Term Project Report

"GOOD" OR "BAD" CLIENT?

Group 1: Miranda Nguyen, Karan Parihar, Phuong Phan, & Julia Schroeder
BSAN-460-001
Dr. Christopher Gaffney
March 20, 2024

*Term Presentation

TABLE OF CONTENTS

*Term Presentation

I. INTRODUCTION

➢ Project Summary & Goal

For this project, we are acting as a bank using the dataset to decide if the client is eligible for a loan based on the variables. The current objective is to examine several variables to determine their significance in predicting client eligibility. Our overall goal is to create two predictive models that can determine whether or not a client is good or bad, allowing us to see if the client is eligible for a loan. Through testing, training, and tweaking of models, our group will seek a highly accurate and reliable decision tree and logistic regression model.

II. DATA ANALYSIS

➢ Data Description

The 14 variables contained in our dataset are as follows, with a total of 1,723 observations:

Variable	Type	Description
month	Categorical	Month of purchase (*1-12 [month #]*)
credit_amount	Numerical	Loan request (*5k - 301k [$]*)
credit_term	Numerical	Loan term (*3-36 [months]*)
age	Numerical	Client age (*18-90 [years-old]*)
sex	Categorical	Client gender (*female, male*)
education	Categorical	Client education (*Higher education, Incomplete higher education, Incomplete secondary education, PhD degree, Secondary education, Secondary special education*)
product_type	Categorical	Type of product (*Cellphones, Household appliance, etc.*)
having_children_flg	Categorical	The presence of children at the client's household (*0,1 [no,yes]*)
region	Categorical	Client location (*0-2 [Pennsylvania, New Jersey, New York]*)
income	Numerical	Client's income (*1k - 401k [$]*)
family_status	Categorical	Family status (*Married, Unmarried, Another*)
phone_operator	Categorical	Mobile operator (*0-4 [operator]*)
is_client	Categorical	Is a client of the bank (*0,1 [no,yes]*) **Note*: *61% of responses are non clients (1),*

**Term Presentation*

		while 39% of responses are from existing clients (0).
bad_client_target	Categorical	Considered a bad client target (*0,1 [no,yes]*)

The "bad_client_target" variable will serve as our target or dependent variable for model building. However, the severe imbalance in this variable poses a challenge due to the scarcity of instances labeled as 1 (bad client target), leading to data imbalance. Before addressing this issue, data cleaning procedures are necessary.

➢ Data Cleaning Efforts

Aside from converting each column into its respective data type, minimal adjustments were required in the dataset. Tests were conducted to detect any missing or duplicate variables, but none were found. Variables such as credit amount, credit term, age, income, and family status are likely to be crucial in assessing a client's creditworthiness. Furthermore, we excluded irrelevant variables and deleted their column, including product_type, as it does not directly contribute to our loan eligibility prediction objective. From here, we conducted an outlier analysis to identify data points that differ significantly from the other observations.

```
#Converting each column into its respective data type
clients[sapply(clients, is.character)]=lapply(clients[sapply(clients, is.character)],as.factor)
clients$month=as.factor(clients$month)
clients$having_children_flg=as.integer((clients$having_children_flg))
clients$is_client=as.factor(clients$is_client)
clients$bad_client_target=as.factor(clients$bad_client_target)
clients$region=as.factor(clients$region)
```

Fig 1. Converting Columns Into Respective Data Types

```
#Removing 'Product Category' column (irrelevant)
clients = clients[,-7]
summary(clients)
```

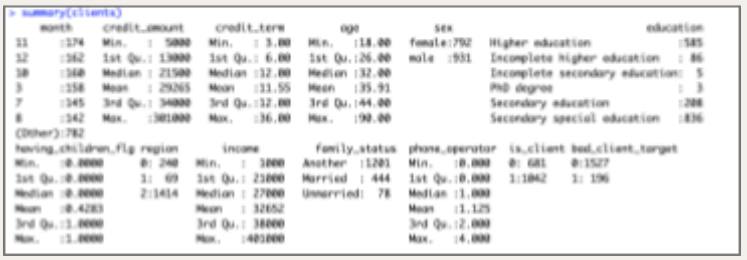


Fig 2. Removing 'product_type' Column

*[Term Presentation]

```
#Identify missing values
is.na(clients)
missing_values <- is.na(clients)
print(missing_values)

#Count of how many missing values are in each column
col.ms <- colSums(missing_values)
print(col.ms)
```

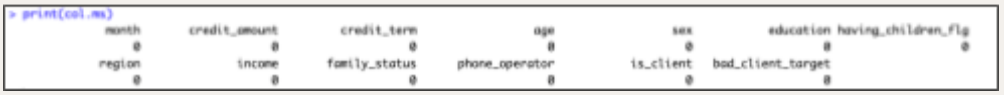


Fig 3. Identifying Any Missing Values

```
#Duplicate
##Identifying rows with duplicate data
duplicated(clients)
##Total number of duplicate data in dataset
sum(duplicated(clients))
```

```
> ##Total number of duplicate data in dataset
> sum(duplicated(clients))
[1] 0
```

Fig 4. Identifying Any Duplicate Values

➢ Outlier Analysis

After running outlier analyses on four main variables, namely Credit Amount, Credit Term, Age, and Income, our group concluded that the outliers were not something we felt enough justification to remove. Each variable has its own practical reasoning for keeping the outliers:

Credit Amount: These outliers could be due to a variety of reasons, such as loans for more expensive products or services. For the sake of our objective, keeping outliers in the credit amount as they reflect real world scenarios and represent instances where clients may be looking to take out small or large loans. In other words, not all loans are going to be within a close range of each other. By keeping the outliers, we are also maintaining the diversity of the customer profiles. When assessing client

creditworthiness, these outliers are important because they are able to represent extremes in loan sizes which carry various levels of risk.

Credit Term: The outliers within the credit term column might represent special loan products designed for specific purposes. Similarly to credit amount, these outliers represent a diverse range of loan types and reflect real-world scenarios in which clients will have different requirements for shorter or longer repayment periods. The credit term outliers play an important role during the evaluation of client repayment capacity and can help the bank make loan eligibility assessments.

Age: Analyzing age outliers helps ensure that the models accurately reflect the diversity of the client base. Retaining the outliers prevents the model from being biased and helps to remain inclusive. Different ages may account for financial independence of the client, thus helping to identify creditworthy clients.

Income: Keeping the outliers of income can allow us to build models that accurately reflect the impact of varying income levels, including those of high earners, and can more effectively predict loan eligibility and risk as a result.

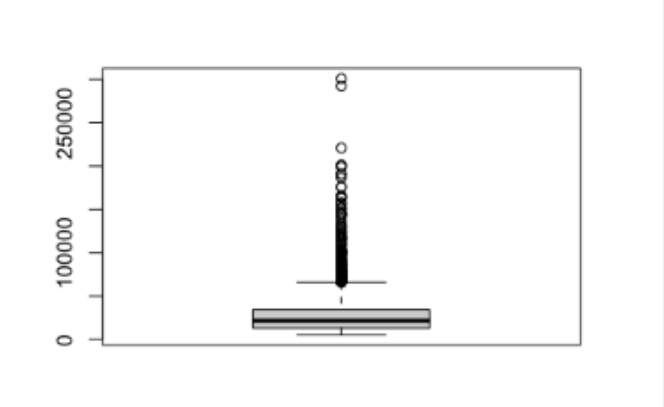
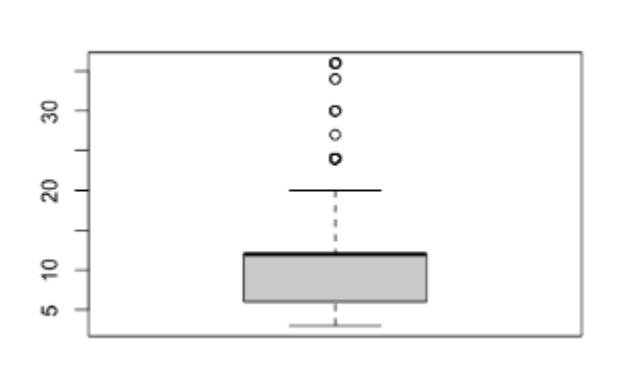


Fig 5. Boxplot for Credit Amount



*[Term Presentation]

Fig 6. Boxplot for Credit Term

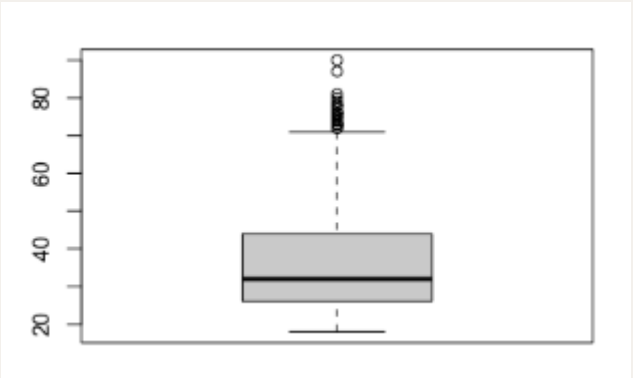


Fig 7. Boxplot for Age

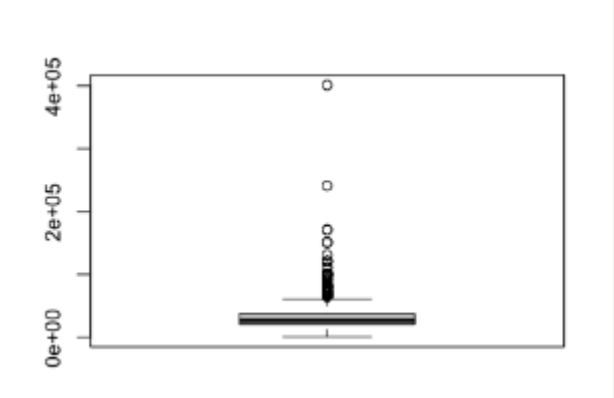


Fig 8. Boxplot for Income

III. SOLUTION APPROACHES

To address the bank's objective of determining clients' eligibility for loans using the provided dataset, we employed classification trees and logistic regression as solution approaches. Classification trees, such as the C45 tree, are advantageous for their ability to represent decision-making processes intuitively. Furthermore, a logistic regression approach allows us to model the probability of a client being good or bad based on the provided variables, which aligns to determine client eligibility for a loan.

➢ First Approach: Classification Tree

Once the classification tree is trained, its structure can be visualized, allowing us to interpret the decision rules and understand the key factors influencing loan eligibility. For

instance, the tree might reveal that clients with certain income levels, ages, or credit amounts are more likely to be classified as bad clients. The insights gained from the tree can guide us in making informed decisions about loan approvals and help us understand the factors contributing to the classification of clients as good or bad candidates for a loan.

Strengths	Weaknesses
• *Relatively simple to understand*, and it provides a straightforward representation of decision-making processes. • *Identifies the most important variables* contributing to the prediction of loan eligibility, helping the bank prioritize which factors to focus on when assessing clients' applications.	• *Imbalanced data can lead to biased models*, where the algorithm may have a tendency to predict the majority class more frequently, thus resulting in poor performance in predicting the minority class. • *Classification trees are prone to overfitting*, which can lead to poor generalization performance on unseen data. • *Decision trees can become complex,* especially with large datasets and many predictors, which may make the model more difficult to interpret.

➢ Second Approach: Logistic Regression

Additionally, logistic regression is also a suitable choice for this problem due to its probabilistic nature and ability to handle imbalanced datasets more gracefully. Logistic regression estimates the likelihood of a client being a bad client based on the given features. To enhance the model's performance, we plan to carefully select relevant variables for both classification trees and logistic regression.

Strengths	Weaknesses
• *Provides probabilities as output*, allowing for easy interpretation of the likelihood of a client being a bad client based on the given features. • *Provide insights into the direction and strength of the relationship between independent variables and the probability of the outcome*, aiding in understanding the factors influencing loan eligibility.	• *May not capture complex interactions between variables effectively*, especially if the relationships are non-linear. • *Outliers in the data can disproportionately influence the estimated coefficients and predictions*, potentially leading to biased results.

*Term Presentation

➢ Evaluating Models

To evaluate the performance of our models, we will employ metrics such as accuracy, precision, and recall. In such cases where the data set has a severe class imbalance, accuracy alone can be misleading, as the model could achieve high accuracy by simply predicting the majority class (0). By using metrics like precision and recall, the evaluation considers both false positives and false negatives, providing a more balanced assessment of the model's performance. By systematically addressing class imbalance, selecting pertinent variables, and employing comprehensive evaluation metrics, we aim to build predictive models that enhance the bank's decision-making process regarding client loan eligibility.

IV. BALANCING THE DATASET

To preface, we will first be focusing our efforts on building our classification tree model. After dividing the total of '1' observations under bad_client_target by total observations, we found that only about 11.3% of the responses were labeled as '1', or 'bad client'. This severe imbalance in the bad_client_target variable poses a challenge, as the scarcity of instances labeled as 1 (bad client target) impedes the classification tree's ability to create meaningful splits related to bad clients. This also affects the logistic regression approach's ability to create unbiased estimates of model parameters, as the model may disproportionately focus on the majority class (instances labeled as 0) and fail to accurately capture the characteristics of the minority class (instances labeled as 1). To mitigate this issue, we plan to first create a classification tree using the original dataset. Running the tree on the original dataset allows for a comprehensive exploration of the data structure and patterns, which is crucial for determining the appropriate strategies to address the imbalance effectively. Furthermore, we will split the data into two groups (about 1,300 responses for training the machine learning algorithms, and 400 responses for testing the algorithm (75%-25%)) to ensure that the test dataset is as unbiased as it can be and reflects a true evaluation of the model.

