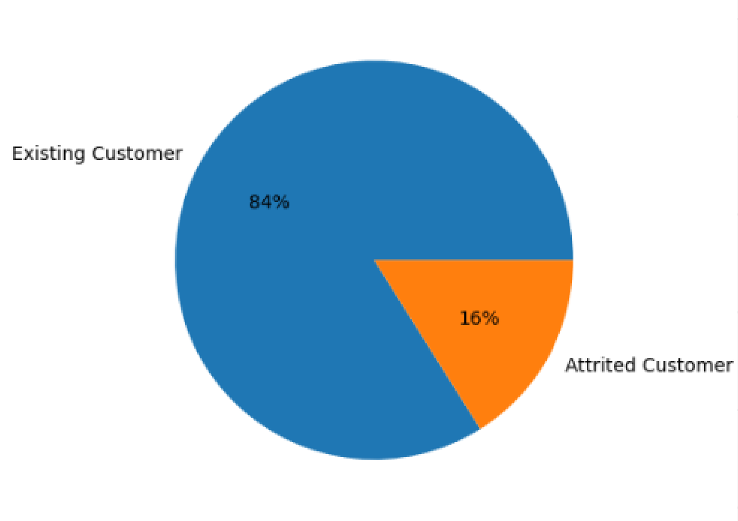
Univariate Analysis (Displot)

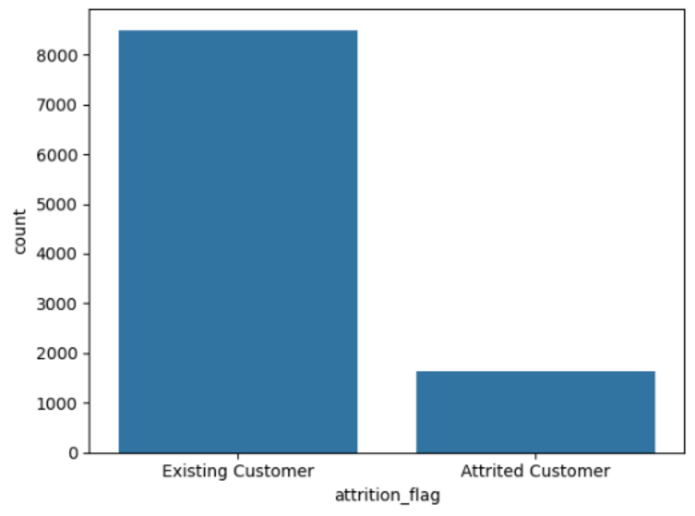


Key takes:

- **avg_utilization_ratio, avg_open_to_buy, total_trans_amt, credit_limit, months_on_book** features have outliers
- Columns with outliers have a right-skewed distribution because they have values that are much larger than the mean

Univariate Analysis (Pie)

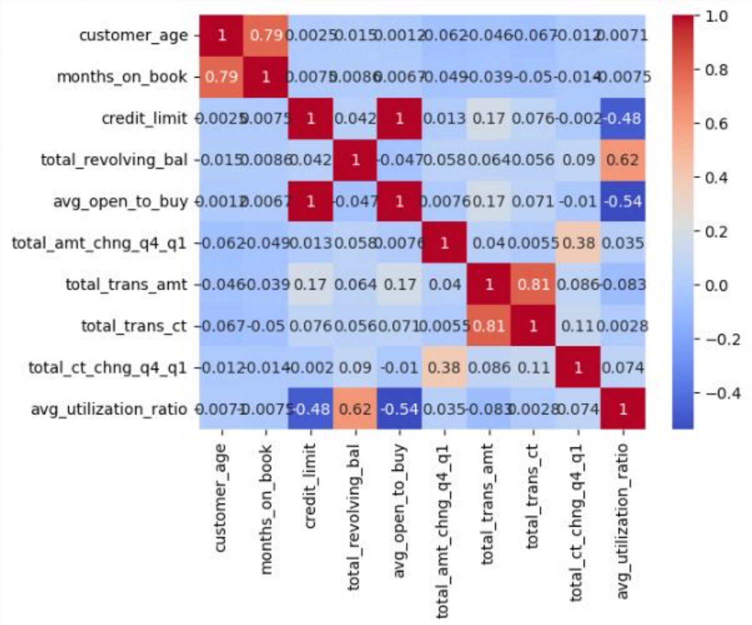




Key takes:

- Data target unbalance, mostly the customer existed and loyal.

