Santander Customer Satisfaction

Utilizing :
General Linear Models

By: Snigda Gedela

Santander

Agenda

1. Problem Statement
2. Literature Review
3. Data Description
4. Data Preprocessing - Feature Engineering
5. Data Preprocessing - Exploratory Data Analysis (EDA)
6. Methodology
7. Result
8. Recommendations

Problem Statement

1 The financial impact of losing a customer would be significant, particularly for a large bank like Santander.

2 Acquiring a new customer can cost 5 to 25 times more than retaining an existing one (Estimated)

Overall Goal

1 Target resources and efforts towards the customers who are most at risk of leaving

2 Increase the revenue by retaining the loyal customers.

Customer satisfaction is important -

- Helps in reducing negative word of mouth
- Increases intention of repurchasing
- Cost of acquisition of new customer is very high compared to retaining the loyal customers.

Literature Review

Predicting Customer's Satisfaction (Dissatisfaction) Using Logistic Regression
Evaluating and predicting whether the existing/new customer is satisfied/ dissatisfied from their current offering.

This project also looks into the features and compare the similarities.

Machine-Learning Techniques for Customer Retention: A Comparative Study
Identify customers who may have negative experiences or are at risk of leaving, allowing the business to take proactive steps to improve their experience and retain their business.

Use of machine learning techniques to build predictive models.

Predicting Employee Attrition Using Machine Learning Techniques
Use of machine learning techniques to predict which employees are most likely to leave the company.

Require feature engineering and careful selection of input features to achieve the best performance.

Data Description

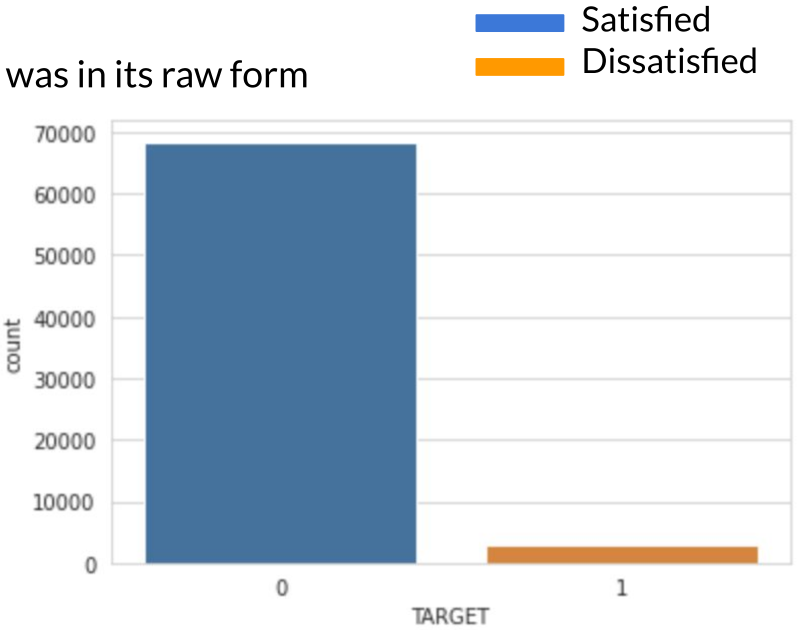
The data we have collected is the Company data , so the data was in its raw form

01 | Number of records - ~76K records, ~370 features

02 | Target Variable - 0 (Satisfied), 1 (Dissatisfied)

03 | The data is highly Sparse.

04 | 56 Categorical Features



Feature Engineering

The data was not in good shape and required good amount of feature engineering before proceeding with modeling

01 | Duplicate records -~ 4800 rows dropped

02 | Constant value features(34) and Duplicate features(29) dropped

03 | Class Imbalance - 97% vs 3% - highly imbalanced (SMOTE)

