Assignment 1

Data Analysis

Dataset Overview

The dataset contains 2,240 customers with 29 features covering demographics, spending behavior, and marketing responses. We are predicting 'Complain'.

Most spending features are highly skewed. Most customers spend little and few spend a lot. It seems that only age seems approximately normal.
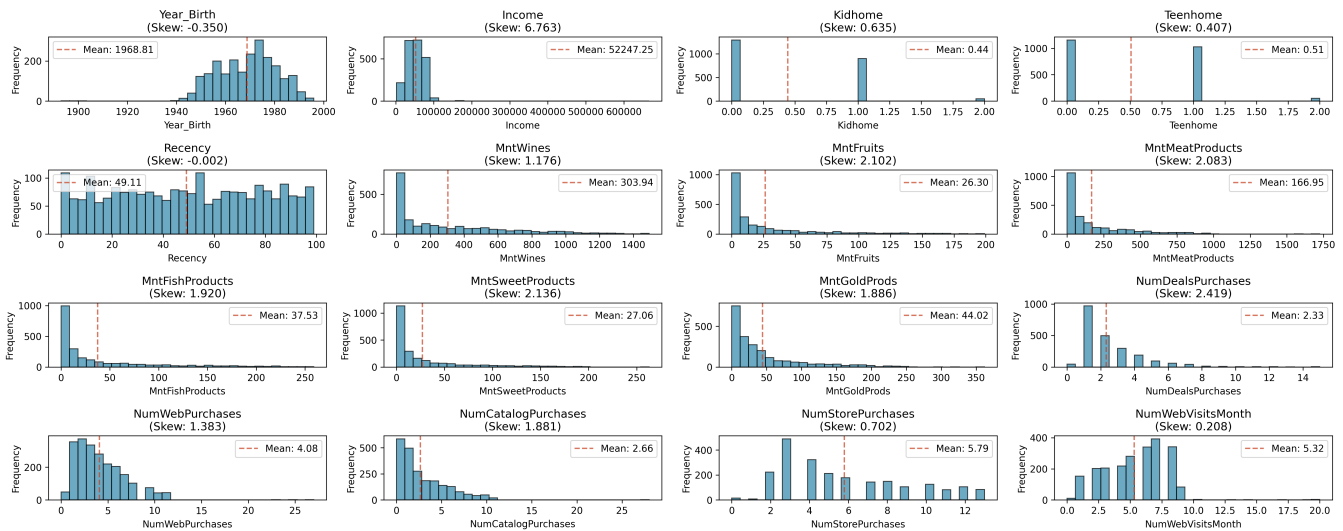

Figure 1: Continuous Features Histograms

Marketing campaigns have low acceptance rates (~5%), showing customers are generally not very responsive to campaigns.
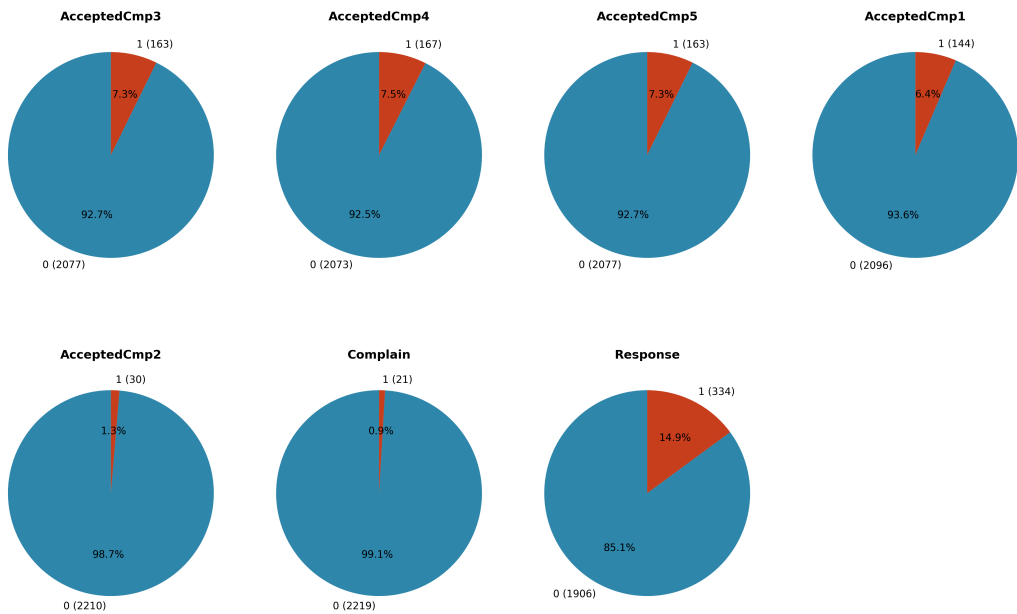

Figure 2: Binary Features Pie Charts

The only feature with missing values is income, with around 1.07% of its entries missing.
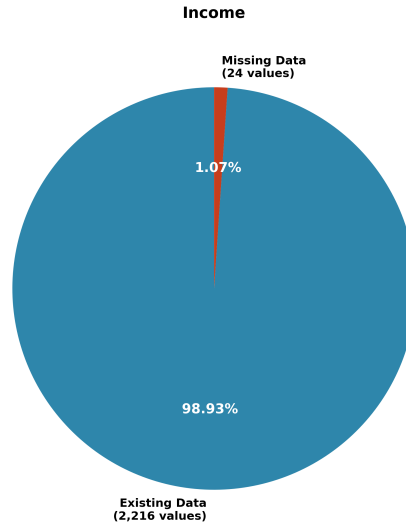
Income

Figure 3: Missing Data Analysis

Feature Engineering

A few simple features were created from the existing data:

Customer_Tenure: Days since registration.