Multivariate Analysis



Key takes:

- **month_on_book** has high positive correlation with **customer_age**
- **total_trans_amt** has high positive correlation with **total_trans_ct**
- **total_revolving_bal** has high positive correlation with **avg_utilization_ratio**
- **avg_open_to_buy** has high negative correlation with **avg_utilization_ratio**

02

Data Pre-processing

Missing value & outlier handling, feature engineering, Standardization

Missing value and Duplicated Rows

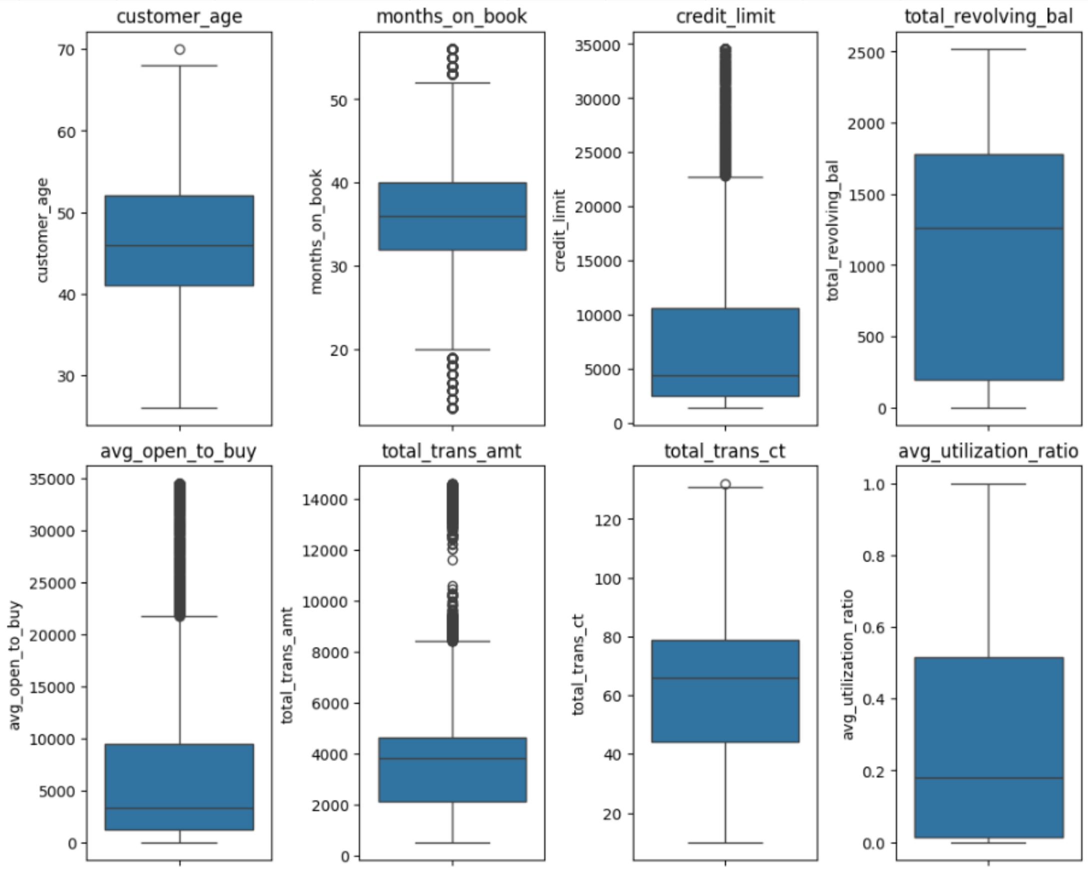
Missing Value

0 missing value

Duplicated Rows

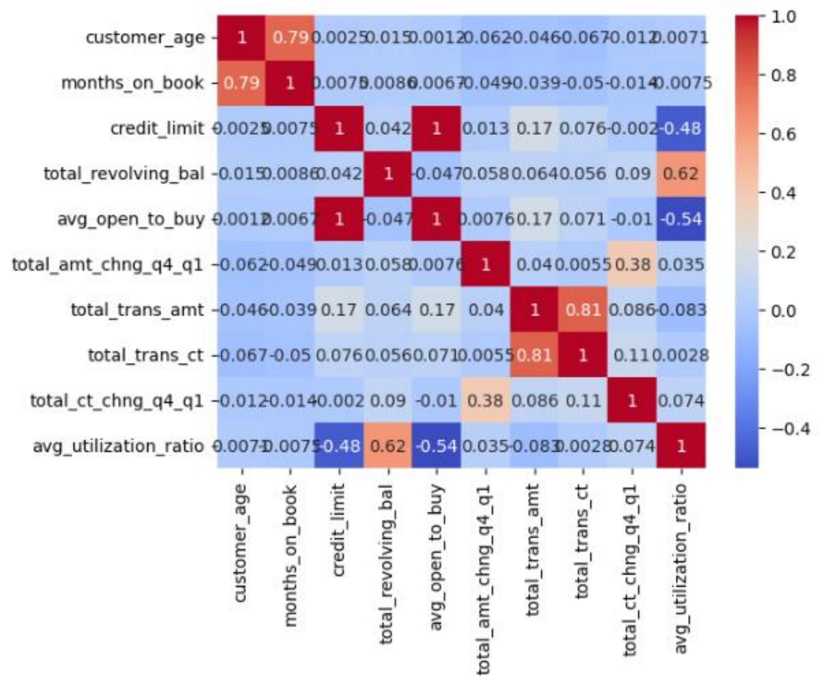
0 duplicated rows

Outlier Handling With Z-Score



Remove outliers using z-score

Feature Engineering (redundant feature)



	feature	vif_score
3	credit_limit	inf
4	total_revolving_bal	inf
