INFO 4604-5604
Applied Machine Learning
Spring 2023
Project Report

Student Name	Bailey Gimpel
Student ID	108659176
Project Title	Term Deposit Classification with Unbalanced Classes
Date Submitted	May 8, 2023

Project Context

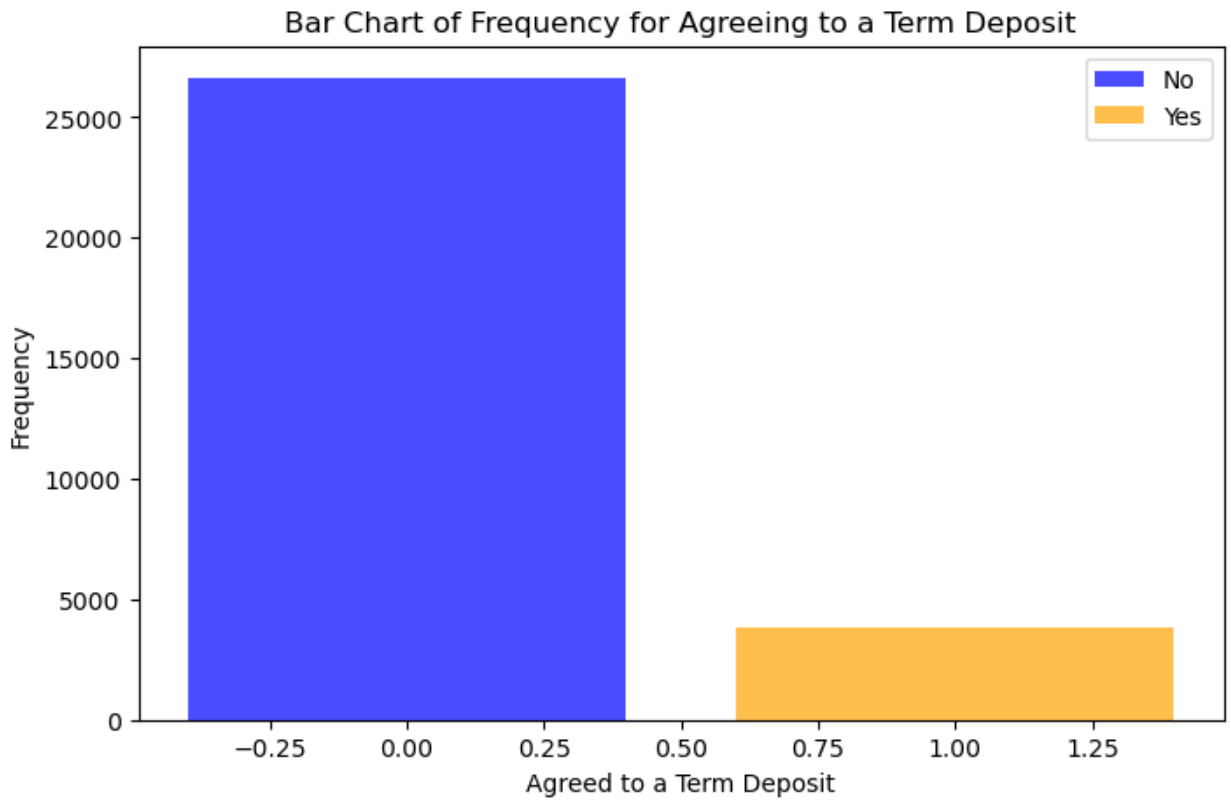
Term deposits are an important financial product for banks, as they provide a stable source of funding and enable the bank to offer lower interest rates on loans. When a customer agrees to a term deposit, they agree to deposit a certain amount of money with the bank for a fixed period of time, usually ranging from one month to several years. In return, the bank pays the customer a fixed interest rate on the deposit. This arrangement is beneficial for both the bank and the customer. For the bank, term deposits provide a stable source of funding that can be used to finance loans and other investments. The bank knows exactly how much money it will have available to lend out, and for how long. This allows the bank to plan its finances more effectively and to offer lower interest rates on loans, which can be a competitive advantage. For customers, term deposits provide a safe and reliable investment opportunity. The fixed interest rate means that the customer knows exactly how much they will earn on their deposit, regardless of market fluctuations. This can be particularly attractive during times of economic uncertainty, when other investment options may be more volatile.