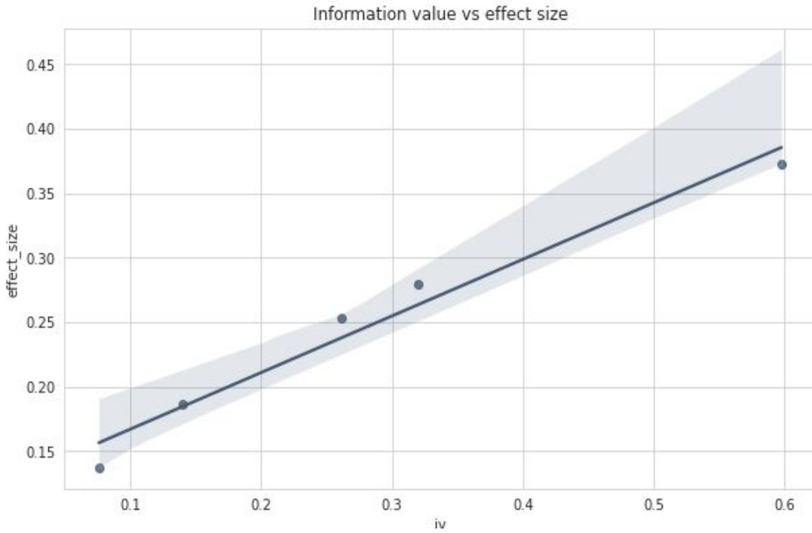
04 | Correlated features > 95% correlated dropped (pearson method)

05 | Standardization - Scaled data

06 | Sparsity - 23 columns dropped having > 99% as 0 observations

07 | Feature Importance-Random Forest -Top 5

08 | Feature Importance - WoE , Information Value (to conform previous results) - RF performed better



Strong relationship between IV and effect size.
Features with high IV have high effect size as well.
Correlation coefficient: 0.98 (Pearson) and 1.0 (Spearman)
Values closer to 0 imply very weak (or lack of) relationship, while higher values suggest stronger relation

IV Interpretation

IV on all variables

IV Interpretation	Number of features
useless	204
weak	9
medium	6
strong	1
suspicious	1

IV on top 5 variables

IV Interpretation	Number of features
weak	1
medium	2
strong	1
suspicious	1

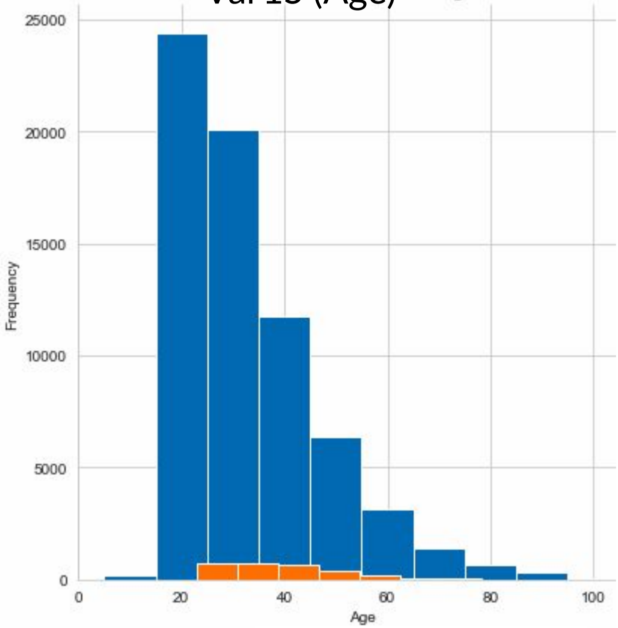
Variables	iv	p-value	effect_size	iv_interpretation	es_interpretation
var15	0.597443	0	0.372537	suspicious	medium
saldo_medio_var5_ult3	0.31962	0	0.278887	strong	medium
var36	0.298604	0	0.269802	medium	medium
num_meses_var5_ult3	0.288765	0	0.265311	medium	medium
saldo_medio_var5_hace3	0.261559	0	0.252906	medium	medium
...	...	...	...	...	...
num_op_var40_ult1	0	1	0	useless	useless
num_op_var41_hace3	0	1	0	useless	useless
num_op_var41_ult1	0	1	0	useless	useless