5	avg_open_to_buy	inf
10	avg_utilization_ratio	2.910105
8	total_trans_ct	2.810709
7	total_trans_amt	2.773756
1	customer_age	2.591489
2	months_on_book	2.584340
9	total_ct_chng_q4_q1	1.181157
6	total_amt_chng_q4_q1	1.115876

Key takes:

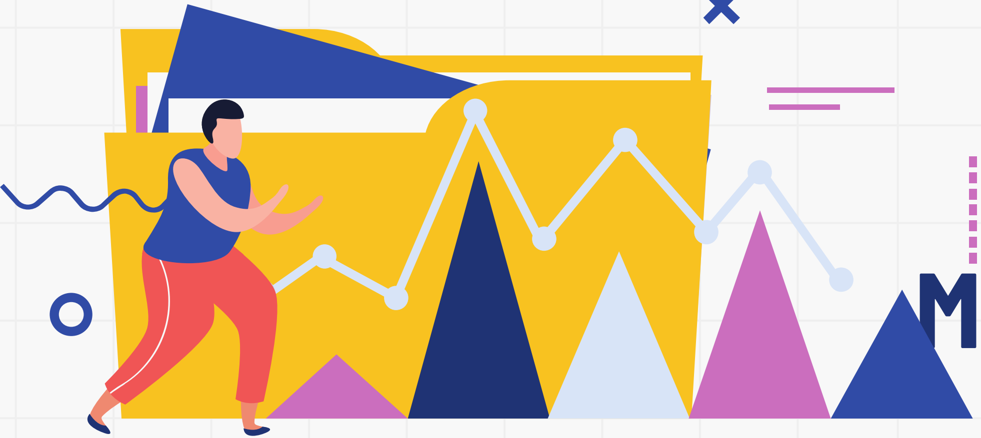
- **credit_limit, total_revolving_bal,avg_open_to_buy** are features with perfectly corelated with another features with correlation score = 1, besides with their own features
- **Remove redundant column**

We remove **credit_limit, total_revolving_bal,avg_open_to_buy** for this analysis.

Feature Engineering (Standardization)

Using ensemble model which build by tree like random forest, XGB, LGBM does not necessary to scale your data. But it is necessary to scale your data if you decide model with distance based like k-means, KNN.