From here, we plan to combat the imbalance by employing techniques like downsampling the majority class (bad_client_target_0) and oversampling the minority class (bad_client_target_1). By either providing more instances of bad clients or reducing the dominance of the majority class in the dataset, the tree can gain a better understanding of the patterns associated with loan ineligibility. This balanced representation ensures that the tree can effectively split the data and generate decision rules that accurately predict whether a client is eligible for a loan.

➢ Preliminary Models to Understand the Imbalance: Model I & Model II

i) Model I Building

*Term Presentation

To identify clients who are risky for a loan, we will focus on the model's ability to correctly predict the bad_client_target_1 group. We fed all 13 predictor variables into the model. We started by performing initial data cleaning and then ran the CART tree and the C45 tree on the original dataset before splitting and balancing. The purpose of these two trees, called model I, is to see how imbalance impacts the tree algorithms.

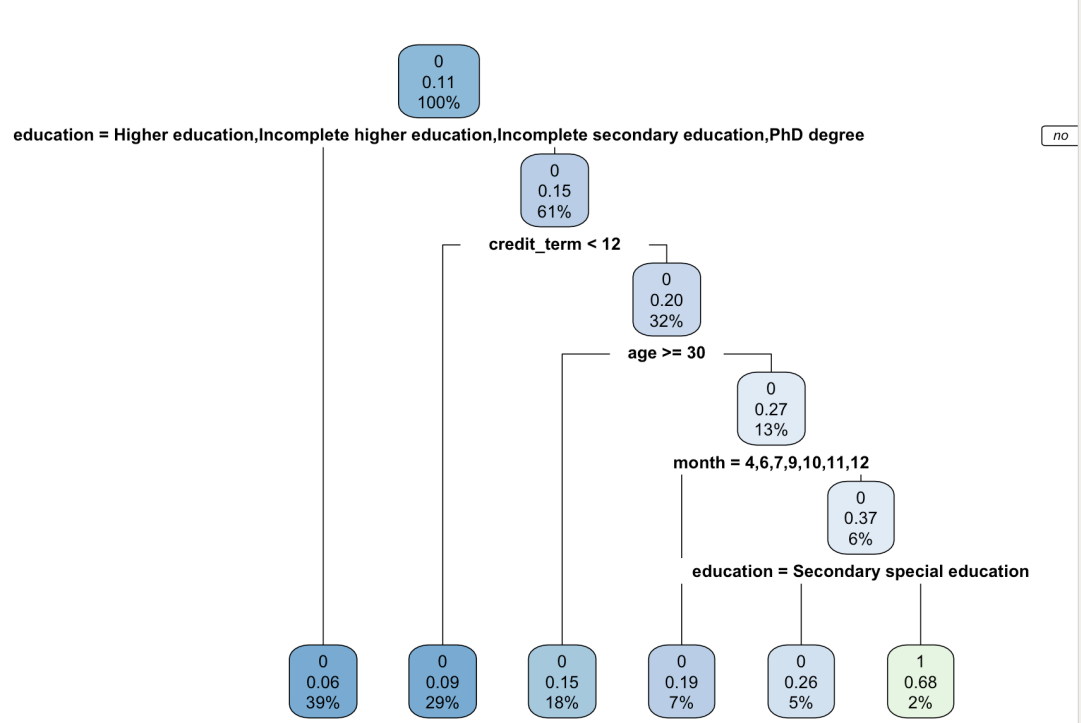


Fig 9. CART tree - Model 1

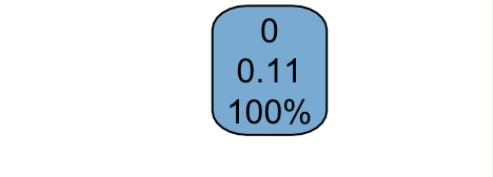


Fig 10. C45 tree - Model 1

ii) Model I Evaluation

The target variable bad_client_target is heavily skewed, with only 196 bad_client_target_1 answers, and 1527 bad_client_target_0 answers. This imbalance causes the CART tree not to be able to split properly, or not split at all with the C45 tree. We need more representation for bad_client_target_1 to balance the data out and provide more information to the tree so that they can predict better.

iii) Model II building:

For Model II, we split the data into a training set and a test set (75% - 25%), used the training set to train the trees, and ran the trees on test set data to evaluate their accuracy. We

fed all 13 predictor variables into the model. We will evaluate them using the confusion matrix and accuracy.

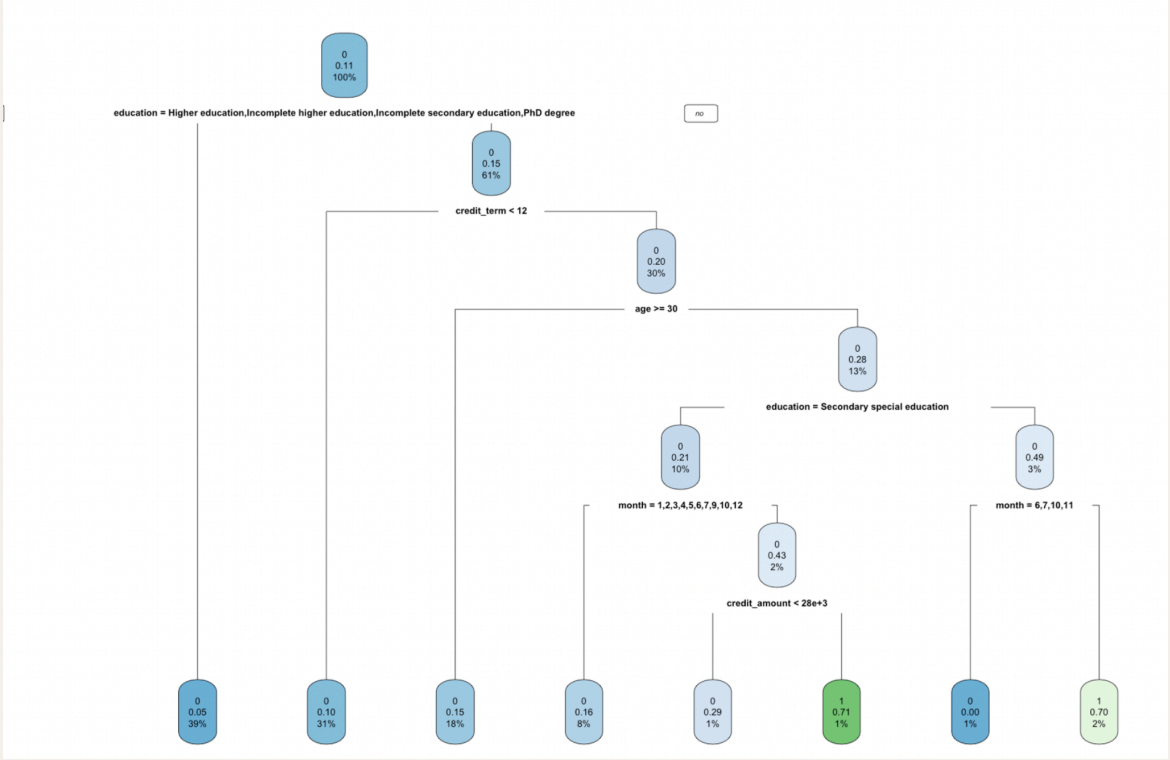


Fig 11. CART tree - Model II

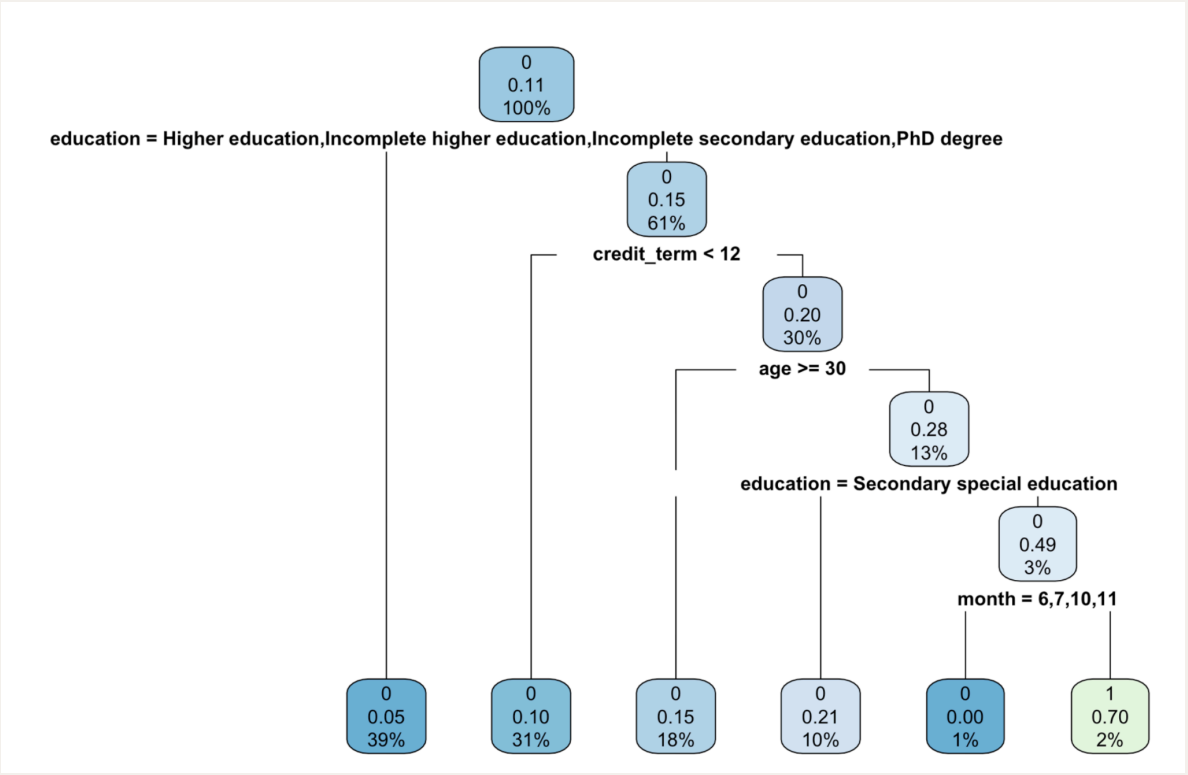


Fig 12. C45 tree - Model II

**Term Presentation*

iv) Model II explanation:

For CART tree

- **Example of a client who's not eligible for a loan:** someone who only has a secondary special education, credit term > 12, age < 30, month = 8, credit amount >= 28,000

For C45 tree

- **Example of a client who's not eligible for a loan:** someone who only has a secondary education, credit term > 12, age < 30, month = 1

The CART tree from both model I and model II seem to have more leaf nodes and more splits than the C45 tree. For the target prediction of our model, the profile of a client who's not eligible for a loan, the CART tree has two leaf nodes leading to this result while the C45 tree has only one. Both of the trees share the same leaf node scenario for bad_client_target_1 with 2% of the training set explained, while the CART tree has another node that accounts for another 1%. The CART tree has 8 leaf nodes in total, and the C45 tree has 6.

Despite some differences in splitting nodes, the variables that go into the tree's splitting criteria are the same: education, credit term, age, month, and credit amount.

v) Model II evaluation

For CART tree

Training set accuracy

	Predicted	
Actual	0	1
0	1,108	10
1	118	24

Test set accuracy

	Predicted	
Actual	0	1
0	1,336	39
1	152	108

- ☐ The CART tree correctly predicted 89.84% of the observations in the training set.
- ☐ The CART tree correctly predicted 87.9% of the observations in the test set.

*Term Presentation

For C45 Tree

Training set accuracy

	Predicted	
Actual	0	1
0	1,108	10
1	118	24

Test set accuracy

	Predicted	
Actual	0	1
0	1,336	39
1	152	108

- ☐ The C45 tree correctly predicted 89.6% of the observations in the training set.
- ☐ The C45 tree correctly predicted 88.12% of the observations in the test set.

Accuracy is a metric that provides an overall assessment of the model's ability to make correct predictions across both yes (1) and no (0) responses. The C45 tree's test accuracy is slightly better than the CART's. The accuracy differences going from training to test set are relatively small, showing that the trees are not prone to overfitting, and the results can be used for future prediction.

At first glance, these two accuracies seem quite high for a first-time model, without any cost training. However, when we put the percentage of bad_clients_target_0 from the original dataset in comparison, which is 88.73%, these two accuracies cannot exceed it. This means that even if we're using a prediction model, the chance of us correctly labeling whether someone is a good or bad client is not as high as just assuming everyone is a good customer with an accuracy of 88.73%, which is equal to the proportion of the bad_clients_target_0 from the original dataset. Therefore, even if the accuracy is relatively high in this initial run, using a predictive model is unnecessary here. This calls for rebalancing the data to better increase the model's accuracy. We will use the CART tree mainly for the upcoming model developments.

