# Predict Customer Responses to Marketing Campaigns

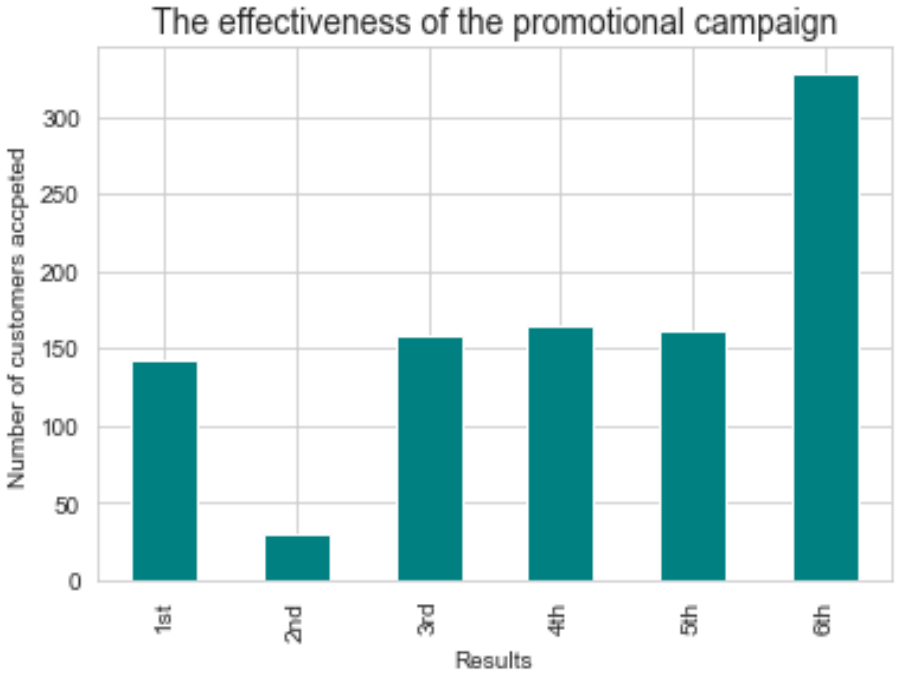
LI LI

APRIL 2022

SPRINGBOARD

DATA SCIENCE CAREER TRACK

# Problem Statement & Context




The effectiveness of 6 marketing campaigns

- Marketing campaigns aim to encourage customers to act in some manner. The main interest lies in how customers respond to the last 6th campaign.
- This project trained five different machine learning models to best predict customer responses:
  - What are key factors based on customers' behaviors?
  - Identify what types of customers the business is trying to reach and optimize new campaigns by refining audience selection.

# Introduction to Data

- The dataset was obtained from Kaggle's competition webpage: (2240 rows and 29 columns)

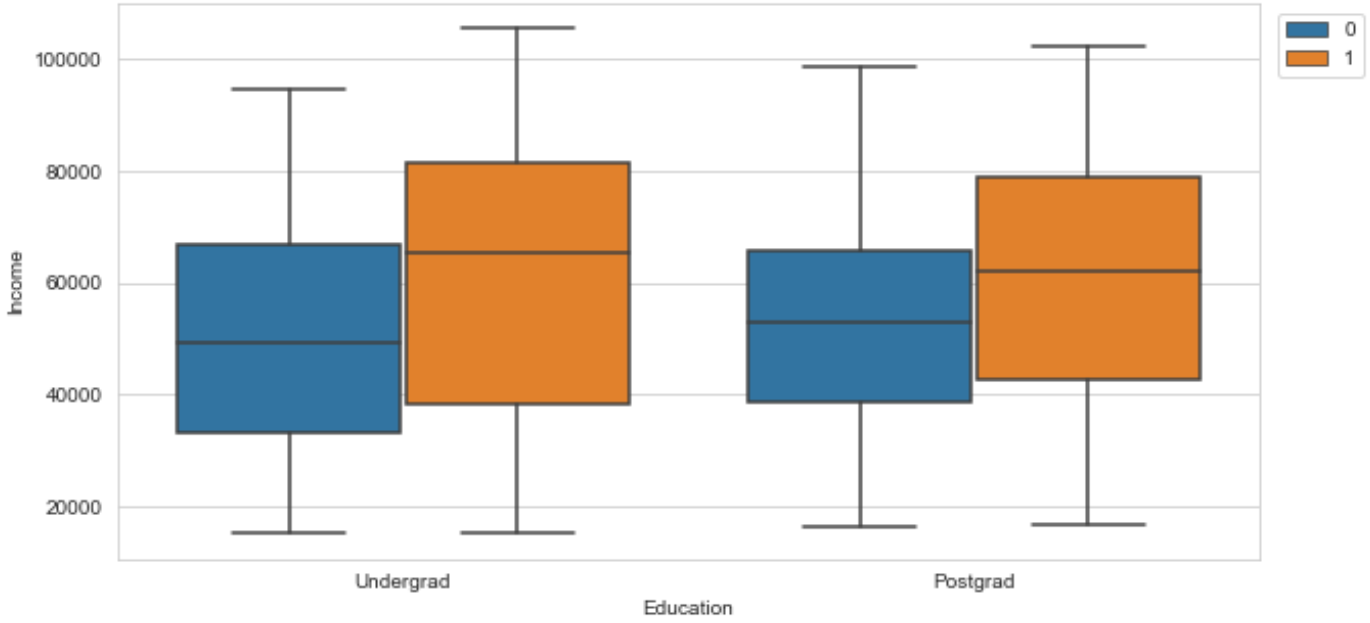
  https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis?select=marketing_campaign.csv