So in this analysis, we were not scaling our data. Just let it be



03

Modelling & Evaluation

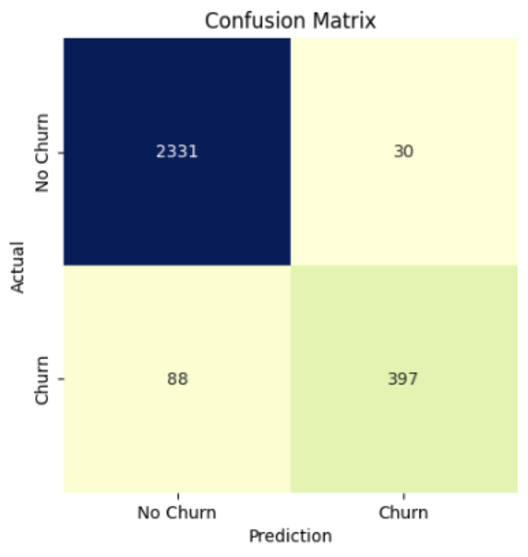
Modeling using Random Forest Classifier, XGboost, LightGBM

Random Forest Classifier

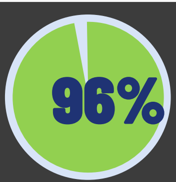
Classification Report

```
RANDOM_FOREST
              precision    recall  f1-score   support

           0       0.96      0.99      0.98      2361
           1       0.93      0.82      0.87       485

    accuracy                           0.96      2846
   macro avg       0.95      0.90      0.92      2846
weighted avg       0.96      0.96      0.96      2846
```

Confusion Matrix



Random Forest Accuracy



Feature Importance

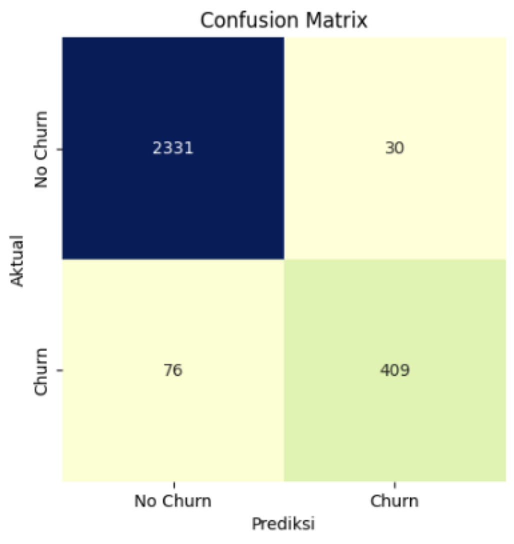
	Feature	Importance
12	total_trans_amt	0.210142
13	total_trans_ct	0.193668
14	total_ct_chng_q4_q1	0.126892
15	avg_utilization_ratio	0.120018
8	total_relationship_count	0.078774
11	total_amt_chng_q4_q1	0.066413
0	customer_age	0.043584
7	months_on_book	0.031163
9	months_inactive_12_mon	0.030556
10	contacts_count_12_mon	0.026377
2	dependent_count	0.015738
3	education_level	0.014140
5	income_category	0.013775
1	gender	0.013478
4	marital_status	0.012100
6	card_category	0.003181

Classification Report

LIGHT_GBM

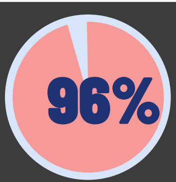
	precision	recall	f1-score	support
0	0.97	0.99	0.98	2361
1	0.93	0.84	0.89	485
accuracy			0.96	2846
macro avg	0.95	0.92	0.93	2846
weighted avg	0.96	0.96	0.96	2846

Confusion Matrix



LGBM Classifier

Random Forest Accuracy



Feature Importance

	Feature	Importance
12	total_trans_amt	694
13	total_trans_ct	383
11	total_amt_chng_q4_q1	351
14	total_ct_chng_q4_q1	292
15	avg_utilization_ratio	237
0	customer_age	218
8	total_relationship_count	164
7	months_on_book	159
9	months_inactive_12_mon	137
10	contacts_count_12_mon	119
4	marital_status	54
3	education_level	50
2	dependent_count	49
1	gender	37
5	income_category	37
6	card_category	19