➢ Balancing data: Downsampling the majority (Model III)

i) Downsampling the dataset

Downsampling refers to removing records from majority classes to create a more balanced dataset. The simplest way of downsampling majority classes is by randomly removing records from that category.

*Term Presentation

In our training data, 11.269% of them are in the minority class. We want to increase the bad_client_target_1 up to 15% by randomly removing 314 observations, or 28.02%, of bad_client_target_0. The new balanced training dataset with 13 variables will be fed into the trees and run similarly to the above models.

ii) Model III building

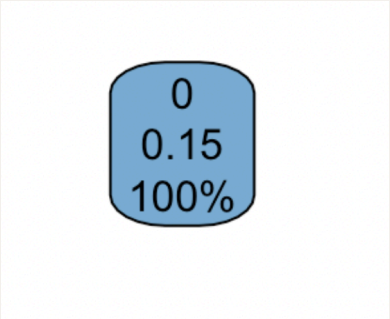


Fig 13. CART tree - Model III

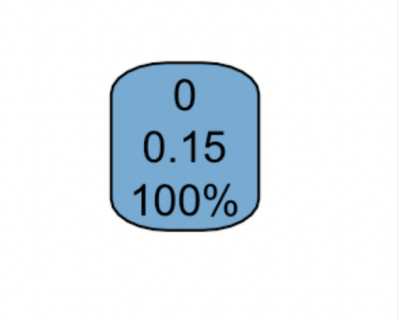


Fig 14. C45 tree - Model III

iii) Model III evaluation

Both of the two trees fail to split, which is similar to the C45 tree - Model 1 (Fig. 6). However, the percentage of positive prediction - or bad_client_target_1 - increases from the original 11% to 15% in both models. This, unfortunately, is exactly the result of the downsampling method.

By downsampling, we unintentionally remove valuable data points that help aid the tree's splitting criteria. 28% is one-third of the pool for bad_client_target_0, which is too big to remove from the training set. This is not an appropriate balancing method.

*Term Presentation

➢ Balancing data: Random Oversampling the Minority

i) Model IV development

Original training set:

bad_client_target_0	bad_client_target_1
1118	**142 (11.27%)**

Random oversampling training set:

bad_client_target_0	bad_client_target_1
1118	**194 (14.7%)**

In order to address the class imbalance in the dataset, 52 more of the bad_client_target_1 were randomly generated using the Random function in Excel. Thus, 14.7% of the dataset are cases where instances were labeled "bad_client_target_1". We then ran the tree and computed other metrics.

ii) Model IV: Random Oversampling Classification Tree:

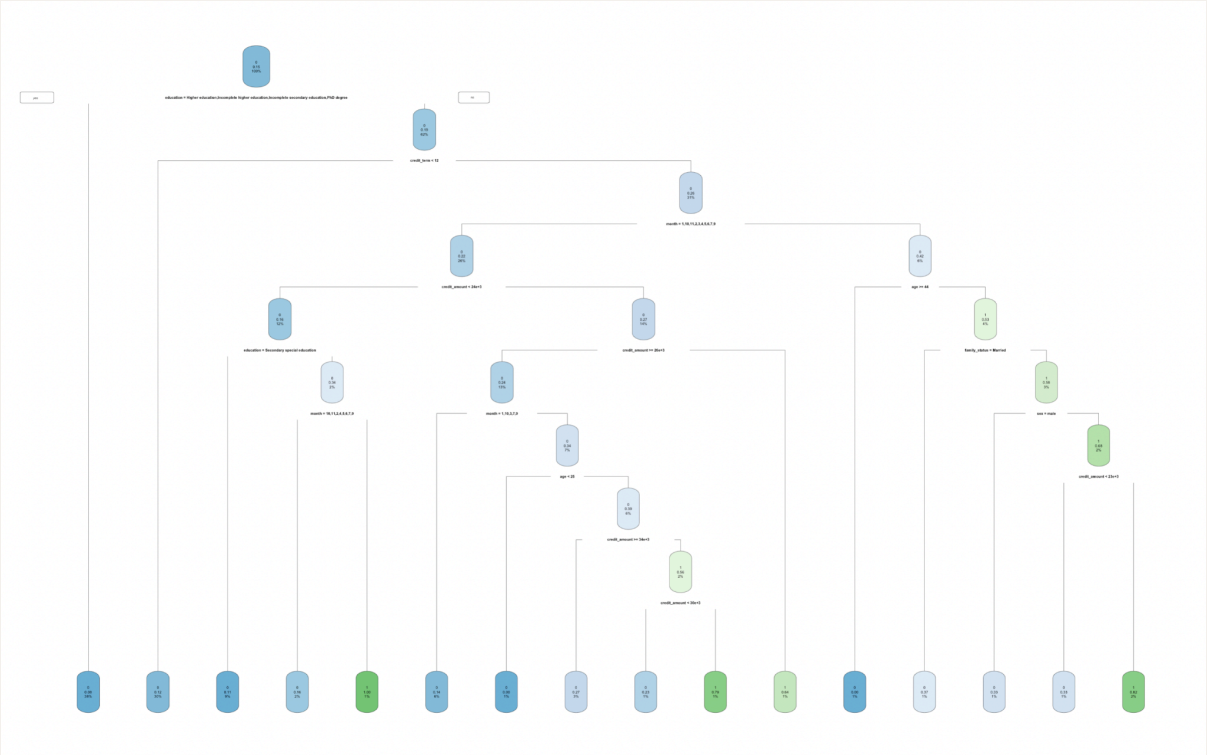


Fig 15. CART tree - Model IV

iii) Model IV Evaluations

Model IV splits, and splits better than *Model II* we run on original and imbalanced data. Compared to *Model III: Downsampling*, which did not split at all, this is the right direction. We evaluate external metrics from it and will continue to develop the model based on this random oversampling training set.

*Term Presentation

Confusion Matrix for Test Set:

Model IV	Predicted	
Actual	0	1
0	395	14
1	**53**	**1**

	Training Set	Testing Set
Accuracy	87.96%	85.53%
Precision	79.03%	6.67%
Recall	25.25%	1.85%

- ☐ While accuracy is quite high for the training set and only decreases a bit for the test set, precision and recall plummet (6.67% and 1.85%). Looking at the confusion matrix, it is easy to understand why: the classification tree only correctly identified one bad client.
- ☐ The false negative number is higher than the false positives, so precision is a bit higher than recall. However, they are too low.
- ☐ Only the training set prediction accuracy is significantly better than no information rate with a p-value < 0.05

iv) Next steps:

While accuracy is promising, this model does not serve our original purpose of better identifying bad client targets for the loan. It is only better at predicting good customers, so accuracy is high. However, one thing it proves is that Random Oversampling is a better-fit data balancing method for our dataset. We will use this new oversampling set forward, and incorporate the cost-sensitivity model to boost its desired metrics for both classification trees and logistic regressions.

V. METHOD 1: CLASSIFICATION TREES

Our data balancing effort goes hand-in-hand with the development of our classification trees. With the Random Oversampling dataset as a base, we continue to build more advanced classification tree models in this section and will pick the best-performing one.

**Term Presentation*

➢ Model Evaluation & Cost-sensitivity Training Overview

i) Model evaluation

Both precision and recall are crucial for information retrieval, where positive class mattered the most compared to negative. Only TP, FP, and FN are used in Precision and Recall.

- ☐ **Precision:** Out of all the positives predicted, what percentage is truly positive?
- ☐ **Recall:** Out of the total positive, what percentage are predicted positive? It is the same as TPR (true positive rate).

In this model, true positives are correctly identified customers who are risky for the loan, while true negatives (the majority of the model prediction) are the normal customers who are qualified for a loan. However, if we fail to identify a risky customer and give them a loan, this will be costly for the bank. False positives, on the other hand, will make the business miss out on potential loan customers. However, after weighing these two, we decided that a False negative would have more negative effects than a false positive.

We want to make sure to get all the risky customers, so **False-Negative should be as low as possible**. In these situations, we can compromise with the low precision, but recall should be high.

ii) Cost-sensitivity training (Or weighted-cost function model)

A basic model assumes a relative cost of 1 for both FN and FP.

	Predicted	
Actual	0	1
0	0	1
1	1	0

To incorporate cost-difference training, we must assume a relative difference in the cost of FN versus FP. We experiment with various cost matrices to observe the changes in the model's prediction. The three assumptions were:

- ☐ FN is double the cost of FP
- ☐ FN is triple the cost of FP
- ☐ FN is quadruple the cost of FP

*Term Presentation

➢ Variable Selection: Random Forests

In order to justify what variables our group would use with the classification tree, a simple variable importance plot was created using the randomForest package. The resulting plot charts are on a "MeanDecreaseAccuracy" and "MeanDecreaseGini" basis. MeanDecreaseAccuracy expresses the loss in accuracy by excluding each variable. Therefore, when interpreting the chart, Credit_Amount is shown to be the greatest loss of model accuracy if removed. MeanDecreaseGini uses the Gini coefficient to represent each variable's part in the homogeneity of nodes and leaves within a random forest. Once again, our highest valued variable was Credit_Amount. Region plays a bigger part in the model than we originally thought, so it helped justify us keeping it. Phone_Operator, also a variable that our group initially questioned the importance of, does not contribute as much to the accuracy, but plays a reasonable role in the homogeneity, as recognized by the Gini plot.

Since this analysis proves that variables we thought might not be important, like Region or Phone_Operator, turn out to be an important factor in the decision tree, we will use all 13 variables (excluding product_type) in the classification tree moving forward.

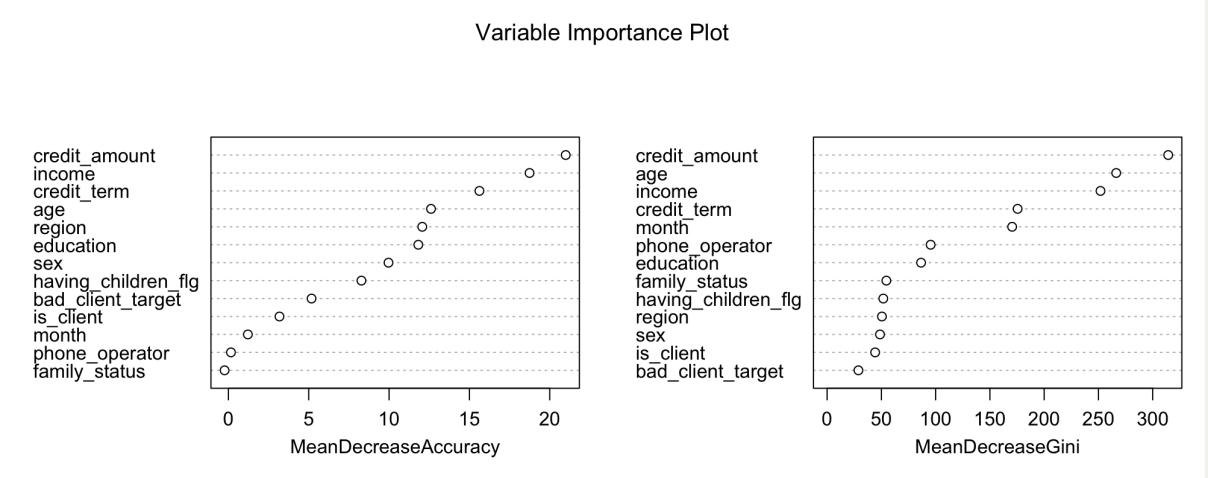


Fig 16. Variable Importance Plot

➢ Model V-VI-VII: Using Cost-Sensitive Training

i) Model developments:

We built three classification trees with the penalty for FN increasing.

