**PREDICTING CAMPAIGN SUCCESS**

Overview: This document provides a comprehensive analysis of a bank marketing dataset, with the goal of predicting the success of marketing campaigns. Two main modeling approaches are employed: random forest classification and logistic regression. The initial random forest model performed well in predicting negative responses but struggled with positive predictions due to class imbalance. To address this, the analysis applied SMOTE (Synthetic Minority Over-sampling Technique), which significantly improved model performance—cutting the error rate and increasing the accuracy of predicting positive responses. Variable importance measures identified call duration as the strongest predictor, though its use is limited for forward-looking models since it is only known after contact. A logistic regression model was also implemented using standardized predictors, allowing for interpretable coefficients and odds ratios. Both models were evaluated using AUC from ROC curves, with high scores (0.9425 for random forest, 0.9291 for logistic regression), indicating strong predictive performance. Overall, the document outlines a robust and methodical approach to understanding and improving predictive marketing outcomes.

1. *Random Forest Regression*

Summary of results----

Call:

randomForest(formula = factor(y\_code) ~ ., data = rf\_train, importance = TRUE)

Type of random forest: classification

Number of trees: 500

No. of variables tried at each split: 4

OOB estimate of error rate: 8.52%

Confusion matrix:

|  |  |  |  |
| --- | --- | --- | --- |
| **Actual/**  **Predicted** | **Predicted**  **No** | **Predicted**  **Yes** | **Class**  **Error** |
| No (1) | 24,714 | 845 | 3.31% |
| Yes (2) | 1,611 | 1,661 | 49.24% |

**Interpretation**:

While the model exhibits a very low Out-of-Bounds error rate and predicts the response of “no” quite well, it incorrectly predicts yes about 50% of the time. This likely stems from an imbalance in “no” versus “yes responses”. A quick look at the counts of each response reveals that imbalance is likely troubling the model. To address this shortcoming of the model, we use SMOTE to handle the imbalance.

A graph with a rectangle and a rectangle

Description automatically generated

*Oversampling*

This method increases the number of instances in the minority class by duplicating existing samples or generating new ones. The most common technique for oversampling is SMOTE (Synthetic Minority Over-sampling Technique). Oversampling is the preferred method, especially since it enables taking advantage of data augmentations for additional diversity.

After adjusting for imbalance via SMOTE, the AUC stays largely the same but the predictive capacity increases substantially:

Call:

randomForest(formula = factor(y\_code) ~ ., data = rf\_train.smote, importance = TRUE)

Type of random forest: classification

Number of trees: 500

No. of variables tried at each split: 4

OOB estimate of error rate: 1.96% 🡨 decreased from last model

Confusion matrix:

|  |  |  |  |
| --- | --- | --- | --- |
| **Actual/**  **Predicted** | **Predicted**  **No** | **Predicted**  **Yes** | **Class**  **Error** |
| No (1) | 129,449 | 4,735 | 3.53% |
| Yes (2) | 535 | 133,649 | 0.4% |

*Determining Variable Importance*

A graph of a model

Description automatically generated with medium confidence

|  |  |
| --- | --- |
| Variable | Description |
| duration | Call duration — highest predictor (known to dominate bank marketing models) |
| euribor3m | 3-month Euribor rate — an economic indicator that indirectly influences banks’ interest rates (and hence purchasing behavior) |
| emp.var.rate | Employment variation rate — proxy for economic conditions, which in turn influence purchasing behavior |
| nr.employed | Number of employees — macroeconomic context |
| cons.conf.idx | Consumer confidence index — behavioral economics signal |
| month | Month when contact was made — seasonal trends |
| contact | Type of communication (e.g., phone, cellular) — strategy-related |
| cons.price.idx | Consumer price index — macroeconomic stability |

While duration is very predictive, it’s not usable for campaign planning — it’s only known after the contact happens. If the goal of this project was prospective modeling (e.g., selecting leads), it may be opportune to train a version excluding duration.

1. *Logistic Regression*

The sigmoid function used in logistic regression maps the linear combination of predictor variables to a probability between 0 and 1. This function's shape helps to moderate the impact of outliers, especially if the extreme values are not severely at odds with the model's overall trend. While outliers can still potentially impact logistic regression, its design and the use of the sigmoid function make it more resistant to their influence compared to methods like least squares regression

The model was robust (Pseudo R-Squared was x and significant at x)

Given that each variable is measured on different scales, the z-scores of each variable were computed to allow for comparison. The table below shows the results of the logistic regression on the standardized data.



To aid the interpretation of the table, reference this key:

|  |  |  |
| --- | --- | --- |
| **Sign of Coefficient β** | **Odds Ratio** | **Adjusted Interpretation** |
| **Positive** (β > 0) | OR > 1 | Increases odds of **success** (more likely to be "yes") |
| **Negative** (β < 0) | OR < 1 | Decreases odds of **success** (more likely to be "no") |

**Summary**

* Strong positive effects:
  + Contact duration increases odds of campaign success by 226%
  + Macroeconomic conditions, as measured through the consumer price and consumer confidence indices in particular, have notable impact on odds of campaign success
  + Month of the year and contact method have modest boosts on odds of campaign success
* Strong negative effects:
  + Employment variation rate more than halves the odds of success (–57%)
  + default\_code (–33%)
  + poutcome\_code (–21%)
  + euribor3m (–15%)

1. *Comparing Model Performance*

The ROC (Receiver Operating Characteristic)curve helps us to visualize the true positive rate or true negative rate of a prediction based on some model. This helps us to assess how well a regression model has fitted the data. The AUC (Area under Curve) of this ROC curve helps us to determine the specificity and sensitivity of the model. The closer the AUC value is to the 1, the better the given model fits the data

Random Forest

Area under the curve: 0.9425

A graph of a logistic regression

Description automatically generated

Logistic Regression Model Validation:

Area under the curve: 0.9291

A graph of a logistic regression

Description automatically generated

**Interpretation: The refined Random Forest model performs better**