- Data cleaning
  - Missing values are filled in with the statistical value (median)
  - Drop less useful columns
  - Replace the names of the columns
  - Group classes within columns
  - Generate new features

# Exploratory Data Analysis (EDA)

EDA includes:

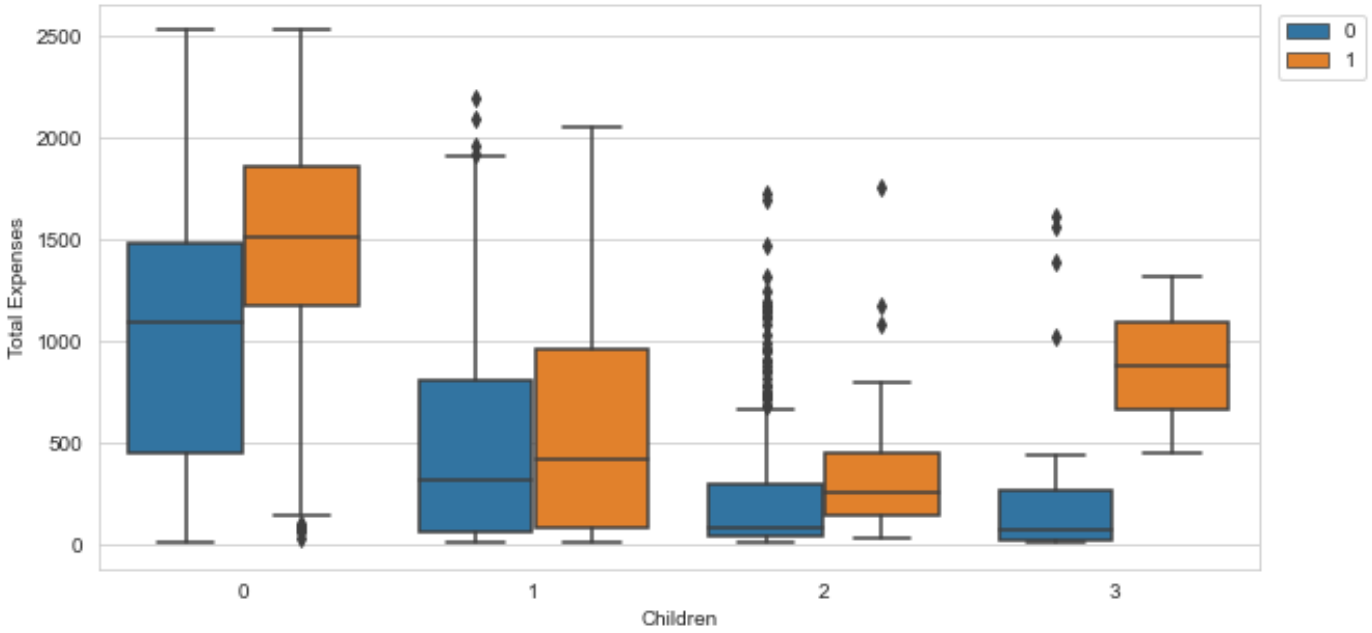
- Visualization by groups:
  - Income
  - Total expenses
  - Days enrolled
- PCA to look at the data from the most informative viewpoint