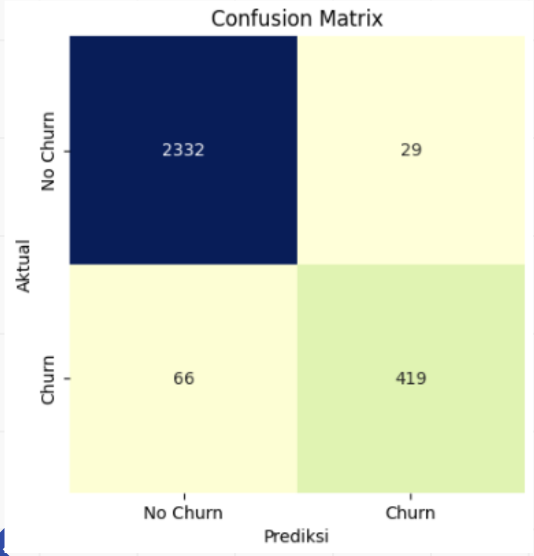
XGBOOST Classifier

Classification Report

XGBOOST

	precision	recall	f1-score	support
0	0.97	0.99	0.98	2361
1	0.94	0.86	0.90	485
accuracy			0.97	2846
macro avg	0.95	0.93	0.94	2846
weighted avg	0.97	0.97	0.97	2846

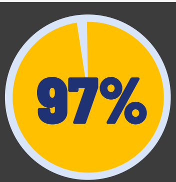
Confusion Matrix



Feature Importance

	Feature	Importance
13	total_trans_ct	0.250166
8	total_relationship_count	0.164546
15	avg_utilization_ratio	0.125956
12	total_trans_amt	0.076421
1	gender	0.064924
9	months_inactive_12_mon	0.060954
14	total_ct_chng_q4_q1	0.059463
0	customer_age	0.043158
11	total_amt_chng_q4_q1	0.027817
10	contacts_count_12_mon	0.027420
6	card_category	0.021734
2	dependent_count	0.021119
4	marital_status	0.018271
7	months_on_book	0.014226
3	education_level	0.012510
5	income_category	0.011317

XGBOOST Accuracy

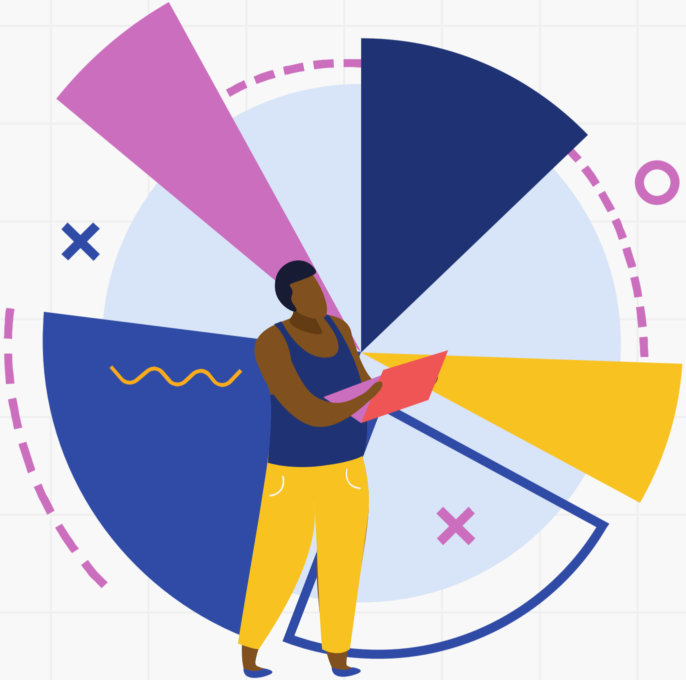


Model Evaluation Summary

Note : we're focusing on class 1, customer churn.

	Precision	Recall	F1-Score	Accuracy
Random Forest	93%	82%	87%	96%
LightGBM	93%	84%	89%	96%
XGBOOST	94%	86%	90%	97%

XGBOOST has better performance than other models. But good performance does not always have a good reasonable feature. We will use explainable ai to interpret the model.