TotalMnt: Total spending across all categories.

TotalPurchases: Total number of purchases.

Purchase Share Ratios: Percent of purchases online vs store vs catalog. Online customers might complain more.

TotalCampaignAccepts: How many marketing campaigns they accepted.

TotalKids: Total children at home.

The engineered features show a few useful patterns. *Customer_Tenure* is relatively spread out with most customers being around 500+ days old. *TotalMnt* seems to be highly skewed and *TotalPurchases* being seemingly bimodal. Purchase shares resemble normality with some skew.
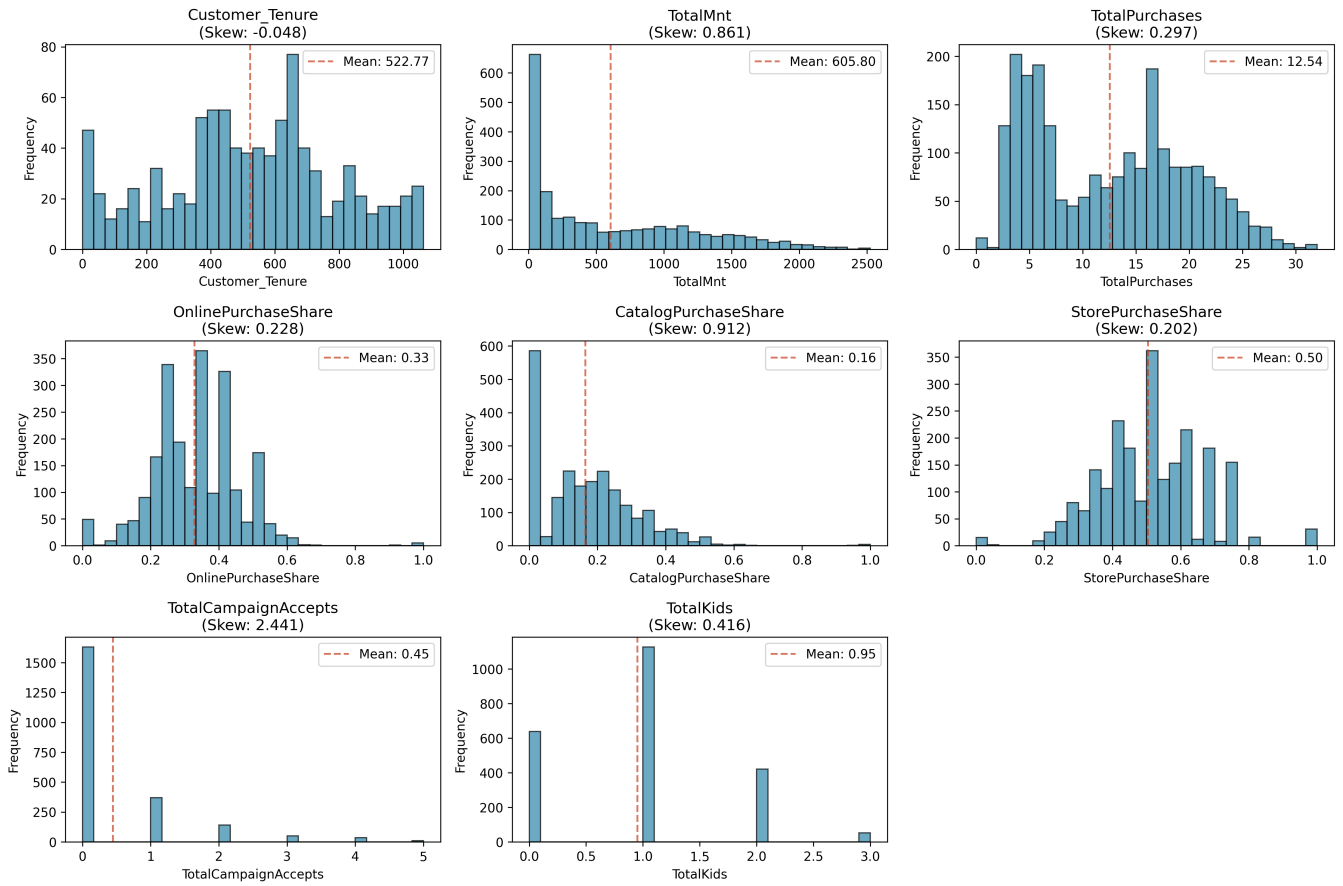
Figure 4: Engineered Features Histograms

