Final project

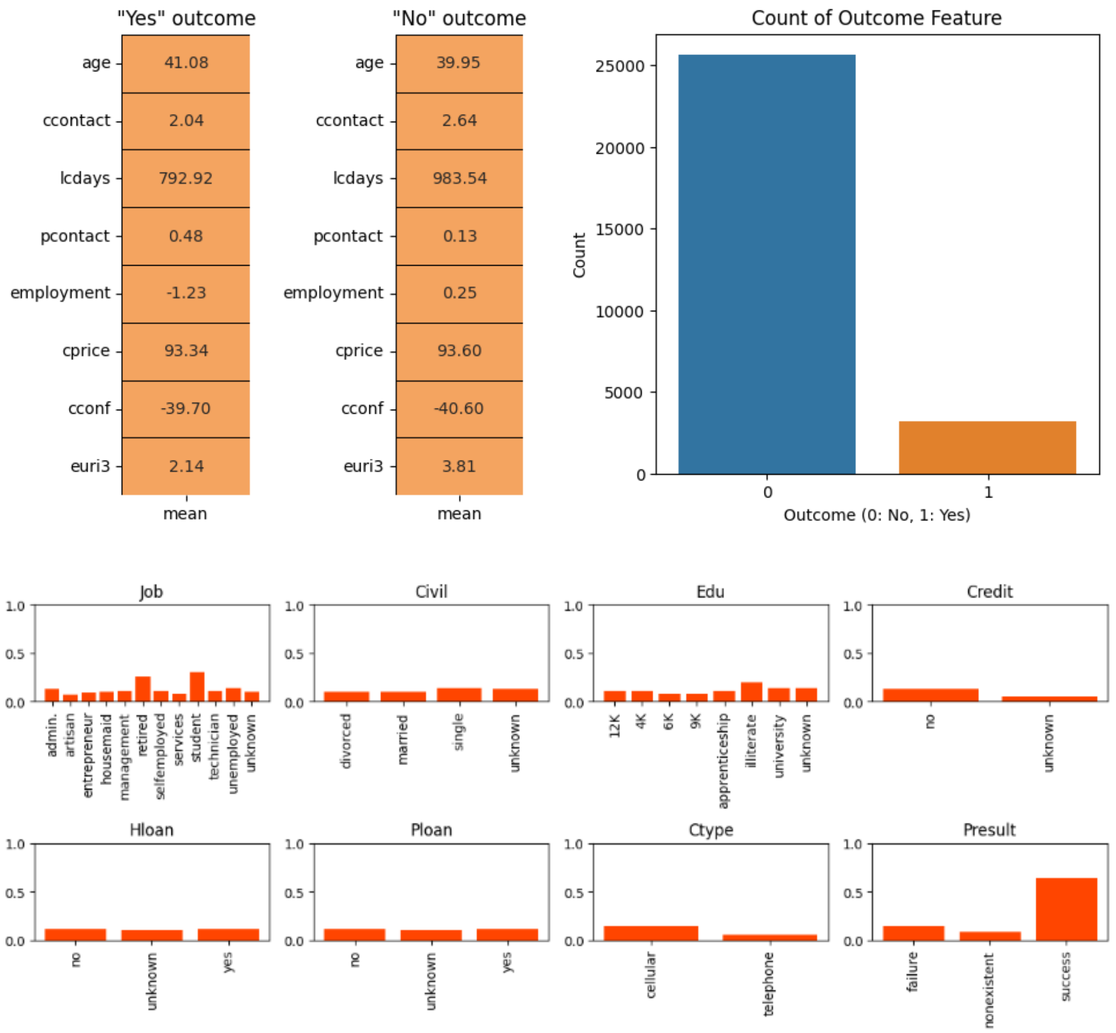
DATA MINING

BANK MKT ANALYSIS

- Dataset overview
- Data Preparation
- Modeling
- Model Evaluation and Use
- Conclusion & Limitations

DATASET OVERVIEW

This dataset provides insights into the factors affecting the opening of savings accounts by customers. The dataset is highly unbalanced, with the majority of the observations having an outcome variable of "0" (no account opened), making it challenging to accurately classify the minority class.