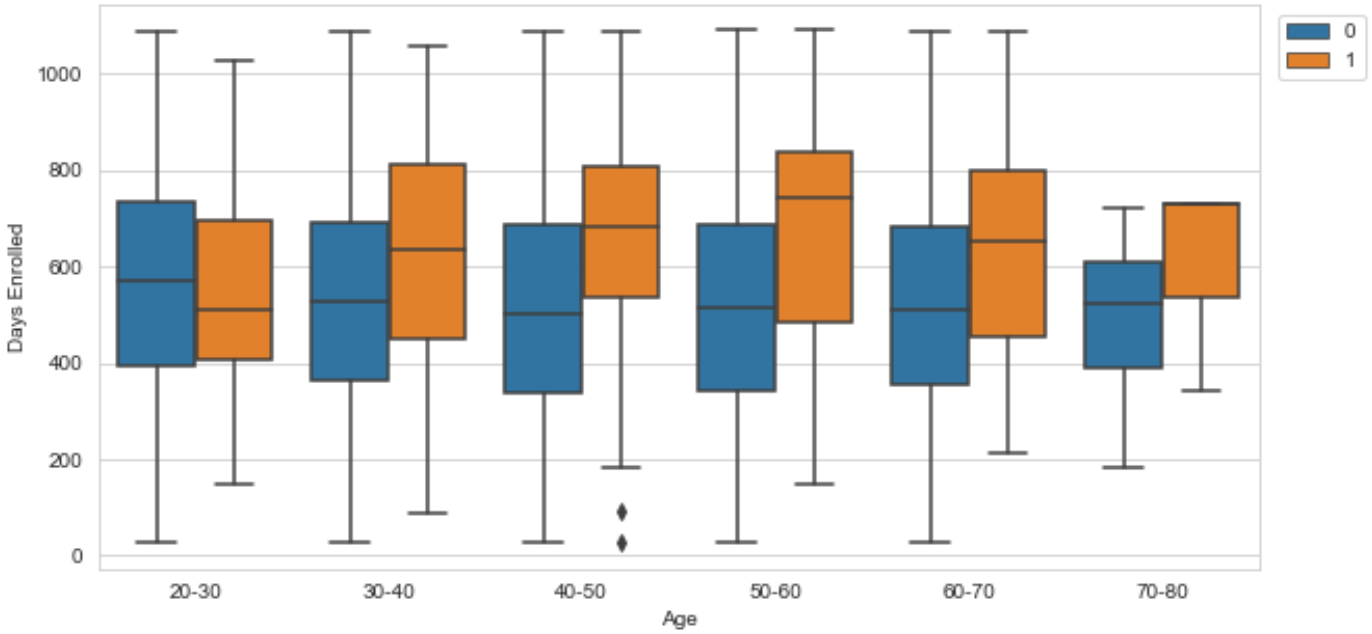

# EDA – by Income

- Customers with higher income & undergrad backgrounds are more active with responses to the last campaign.

# EDA – by Total Expenses




- Customers with '0' and '1' children are more active in the last campaign than those of '2' and '3'.

