PREDICTIVE MODEL FOR TERM DEPOSIT SUBSCRIPTION

Team Members:

- *Manoj Velu (Team Manager) – mvelu1@student.gsu.edu*
- *Bharath Badri Venkata – bbadrivenkata1@student.gsu.edu*
- *Nivethitha Avarampalayam Manoharan – navarampalayammanoh1@student.gsu.edu*
- *Varshini Vaisnavi Srinivasan – vsrinivasan3@student.gsu.edu*

1. Introduction

1.1 Motivation and Problem Statement

In the fast-paced world of banking, our challenge is to smartly handle term deposit subscriptions. The significance of this problem extends beyond immediate financial gains. The effective management of term deposit subscriptions is integral to a bank's financial stability and growth. Successfully predicting and optimizing term deposit subscriptions enables the bank to allocate resources efficiently, enhance customer satisfaction, and maintain a robust relationship with clients

1.2 Solution

The aim is to create a powerful tool - a predictive model using telemarketing data. The key challenge is decoding complex data patterns to accurately predict and improve term deposit subscriptions for the bank. The goal is to extract useful insights to guide strategic marketing efforts. These efforts are designed to boost subscription rates and strengthen the bank's financial position in the competitive market.

2 Dataset

2.1 Data Collection

The data pertains to the direct marketing initiatives of a Portuguese banking institution, specifically focusing on phone-based campaigns. These campaigns frequently involved multiple contacts with the same client to ascertain whether they subscribed to the product, a bank term deposit, or not.

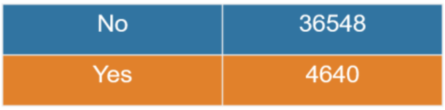
2.3 Data Preprocessing

The dataset is inherently clean and devoid of the need for preprocessing. Rigorous cleaning procedures have been previously applied, ensuring the data's integrity, and eliminating the necessity for further preprocessing steps

3. Exploratory Data Analysis:

3.1 Class Imbalance:

The initial exploration of the dataset uncovered a significant class imbalance in the target variable, with a substantial preponderance of instances categorized as 'no' in comparison to 'yes.' Recognizing the potential impact on subsequent analyses and model development, proactive measures were implemented to rectify this imbalance and ensure a more equitable representation of both classes.

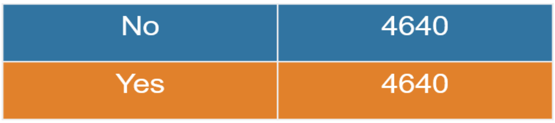


No	36548
Yes	4640

3.2 Handling Class Imbalance:

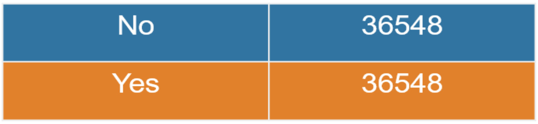
To mitigate the significant class imbalance observed in the dataset, three distinct strategies were implemented:

Random Under sampling: Instance from the majority class were randomly removed to achieve a more balanced distribution, thereby preventing the model from being biased towards the majority class.



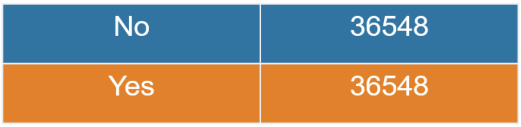
No	4640
Yes	4640

Random Oversampling: To address the scarcity of instances in the minority class, additional instances were randomly duplicated, ensuring a more equitable representation of both classes in the dataset.



No	36548
Yes	36548

Synthetic Minority Oversampling Technique (SMOTE): Synthetic instances were generated for the minority class using SMOTE, creating a more balanced dataset by introducing synthetic instances while preserving the characteristics of the existing minority class instances.



No	36548
Yes	36548

These resampling techniques were instrumental in preparing a balanced dataset for subsequent analyses and model development, mitigating the challenges associated with class imbalance.

4. Model Development and Evaluation:

4.1. Objective:

Developed a machine learning model to predict term deposit subscriptions using telemarketing data.

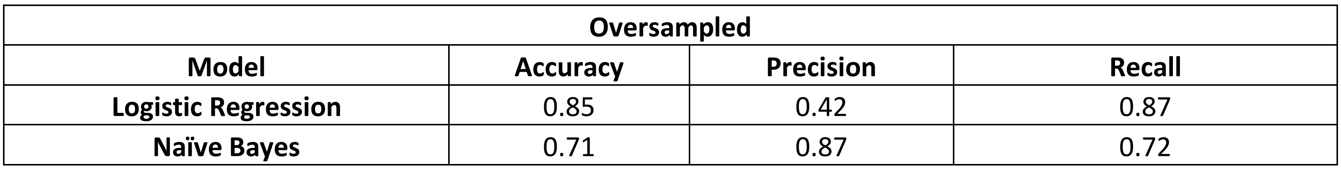
4.2. Data Preparation:

Applied a rigorous train-test split with a 70-30 ratio to assess model performance systematically.