The correlation matrix shows relationships between engineered features and complaints. *Customer_Tenure* has the strongest correlation with complaints (0.074) which can help indicate that longer customers are more likely to complain. *TotalMnt* surprisingly shows negative correlation (-0.037), so higher spenders complain less.
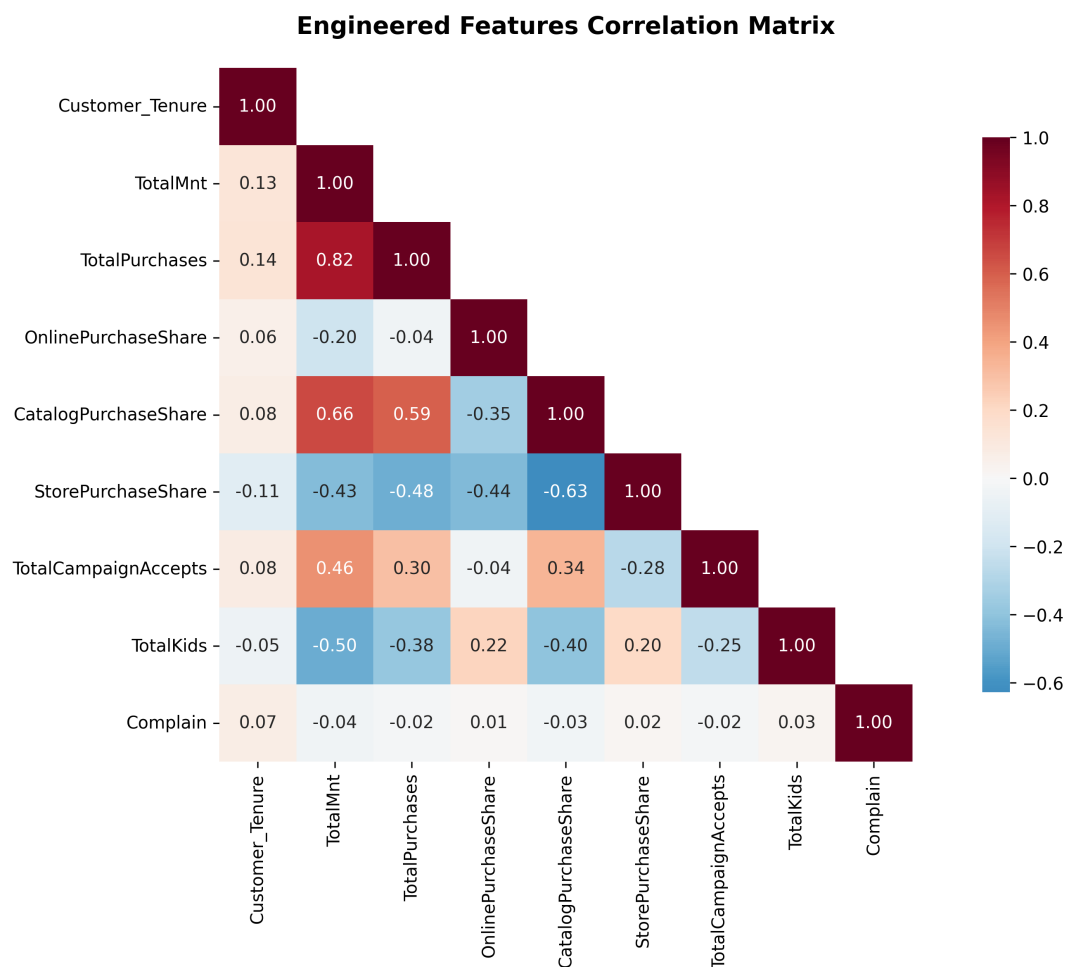
Engineered Features Correlation Matrix

Figure 5: Engineered Features Correlation Matrix

Data Preprocessing

Only Income has missing values (24 missing, 1.07%) and since it is highly skewed, we use median instead of mean to avoid bias from outliers.

The numeric features use the StandardScaler and the categorical features are one-hot encoded.

Supervised Learning

Models

In this task I have used the 6 models in the prediction task.

Logistic Regression (LR): A simple linear classifier for binary classification.

Decision Tree (DT): A non-parametric classifier that can capture non-linear relationships but can often overfit depending on its depth.

Support Vector Machine (SVM): A classifier used to find an optimal decision boundary by maximizing the margins.

k-Nearest Neighbor (k-NN): A classifier that classifies similar elements usually based on distance.

Random Forest (RF): An ensemble of decision trees that better handle overfitting.

Gradient Boosting (GB): A sequential ensemble that learns from previous steps.

Parameter Tuning and Model Selection

I used Bayesian optimization instead of grid search as it's more efficient and gives me less room to guess. I also used 2-fold cross-validation due to the performance requirements of higher amount of folds and also due to the extreme class imbalance. For the final evaluation, I ran 5-fold cross-validation on the best models to meet the assignment requirements.