However, convincing customers to commit to a term deposit can be a challenge for banks. This project aims to use machine learning techniques to predict which customers are most likely to agree to a term deposit, based on their demographic information and banking history. By analyzing the results of the machine learning model, the goal is to provide insights into which demographic factors are most strongly associated with a customer's likelihood of agreeing to a term deposit. Banks can use this information to target their marketing campaigns more effectively and increase their success rates in selling term deposits.

To achieve this goal, the data used was from a marketing campaign conducted by a bank in Portugal. The dataset includes information about 45,211 customers who were contacted during the campaign, as well as their response to the bank's offer of a term deposit. The dataset also includes a variety of demographic and banking variables, such as age, job, marital status, education level, balance, and number of previous contacts with the bank. To predict which customers are most likely to agree to a term deposit, use a Random Forest Classifier. This is a popular machine learning algorithm that works by constructing multiple decision trees and combining their predictions to make a final prediction. Banks can use this information to target their marketing campaigns more effectively and increase their success rates in selling term deposits.

Problem Definition

The goal of this project is to predict whether a customer will agree to a term deposit, based on their demographic information and banking history. This is a binary classification problem, where the two classes are "yes" and "no" (indicating whether the customer agreed or did not agree to the term deposit). The dataset used for this project is imbalanced, with a significantly larger number of customers in the "no" class than in the "yes" class. Specifically, only 12.3% of the customers in the dataset agreed to the term deposit. This can be a problem for machine learning algorithms, as they tend to be biased towards the majority class and may have difficulty accurately predicting the minority class.



To address this imbalance, random over-sampling was used to increase the number of instances in the minority class. Random over-sampling involves randomly duplicating cases in the minority class until the number of instances in both classes is roughly equal. The imblearn library in Python was used to perform random over-sampling before training in the Random Forest Classifier. The machine learning model's performance was evaluated using standard machine learning metrics such as accuracy, precision, recall, and F1 score. The results of this analysis provide insights into which demographic factors are most strongly associated with a customer's likelihood of agreeing to a term deposit, and can be used by banks to target their marketing campaigns more effectively.

Project Motivation

The motivation for this project comes from the importance of term deposits for banks, and the challenge of convincing customers to commit to them. Machine learning techniques can be used to predict which customers are most likely to agree to a term deposit, which can help banks target their marketing campaigns more effectively and increase their success rates in selling term deposits. This can have significant financial benefits for the bank, as it provides a stable source of funding and enables them to offer lower interest rates on loans. At the same time, customers benefit from the safety and reliability of term deposits as an investment opportunity.

However, imbalanced datasets can present a challenge for classification algorithms, as they tend to be biased towards the majority class and may have difficulty accurately predicting the minority class. In this project, random over-sampling was used to increase the number of instances in the minority class. Other strategies for dealing with imbalanced datasets include under-sampling the majority class, using ensemble methods like Boosting or Bagging or using cost-sensitive learning algorithms that assign higher misclassification costs to the minority class. By analyzing the results of the balanced Random Forest model and comparing it to the results of the Random Foreset Classification adapted for imbalanced classes, insights can be provided into the most effective strategies for predicting term deposit agreements and contributing to the development of more effective marketing strategies for banks, with unbalanced marketing data.

Data Source

The data source used in the Random Forest Classifier is entitled Bank Marketing (with social and economic context) and was created from term deposit campaigns conducted by a Portuguese bank. The data set consists of 20 features used for prediction and one target variable. The target variable y contains two values a “yes,” or “no,” which indicates whether the person contacted agreed to a term deposit. The following list consists of features that were used in the test data as described by [Moro et al., 2014]:

Bank client data:

1. Age (numeric)
2. Job: type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
3. Marital: marital status (categorical: 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed)
4. Education: (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
5. Default: has credit in default? (categorical: 'no', 'yes', 'unknown')
6. Housing: has housing loan? (categorical: 'no', 'yes', 'unknown')
7. Loan: has personal loan? (categorical: 'no', 'yes', 'unknown')

Social and economic context attributes :

1. Emp.var.rate: employment variation rate - quarterly indicator (numeric)
2. Cons.price.idx: consumer price index - monthly indicator (numeric)
3. Cons.conf.idx: consumer confidence index - monthly indicator (numeric)
4. Euribor3m: euribor 3-month rate - daily indicator (numeric)
5. Nr.employed: number of employees - quarterly indicator (numeric)

Project Methodology

In order to preprocess the data to manipulate it into the optimal inputs for the Random Forest Classifier algorithm the data must be encoded into binary and multi-classes. First, the data frame was filtered to exclude any observations containing an “unknown” value because unknown demographics will add unwanted noise to the Random Forest Regressor. The Default, Housing, Loan, and Target Variables were binarily encoded mapping [yes, no] to the corresponding values [1,0]. The social and economic attributes were not scaled or transformed because the Random Forest Classifier Algorthimn is not vulnerable to unscaled values.