KEY FINDINGS

- **Class Imbalance:** The dataset's outcome variable is highly unbalanced, potentially leading to a bias towards the majority class during model training. This issue may affect the accurate classification of the minority class.
- **Important Features:** Age, lcdays, cconf, euri3, and employees are found to have higher distributions within the "1" class, suggesting that these features may be crucial predictors for the model's outcome.
- **Demographics and Behavior:** Customers opening savings accounts are typically older and have a lower consumer price index, indicating a more stable financial position. These customers also tend to contact the bank more frequently, suggesting higher engagement with their finances and interest in the bank's products and services.
- **Employment and Financial Stability:** A lower employment rate among customers opening bank accounts implies they may have more disposable income or financial stability, making them more likely to save money in a bank account.
- **Risk Aversion and Interest Rates:** Customers opening savings accounts exhibit a lower Euribor 3-month rate, suggesting that they may be more risk-averse and prefer low-risk accounts. This observation indicates that the bank's attractive interest rates for savings accounts could serve as an effective marketing strategy to attract more customers.

DATA PREPARATION

PREPROCESSING STEP

In this analysis, I performed several data preprocessing steps to ensure high-quality inputs for our predictive models. These steps included:

- **Dropping variables:** Based on the dataset overview, I defined some important features such as **age, lcday, cconf, euri3**, in which employees is highly correlated to **employment** and **euri3**. In addition, the **civil** feature plot shows the equal distribution of four elements divorced, married, single and unknown. I decided to drop **id, employees, day** columns.
- **Transform variables:** Transforming variables involved standardizing and scaling numerical variables as well as encoding categorical input variables using one-hot encoding. I used the **StandardScaler** function from scikit-learn to standardize the numerical variables, and I used the **get_dummies** function to **one-hot encode** the categorical variables. This transformed the original variables into a format that can be used in our models.
- **PCA:** Using PCA to transform these variables. In this case, I did not use PCA to reduce dimensionality of numerical variables. These step increase the performance of the model compared to "not using PCA" and "using PCA to reduce dimensionality.
- **Dividing dataset into train and validation sets:** I divided the dataset into a training set and a validation set using an 80-20 split. The training set is used to train the models, while the validation set serves as an unseen dataset to evaluate the models' performance. This process ensures that the models' performance metrics, specifically the area under the ROC curve (AUC).
- **Addressing class imbalance:** I utilized the Synthetic Minority Over-sampling Technique (SMOTE) to create synthetic observations of the minority class. I considered two: (a) oversampling class 1 until it equaled the size of class 0, and (b) oversampling class 1 in a specific ratio to reduce the severity of imbalance betIen the two classes. These approaches are used to experiment the change in AUC when predicting on validation set and test data.

FEATURES SELECTION

As mentioned above, I fine some important features and decide to drop features with less impact or no impact on class prediction. However, this is not an accurate way to define features with highest impact. Therefore, I decided to perform feature selection process using a random search approach in conjunction with a Random Forest classifier. The primary objective of this process is to identify the most relevant features contributing to the prediction of the target variable by maximizing the area under the ROC curve (AUC).

- Define the variables: features stores all available feature names, **features_selected_list** stores the selected features, **select_k** defines the number of features to be randomly sampled in each iteration, **auc** stores the best AUC score achieved, **it** is the iteration counter, and **max_iter** is the maximum number of iterations allowed.
- The while loop iterates until the maximum number of iterations (**max_iter**) is reached or all features have been selected. Within the loop:
- Calculate the set difference between all features and the currently selected features to obtain the available features.
- Randomly sample one or more features (based on the value of **select_k**) from the available features.