04

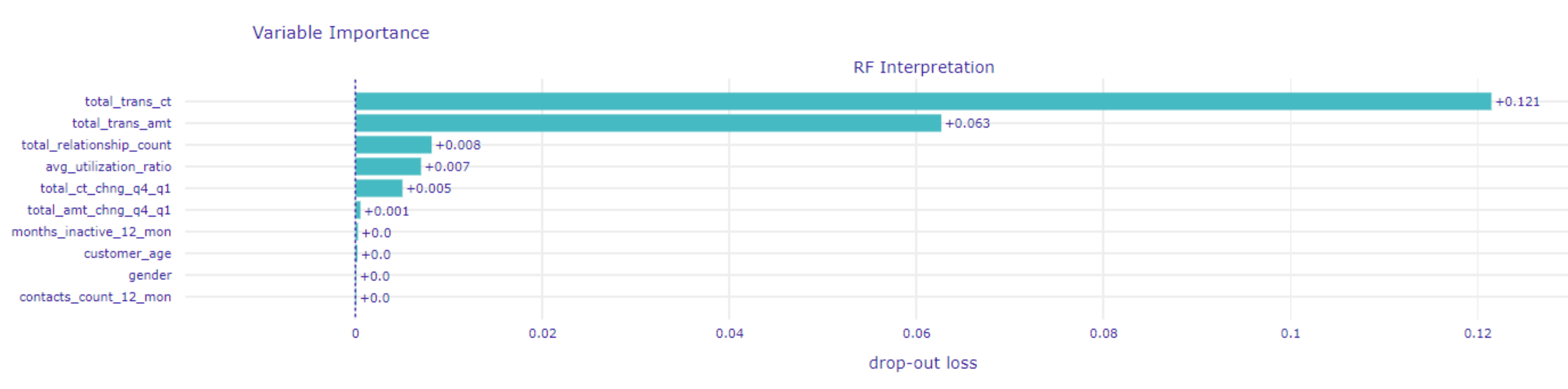
Explanation AI

Explanation AI using Permutation Feature Importance

Permutable Feature Importance

Random Forest Classifier

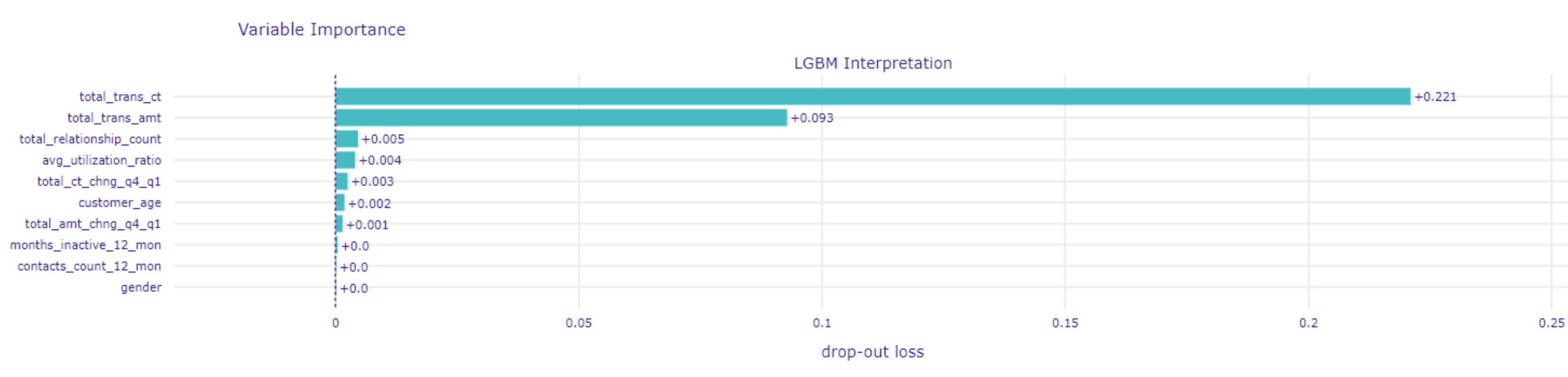
The features are ranked based on their importance to the model. The importance reflects how much each feature contributes to predicting customer churn. A higher value indicates a more significant impact on the prediction



total_trans_ct(Total Transactions Count)is the most influential feature (+0.121), meaning it has the highest impact on the model's prediction. total_trans_amt (Total Transaction Amount) is the second most important feature (+0.063). Other feature have lower importance