Santander

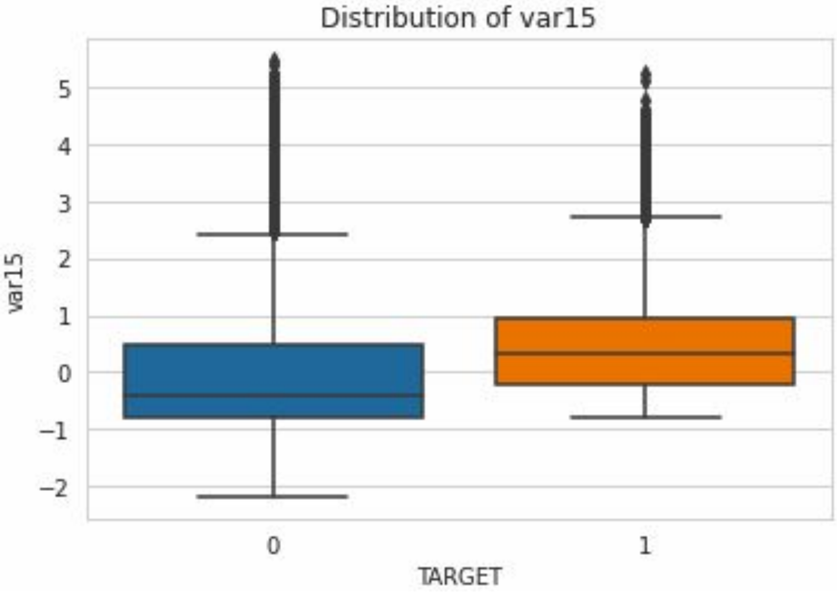
Exploratory Data Analysis

0: Satisfied
1: Dissatisfied

Var15 (Age)



Observation: The most of the younger people(<23) are satisfied

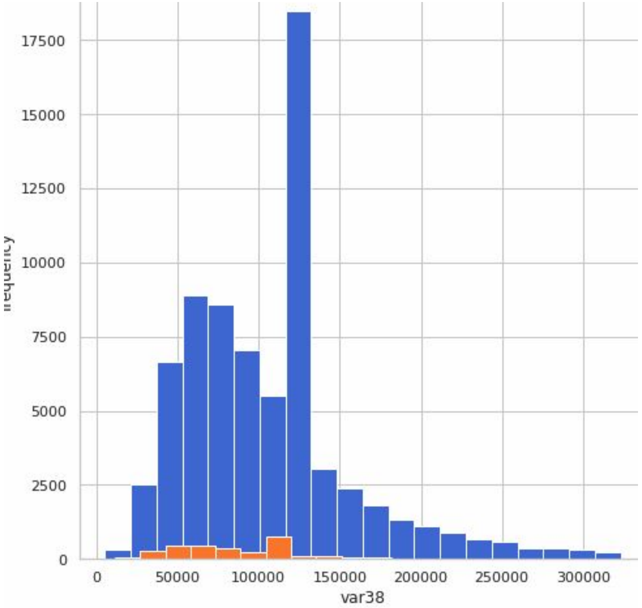


Observation: The distribution is almost similar for the two target classes

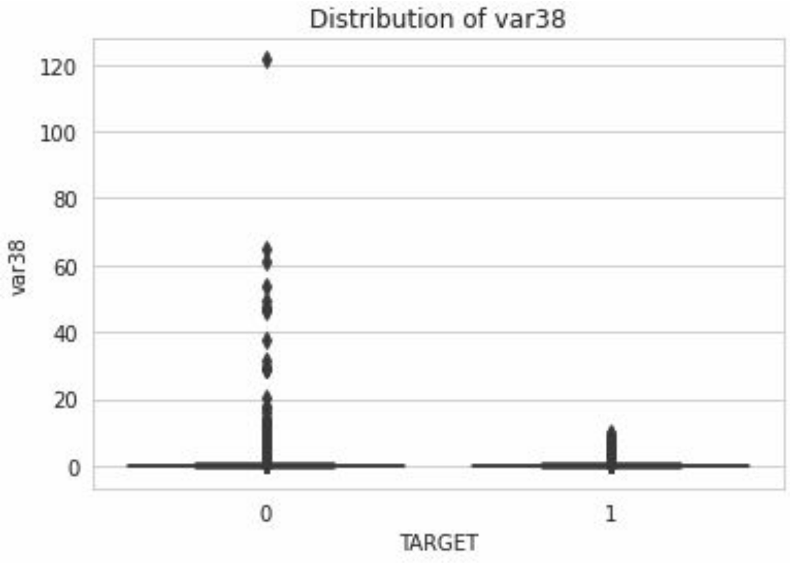
Exploratory Data Analysis

0: Satisfied
1: Dissatisfied

Var38 (Mortgage Value)