*Term Presentation

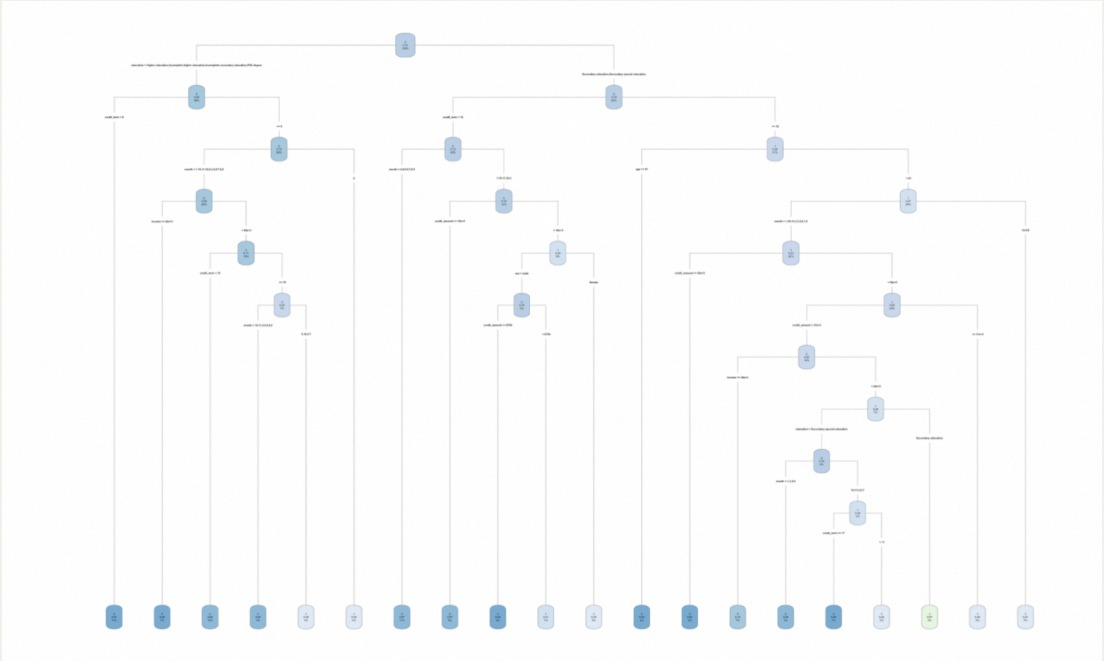
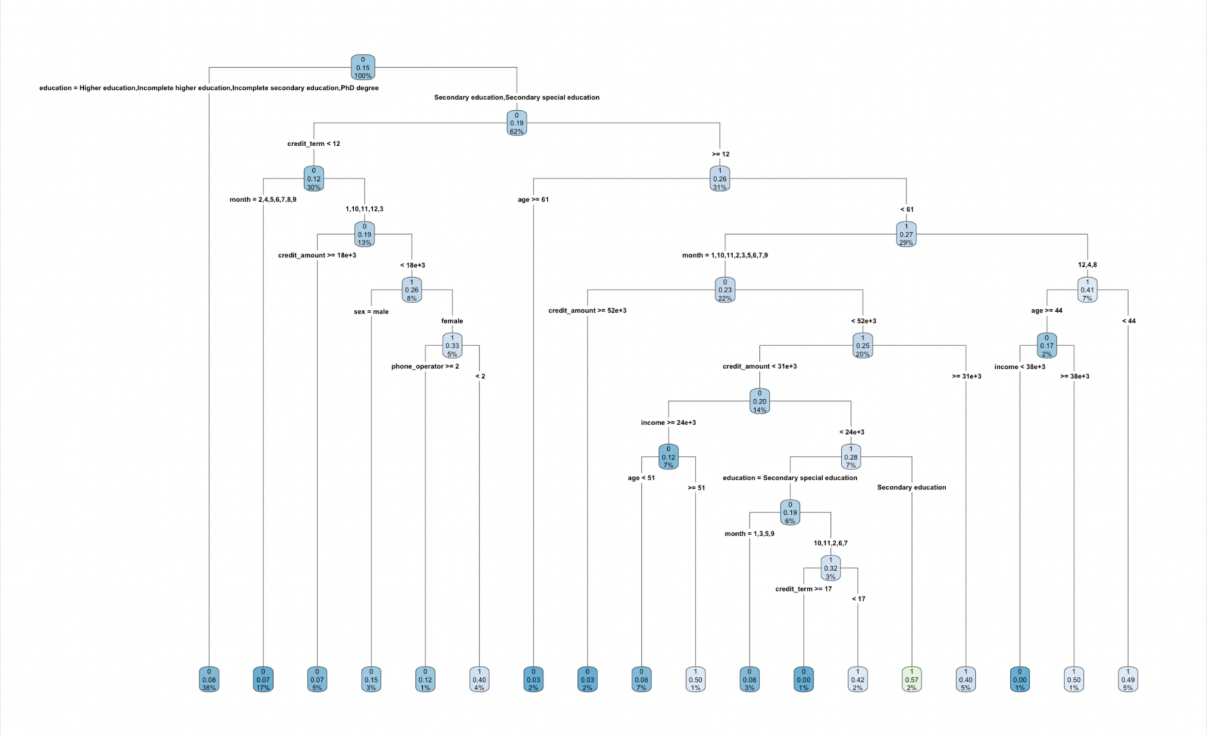


Fig 17. CART tree - Double-cost Model V

Double-cost cost matrix	Predicted	
Actual	0	1
0	0	**1**
1	**2**	0



*[Term Presentation]

Fig 18. CART tree - Triple-cost Model VI

Triple-cost cost matrix	Predicted	
Actual	0	1
0	0	1
1	3	0

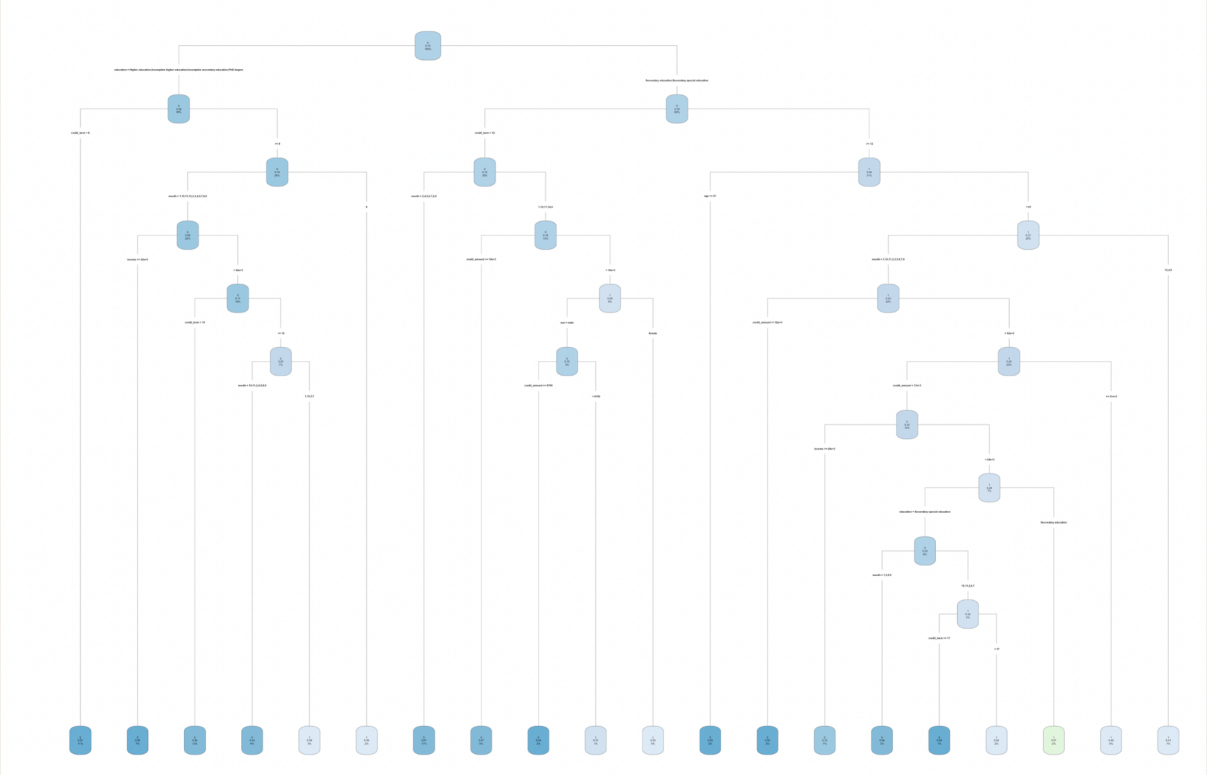


Fig 19. CART tree - Quadruple-cost Model VII

Quadruple-cost cost matrix	Predicted	
Actual	0	1
0	0	1
1	4	0

*[Term Presentation]

All three models have very complex splitting systems, with approximately 18-20 leaf nodes, taking into account a lot of variables: education, credit_amount, credit_term, month, age, sex, phone_operator, income, etc.

ii) Model evaluations

Model V: DOUBLE COST	Training Set	Testing Set
Accuracy	87.42%	80.78%
Precision	**59.47%**	**23.08%**
Recall	**46.91%**	**27.78%**

Model VI: TRIPLE COST	Training Set	Testing Set
Accuracy	83.23%	76.03%
Precision	**44.92%**	**21.21%**
Recall	**59.28%**	**38.89%**

Model VII: QUADRUPLE COST	Training Set	Testing Set
Accuracy	79.42%	69.11%
Precision	**39.44%**	**17.98%**
Recall	**73.19%**	**46.29%**

As the penalty increases, accuracy diminishes across the models. This decrease is the result of penalizing FN instead of keeping the normal cost of the model. However, precision and recall start to pick up. The highest precision among the three is 21.21%, and the highest recall is 46.29%. The model is getting better at detecting bad_client_target, but it comes with compromising overall accuracy.

However, it is worth noting that we do not want accuracy to drop below 70%. While recall increases as the penalty goes up, precision starts to drop at the quadruple model. It seems that in Model VII, the penalty on FN is too heavy on the overall effect. With those two criteria in mind, Model VI: Triple Cost is the best-performing model for the classification tree method.

**Term Presentation*

➢ Best-Performing Classification Tree Model

Model VI-Triple Cost tree is the tree that we pick as the best performing classification tree model.

Confusion Matrix for Test Set: Cost-Sensitivity Model VI versus Random Oversampling Model IV

Model VI	Predicted	
Actual	0	1
0	331	**78**
1	**33**	21

Model IV	Predicted	
Actual	0	1
0	395	14
1	**53**	**1**

The model is better at predicting true positive, or bad_client_target_1. FN has gone down, while TP has gone up. The downside is that FP is going up while TN is going down. This changes in the confusion matrix is in accordance to the changes in accuracy, precision, and recall.

```
node), split, n, loss, yval, (yprob)
      * denotes terminal node

   1) root 1312 582 0 (0.85213415 0.14786585)
     2) education=Higher education,Incomplete higher education,Incomplete secondary education,PhD degree 503
117 0 (0.92246521 0.07753479) *
     3) education=Secondary education,Secondary special education 809 465 0 (0.80840544 0.19159456)
       6) credit_term< 11.5 397 147 0 (0.87657431 0.12342569)
        12) month=2,4,5,6,7,8,9 220  45 0 (0.93181818 0.06818182) *
        13) month=1,10,11,12,3 177 102 0 (0.80790960 0.19209040)
          26) credit_amount>=17750 67  15 0 (0.92537313 0.07462687) *
          27) credit_amount< 17750 110  81 1 (0.73636364 0.26363636)
            54) sex=male 41  18 0 (0.85365854 0.14634146) *
            55) sex=female 69  46 1 (0.66666667 0.33333333)
             110) phone_operator>=1.5 17   6 0 (0.88235294 0.11764706) *
             111) phone_operator< 1.5 52  31 1 (0.59615385 0.40384615) *
       7) credit_term>=11.5 412 306 1 (0.74271845 0.25728155)
        14) age>=60.5 30   3 0 (0.96666667 0.03333333) *
        15) age< 60.5 382 277 1 (0.72513089 0.27486911)
          30) month=1,10,11,2,3,5,6,7,9 287 198 0 (0.77003484 0.22996516)
            60) credit_amount>=51500 30   3 0 (0.96666667 0.03333333) *
            61) credit_amount< 51500 257 192 1 (0.74708171 0.25291829)
             122) credit_amount< 30750 190 114 0 (0.80000000 0.20000000)
               244) income>=24500 94  33 0 (0.88297872 0.11702128)
                 488) age< 50.5 86  21 0 (0.91860465 0.08139535) *
                 489) age>=50.5 8   4 1 (0.50000000 0.50000000) *
               245) income< 24500 96  69 1 (0.71875000 0.28125000)
                 490) education=Secondary special education 73  42 0 (0.80821918 0.19178082)
                   980) month=1,3,5,9 39   9 0 (0.92307692 0.07692308) *
                   981) month=10,11,2,6,7 34  23 1 (0.67647059 0.32352941)
                    1962) credit_term>=16.5 8   0 0 (1.00000000 0.00000000) *
                    1963) credit_term< 16.5 26  15 1 (0.57692308 0.42307692) *
                 491) education=Secondary education 23  10 1 (0.43478261 0.56521739) *
             123) credit_amount>=30750 67  40 1 (0.59701493 0.40298507) *
          31) month=12,4,8 95  56 1 (0.58947368 0.41052632)
            62) age>=43.5 23  12 0 (0.82608696 0.17391304)
             124) income< 38500 15   0 0 (1.00000000 0.00000000) *
             125) income>=38500 8   4 1 (0.50000000 0.50000000) *
            63) age< 43.5 72  37 1 (0.51388889 0.48611111) *
```

Fig 20. CART tree - Triple-cost Model VI

**Term Presentation*

Using this decision tree, the bank can identify the potential profile of bad_client_target. Further background checks should be considered before these profiles can be qualified for a loan:

- ☐ Client who is: secondary education/ secondary special education, credit_term < 12, month = 1,10,11,12,3, credit_amount < 18,000, female, phone_operator <2
- ☐ Client who is: secondary education/ secondary special education, credit_term >= 12, 51 <= age < 61, month=1,10,11,2,3,5,6,7,9, credit_amount < 31,000, income >= 24,000
- ☐ Client who is: secondary special education, 12 <= credit_term < 17, 51 <= age < 61, month=10,11,2,6,7, credit_amount < 31,000, income < 24,000
- ☐ Client who is: secondary education, credit_term >= 12, 51 <= age < 61, month=1,10,11,2,3,5,6,7,9, credit_amount < 31,000, income < 24,000
- ☐ Client who is: secondary education/ secondary special education, credit_term >= 12, age < 61, month=1,10,11,2,3,5,6,7,9, 31,000 <= credit_amount < 52,000,
- ☐ Client who is: secondary education/ secondary special education, credit_term >= 12, 44 <= age < 61, month=12,4,8, income >= 38,000
- ☐ Client who is: secondary education/ secondary special education, credit_term >= 12, age < 44, month=12,4,8

True Positive Rate, or Recall is 38.8%. Out of all actual positive, bad_client_target_1, 38.8% of them is correctly identified by our model.

