Predicting the Success of Bank Telemarketing

By:
Boyuan Chen
Qingyuan Chen
Ruiqi Jiang
Yumeng Tang

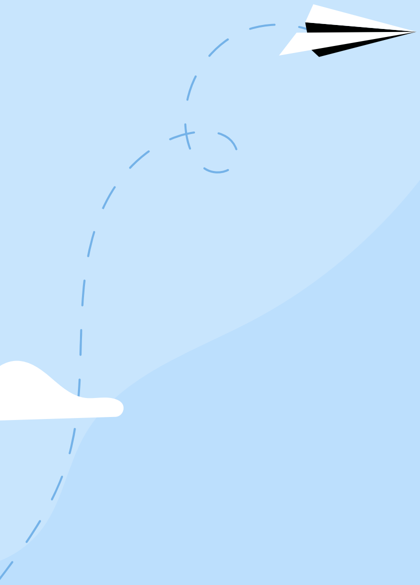
Table of Contents

01. Background

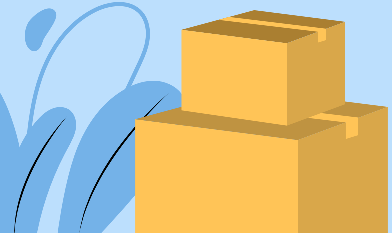
02. Data Exploration

03. Modeling

04. Business Insights & Conclusion



01. Background



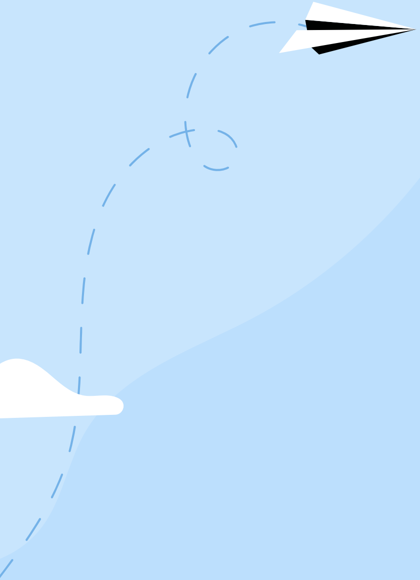
Problem Definition & Data Source

Demographic Information	***Business-related features***
Job	Contact
Marital	day_of_month
Education	month
Default	p_outcome
Housing	duration
loan	campaign
age	pdays
balance	previous

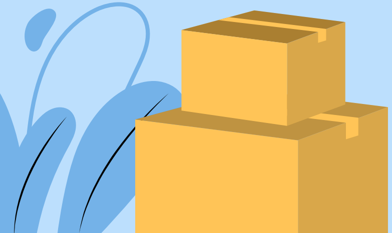
This dataset pertains to banking marketing initiatives conducted by a bank in Portugal.

The objective of this classification task is to predict whether a client will opt in ('yes') or opt out ('no') of a term deposit (denoted as variable y).