Observation: There is an outlandish peak between 100,000 and 150,000



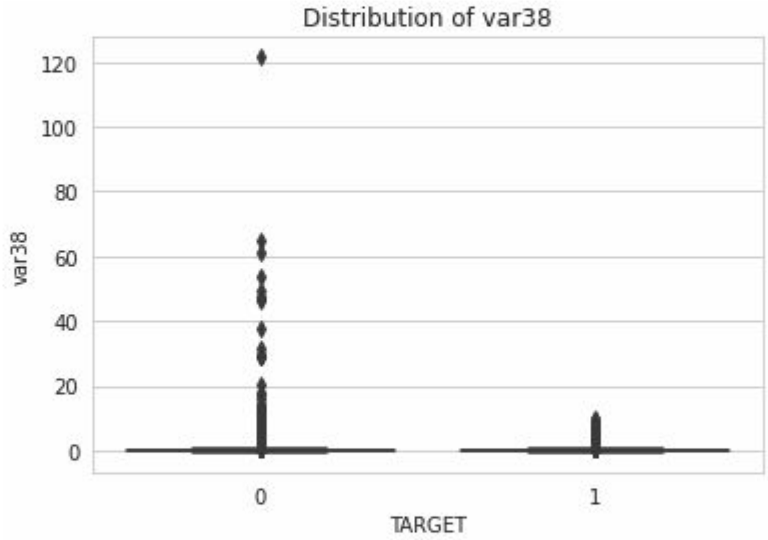
Observation: The outliers are much dense in case of satisfied customers than unsatisfied customers.

Santander

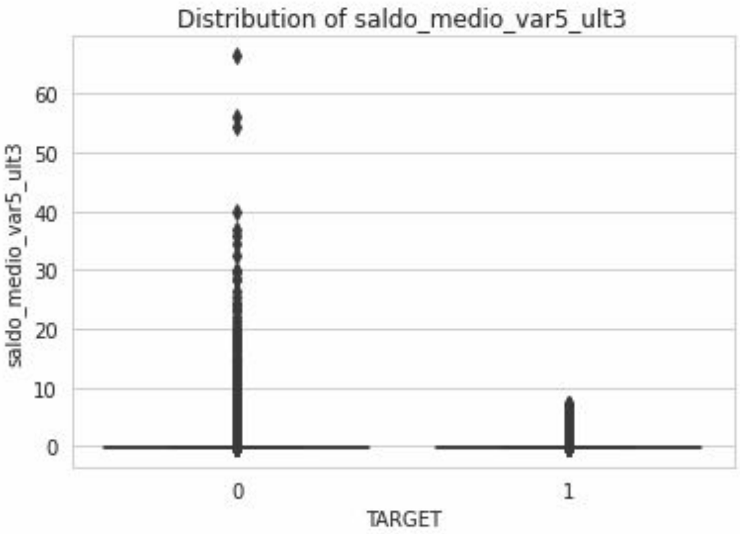
Exploratory Data Analysis

0: Satisfied
1: Dissatisfied

Var38 (Mortgage Value)



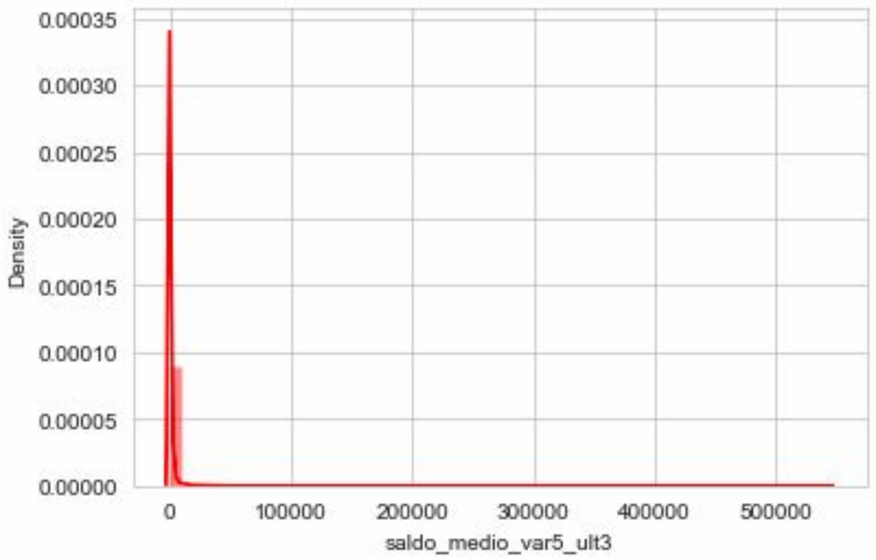
saldo_medio_var5_ult3



Observation: In case of both "var38" and "saldo_medio_var5_ult3" the distribution lies mostly in the outlier region