The input predictors of Job, Martial, and Education could have been encoded using two separate strategies. The first is a one-hot encoding strategy that uses the pandas function get_dummies(), which creates columns with a binary encoding representing a yes or no for each label within the feature. Although this method of encoding proves useful, features like a person's job, education, and marital status have a direct impact on their income. Which in turn would have an influence on whether that person is able to afford to agree to a term deposit. Therefore an ordinal encoding must be applied that reflects how each label impacts wealth, which corresponds with the ability to make a term deposit. The ordinal encoding for these variables was determined based on domain knowledge, ordering education by the number of years, ordering marital by married to divorce, and jobs by income.

This data will serve as the input for two Random Forest Regressors, one that uses the RandomOverSampling function to balance minority classes, and another model that does not account for the unbalanced classes. By doing so the outputs and evaluations of each model can be compared to see if the RandomOverSampling function had an impact on the classifier's ability to predict the minority class while still accurately classifying the majority class. Data visualization and exploration will be used to determine the best demographics to market term deposits to based on the model outputs. In addition to demographics, the economic and social variables will also be assessed through visualizations based on the output of the classification algorithm.

ML Model Design

The machine learning model used in this project is a Random Forest Classifier. Random Forest is an ensemble learning algorithm that works by creating multiple decision trees and combining their results to produce a final prediction. Random Forest is particularly useful for classification tasks because it is robust to noise and able to handle both continuous and categorical variables. To address the issue of imbalanced classes in the dataset, we used a Random Over Sampling technique to increase the number of instances in the minority class.

This technique involves randomly duplicating instances in the minority class until it is balanced with the majority class. We applied the Random Over Sampling technique to the training data only, to prevent data leakage into the testing set.

The model was then evaluated using a 10-fold cross-validation to gain insight into whether Random Over Sampling contributed to overfitting. Additionally, the classification report from the sklearn library was used to obtain the accuracy, recall, and f1 scores to be compared between the two Random Forest Classification algorithms. The hyperparameters of each algorithm were not tuned in any way because the results of the algorithm provided satisfactory results while keeping computational complexity to a minimum. The ROC AUC score was also calculated and a confusion matrix was used to visually display the true positive and false positive rates.