*[*Term Presentation*]

VI. METHOD 2: LOGISTIC REGRESSION MODEL

➢ Model I: Using Unbalanced Data Set

In our endeavor to refine the bank's loan eligibility assessment process, we embarked on the development of a second predictive model utilizing the original, unbalanced dataset. This approach was strategic, aimed at capturing the inherent complexities and real-world distribution of client profiles within our dataset. By choosing not to artificially balance the data through oversampling the minority class, we sought to understand the natural implications of our dataset's skewness on machine learning model performance. Through meticulous testing, training, and evaluation, our goal was to craft a logistic regression model that can accurately predict client loan eligibility, thus enabling more informed and nuanced decision-making in loan approval processes. This approach acknowledges the inherent challenges posed by unbalanced datasets in machine learning, particularly in the context of predicting loan eligibility, where the stakes of accurate classification are high.

Wald Test & Log-Odds Results

Code Output

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2071  -0.5357  -0.3806  -0.2656   2.6538

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                         -2.082e+00  6.349e-01  -3.279  0.00104 **
month                                3.058e-02  2.695e-02   1.135  0.25649
credit_amount                       -2.622e-06  4.998e-06  -0.525  0.59980
credit_term                          5.268e-02  1.627e-02   3.237  0.00121 **
age                                 -2.431e-02  8.231e-03  -2.954  0.00314 **
sexmale                             -4.786e-01  2.011e-01  -2.380  0.01732 *
educationIncomplete higher education -9.684e-01  7.597e-01  -1.275  0.20241
educationIncomplete secondary education -1.255e+01  7.066e+02  -0.018  0.98583
educationPhD degree                 -1.291e+01  8.234e+02  -0.016  0.98749
educationSecondary education         1.297e+00  3.105e-01   4.177 2.95e-05 ***
educationSecondary special education  7.351e-01  2.542e-01   2.892  0.00383 **
having_children_flg                 -2.189e-01  2.018e-01  -1.084  0.27815
region                              -6.014e-02  1.604e-01  -0.375  0.70776
income                              -1.433e-05  7.921e-06  -1.809  0.07045 .
family_statusMarried                 2.201e-01  2.070e-01   1.063  0.28759
family_statusUnmarried               2.212e-01  4.389e-01   0.504  0.61424
phone_operator                      -5.817e-02  9.166e-02  -0.635  0.52567
is_client                            5.338e-01  2.106e-01   2.534  0.01127 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 887.34  on 1259  degrees of freedom
Residual deviance: 802.07  on 1242  degrees of freedom
AIC: 838.07

Number of Fisher Scoring iterations: 14
```

	Δ.Log.Odds.Ratio	Δ.Odds.Ratio	Percent.Δ.Odds
(Intercept)	-2.0817	0.1247	-0.8753
month	0.0306	1.0311	0.0311
credit_amount	0.0000	1.0000	0.0000
credit_term	0.0527	1.0541	0.0541
age	-0.0243	0.9760	-0.0240
sexmale	-0.4786	0.6196	-0.3804
educationIncomplete higher education	-0.9684	0.3797	-0.6203
educationIncomplete secondary education	-12.5457	0.0000	-1.0000
educationPhD degree	-12.9065	0.0000	-1.0000
educationSecondary education	1.2973	3.6594	2.6594
educationSecondary special education	0.7351	2.0858	1.0858
having_children_flg	-0.2189	0.8034	-0.1966
region	-0.0601	0.9416	-0.0584
income	0.0000	1.0000	0.0000
family_statusMarried	0.2201	1.2462	0.2462
family_statusUnmarried	0.2212	1.2476	0.2476
phone_operator	-0.0582	0.9435	-0.0565
is_client	0.5338	1.7055	0.7055

*[*Term Presentation*]

Code Interpretation

Variable	Interpretation
month	• **Log-Odds:** A coefficient of 0.0306 suggests a minor increase in the log-odds of being a bad client with each passing month. • **Odds Ratio:** The odds increase by a factor of 1.0311 for each month. • **Significance:** Not significant ($p > 0.05$).
credit_amount	• **Log-Odds:** The coefficient is practically zero, indicating no direct relation to being a bad client. • **Odds Ratio:** The odds ratio of 1.0000 suggests no change. • **Significance:** Not significant ($p > 0.05$).
credit_term	• **Log-Odds**: A coefficient of 0.0527 indicates a slight increase in the log-odds of being a bad client for longer credit terms. • **Odds Ratio:** The odds of being a bad client increase by a factor of 1.0541 with each unit increase in the credit term. • **Significance:** Not significant ($p > 0.05$).
age	• **Log-Odds**: Each additional year decreases the log-odds of being a bad client by 0.0243. • **Odds Ratio:** The odds of being a bad client decrease by 0.9760 for each additional year. • **Significance:** Significant ($p < 0.05$).
sex (male)	• **Log-Odds:** Being male decreases the log-odds of being a bad client by 0.4786 compared to females. • **Odds Ratio:** Males have 0.6196 times the odds of being a bad client compared to females. • **Significance**: Significant ($p < 0.05$).
education	• **Incomplete Higher Education:** Log-odds of -0.9684; not significant. • **Incomplete Secondary Education:** Log-odds of -12.5457; indicates no instances or a zero-count for bad clients in this category. • **PhD Degree:** Log-odds of -12.9065; similar interpretation as incomplete secondary education. • **Secondary Education:** Log-odds of 1.2973; highly significant ($p < 0.01$) with an odds ratio of 3.6594. • **Secondary Special Education:** Log-odds of 0.7351; significant ($p < 0.05$) with an odds ratio of 2.0858.
having_children_flg	• **Log-Odds:** -0.2189; indicates that having children is associated with a slight decrease in the log-odds of being a bad client. • **Odds Ratio:** 0.8034; slightly lower odds for those with children, but not significant.
region	• **Log-Odds:** -0.0601; a minor decrease in log-odds associated with different regions. • **Odds Ratio:** 0.9416; slight, non-significant decrease in odds.

*[*Term Presentation*]

income	• **Log-Odds:** Practically zero, indicating no significant effect on being a bad client. • **Odds Ratio:** The odds are unchanged across different income levels.
family_status	• **Married (family_statusMarried):** Log-odds of 0.2201; not significant. • **Unmarried (family_statusUnmarried):** Log-odds of 0.2212; not significant.
phone_operator	• **Log-Odds:** -0.0582; suggests that being with a specific phone operator is associated with a slight decrease in the log-odds of being a bad client, but not significantly.
is_client	• **Log-Odds:** The log-odds of being a bad client increases by 0.5338 if the person is already a client. • **Odds Ratio:** Clients are 1.7055 times more likely to be classified as bad clients than non-clients. • **Significance:** Significant ($p < 0.05$). *_**NOTE**_: It's important to acknowledge that 61% of responses are non clients (1), while 39% of responses are from existing clients (0). This, in effect, raises the log-odds of is_client.

Null Deviance	Residual Deviance
1099.43	995.96
The reduction suggests that adding the predictor variables into the model **significantly improves** its fit compared to the null model.	

Akaike Information Criterion (AIC)
838.07
Meaningless without comparison.

The logistic regression's Wald test and log-odds results reveal intricate patterns within the variables affecting client loan eligibility. The model scrutinizes various predictors, including demographic, financial, and client history variables, to determine their impact on the likelihood of being classified as a "bad" client. Notably, significant predictors such as age, sex (male), and client status (is_client) emerge with statistical significance, suggesting these factors are crucial in assessing loan eligibility. For instance, older clients and males show a

*[*Term Presentation*]

reduced risk of default, indicated by negative coefficients, while existing clients present a higher risk, as evidenced by their positive coefficients and significance levels.

The evaluation of model fit through null and residual deviance indicates that the inclusion of these variables significantly improves model accuracy over a null model, which does not consider these predictors. This improvement is crucial for predictive accuracy but must be weighed against the model's complexity, as indicated by the Akaike Information Criterion (AIC). Although the AIC provides a measure for model comparison, its value alone does not dictate the final model choice but rather emphasizes the need for a balanced approach between model complexity and predictive power.

Performance Metrics

Training Set

	Predicted	
Actual	0	1
0	1,117	1
1	142	0

Testing Set

	Predicted	
Actual	0	1
0	408	54
1	1	0

	Training Set	Testing Set
Accuracy	**88.65%**	**88.12%**
Precision	**0%** (No TPs)	**0%** (No TPs)
Recall	**0%** (Failed to identify any actual bad clients)	**0%** (Failed to identify any actual bad clients)
Specificity	**99.91%**	**88.31%**
F1 Score	**NA** (Due to 0% precision & recall)	**NA** (Due to 0% precision & recall)

Training the model on a 75% subset of the data and then evaluating it on the remaining 25% aimed to ensure robustness and generalizability. The training set results highlight a high overall accuracy (88.65%) but reveal a critical shortfall in the model's

*Term Presentation

precision and recall, both registering at 0%. This indicates the model's inability to correctly identify bad clients, a significant concern given the bank's emphasis on minimizing false negatives due to their higher theoretical cost compared to false positives. This model's tendency to err on the side of caution (high specificity at 99.91%) aligns with the bank's conservative stance but does not capitalize on the opportunity to identify all potential risks accurately.

Applying the model to the testing set mirrors this pattern, with accuracy slightly decreasing to 88.12% and specificity remaining high. The consistency of these metrics between training and testing sets confirms the model's stability across different data samples. However, the persistent lack of precision and recall underscores the model's limitation in detecting bad clients, highlighting a gap between the model's theoretical design and its practical application.

This analysis underscores the challenge of balancing the desire for high model accuracy with the practical need to identify bad clients effectively, especially under the bank's operational paradigm where false negatives carry a heavier penalty than false positives. The current model's inability to address this balance points to the necessity for further model refinement that can better accommodate the bank's risk tolerance and operational objectives. The ultimate goal remains to enhance the predictive model to a point where it not only achieves high accuracy but also aligns more closely with the bank's strategic emphasis on minimizing the risk of undetected bad loans.

VS

**Term Presentation*

Training Set vs Testing Set Comparison Insights

Consistency in Accuracy: The model shows a slight decrease in accuracy from the training set to the test set, which is expected as models generally perform slightly worse on unseen data. However, the consistency in high accuracy between the sets indicates that the model is stable across different data samples.

Issue with Predicting Positive Cases: Both the training and test sets reveal a significant issue in the model's ability to predict positive cases (bad clients). The precision and recall values of 0 in both datasets highlight a model skewed towards predicting the majority class (good clients), likely due to the imbalanced nature of the dataset.

Specificity: The model maintains high specificity across both sets, although slightly lower in the test set. This indicates a consistent ability to identify true negatives, which, while useful, emphasizes the model's imbalance towards predicting clients as not being bad.

Model Generalization: The similarity in performance metrics (accuracy, precision, recall, specificity) between the training and test sets suggests that the model generalizes well to unseen data, at least in terms of predicting negative outcomes. The lack of overfitting to the training data is a positive sign of model robustness.

Need for Improvement: The consistent inability to predict positive cases across both datasets underscores the need for model improvement. Strategies might include addressing the class imbalance, or incorporating different predictors that could help in identifying positive cases more accurately.

➢ Model II: Using Balanced Data Set

In addressing the challenge of loan eligibility prediction, we then developed a logistic regression model using an oversampled, more balanced dataset, aiming to mitigate the inherent biases of the class imbalance.

*Term Presentation

Wald Test & Log-Odds Result

Code Output

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3242  -0.6131  -0.4375  -0.2996   2.4618

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                         -2.011e+00  5.566e-01  -3.614 0.000302 ***
month                                2.758e-02  2.352e-02   1.173 0.240943
credit_amount                       -3.489e-06  4.366e-06  -0.799 0.424205
credit_term                          5.955e-02  1.410e-02   4.223 2.41e-05 ***
age                                 -2.074e-02  7.109e-03  -2.918 0.003524 **
sexmale                             -4.523e-01  1.768e-01  -2.558 0.010534 *
educationIncomplete higher education -8.605e-01  6.295e-01  -1.367 0.171646
educationIncomplete secondary education -1.285e+01  7.064e+02  -0.018 0.985489
educationPhD degree                 -1.323e+01  8.199e+02  -0.016 0.987121
educationSecondary education         1.252e+00  2.716e-01   4.611 4.01e-06 ***
educationSecondary special education 7.035e-01  2.176e-01   3.232 0.001229 **
having_children_flg                 -1.241e-01  1.747e-01  -0.710 0.477587
region                              -3.415e-02  1.408e-01  -0.242 0.808393
income                              -1.288e-05  6.790e-06  -1.897 0.057867 .
family_statusMarried                 2.257e-01  1.825e-01   1.236 0.216310
family_statusUnmarried               4.110e-01  3.586e-01   1.146 0.251731
phone_operator                      -7.076e-02  8.137e-02  -0.870 0.384523
is_client                            5.012e-01  1.822e-01   2.751 0.005948 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1099.43  on 1311  degrees of freedom
Residual deviance:  995.96  on 1294  degrees of freedom
AIC: 1032

Number of Fisher Scoring iterations: 14
```