Exploratory Data Analysis

saldo_medio_var5_ult3



Percentage value counts(top 5 only) in the data for 'saldo_medio_var5_ult3':

Value	Count%
0.00	30.835662
2.88	1.391600
2.34	1.273644
2.85	1.213262
2.07	1.202028

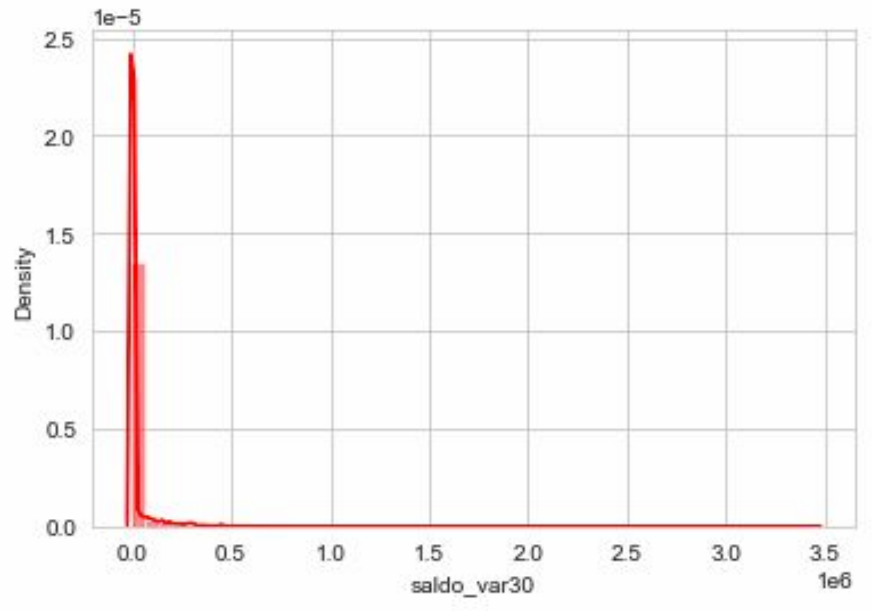
Percentage value counts(bottom 5 only) in the data for 'saldo_medio_var5_ult3':

Value	Count%
516.36	0.001404
1229.40	0.001404
82.35	0.001404
1750.17	0.001404
1025.37	0.001404

Observation: ~30% of the "saldo_medio_var5_ult3" feature has a value of 0.

Exploratory Data Analysis

saldo_var30



Percentage value counts(top 5 only) in the data for 'saldo_var30':

Value	Count%
0.0	24.725823
3.0	23.318776
90.0	6.868128
30.0	2.222909
15.0	2.023507

Percentage value counts(bottom 5 only) in the data for 'saldo_var30':

Value	Count%
1107.75	0.001404
31681.80	0.001404
581.61	0.001404
30276.54	0.001404
48191.22	0.001404

Observation: ~25% of the "saldo_var30" feature has a value of 0.

- 23% of the "saldo_var30" feature has a value of 3.

Santander

Methodology

Why Logistic Regression ?

- Logistic regression is a statistical technique used for binary classification problems, where the outcome of interest is binary, such as yes/no, true/false, or 0/1.

- Especially important in the **banking industry,** where interpretability and transparency are highly valued.

- **Computationally efficient** and can handle both categorical and continuous predictor variables.

- The output of logistic regression is a probability score between 0 and 1 that tells us how likely the outcome is to occur.
  Also helps us understand which independent variables are most important in predicting the outcome.

Methodology

1

The adopted approach for modeling customer satisfaction takes a dual path: one employs imbalanced data, while the other utilizes balanced data.