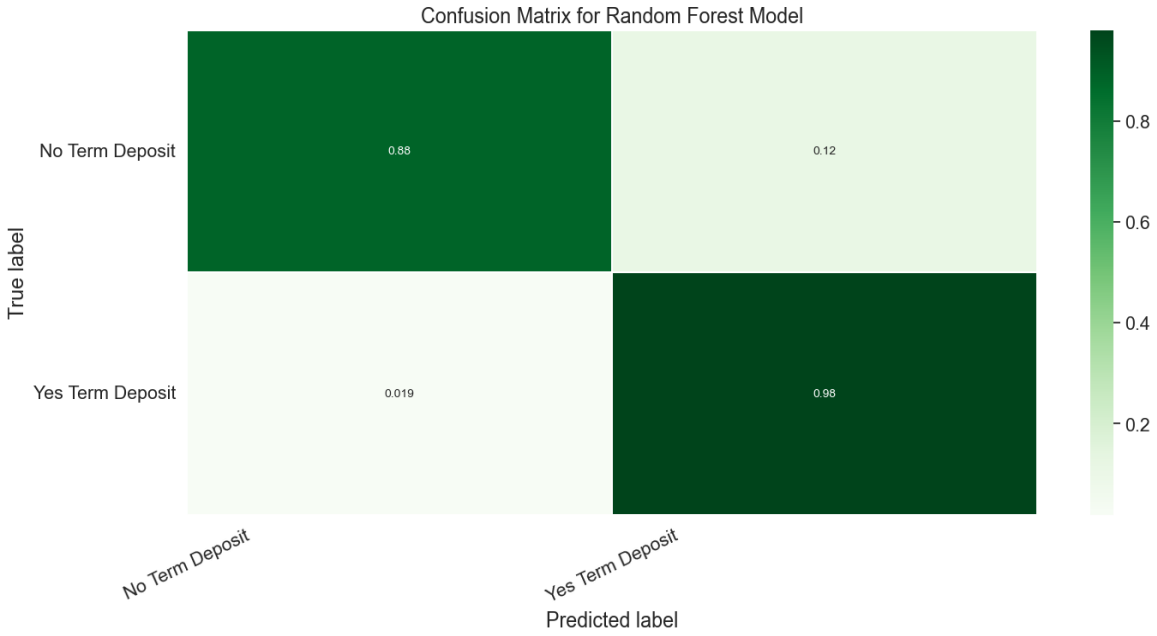
Model Evaluation

	precision	recall	f1-score	support
0	0.90	0.95	0.92	5304
1	0.45	0.27	0.34	794
accuracy			0.86	6098
macro avg	0.67	0.61	0.63	6098
weighted avg	0.84	0.86	0.85	6098

	precision	recall	f1-score	support
0	0.98	0.88	0.93	5308
1	0.87	0.98	0.92	4279
accuracy			0.93	9587
macro avg	0.93	0.93	0.93	9587
weighted avg	0.93	0.93	0.93	9587

The classification reports of the Random Forest Classifiers trained on the unbalanced data (top) and the RandomOverSampled data (bottom) are given above. It can be seen that after the RandomOverSampler() function was used the precision, f1, and recall scores for the minority class have greatly improved. Additionally, the precision and f1-scores for the majority class also had increased, accompanied by a natural decrease in recall. This decrease in the recall is due to the fact that there are more minority cases which makes it harder for the classifier to correctly identify those who did not agree to a term deposit. This points to an overall improvement in the ability of the classifier to accurately predict both the minority and majority classes, even though recall in the unbalanced model decreased. Furthermore, the ROC AUC scores and results of the 10-fold cross-validation further corroborate the accuracy of this model, without overfitting.

```
Cross-validation Scores: [0.91890482 0.92333768 0.92568449 0.92151239 0.93116037 0.92723005
 0.92801252 0.9246218  0.92357851 0.92279604]
Mean Cross-validation Score: 92.47%
```