	Δ.Log.Odds.Ratio	Δ.Odds.Ratio	Percent.Δ.Odds
(Intercept)	-2.0112	0.1338	-0.8662
month	0.0276	1.0280	0.0280
credit_amount	0.0000	1.0000	0.0000
credit_term	0.0595	1.0614	0.0614
age	-0.0207	0.9795	-0.0205
sexmale	-0.4523	0.6362	-0.3638
educationIncomplete higher education	-0.8605	0.4230	-0.5770
educationIncomplete secondary education	-12.8476	0.0000	-1.0000
educationPhD degree	-13.2344	0.0000	-1.0000
educationSecondary education	1.2525	3.4989	2.4989
educationSecondary special education	0.7035	2.0207	1.0207
having_children_flg	-0.1241	0.8833	-0.1167
region	-0.0342	0.9664	-0.0336
income	0.0000	1.0000	0.0000
family_statusMarried	0.2257	1.2532	0.2532
family_statusUnmarried	0.4110	1.5083	0.5083
phone_operator	-0.0708	0.9317	-0.0683
is_client	0.5012	1.6506	0.6506

Code Interpretation

Variable	Interpretation
month	• **Log-Odds:** A coefficient of 0.02758 suggests a slight increase in the log-odds of being a bad client as the month number increases. • **Odds Ratio:** The odds increase by a factor of 1.0280 for each increment in month. • **Significance:** Not statistically significant (p > 0.05).
credit_amount	• **Log-Odds**: The coefficient is nearly zero, indicating no association with the log-odds of being a bad client. • **Odds Ratio:** An odds ratio of 1.0000 suggests no change in odds with different credit amounts. • **Significance:** Not statistically significant (p > 0.05).
credit_term	• **Log-Odds:** Each unit increase in credit term increases the log-odds of being a bad client by 0.0595. • **Odds Ratio:** For each unit increase, the odds of being a bad client increase by

**Term Presentation*

	1.0614. • **Significance:** Not statistically significant ($p > 0.05$).
age	• **Log-Odds:** Each year increase in age decreases the log-odds of being a bad client by 0.0274. • **Odds Ratio:** Each additional year of age decreases the odds of being a bad client by a factor of 0.9735. • **Significance:** Significant ($p < 0.05$).
sex (male)	• **Log-Odds**: Being male decreases the log-odds of being a bad client by 0.4523 compared to females. • **Odds Ratio:** Males have 0.6362 times the odds of being a bad client compared to females. • **Significance**: Significant ($p < 0.05$).
education	• **Incomplete Higher Education:** Log-Odds of -0.8605; not significant. • **Incomplete Secondary Education:** Log-Odds of -12.8476; no effect likely due to no observations in this category. • **PhD Degree:** Log-Odds of -13.2344; similarly likely no observations in this category. • **Secondary Education:** Log-Odds of 1.2525; significant ($p < 0.01$) with an odds ratio of 3.4989, indicating higher odds of being a bad client. • **Secondary Special Education:** Log-Odds of 0.7035; significant ($p < 0.05$) with an odds ratio of 2.0207.
having_children_flg	• **Log-Odds:** -0.1241; not significant. • **Odds Ratio**: 0.8833; suggests having children slightly decreases the odds of being a bad client, but not significantly so.
region	• **Log-Odds:** -0.0342; not significant. • **Odds Ratio**: 0.9664; a slight, non-significant decrease in odds associated with different regions.
income	• **Log-Odds**: -0.00012885; indicates that higher income slightly decreases the log-odds of being a bad client. • **Odds Ratio:** 1.0000; suggests no significant change in odds with income.
family_status	• **Married (family_statusMarried):** Log-Odds of 0.2257; not significant. • **Unmarried (family_statusUnmarried):** Log-Odds of 0.4110; not significant.
phone_operator	• **Log-Odds:** -0.07076; indicates that being with a specific phone operator decreases the log-odds of being a bad client, but not significantly.
is_client	• **Log-Odds:** 0.5012; suggests being a client increases the log-odds of being a bad client. • **Significance:** With a p-value of 0.005948, this is statistically significant, suggesting that current clients are more likely to be seen as bad credit risks.

*[*Term Presentation*]

Null Deviance	Residual Deviance
887.34	802.07
The decrease upon adding the predictors suggests that the model with predictors fits the data **significantly better** than a model without any predictors.	

Akaike Information Criterion (AIC)
1032
While the AIC has increased from model I (838.07)- meaning that model I does a better job at explaining the variance in the data- the context of the objective is crucial to remember. Model II's approach to addressing class imbalance might still offer value despite its higher AIC, as it shows a slightly improved ability to detect positive cases, which is crucial for the task at hand.

The model, trained on a balanced dataset, evaluates various predictors including age, sex, education level, and other factors to estimate the odds of a client being a poor credit risk. Notably, variables such as age and sex (male) were found to have significant effects on the log-odds of being a bad client, with age decreasing the risk and being male significantly reducing the odds compared to females. Education levels, specifically secondary and secondary special education, significantly increased the odds of being a bad client, pointing to the impact of education on creditworthiness. Other variables like month, credit amount, and credit term, though analyzed, did not show statistically significant effects on the likelihood of being a bad client.

Performance Metrics

Training Set

	Predicted	
Actual	0	1
0	1,115	3
1	190	4

*Term Presentation

Testing Set

	Predicted	
Actual	0	1
0	362	50
1	1	1

	Training Set	Testing Set
Accuracy	**85.29%**	**87.61%**
Precision	**57.14%**	**1.85%**
Recall	**2.06%**	**50%**
Specificity	**99.73%**	**87.78%**
F1 Score	**3.98%**	**3.57%**

The logistic regression model's performance metrics reveal a divergence in accuracy and precision between the training set (75% of the dataset) and the test set (25% of the dataset). Both sets show high accuracy and specificity, indicating the model's effectiveness in identifying good clients. However, there's a notable struggle with low precision and F1 score in predicting bad clients, which is particularly exacerbated in the test set where precision significantly drops, indicating a high rate of false positives. This discrepancy highlights the model's conservative approach, prioritizing minimizing false positives at the expense of failing to identify a considerable number of bad clients, evidenced by the minimal number of true positives and a considerable amount of false negatives, especially in the training set.

The improved recall in the testing set suggests the model may have a slightly better ability to identify bad clients in new, unseen data, despite a decrease in precision leading to more false positives. This trade-off underscores the challenge of balancing the need to accurately identify bad clients without overly penalizing potentially good clients, a critical consideration for practical applications in the banking sector. The model's higher AIC in comparison to its predecessor indicates a less parsimonious model, yet the context of improving detection of positive cases justifies the model's utility given the project's objectives.

To conclude, while this logistic regression model demonstrates a strong capability to accurately predict non-bad clients, its limited success in correctly identifying bad clients, despite better than the previous logistic regression model, underscores the need for further model optimization. This involves refining the model's sensitivity to false negatives, which are considered most costly in this context.

*Term Presentation

VS

Training Set vs Testing Set Comparison Insights

Consistency in Predicting Non-bad Clients: Both the training and testing set demonstrate the model's strong capability to accurately predict non-bad clients, as indicated by the large number of true negatives. This suggests the model is highly conservative, prioritizing the minimization of false positives.

Struggle with Identifying Bad Clients: Despite the model's high accuracy in predicting true negatives, it struggles significantly to correctly identify bad clients across both sets. This is evidenced by the minimal number of true positives and a considerable amount of false negatives, particularly in the training set.

Improved Recall in the Testing Set: While the training set showed extremely low recall, indicating almost no correct predictions of bad clients, the testing set exhibited improved recall. This suggests that while the model may perform poorly in recognizing bad clients within the data it was trained on, it could potentially identify them slightly better in unseen data. However, this improvement in recall comes with a decrease in precision, leading to more false positives.

Low Precision and F1 Score: Both sets suffer from low precision and F1 scores, but the testing set, in particular, showcases a slight improvement in the F1 score. This increment

*[*Term Presentation*]

indicates a better balance between precision and recall when the model is applied to new data, although the overall effectiveness in predicting bad clients remains poor.

Generalization to New Data: The slight differences in performance metrics between the training and testing sets hint at the model's ability to generalize to new data. The improved recall in the testing set suggests that the model, despite its limitations, might still capture relevant patterns that apply to unseen data. However, the significant drop in precision indicates a need for careful consideration of the trade-offs between identifying more true positives and the cost of increased false positives.

Need for Model Improvement or Alternative Approaches: The comparison underscores the need for model improvement or exploring alternative modeling approaches. The current model's conservative nature towards predicting non-bad clients does not align well with the objective of effectively identifying bad clients.

➢ Variable Selection: Stepwise Regression

The stepwise regression analysis using the original, unbalanced dataset has provided critical insights for developing predictive models to assess loan eligibility for a banking institution. This approach meticulously selected variables based on their significance in predicting a client's loan eligibility, with the goal of minimizing the Akaike Information Criterion (AIC) to ensure the model's predictive accuracy without overfitting.

```
Call:
glm(formula = bad_client_target ~ credit_term + age + sex + education +
    income + family_status + is_client, family = binomial(),
    data = clients)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0783  -0.5424  -0.3902  -0.2820   2.5923

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                           -2.192e+00  4.201e-01  -5.217 1.81e-07 ***
credit_term                            4.444e-02  1.130e-02   3.933 8.38e-05 ***
age                                   -2.813e-02  6.982e-03  -4.029 5.60e-05 ***
sexmale                               -4.603e-01  1.671e-01  -2.755 0.005872 **
educationIncomplete higher education  -3.289e-01  5.012e-01  -0.656 0.511775
educationIncomplete secondary education -1.249e+01  6.411e+02  -0.019 0.984461
educationPhD degree                   -1.291e+01  8.186e+02  -0.016 0.987412
educationSecondary education           1.206e+00  2.635e-01   4.576 4.74e-06 ***
educationSecondary special education   7.591e-01  2.134e-01   3.557 0.000375 ***
income                                -8.623e-06  5.674e-06  -1.520 0.128598
family_statusMarried                   3.694e-01  1.724e-01   2.142 0.032167 *
family_statusUnmarried                 1.642e-01  3.802e-01   0.432 0.665792
is_client                              5.397e-01  1.761e-01   3.066 0.002173 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1220.9  on 1722  degrees of freedom
Residual deviance: 1118.8  on 1710  degrees of freedom
AIC: 1144.8

Number of Fisher Scoring iterations: 14
```

The final model, as a result of stepwise regression, incorporates variables such as 'credit_term', 'age', 'sex', 'education', 'income', 'family_status', and 'is_client'. These variables have been identified as pivotal in assessing a client's eligibility for a loan, suggesting a robust framework for understanding the factors that influence loan approval decisions. The exclusion of variables like 'region' and 'credit_amount' early in the iterations indicates their

limited impact on the model's ability to predict bad clients, highlighting the effectiveness of stepwise regression in refining the model by removing non-contributory predictors.

Key Observations from the Stepwise Regression Output

Region and Credit Amount: Early iterations suggest the removal of variables like 'region' and 'credit_amount,' indicating they may not significantly impact the model's ability to predict bad clients.

Significant Variables: Variables like 'credit_term,' 'age,' 'sex,' 'education,' 'income,' 'family_status,' and 'is_client' remain in the model, suggesting they have significant predictive power regarding client's loan eligibility.

Education Levels: Different levels of education ('Incomplete higher education,' 'PhD degree,' 'Secondary education,' 'Secondary special education') are considered, with some levels showing more significance than others in predicting bad clients.

Marital Status: Variables like 'family_statusMarried' and 'family_statusUnmarried' indicate that family status has a role in loan eligibility prediction, with different statuses contributing differently to the risk assessment.