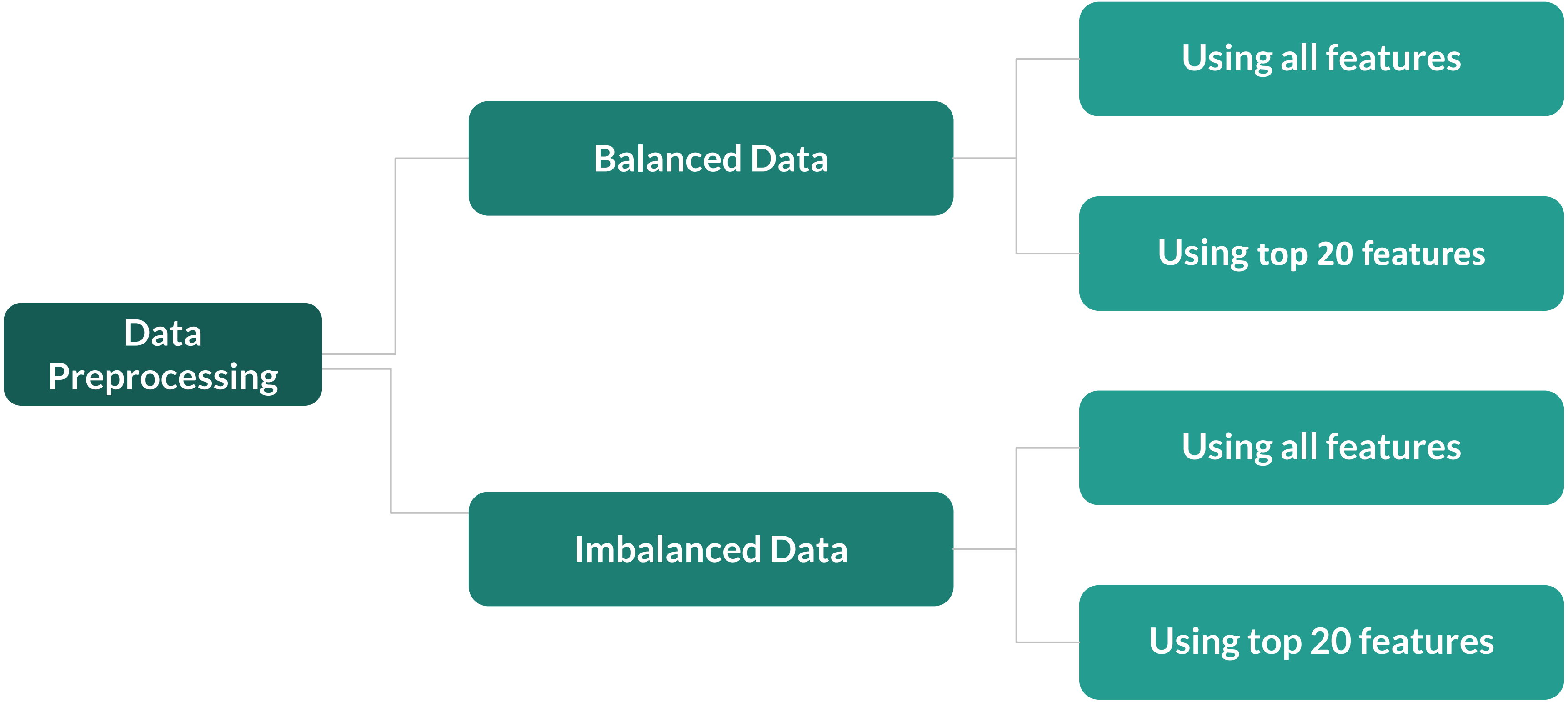
2

In the context of addressing imbalanced and balanced data, an analysis is conducted on both the complete dataset and the leading 20 features.

3

A comprehensive analysis of all four models was undertaken to avoid overlooking crucial features and to validate the findings.