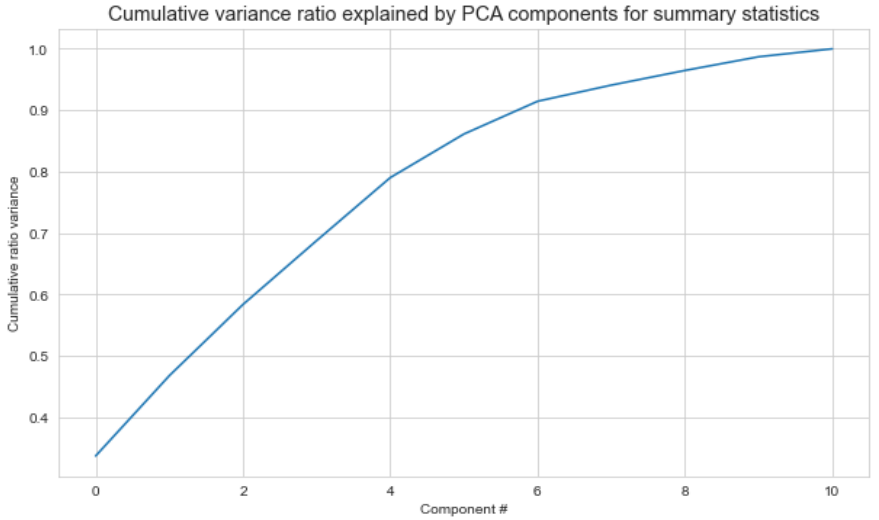


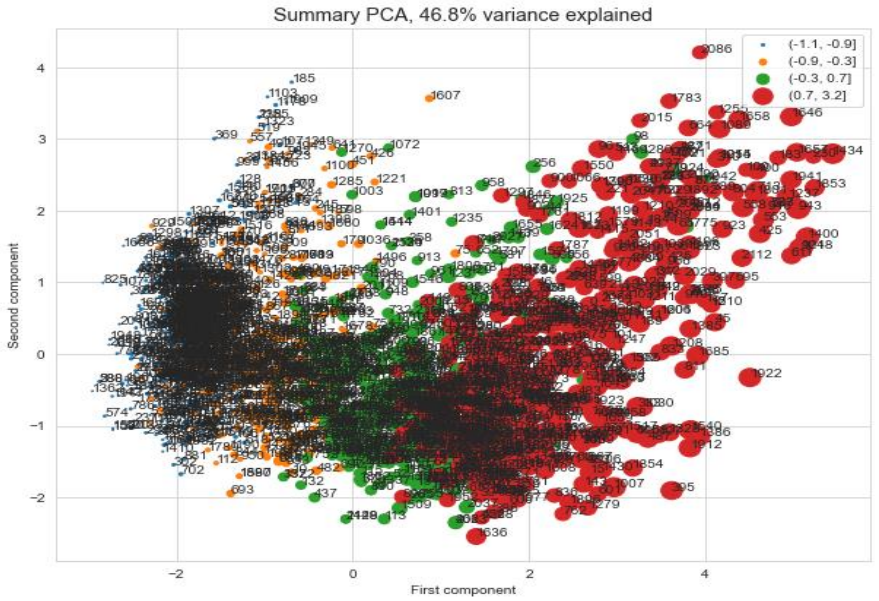

# EDA – by Days Enrolled

- Customers in different age groups with longer enrolled days are more active in the last campaign. Those of age 20- 30 are slightly different.

# PCA

- The first two components of data in the new dimension seem to account for about 50% of the variance, and the first five for over 80%.
- The red points represent the upper quartile of 'TotalExpense' and they spread across the first dimension (>0). There's also a spread of the other quartiles as well.



# Data Preprocessing

- Final data includes:
  - One-hot encoding to convert two categorical variables ‘education’ & ‘marital status’ to binary data (0/1)
  - Target variable ‘AcceptedCmp6’: the 0/1 indicators for customers who responded to a given offer or not. It is highly unbalanced between the number of responsive customers and that with have no responses.

# Approach

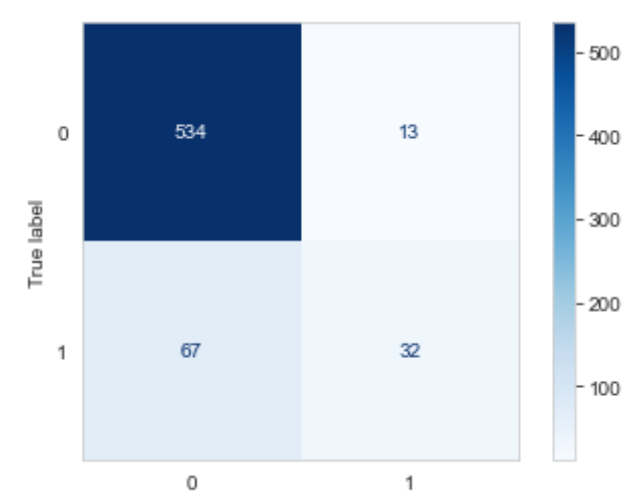
- Five different machine learning algorithms are used to train on the data in order to build an effective classifier that can identify the active customers:
  - Non-ensemble algorithms:
    - Logistic regression
    - KNN
  - Ensemble Algorithms:
    - Random forest
    - Gradient boosting
    - XGBoost
- Each ML model conducted hyperparameter tuning and cross-validation to improve model accuracy
- Evaluation of the model performance