➢ Model III: Using Cost Sensitive Training

This approach, which incorporates cost-sensitive training to address the class imbalance, is particularly valuable in this scenario, where the cost of false negatives (not identifying a bad client) is higher than the cost of false positives (incorrectly identifying a good client as bad). This model prioritizes the correct identification of bad clients over the misclassification of good ones by increasing the cost of false negatives. This approach is particularly relevant in the banking sector, where failing to identify a bad client can have dire financial consequences. By increasing the weight of bad client instances (bad_client_target == 1), the model is encouraged to pay more attention to correctly identifying bad clients over good ones. Below are the double cost and triple cost models. To reiterate, the assumptions are as followed:

- ☐ Double Cost: FN is double the cost of FP
- ☐ Triple Cost: FN is triple the cost of FP

DOUBLE THE COST

[Wald Test & Log-Odds Results]

Code Output

```
Call:
glm(formula = bad_client_target ~ credit_term + age + sex + education +
    income + family_status + is_client, family = binomial(link = "logit"),
    data = clients_train3, weights = clients_train3$weights)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5934  -0.8283  -0.6065  -0.4181   3.0316

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -1.258e+00  3.289e-01  -3.825 0.000131 ***
credit_term                        5.662e-02  9.184e-03   6.165 7.03e-10 ***
age                               -2.175e-02  5.282e-03  -4.118 3.82e-05 ***
sexmale                           -5.114e-01  1.339e-01  -3.819 0.000134 ***
educationIncomplete higher education -8.211e-01  4.588e-01  -1.790 0.073586 .
educationIncomplete secondary education -1.254e+01  4.237e+02  -0.030 0.976385
educationPhD degree               -1.289e+01  4.898e+02  -0.026 0.979006
educationSecondary education       1.256e+00  2.104e-01   5.970 2.37e-09 ***
educationSecondary special education  7.284e-01  1.625e-01   4.481 7.41e-06 ***
income                            -1.474e-05  4.598e-06  -3.207 0.001343 **
family_statusMarried               2.428e-01  1.412e-01   1.719 0.085604 .
family_statusUnmarried             3.884e-01  2.837e-01   1.369 0.170960
is_client                          4.695e-01  1.372e-01   3.422 0.000621 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1718.6  on 1311  degrees of freedom
Residual deviance: 1542.8  on 1299  degrees of freedom
AIC: 1568.8

Number of Fisher Scoring iterations: 13
```

	Δ.Log.Odds.Ratio	Δ.Odds.Ratio	Percent.Δ.Odds
(Intercept)	-1.2580	0.2842	-0.7158
credit_term	0.0566	1.0583	0.0583
age	-0.0218	0.9785	-0.0215
sexmale	-0.5114	0.5997	-0.4003
educationIncomplete higher education	-0.8211	0.4399	-0.5601
educationIncomplete secondary education	-12.5430	0.0000	-1.0000
educationPhD degree	-12.8901	0.0000	-1.0000
educationSecondary education	1.2559	3.5111	2.5111
educationSecondary special education	0.7284	2.0718	1.0718
income	0.0000	1.0000	0.0000
family_statusMarried	0.2428	1.2748	0.2748
family_statusUnmarried	0.3884	1.4746	0.4746
is_client	0.4695	1.5992	0.5992

Code Interpretation

Variable	Interpretation
credit_term	• **Log-Odds:** A one-unit increase in the credit term increases the log-odds of being a bad client by 0.0566. • **Odds Ratio:** The odds of being a bad client increase by a factor of 1.0583 for each unit increase in credit term. • **Significance:** Credit term is statistically significant ($p < 0.05$).
age	• **Log-Odds:** Each additional year of age decreases the log-odds of being a bad client by 0.0218. • **Odds Ratio:** The odds of being a bad client decrease by a factor of 0.9785 for

**Term Presentation*

	each additional year of age. • **Significance:** Age is statistically significant ($p < 0.001$).
sex (male)	• **Log-Odds:** Being male decreases the log-odds of being a bad client by 0.5114 compared to females. • **Odds Ratio:** Males have 0.5997 times the odds of being a bad client compared to females. • **Significance:** Gender is statistically significant ($p < 0.001$).
education	• **Incomplete Higher Education:** Decreases the log-odds of being a bad client by 0.8211 with an odds ratio of 0.4399, which is not statistically significant. • **Incomplete Secondary Education:** The large negative coefficient and an odds ratio of 0 indicate that there were likely no observations or no bad clients in this category. Therefore, it's not statistically significant. • **PhD Degree**: Similar to incomplete secondary education, an odds ratio of 0 and a large negative coefficient indicate no observations or a zero-count for bad clients with PhDs. • **Secondary Education:** Increases the log-odds of being a bad client by 1.2559 with an odds ratio of 3.5111, which is statistically significant ($p < 0.001$). • **Secondary Special Education:** Increases the log-odds of being a bad client by 0.7284 with an odds ratio of 2.0718, which is statistically significant ($p < 0.05$).
income	• **Log-Odds**: The coefficient is -1.474e-05, indicating that an increase in income slightly decreases the log-odds of being a bad client. • **Odds Ratio**: An odds ratio of 1.000 (when rounded) indicates almost no change in odds per unit increase in income. • **Significance:** Income is statistically significant ($p < 0.01$).
family_status	• **Married (family_statusMarried):** Increases the log-odds of being a bad client by 0.2428 with an odds ratio of 1.2748, but it's not statistically significant. • **Unmarried (family_statusUnmarried)**: Increases the log-odds of being a bad client by 0.3884 with an odds ratio of 1.4746, but it's not statistically significant.
is_client	• **Log-Odds:** Being a client increases the log-odds of being a bad client by 0.4695. • **Odds Ratio:** Clients have 1.5992 times the odds of being bad clients compared to non-clients. • **Significance:** The 'is_client' variable is statistically significant ($p < 0.001$).

Null Deviance	Residual Deviance
1718.6	1542.8

The reduction suggests that adding the predictor variables into the model **significantly improves** its fit compared to the null model.

*[*Term Presentation*]

Akaike Information Criterion (AIC)
1568.8
Meaningless without comparison.

The results of the Wald test and log-odds ratios from this logistic regression model offer insightful observations into factors influencing client eligibility for loans. Notably, variables such as 'credit_term', 'age', 'sex (male)', 'education', 'income', 'family_status', and 'is_client' have shown statistical significance, shedding light on their importance in predicting loan eligibility. For instance, an increase in 'credit_term' enhances the odds of being labeled a bad client, suggesting a higher risk associated with longer loan terms. Conversely, age and being male decrease the likelihood of being considered a bad client, indicating these demographics are viewed as lower risk. Education level also plays a crucial role, with secondary education significantly increasing the risk of being a bad client, contrasting with higher education levels which do not show significant effects due to lack of observations or the absence of bad clients in those categories. The 'income' variable, while statistically significant, shows a minimal impact on the odds, and 'is_client' status significantly increases the risk of being classified as bad, pointing to previous client relationships as a risk factor.

Performance Metrics

Training Set

	Predicted	
Actual	0	1
0	1,061	57
1	167	27

Testing Set

	Predicted	
Actual	0	1
0	362	44
1	14	9

*Term Presentation

	Training Set	Testing Set
Accuracy	**82.93%**	**83.68%**
Precision	**32.14%**	**18.18%**
Recall	**13.92%**	**27.78%**
Specificity	**94.90%**	**88.72%**
F1 Score	**19.42%**	**21.98%**

Performance metrics on the training set demonstrate an overall accuracy of 82.93%, with a notable emphasis on specificity (94.90%) over recall (13.92%), indicating the model's strength in correctly identifying good clients but at the cost of missing a significant number of bad clients. However, when this model is applied to the testing set, a slight improvement in accuracy to 83.68% is observed, alongside an increase in recall to 27.78% and a decrease in precision to 18.18%. This shift suggests the model becomes slightly better at identifying bad clients it previously missed but at the cost of a higher rate of false positives. The specificity remains high across both sets, maintaining the model's ability to identify good clients accurately.

This comparative analysis reveals that while the logistic regression model shows a promising increase in accuracy and recall in the testing set, the trade-off between recall and precision becomes more pronounced. The model's enhanced sensitivity to bad clients is achieved at the expense of falsely identifying more good clients as bad, reflecting the cost-sensitive training's influence. Despite these challenges, the slight improvement in the F1 score in the testing set indicates a better balance between precision and recall, albeit both metrics remain relatively low. This outcome underscores the complexity of achieving high model performance while also addressing the bank's strategic priorities of minimizing false negatives.

Training Set vs Testing Set Comparison Insights

Consistency in Predicting Non-bad Clients: Both the training and testing set underscore the model's robust ability to accurately identify non-bad clients, evidenced by a high number of true negatives. This indicates the model's conservative nature, prioritizing the avoidance of false positives. Such consistency suggests that the model is reliable in identifying clients who are less likely to default, thus minimizing the risk of financial losses due to incorrect loan approvals.

Struggle with Identifying Bad Clients: Across both sets, the model shows a notable difficulty in accurately pinpointing bad clients, as shown by the low number of true positives and a high volume of false negatives, especially pronounced in the training set. This struggle highlights the model's limitations in detecting potentially risky loan applicants, which is a critical area for improvement, considering the bank's emphasis on minimizing false negatives due to their higher cost implication compared to false positives.

*Term Presentation

Improved Recall in the Testing Set: While the model demonstrates a struggle in identifying bad clients in the training set, there is a noticeable improvement in recall in the testing set. This suggests that the model becomes slightly better at catching bad clients that it previously missed, a positive development towards reducing the number of false negatives. However, this improvement in recall comes with a trade-off, as discussed below.

VS

Precision and Recall Trade-off: The testing set shows a significant decline in precision, indicating a higher rate of false positives—good clients incorrectly classified as bad. This trade-off between recall and precision is a common challenge in predictive modeling, especially in scenarios where the cost of misclassification varies between classes. The model's increased sensitivity to identifying bad clients in the testing set, therefore, comes at the cost of accuracy in predicting good clients.

Slight Improvement in Overall Accuracy and F1 Score: The testing set shows a minor improvement in overall accuracy and a slightly better F1 score compared to the training set. This indicates a somewhat better balance between precision and recall in the testing set, suggesting that the model, while not without its flaws, has a degree of generalizability to unseen data.

*[*Term Presentation*]

TRIPLE THE COST

Wald Test & Log-Odds Results

Code Output

```
Call:
glm(formula = bad_client_target ~ credit_term + age + sex + education +
    income + family_status + is_client, family = binomial(link = "logit"),
    data = clients_train4, weights = clients_train4$weights)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8059  -0.9791  -0.7264  -0.5013   3.4318

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                           -8.741e-01  2.867e-01  -3.049 0.002297 **
credit_term                            5.974e-02  8.390e-03   7.121 1.07e-12 ***
age                                   -2.162e-02  4.617e-03  -4.684 2.82e-06 ***
sexmale                               -5.253e-01  1.176e-01  -4.465 8.00e-06 ***
educationIncomplete higher education  -8.264e-01  3.865e-01  -2.138 0.032517 *
educationIncomplete secondary education -1.293e+01  4.286e+02  -0.031 0.975477
educationPhD degree                   -1.327e+01  4.858e+02  -0.027 0.978209
educationSecondary education           1.255e+00  1.853e-01   6.769 1.29e-11 ***
educationSecondary special education   7.398e-01  1.402e-01   5.276 1.32e-07 ***
income                                -1.548e-05  4.002e-06  -3.869 0.000109 ***
family_statusMarried                   2.597e-01  1.242e-01   2.090 0.036615 *
family_statusUnmarried                 3.770e-01  2.525e-01   1.493 0.135408
is_client                              4.668e-01  1.190e-01   3.922 8.77e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2184.8  on 1311  degrees of freedom
Residual deviance: 1948.7  on 1299  degrees of freedom
AIC: 1974.7

Number of Fisher Scoring iterations: 13
```

	Δ.Log.Odds.Ratio	Δ.Odds.Ratio	Percent.Δ.Odds
(Intercept)	-0.8741	0.4173	-0.5827
credit_term	0.0597	1.0616	0.0616
age	-0.0216	0.9786	-0.0214
sexmale	-0.5253	0.5914	-0.4086
educationIncomplete higher education	-0.8264	0.4376	-0.5624
educationIncomplete secondary education	-12.9289	0.0000	-1.0000
educationPhD degree	-13.2701	0.0000	-1.0000
educationSecondary education	1.2546	3.5063	2.5063
educationSecondary special education	0.7398	2.0955	1.0955
income	0.0000	1.0000	0.0000
family_statusMarried	0.2597	1.2965	0.2965
family_statusUnmarried	0.3770	1.4579	0.4579
is_client	0.4668	1.5949	0.5949

Code Interpretation

Variable	Interpretation
credit_term	• **Log-Odds:** A one-unit increase in credit term is associated with a 0.0597 increase in the log-odds of being a bad client. • **Odds Ratio:** The odds of being a bad client are 1.0616 times higher with each additional unit increase in credit term. • **P-value:** The p-value is less than 0.05, indicating that credit term is a statistically significant predictor.
age	• **Log-Odds:** A one-year increase in age is associated with a 0.0216 decrease in the log-odds of being a bad client. • **Odds Ratio:** For each additional year of age, the odds of being a bad client decrease by a factor of 0.9786. • **P-value:** The p-value is less than 0.001, indicating that age is a statistically

**Term Presentation*

	significant predictor.
sex (male)	• **Log-Odds:** Being male is associated with a 0.5253 decrease in the log-odds of being a bad client compared to females. • **Odds Ratio:** The odds of being a bad client are 0.5914 times lower for males compared to females. • **P-value:** The p-value is less than 0.001, indicating that sex is a statistically significant predictor.
education	• Categories like educationPhD degree have a large negative coefficient, meaning that having a PhD is associated with a significant decrease in the log-odds of being a bad client. However, the odds ratio is 0, and the p-value is not significant (greater than 0.05), indicating no clear evidence of the effect of having a PhD on being a bad client in this sample. • educationSecondary education is a significant predictor with an increase in the log-odds and an odds ratio greater than 1, indicating higher odds of being a bad client.
income	• The coefficient for income is zero, indicating no change in the log-odds of being a bad client with changes in income. • This is supported by the odds ratio of 1 and a significant p-value, which indicates that income is not a statistically significant predictor in this model.
family_status (Married)	• **Log-Odds:** Both married and unmarried coefficients are positive, indicating that being married or unmarried is associated with an increase in the log-odds of being a bad client. • **Odds Ratio:** The odds ratios are greater than 1 for both, suggesting higher odds of being a bad client. • **P-value:** The p-value for married is significant, indicating that marital status has a significant effect on the prediction.
is_client	• **Log-Odds:** The positive coefficient indicates that being an existing client is associated with an increase in the log-odds of being a bad client. • **Odds Ratio:** The odds ratio of approximately 1.5949 indicates that existing clients have higher odds of being bad clients. • **P-value:** The p-value is significant, suggesting that this is a statistically significant pred}

Null Deviance	Residual Deviance
2184.8	1948.7
The reduction suggests that adding the predictor variables into the model **significantly improves** its fit compared to the null model.	