Optimized Parameters: - LR: C (regularization), penalty (l1/l2/elasticnet), max_iter, l1_ratio, class_weight - DT: criterion (gini/entropy), max_depth, min_samples_split, min_samples_leaf, max_features, class_weight
- SVM: kernel (linear/rbf), C, gamma, class_weight - k-NN: n_neighbors, weights (uniform/distance), metric (euclidean/manhattan) - RF: n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features, bootstrap, min_impurity_decrease - GB: n_estimators, learning_rate, max_depth, subsample, min_samples_split, min_samples_leaf, max_features, loss

These parameters were mainly chosen to be explored by gut feeling from reading their documentation.

Evaluation

Confusion Matrix

No Sampling   No sampling shows really bad performance across all models. They barely catch complaints probably due to the extreme class imbalance.
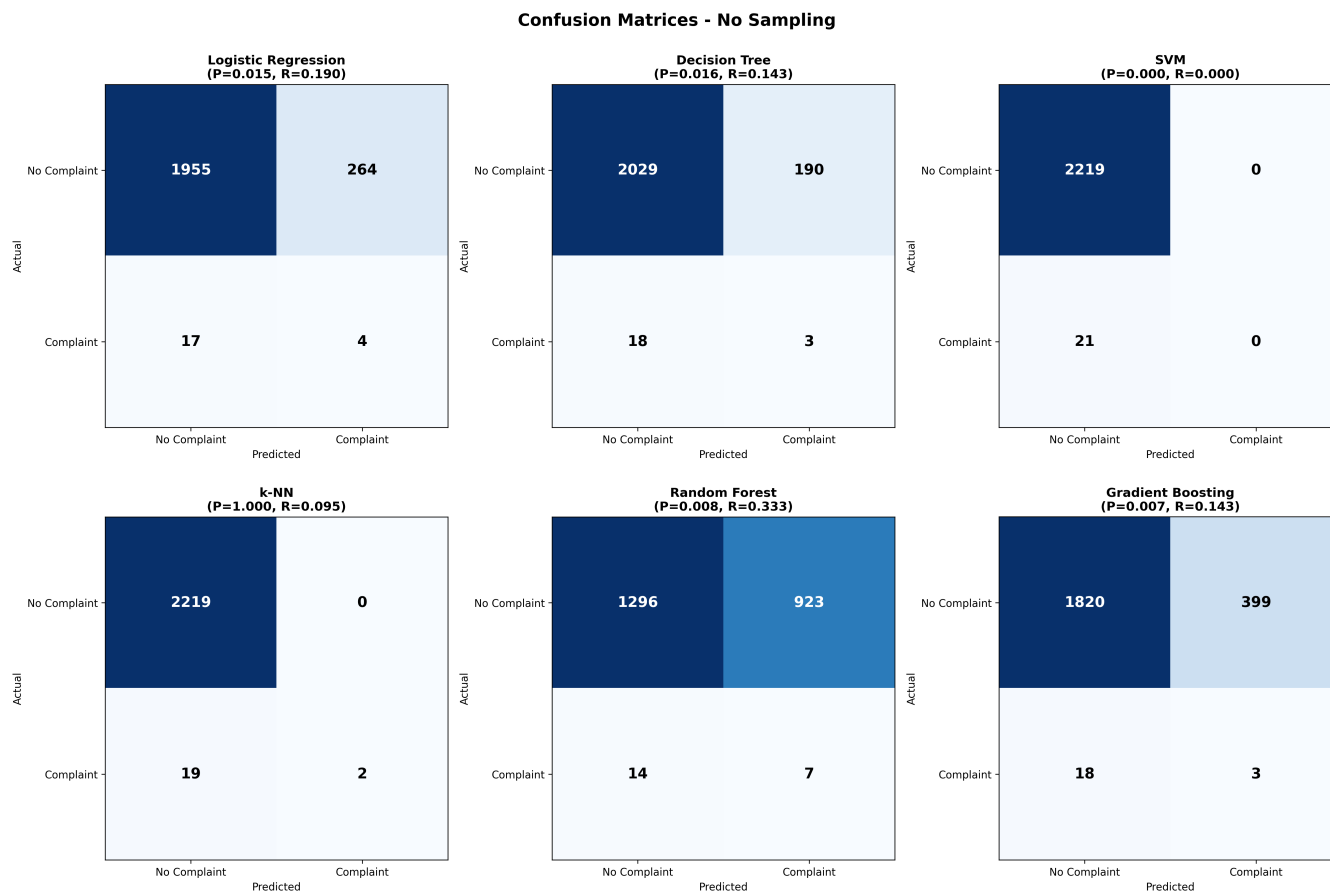