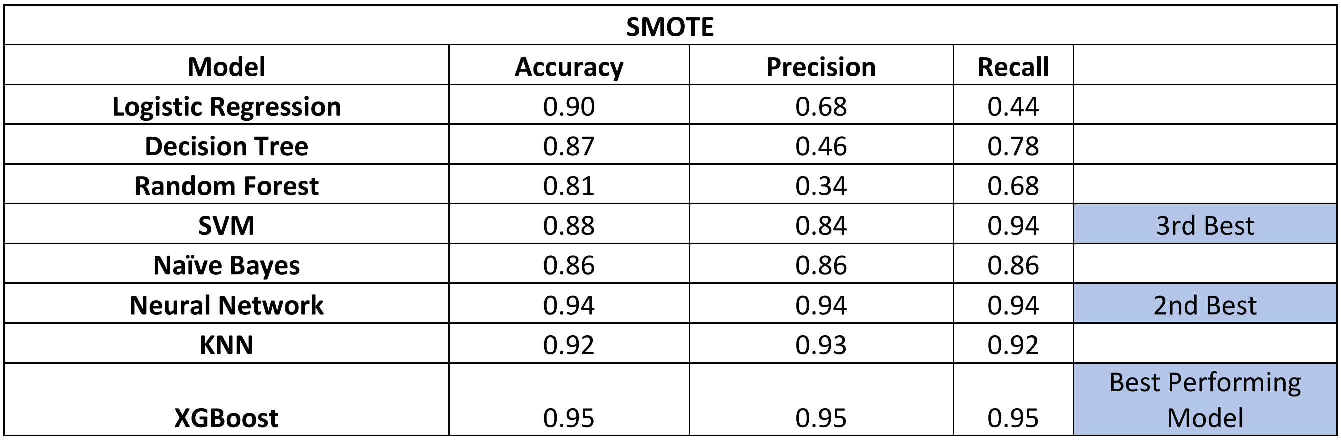
4.3. Evaluation Metrics:

Unsampled			
Model	**Accuracy**	**Precision**	**Recall**
Logistic Regression	0.91	0.68	0.41
Decision Tree	0.91	0.64	0.55
Random Forest	0.78	0.29	0.71
SVM	0.86	0.40	0.46
Naïve Bayes	0.83	0.86	0.83
Neural Network	0.89	0.88	0.90
KNN	0.89	0.89	0.89
XGBoost	0.91	0.91	0.92

Undersampled			
Model	**Accuracy**	**Precision**	**Recall**
Logistic Regression	0.86	0.45	0.87
Naïve Bayes	0.71	0.86	0.71



Oversampled			
Model	**Accuracy**	**Precision**	**Recall**
Logistic Regression	0.85	0.42	0.87
Naïve Bayes	0.71	0.87	0.72



SMOTE				
Model	**Accuracy**	**Precision**	**Recall**	
Logistic Regression	0.90	0.68	0.44	
Decision Tree	0.87	0.46	0.78	
Random Forest	0.81	0.34	0.68	
SVM	0.88	0.84	0.94	3rd Best
Naïve Bayes	0.86	0.86	0.86	
Neural Network	0.94	0.94	0.94	2nd Best
KNN	0.92	0.93	0.92	
XGBoost	0.95	0.95	0.95	Best Performing Model

4.4. Key Achievements:

Effectively managed imbalanced data, ensuring the development of a robust predictive model. Demonstrated proficiency in precision, recall, and accuracy assessments.

5. Feature Importance Analysis and Model Optimization

Developed and fine-tuned a predictive model for term deposit subscriptions through comprehensive feature importance analysis and strategic feature selection.

5.1. Feature Importance Analysis:

A comprehensive array of features was established, covering a wide spectrum of demographic, economic, and marketing variables. These variables were meticulously selected to encompass diverse aspects relevant to the analysis. Subsequently, an XG Boost classifier, renowned for its efficiency in predictive modelling, was employed to train and evaluate the data, discerning the importance of each feature. Through this process, the classifier identified the relative significance of each feature in predicting the target outcome. To communicate these findings effectively, a visual representation, such as feature importance plots or graphs, was meticulously crafted. This visual presentation succinctly showcased the hierarchy of feature importance, providing insights into which variables wielded the most substantial impact on the predictive model's outcomes.

5.2. Highly Correlated Feature Removal:

Highly correlated columns were meticulously removed to enhance dataset efficiency. Subsequent evaluations using XG Boost and Neural Network models showcased their performance on the optimized dataset. This analysis revealed insights into each model's strengths and weaknesses in predictive accuracy and generalization. Removing correlated features maintained consistent accuracy, precision, and recall, highlighting the ability to create efficient models with fewer features, saving computational resources significantly.

6. Model Persistence Strategy for Future Deployment

Once the training of a scikit-learn model is concluded, it proves beneficial to establish a systematic method for storing the model's parameters. This practice greatly aids in future use, as it eliminates the necessity of retraining the model from scratch. Storing the model's parameters allows for quick and efficient access to its learned patterns and configurations, enabling seamless deployment and application in various scenarios without the overhead of repeating the training process. This strategy not only saves computational resources but also ensures consistency and reproducibility in utilizing the trained model across different environments or for future predictions.

With the constructed model, clients have the capability to input their customer data directly, enabling the generation of predictions that aid in effectively targeting the appropriate audience. By leveraging the insights derived from feature importance analysis, the bank can strategically emphasize and prioritize features that significantly contribute to the accuracy of the predictions.

7. Conclusion

This project built a predictive model for term deposit subscriptions using tel-marketing data. Techniques to handle class imbalance were applied, and XGBoost achieved the highest accuracy of 95% among models like Logistic Regression, Decision Trees, SVM, and Neural Networks. Feature importance analysis and managing correlated features improved model efficiency without compromising accuracy. The final XGBoost and Neural Network models maintained strong accuracy using fewer, more impactful features. Overall, this project aligns business context, data science, and machine learning to create a high-performance predictive engine, providing actionable insights for better customer acquisition and business growth via term deposit subscriptions.