Permutable Feature Importance

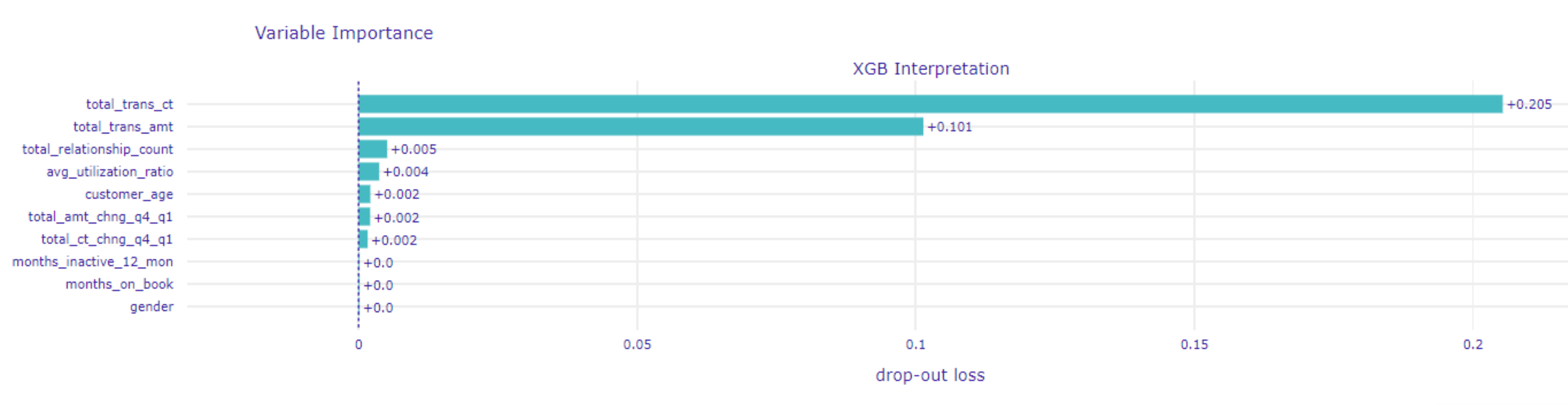
LGBM Classifier



total_trans_ct(Total Transactions Count)is the most influential feature (+0.221), meaning it has the highest impact on the model's prediction. total_trans_amt (Total Transaction Amount) is the second most important feature (+0.093). Other feature have lower importance

Permutable Feature Importance

XGBOOST Classifier



total_trans_ct(Total Transactions Count)is the most influential feature (+0.205), meaning it has the highest impact on the model's prediction. total_trans_amt (Total Transaction Amount) is the second most important feature (+0.101). Other feature have lower importance

05

Analysis and Recommendation

Actionable Insight and Business Recommendation

Model Analysis

XGBOOST models have superior performance compared to other models. This is proven by the better precision, recall, f-1 score and accuracy metric values. In its interpretation, the XGBOOST model takes into account 7 features which are considered quite important with drop loss above +0.002 which indicates model good enough to generalize when applied to new or unseen data.

XGBOOST

- Accuracy (overall prediction correctness) = This model can predict overall 97% correct Customer who existed and churn.
- Precision (proportion of predicted churns that are correct) = This model can predict 94% correct Customer who existed and churn when observed.
- Recall/Sensitivity (proportion of actual churns correctly identified) = This model can predict 86% Customer who are actually churned.
- F1-Score (balance of precision and recall) = F-1 Score 90%, The harmony between recall and precision, while considering False Positive and False Negative

Business Impact

- **Customer Behavior:**

Customers with fewer transactions (total_trans_ct) or lower transaction amounts (total_trans_amt) are likely at a higher risk of churning. These two features indicate customer engagement and financial activity. Focusing on these metrics allows businesses to monitor customer health effectively.

- **Resource Allocation:**

Marketing and retention efforts should be directed toward customers showing a decline in transaction count and transaction amounts. Features such as avg_utilization_ratio (credit utilization) or relationship_count (interactions with the business) can guide tailored interventions.

Business Recommendation

- Improve Engagement

For customers with low total_trans_ct, Offer personalized rewards or loyalty programs to incentivize more frequent transactions. For customers with low total_trans_amt, Introduce tiered pricing or discounts for higher spending.

- Monitor Trends:

Continuously track total_ct_chng_q4_q1 (total transaction between quarters) and identify customers showing declining trends. Proactively reach out to these customers.

- Segmented Communication

Use total_relationship_count to identify customers with weaker ties to the business. Deploy targeted campaigns, such as one-on-one consultations or priority services, to strengthen the relationship.

- Simplify Monitoring

Since features like gender and contacts_count_12_mon are of negligible importance, you might focus data collection efforts on more impactful metrics.

Thanks!

Do you have any questions?

mhaekaal73@gmail.com

https://www.linkedin.com/in/mhaekalakiyat/