Figure 6: Confusion Matrices - No Sampling

Undersampling   Random undersampling improved recall across all models. Most models now catch more complainers, but precision dropped due to more false positives.
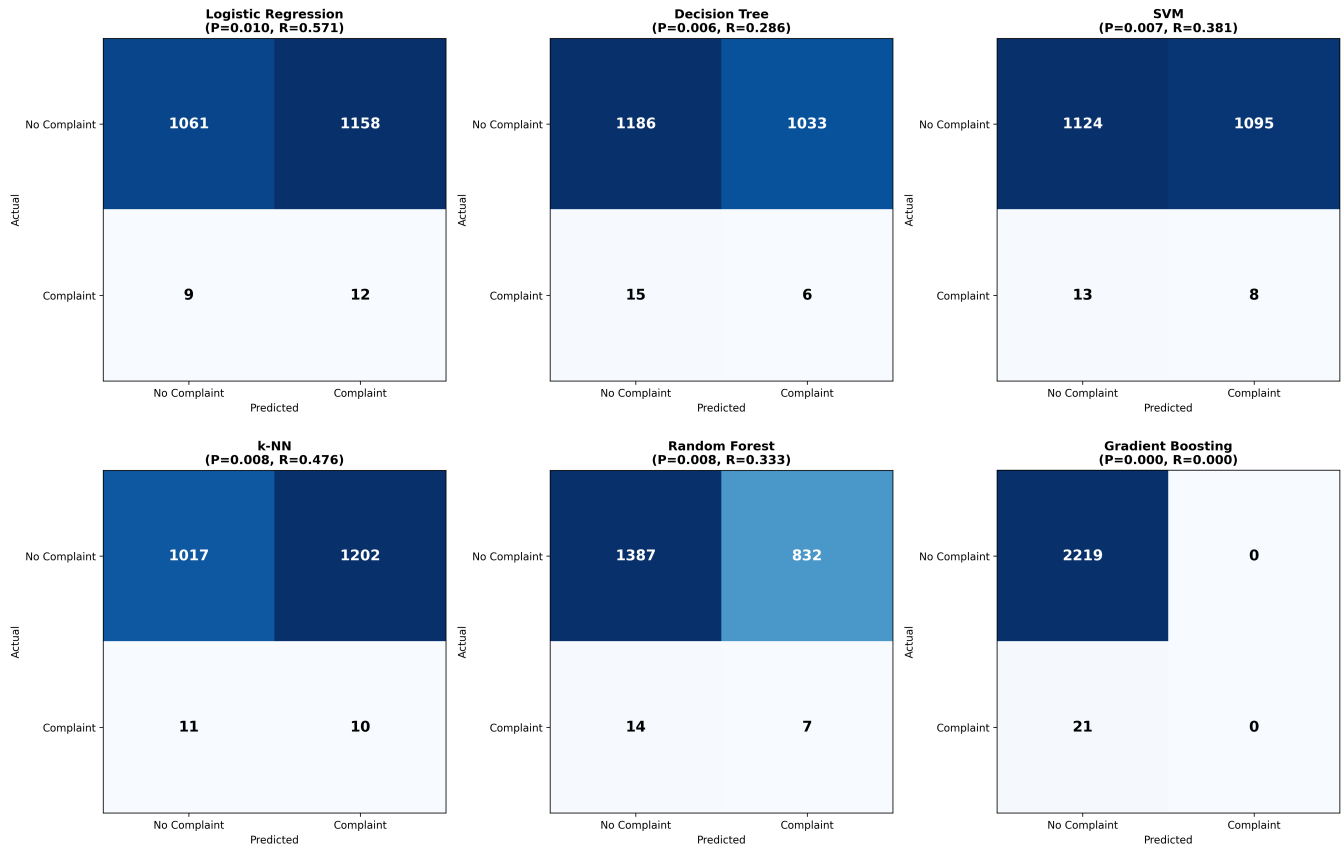
Figure 7: Confusion Matrices - Undersampling