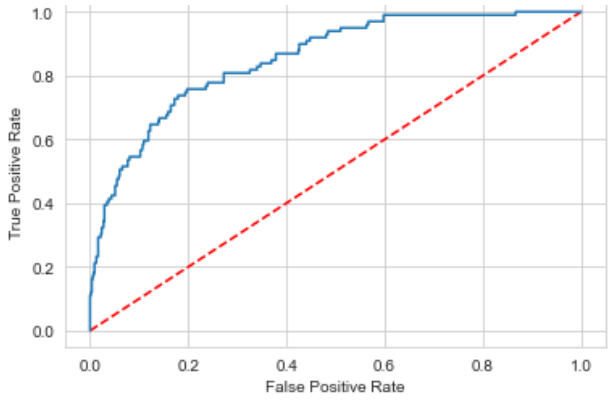
# Hyperparameter Tuning

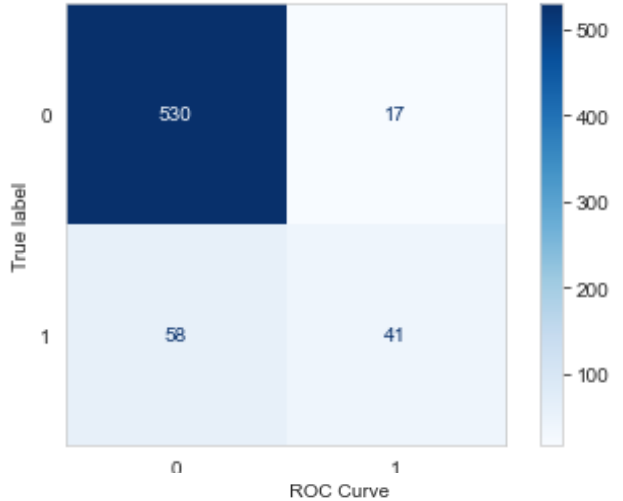
- Logistic Regression
  - Grid search over to find the best regularization parameter $C$ based *only* on the training set; Cs = [0.001, 0.1, 1, 10, 100]
  - C controls the inverse of the regularization strength. A large C can lead to an overfit model, while a small one can lead to an underfit model.
- Random Forest
  - Grid search for the best number of estimators. (n_est=50)
  - Max_depth (10)
- KNN
  - Find the best number of neighbors k=2 using the elbow method
- Gradient Boosting
  - find the best parameters: `{'learning_rate': 0.25, 'max_depth': 2, 'max_features': 2, 'n_estimators': 20}`
- XGBoost
  - "objective":'binary:logistic', "max_depth":3

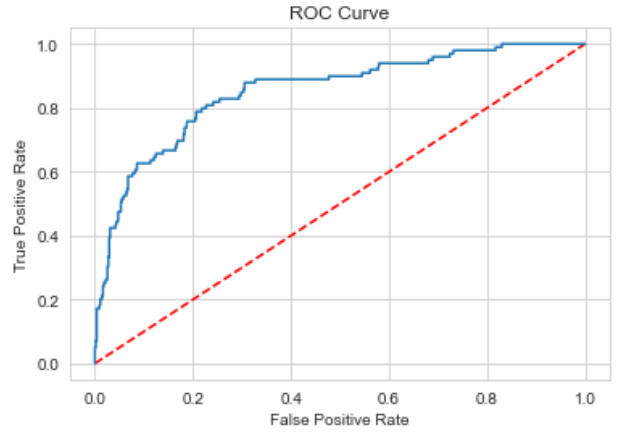
# Metrics for Binary Classification

- Confusion matrix
- Roc_auc score
- left: logistic regression
- right: random forest



# Model Performance

- The results from implementing each model are as the following:

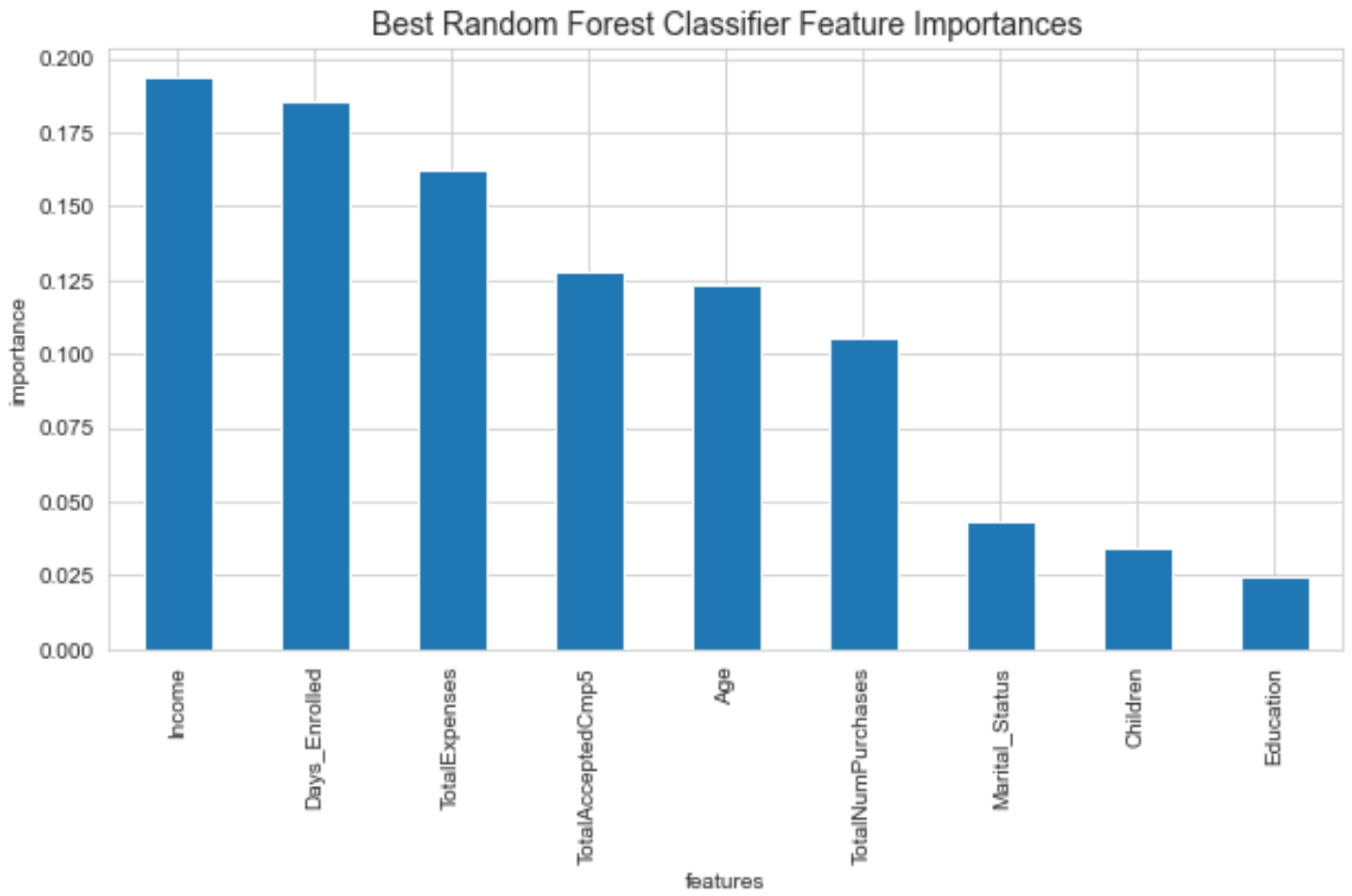
| | name | ROC_AUC |
|---|---|---|
| 0 | LogisticRegression | 0.853120 |
| 1 | RandomForestClassifier | 0.852898 |
| 2 | KNeighborsClassifier | 0.808976 |
| 3 | GradientBoostingClassifier | 0.821229 |
| 4 | XGBClassifier | 0.827286 |

# Model Selection

- Final optimized models are evaluated based on their predictions, confusion matrix, and roc_auc scores.
- Both Logistic Regression and Random forest classifiers have very strong AUC of 0.85, but the Random forest model has a higher True Positive Rate (TP/P) based on the confusion matrix. The model performance on the test set is consistent with the cross-validation results.
- The best model chosen for this dataset is the random forest classifier, which is typically fast to train, easy to tune, and less likely to overfit.

# Important Features

- ‘Income’
- 'Days_Enrolled’
- 'TotalExpenses’
- 'TotalAcceptedCmp5' - total except the last champaign
- 'Age'



# Summary

- Final optimized models are evaluated based on their predictions, confusion matrix and roc_auc scores.

- Both Logistic Regression and Random forest classifiers have very strong AUC of 0.85, but the Random forest model has a higher True Positive Rate (TP/P) based on the confusion matrix. The performance on the test set is consistent with the cross-validation results. The best model chosen is the random forest classifier, which is fast to train, easy to tune, and less likely to overfit.