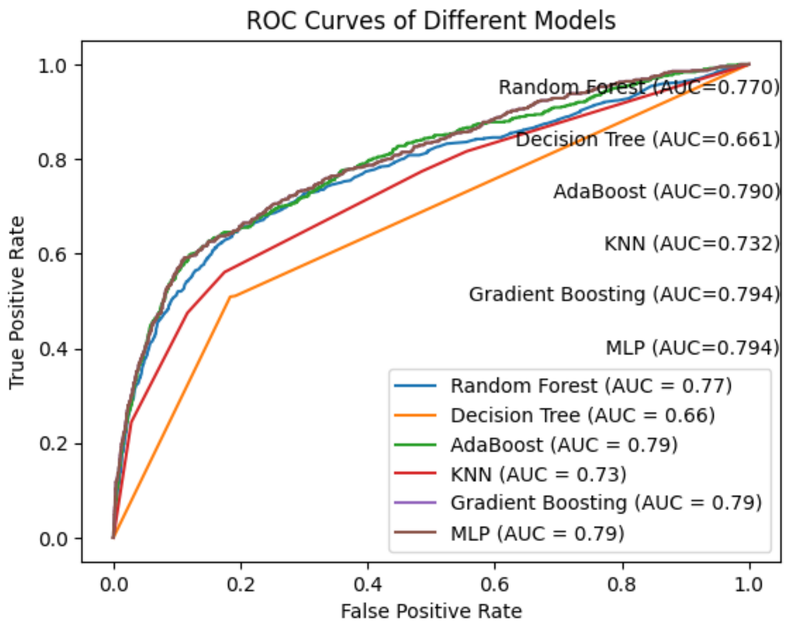
Based on multiple recommendation from each time I ran this feature selection step, I can test several ways to define best features with high impact on class prediction. Furthermore, by removing irrelevant or redundant features, the model becomes less likely to fit noise in the data, which can result in overfitting and to generalize well to unseen data.

MODELING - USE TRAINED MODELS AND KAGGLE AUC

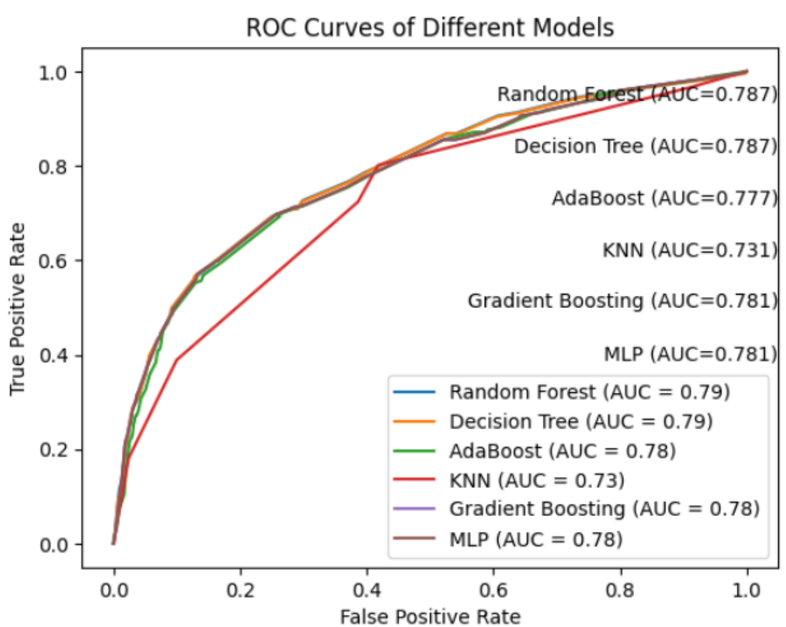
I use Machine Learning Algorithms for Classification to predict discrete class labels, including:

- Random Forest Classifier
- Gradient Boosting Classifier
- Decision Tree Classifier
- AdaBoost Classifier
- K-Nearest Neighbors Classifier
- Multi-Layer Perceptron (MLP) Classifier

As I mentioned on data preparation slide, I will use these model to train on train set in which I train model with full features and selected-features. I will compare the AUC of each model in each approach to find the model with highest AUC.