SMOTE   SMOTE gave me mixed results. Some models improved while others performed worse.
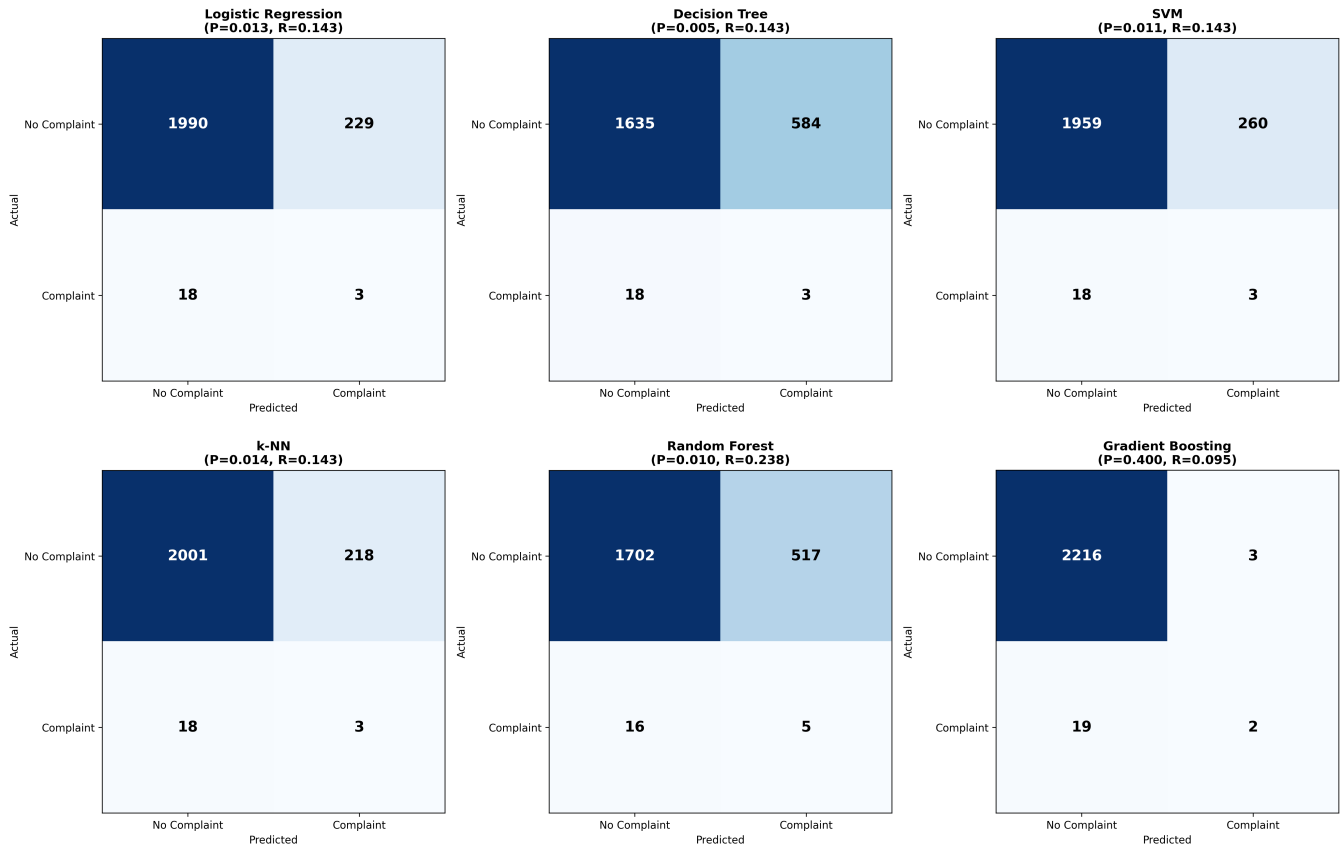
Figure 8: Confusion Matrices - SMOTE

ROC Curves

No Sampling   Most models show poor discriminative ability with AUC values barely above random (0.5). SVM achieved the highest AUC (0.555) but completely failed to predict any complaints, making it practically useless.
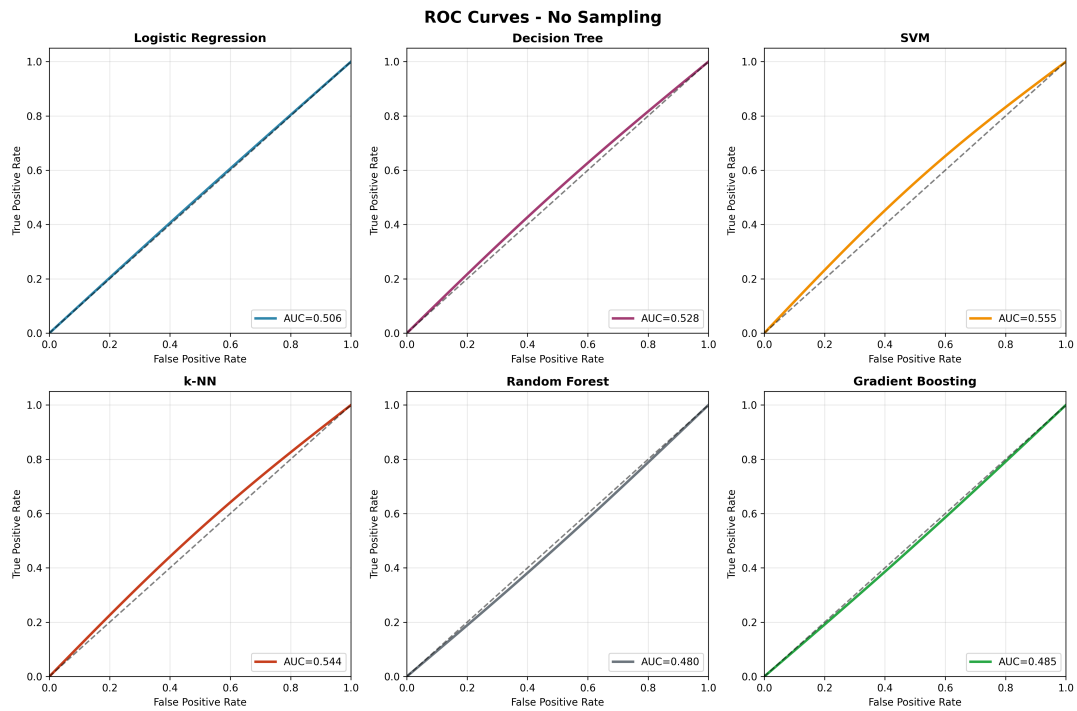
Figure 9: ROC Curves - No Sampling