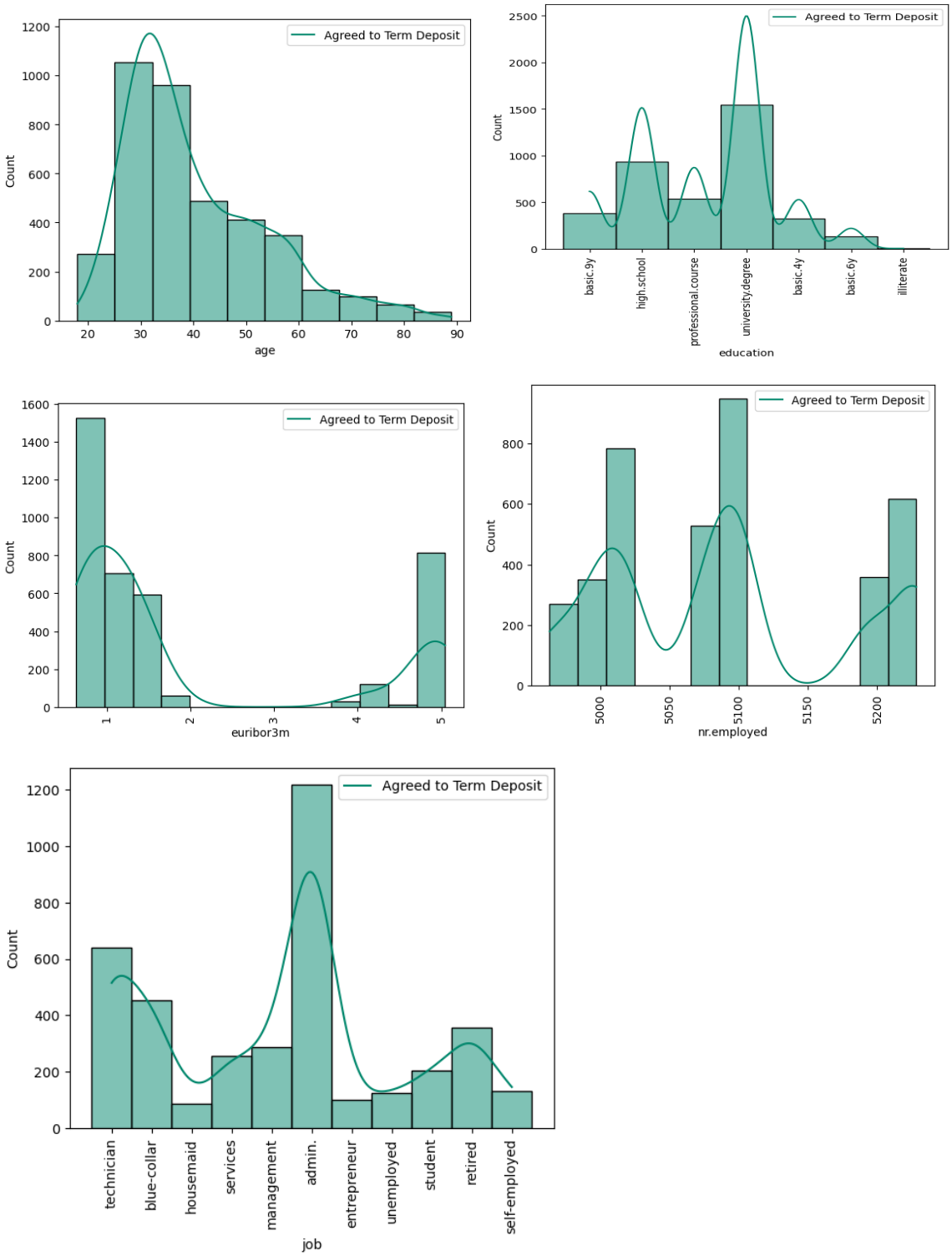


Findings

After evaluating the model and validating its accuracy without overfitting, the feature importance in combination with data visualization and analysis was used to identify the most important demographics and economic factors. Below are the feature importances of the model that used RandomOverSampling:

```
The importance of feature 'age' is 26.31%
The importance of feature 'job' is 9.66%
The importance of feature 'marital' is 3.85%
The importance of feature 'education' is 7.28%
The importance of feature 'default' is 0.0%
The importance of feature 'housing' is 3.29%
The importance of feature 'loan' is 2.3%
The importance of feature 'emp.var.rate' is 4.6%
The importance of feature 'cons.price.idx' is 3.0%
The importance of feature 'cons.conf.idx' is 4.69%
The importance of feature 'euribor3m' is 23.44%
The importance of feature 'nr.employed' is 11.58%
```

According to the feature importances given above the focus of exploratory data analysis will be conducted on the variables [age, job, education, euribor3mn, nr. Employed] which make up around 80% of the explanatory variables. Below are histograms displaying the frequency of agreeing to a term deposit based on these predictor variables.



As shown above, the optimal age demographic for targeted bank marketing for term deposits is 30 to 40 years old, with a sharp drop in the number of agreed term deposits beginning after the age of 60. Additionally, people obtaining the education levels of a college degree and high school diploma are more likely to agree to a term deposit. The jobs that should

be at the center of bank marketing efforts should be admin, technician, and blue-collar. Term deposit campaigns should also be conducted when the Euribor 3-month interest rates are between 0.5 and 1.5, and even when the Eribor interest rate is around 5. Lastly, the banks should conduct their bank marketing campaigns when their number of employees is at its highest to ensure the most return on their efforts.

Potential Real-World Application

One potential application of a random forest classifier for predicting customer behavior is in the field of e-commerce. E-commerce companies often use data-driven approaches to understand their customers and improve their shopping experience. By analyzing customer data, such as past purchase history, browsing behavior, and demographic information, e-commerce companies can predict which customers are most likely to make a purchase and tailor their marketing strategies accordingly.

For example, an e-commerce company could use a random forest classifier to predict whether a customer is likely to make a purchase based on their past behavior on the website. The classifier could take into account features such as the number of items the customer has added to their cart, the time spent on the website, the number of previous purchases, and the customer's demographic information. Using this information, the e-commerce company could personalize its website experience for each customer, displaying products that are most likely to appeal to them and providing tailored promotions and discounts. By improving the shopping experience for its customers and increasing the likelihood of a purchase, the company could increase its revenue and improve customer satisfaction. In this application, the random forest classifier provides a powerful tool for understanding customer behavior and making data-driven decisions to improve the e-commerce experience.

Work Cited

- [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, In press, http://dx.doi.org/10.1016/j.dss.2014.03.001
  Available at: [pdf] http://dx.doi.org/10.1016/j.dss.2014.03.001