Methodology

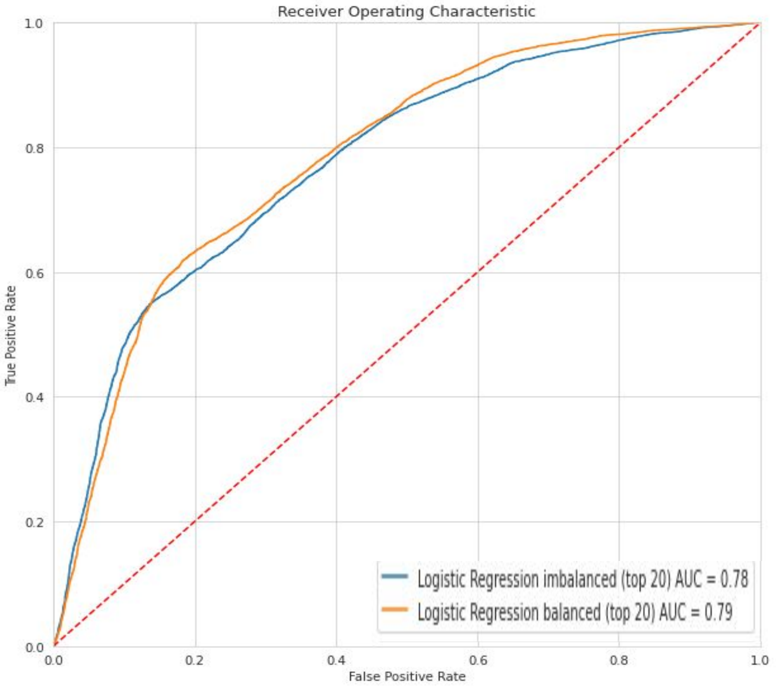


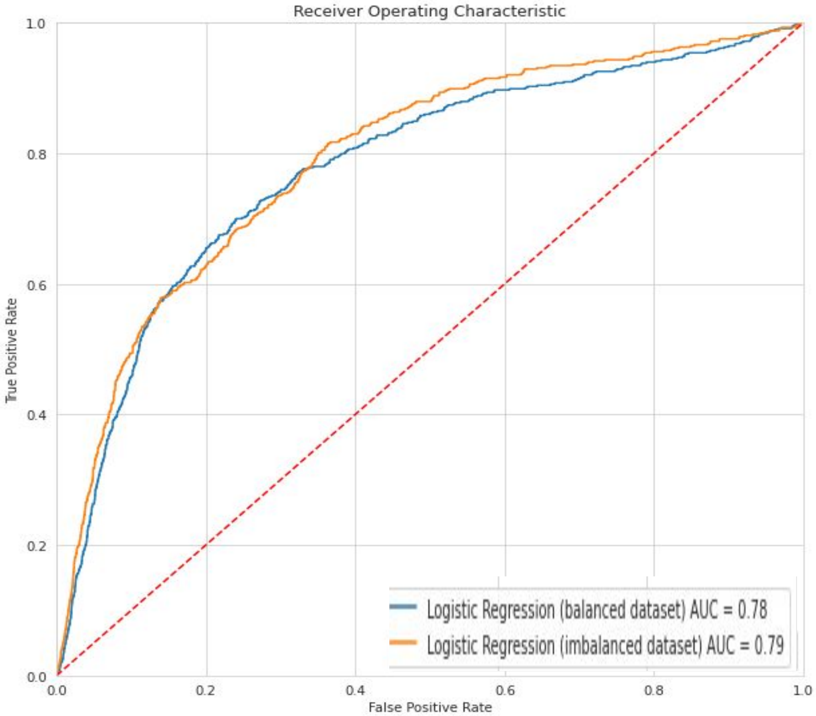
Results

Strategy	Accuracy	F1 Score	Recall	Precision	AUC score
Logistic Regression - with class imbalance dataset	96%	1%	0.5%	1%	79%
Logistic Regression - with class balanced dataset (SMOTE)	72%	17%	73%	2%	78%
Logistic Regression - with top 20 features (imbalanced)	69%	15%	71%	8.9%	79%
Logistic Regression - with top 20 features (balanced)	70%	71%	72%	71%	78%

- It can be concluded that accuracy is not the most reliable parameter to assess the model's performance. This is because accuracy does not consider the presence of class imbalances.
- More appropriate metrics to evaluate the model's performance are the f1-score, AUC score and Precision.
- By looking at these metrics, it is evident that the Logistic Regression performs well **when class is balanced and important features are considered in the modeling**

ROC Curve


Receiver Operating Characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression imbalanced (top 20) AUC = 0.78
Logistic Regression balanced (top 20) AUC = 0.79


Receiver Operating Characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression (balanced dataset) AUC = 0.78
Logistic Regression (imbalanced dataset) AUC = 0.79

Recommendations

- Improve model evaluation metric using Random Forest and XGboost classifier
- Prioritize focus on strengths & work on weakness identified using IV and feature importance results
- Identify causal factors behind the high impact predictors using RCA and use to improve customer satisfaction
- Use methods like PCA, SHAP or LIME on anonymized features for Feature Selection

Thank you