Undersampling Undersampling actually made most models worse. Gradient Boosting improved slightly (0.512) but most others got way worse.
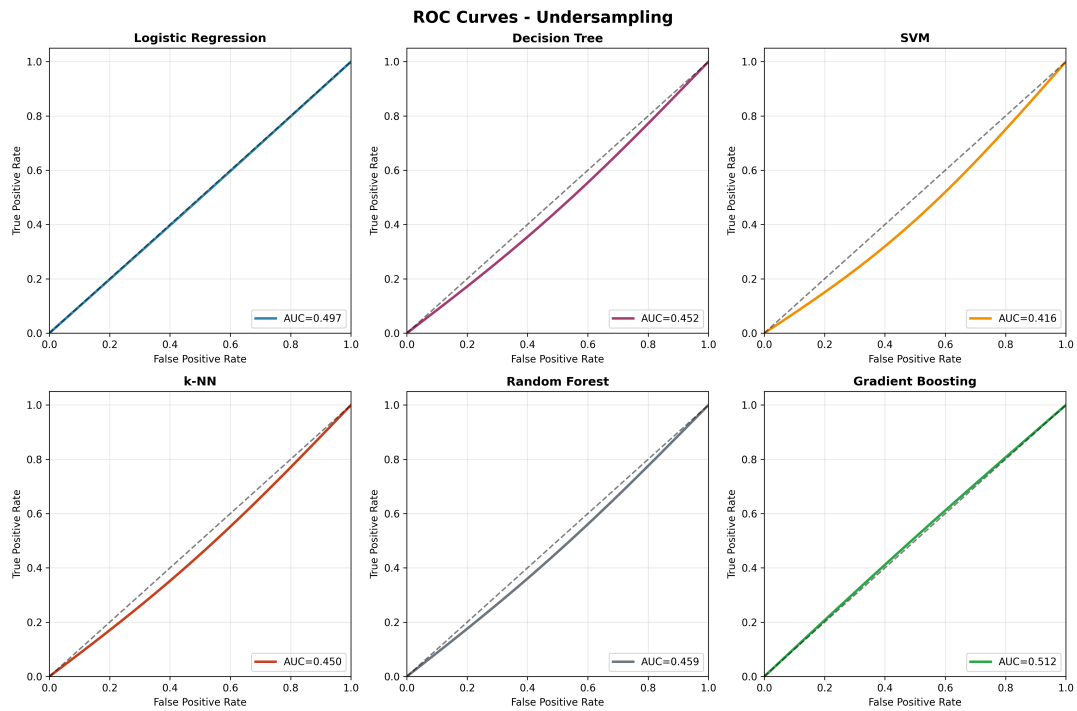

Figure 10: ROC Curves - Undersampling

SMOTE SMOTE, as previously, shows mixed results. SVM improved to 0.527 and Logistic Regression to 0.522, but Decision Tree got worse (0.430).
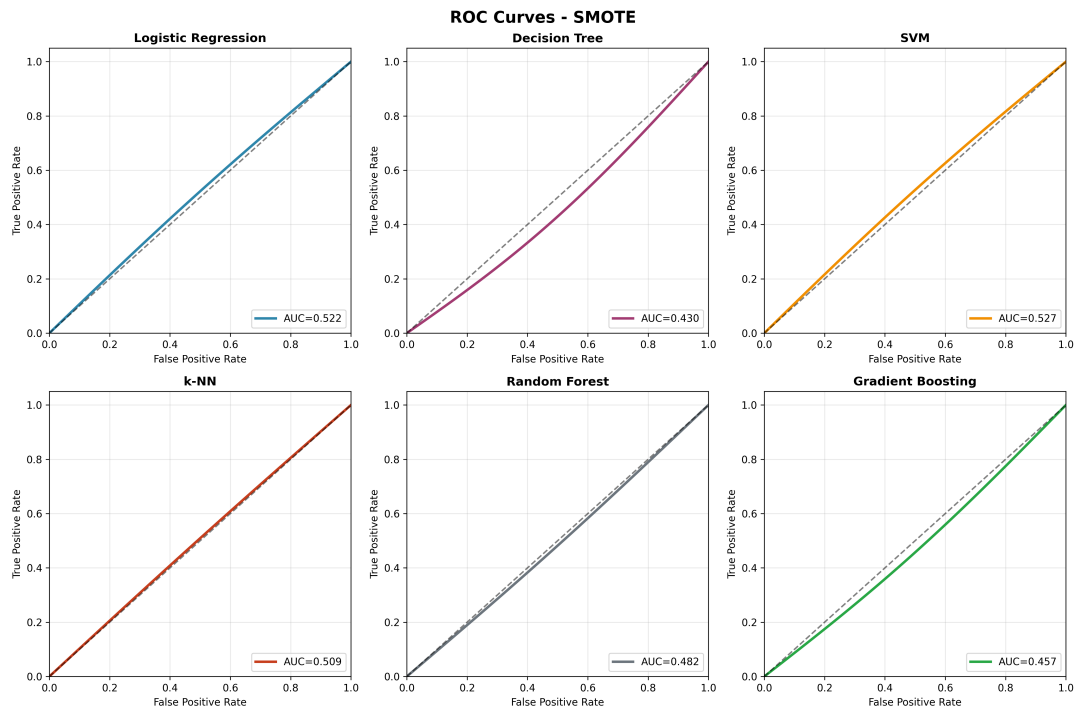
Figure 11: ROC Curves - SMOTE

The synthetic samples help some models but not others.