Full features



Selected features

Compare AUC of different models in each approach (full features and selected features, I concluded that step to select important features enhance the performance of models. In addition, Gradient Boosting, Decision Tree and Random Forrest are models with higher AUC. Therefore, I selected these model to test on test set.

On Kaggle, the predictions when using **Gradient Boosting** have highest AUC. Therefore, this model has the highest performance on both validation set and test dataset. Another thing I consider is that over resampling with **ratio_scaler = 3** and **oversample_ratio = 1**. This ratio for over resample minority class is the most effective approach.

RECOMMENDATIONS & LIMITATIONS

RECOMMENDATION

The selected features which create highest AUC are:

['month_dec', 'civil_married', 'month_oct', 'presult_success', 'job_unknown', 'month_nov', 'month_aug', 'month_jun', 'cprice', 'credit_unknown', 'employment', 'job_management', 'pcontact', 'age', 'edu_6K', 'pca_1', 'pca_6', 'pca_3', 'job_management', 'pca_5', 'pca_7', 'job_selfemployed', 'presult_failure', 'hloan_yes']

I provide the following recommendations for the bank to enhance their marketing campaigns and attract more customers to open saving accounts:

- **Target specific months:** Focus marketing efforts during the months of August, October, November, and December, as customers are more likely to open saving accounts during these periods. This could be due to various factors such as end-of-year bonuses or holiday spending patterns.
- **Focus on married customers:** Married individuals are more likely to open saving accounts. Tailor marketing campaigns to emphasize the benefits of saving accounts for couples and families, such as joint accounts, financial planning, and long-term savings goals.
- **Emphasize successful past campaigns:** Highlight previous marketing campaigns that resulted in successful account openings. This could involve showcasing testimonials or case studies from satisfied customers who have benefited from the bank's saving accounts.
- **Target specific job sectors:** Focus on customers working in management positions or individuals with unknown job titles, as they are more likely to open saving accounts. Develop targeted campaigns that cater to the financial needs and aspirations of these specific groups.
- **Leverage preferred contact methods:** Utilize the most effective communication channels to reach potential customers. For example, if phone contact has proven successful in the past, allocate more resources to telemarketing campaigns.
- **Consider age as a factor:** Since age is an important feature, target older individuals who may be more financially stable and looking for secure saving options. Emphasize the benefits of saving accounts for retirement planning or long-term financial goals.
- **Appeal to specific education levels:** Target customers with specific education backgrounds, tailor marketing messages to resonate with their financial goals and aspirations.

LIMITATIONS

While this analysis project has provided valuable insights into factors affecting the opening of saving accounts, there are several limitations that should be acknowledged:

- **Imbalanced dataset:** The dataset's outcome variable is highly imbalanced, which can lead to a biased model that favors the majority class. Although I have applied the SMOTE technique to mitigate this issue, there may still be some residual bias affecting the model's performance.
- **Feature selection process:** The feature selection method used in this study may not have identified all the relevant features or may have excluded some important predictors. Different feature selection techniques could yield alternative sets of important features, which could affect the conclusions drawn from the analysis.
- **Model selection:** I evaluated a variety of classification models to identify the best-performing ones. However, there may be other models or model configurations that could perform better. Furthermore, model performance is sensitive to hyperparameter tuning, and a more exhaustive search for optimal hyperparameters might result in improved performance.
- **Temporal factors:** The dataset is based on historical data, and the analysis assumes that past patterns will continue to hold in the future. However, customer behavior, market conditions, and bank offerings may change over time, which could affect the relevance and applicability of the findings.
- **Causality versus correlation:** The analysis identifies correlations between the features and the target variable but does not establish causality. Further research is required to determine causal relationships and understand the underlying mechanisms driving these associations.

In conclusion, while the findings of this analysis project offer useful insights for marketing campaigns and customer targeting, it is essential to consider the limitations mentioned above. Future research could address these limitations by employing alternative feature selection techniques, exploring additional models, and conducting more in-depth investigations of causal relationships. This would further enhance the understanding of factors influencing customers' decisions to open saving accounts and inform more effective marketing strategies.