*[*Term Presentation*]

Akaike Information Criterion (AIC)
1974.7
Even though this AIC is higher than the double cost model, the triple cost model should be considered since the priority is strongly on minimizing the risk of granting loans to bad clients, accepting the trade-off of potentially higher complexity and a slight overfitting risk as indicated by its higher AIC. Furthermore, despite having the highest AIC, indicating a less efficient model at explaining variance compared to Models I and II, Model III's emphasis on detecting bad clients aligns more closely with the bank's operational needs.

The Wald test results, complemented by the log-odds ratios, emphasize the importance of variables like credit term, age, gender, and education in predicting loan eligibility. The model's focus on minimizing false negatives, especially with the triple cost adjustment, shifts the balance towards identifying potential risks more conservatively. This strategy aligns with the bank's goal to reduce financial losses from bad loans, accepting the trade-offs in model complexity and the slight increase in false positives as indicated by the higher AIC. The reduction in null to residual deviance indicates a significant improvement in the model's fit with the introduction of predictor variables, underscoring the variables' importance in distinguishing between good and bad clients. The AIC, though higher in the triple cost model compared to its double cost counterpart, is deemed acceptable given the bank's prioritization of reducing false negatives—even at the expense of model simplicity and a potential increase in false positives.

Performance Metrics

Training Set

	Predicted	
Actual	0	1
0	979	139
1	130	64

Testing Set

	Predicted	
Actual	0	1
0	324	31
1	41	15

*Term Presentation

	Training Set	Testing Set
Accuracy	**79.50%**	**79.54%**
Precision	**32.53%**	**32.14%**
Recall	**33.00%**	**26.09%**
Specificity	**87.57%**	**89.62%**
F1 Score	**32.24%**	**28.80%**

Performance metrics between the training and testing sets highlight several crucial aspects of the model's applicability. In the training set, there's an accuracy of 79.50%, a precision of 32.53%, and a recall of 33.00%, which slightly shifts in the testing set to an accuracy of 79.54%, a precision of 32.14%, and a recall of 26.09%. These metrics illustrate the model's consistent ability to predict non-bad clients accurately, as seen in the high specificity across both sets. However, the model struggles with identifying bad clients, a challenge that is somewhat mitigated in the testing set but still remains a concern due to the trade-off between recall and precision.

VS

*[*Term Presentation*]

Training Set vs Testing Set Comparison Insights

Consistency in Predicting Non-bad Clients: Both the training and testing sets reinforce the model's effectiveness in correctly identifying non-bad clients, as shown by the substantial number of true negatives. This indicates the model's conservative approach, where it leans towards minimizing false positives to avoid the riskier scenario of granting loans to potentially bad clients. Such consistency is crucial for the bank, ensuring a reliable baseline for identifying clients who are likely to repay their loans.

Struggle with Identifying Bad Clients: Despite the model's adeptness at recognizing true negatives, it faces challenges in accurately pinpointing bad clients. This issue is mirrored in both sets, characterized by a relatively low number of true positives alongside a noticeable volume of false negatives. The training set, in particular, exhibits this struggle, underscoring the difficulty in capturing all potential bad clients within the constraints of the model's parameters and the chosen triple cost emphasis.

Slight Improvement in Recall in the Testing Set: The testing set shows a marginal increase in recall compared to the training set. This improvement suggests that the model, when applied to unseen data, may have a somewhat enhanced ability to identify bad clients that it previously missed. However, the increase is modest, highlighting the ongoing challenge of balancing the model's sensitivity to bad clients without disproportionately increasing false positives.

Precision and Recall Trade-off: The model demonstrates a trade-off between recall and precision, particularly evident in the testing set. While aiming to improve recall by reducing false negatives, the model concurrently experiences a dip in precision—indicating a higher rate of false positives. This trade-off reflects the model's adjusted focus under the triple cost scheme, prioritizing the detection of bad clients at the cost of mistakenly classifying some good clients as bad.

Maintained Specificity with a Slight Decline in Accuracy: Specificity remains relatively high across both sets, indicating the model's continued proficiency in identifying true negatives. However, there is a slight decline in overall accuracy from the training to the testing set. This slight shift underscores the model's complex balancing act between maintaining specificity and striving to improve other performance metrics like recall.

Overall Balanced Performance with Trade-offs: The triple cost model exhibits a balanced performance, considering the bank's strategic emphasis on minimizing false negatives. While there are trade-offs, particularly with precision and a slight decrease in accuracy, these are deemed acceptable within the context of the bank's objectives. The model's performance, especially in terms of improved recall and maintained specificity, aligns with the need to accurately identify bad clients to prevent financial losses.

**Term Presentation*

➢ Best-performing Logistic Regression Model

Comparison of Logistic Regression Models: Model I, III, & III

Model I: Logistic Regression Using Original Unbalanced Dataset	**Model II:** Logistic Regression Using Oversampled Dataset	**Model III:** Logistic Regression Using Cost-Sensitive Approach
Strengths		
Real-World Implications: By utilizing the original unbalanced dataset, this model captures the inherent complexities and distribution of client profiles, providing a realistic foundation for prediction. **High Specificity**: Exhibits a strong capability to accurately predict non-bad clients, as indicated by a high number of true negatives, aligning with the conservative banking approach to minimize false positives.	**Improved Handling of Class Imbalance:** Addresses the bias introduced by class imbalance through oversampling, potentially offering a more balanced perspective on client risk. **Better Detection of Positive Cases:** Shows a slightly improved ability to detect bad clients, crucial for the task at hand, despite higher AIC indicating a less parsimonious model.	**Prioritization of Minimizing False Negatives:** Incorporates a cost-sensitive approach that aligns closely with the bank's emphasis on accurately identifying bad clients, by increasing the cost of false negatives. **Balanced Performance with Trade-offs:** Exhibits a balanced performance, maintaining specificity while slightly improving recall in the testing set, indicating a degree of generalizability to unseen data.
Weaknesses		
Poor Recall: Struggles significantly to correctly identify bad clients across both sets, evidenced by minimal true positives and a considerable number of false negatives. **Limited Predictive Power for Bad Clients:** The model's inability to accurately detect bad clients highlights a critical gap, necessitating further refinement to meet the bank's operational needs.	**Still Limited in Precision and Recall:** Despite high accuracy and specificity, the model struggles with low precision and recall in predicting bad clients, especially evident in the testing set where precision drops significantly. **Trade-offs in Model Complexity:** The increase in AIC suggests a trade-off between model complexity and the ability to address class imbalance effectively, without a clear advantage in predicting bad clients accurately.	**Trade-off Between Recall and Precision:** While aiming to reduce false negatives, the model experiences a significant decline in precision, leading to a higher rate of false positives, particularly in testing sets. **Slight Decline in Overall Accuracy:** Observes a slight decline in accuracy from the training to the testing set, reflecting the complexities of balancing model sensitivity with the objective of minimizing financial losses from bad loans.

In summary, each logistic regression model presents a nuanced approach to predicting loan eligibility, with Model I focusing on understanding the data set, Model II addressing class imbalance, and Model III prioritizing the minimization of false negatives. However, all models exhibit a struggle in accurately identifying bad clients—a critical objective for the

*Term Presentation

bank—highlighting the need for further optimization to enhance predictive accuracy while aligning with strategic goals.

Given the bank's objective of minimizing financial losses by accurately identifying bad clients, Model III with Triple Weights appears to be the best fit. Despite slightly lower accuracy and specificity compared to the double weights version, it offers a balanced improvement in recall (26.09%) and the highest F1 score (32.24%) among the variants tested. This model demonstrates a better capability to identify bad clients without significantly compromising the identification of good clients. Further tuning and testing could refine this model to better balance precision and recall, optimizing it for the bank's specific lending risk tolerance and objectives.

Based on the analysis of the triple weights model, the potential profile of a "bad client" that the bank can identify includes individuals with certain characteristics that statistically correlate with a higher likelihood of being a bad client according to the model's output and interpretations. Here is a summary of the key findings that can be used to identify potential bad clients:

- ☐ **Credit Term:** Individuals with longer credit terms are more likely to be bad clients. Specifically, each additional unit increase in credit term is associated with a 6.16% higher odds of being a bad client.
- ☐ **Age:** Younger individuals have a higher likelihood of being bad clients. Each one-year increase in age is associated with a 2.14% decrease in the odds of being a bad client.
- ☐ **Gender:** Females are more likely to be bad clients compared to males. Being male is associated with a 40.86% lower odds of being a bad client.
- ☐ **Education:** Individuals with only secondary education are more likely to be bad clients compared to those with higher levels of education, such as a PhD, which does not show a clear effect due to the insignificance of its p-value.
- ☐ **Income:** Income does not have a significant effect on being a bad client, as its coefficient and odds ratio indicate no change in the likelihood.
- ☐ **Marital Status:** Both married and unmarried individuals have a higher likelihood of being bad clients.
- ☐ **Existing Clients:** Existing clients of the bank have higher odds of being bad clients compared to new or potential clients.

*Term Presentation

VII. COMPARISON OF MODELS: CF Model III vs LR Model III

Comparison of Performance Metrics

	Classification Trees Model III: Cost-Sensitive Approach Using Triple Cost	
	Training Set	**Testing Set**
Accuracy	**83.23%**	**76.03%**
Precision	**44.92%**	**21.21%**
Recall	**59.28%**	**38.89%**

	Logistic Regression Model III: Cost-Sensitive Approach Using Triple Cost	
	Training Set	**Testing Set**
Accuracy	**79.50%**	**79.54%**
Precision	**32.53%**	**32.14%**
Recall	**33.00%**	**26.09%**

Accuracy: The Classification Tree Model shows a significant drop in accuracy when moving from training to testing, suggesting potential overfitting. The Logistic Regression Model's accuracy remains stable, indicating better generalization to unseen data.

Precision: The Classification Tree Model shows a dramatic drop in precision in the testing set, which might be due to overfitting during training. The Logistic Regression Model has a slight increase in precision, which suggests it maintains its performance on unseen data better than the tree model.

Recall: Both models show a decrease in recall from training to testing. However, the drop is more pronounced for the Classification Tree Model.

**Term Presentation*

Conclusions

For the objective of predicting loan eligibility and prioritizing the reduction of false negatives, while both models have their strengths and weaknesses, the Logistic Regression Model seems more robust and consistent, especially in an operational environment where it will encounter data it has not been trained on.

The Classification Tree Model's significant drop in performance from training to testing indicates it may not be as reliable in practice, despite its higher recall in the training set. If we value an absolute recall rate, then this might be another choice.

VIII. RECOMMENDATIONS

If we were to move forward with this project, we would employ several recommendations to improve the effectiveness and performance of our predictive models:

Utilize Synthetic Minority Oversampling Technique: This recommendation is driven by the need to balance the dataset due to the presence of a minority class. SMOTE can generate synthetic samples for the minority class, potentially improving the model's ability to identify true positives within that class. Since the Classification Trees Model shows a significant drop in recall from training to testing, using SMOTE might provide more varied cases for the model to learn from, thus hoping to bridge the gap between training and testing performance.

Cost Sensitivity Tuning: The models' current penalty structure is stringent regarding false negatives. By tuning the cost sensitivity, we aim to achieve a better balance between the loss of potential business due to false positives and the operational risks of false negatives. This tuning should be designed in consideration of the operational environment where the models will be deployed. The intention is to reduce the heavy penalties on false negatives, as indicated by the drastic drop in precision from the training to the testing sets for both models.

Threshold Adjustment: Adjusting the decision threshold can provide a more refined control over the classification of clients as high risk. The default threshold of 0.5 might not be optimal, especially in a triple cost scenario where reducing false negatives is crucial. By experimenting with different thresholds, we could improve the models' specificity to the operational context. This step is particularly pertinent for the Classification Trees Model, which saw a considerable reduction in recall upon testing, suggesting that it may benefit from a lowered threshold to maintain its recall rate.

Employ Cross-Validation Techniques: The use of k-fold cross-validation is recommended to avoid overfitting and to ensure the model's robustness. This approach divides the dataset into k smaller sets, and the model is trained and validated k times, with each of the subsets used exactly once as the validation data. This method helps in assessing the model's performance across various splits of the dataset, ensuring the performance metrics are not overly reliant on any particular division of data. Given the observed performance

*Term Presentation

discrepancies between training and testing datasets, cross-validation might help in diagnosing and mitigating overfitting issues.

**Term Presentation*