Model Comparison Best at catching complainers: Logistic Regression with undersampling (0.571 recall) and k-NN with undersampling (0.476 recall).

Best precision: k-NN with no sampling (1.000 precision, 0.095 recall) - only predicts 2 complaints.

Best discriminator: SVM with no sampling (0.555 AUC) but completely useless since it predicts zero complaints.

Most models handle this class imbalance pretty badly. Undersampling helped catching more complainers but hurts on discriminating positives.

Class Imbalance

The dataset has extreme class imbalance: 21 complainers vs 2,219 non-complainers. The impact of this is very noticeable due to the low recall and low precision. The models can achieve 99% accuracy by simply predicting "no complaint" every time making accuracy a pretty misleading metric.
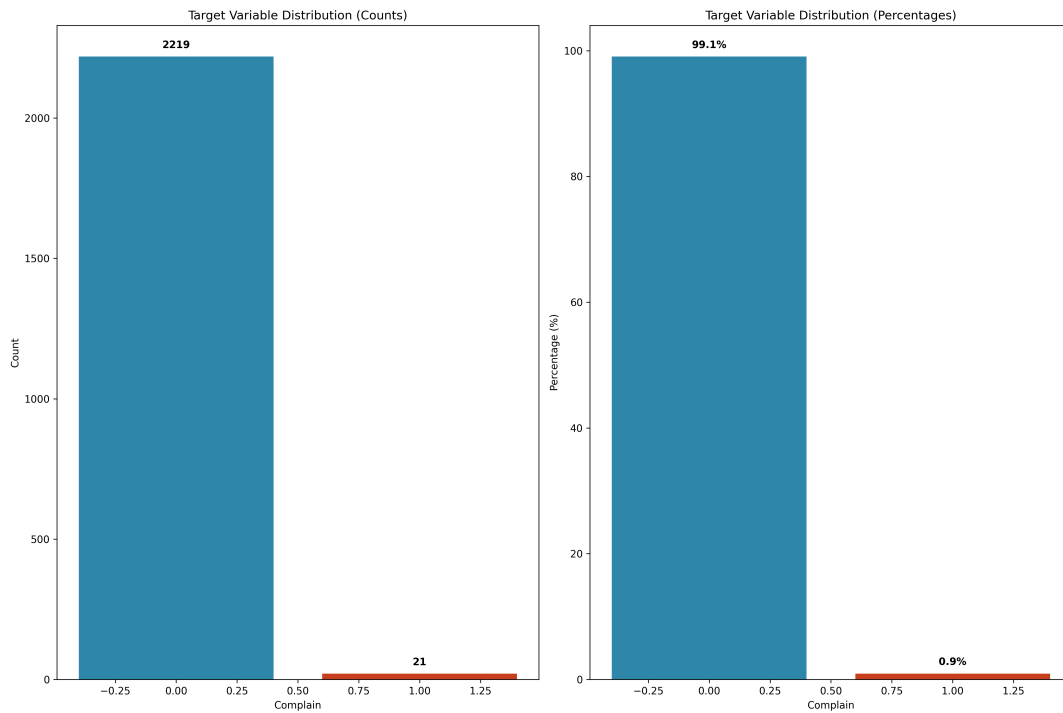
Figure 12: Target Distribution

Summary

2-Fold CV

Model	Sampling	Precision	Recall	Specificity	AUC
LR	under	0.0103	0.5714	0.4781	0.4969
DT	none	0.0155	0.1429	0.9144	0.5284
SVM	none	0.0000	0.0000	1.0000	0.5550
k-NN	under	0.0083	0.4762	0.4583	0.4498
RF	none	0.0075	0.3333	0.5840	0.4804
GB	under	0.0000	0.0000	1.0000	0.5120

5-Fold CV

Model	Sampling	Precision	Recall	Specificity	AUC
LR	under	0.0116	0.6190	0.5400	0.5400
DT	under	0.0087	0.4762	0.4794	0.4794
SVM	under	0.0124	0.7619	0.6069	0.6069
k-NN	under	0.0084	0.4762	0.4589	0.4589
RF	under	0.0115	0.5714	0.5633	0.5633
GB	none	0.0085	0.0952	0.4935	0.4935

SVM achieved the highest recall (0.7619) for catching complainers with random undersampling.

Appendix

The confusion matrices and ROC curves were generated using 2-fold cross-validation results from the tuning process. The 5-fold cross-validation was performed afterwards.