Data Source: Bank Marketing - UCI Machine Learning Repository



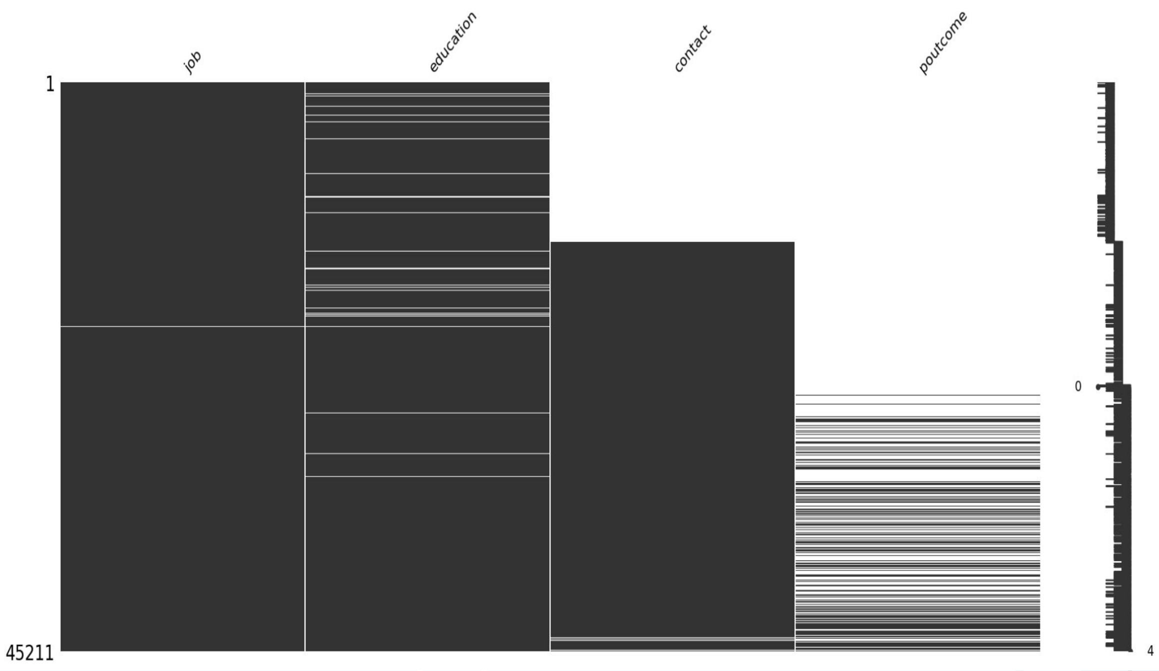
02. Data Exploration



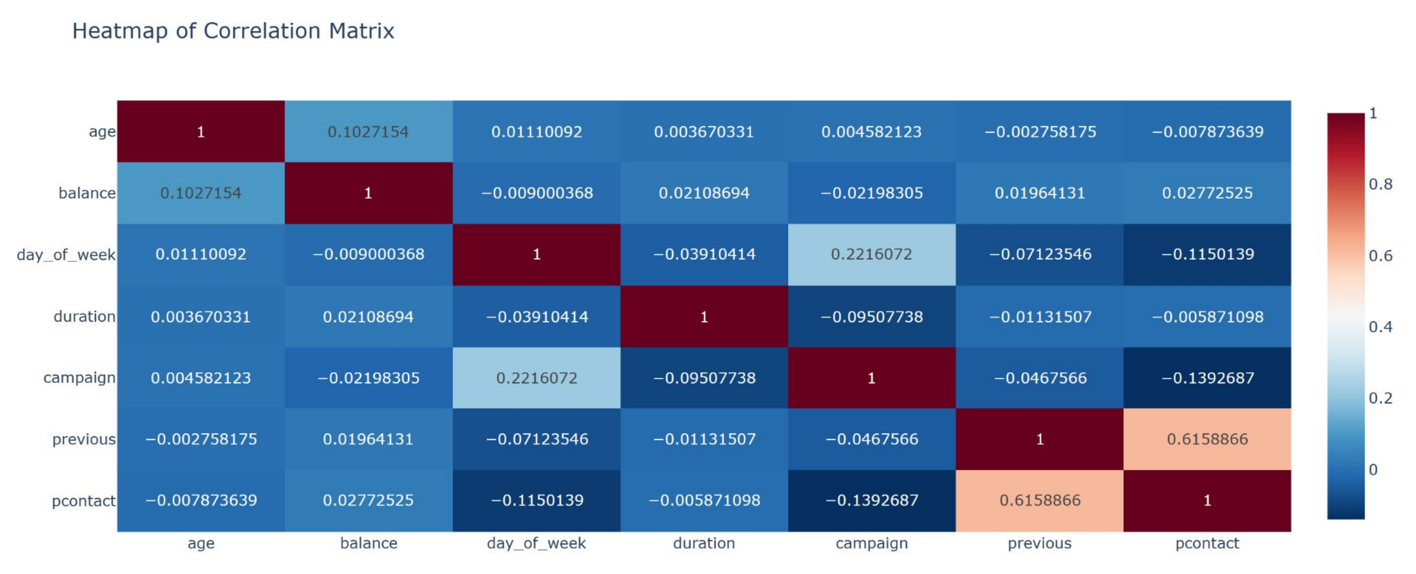
Data cleaning

If null value has no analytical meaning and can not be converted, we drop rows with null (job, education, contact);

If null value can be converted meaningfully, we replace any value except success to failure (poutcome).



Data Exploration

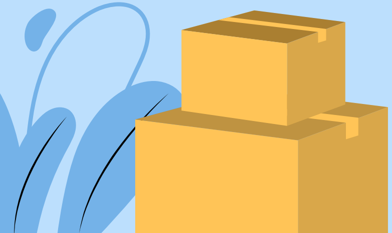


Heatmap of Correlation Matrix

	age	balance	day_of_week	duration	campaign	previous	pcontact
age	1	0.1027154	0.01110092	0.003670331	0.004582123	−0.002758175	−0.007873639
balance	0.1027154	1	−0.009000368	0.02108694	−0.02198305	0.01964131	0.02772525
day_of_week	0.01110092	−0.009000368	1	−0.03910414	0.2216072	−0.07123546	−0.1150139
duration	0.003670331	0.02108694	−0.03910414	1	−0.09507738	−0.01131507	−0.005871098
campaign	0.004582123	−0.02198305	0.2216072	−0.09507738	1	−0.0467566	−0.1392687
previous	−0.002758175	0.01964131	−0.07123546	−0.01131507	−0.0467566	1	0.6158866
pcontact	−0.007873639	0.02772525	−0.1150139	−0.005871098	−0.1392687	0.6158866	1

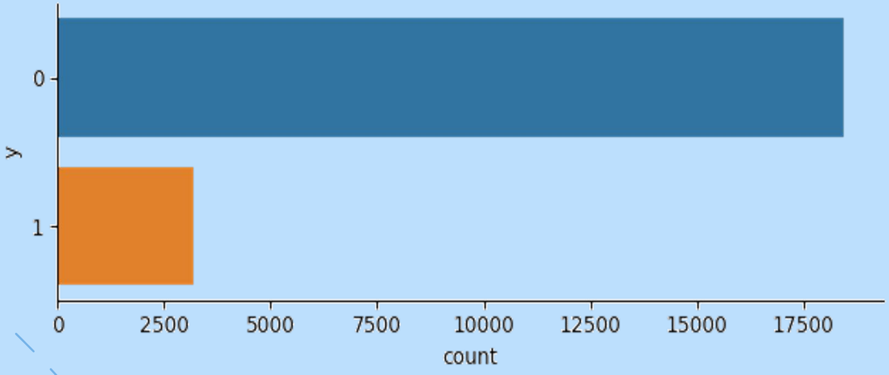
Low Dependence & Correlation between Numeric Features

03. Modeling



Dealing with imbalanced data

Since imbalanced dataset leads to model bias and misleading metrics, we applied the Adaptive Synthetic Sampling on the training dataset.



Class imbalance in training dataset

	TRAIN (70%)		TEST (30%)	
Successful? (Y/N)	Y	N	Y	N
Before ADASYN	3186	18448	1327	7946
After ADASYN	18641	18448	1327	7946

Stacking System Design

Why Stacking System?

Aggregate 4 base models with different underlying assumptions and strengths to make more informed predictions.

Stacking System with Four Base Model Candidates:

- Logistic Regression
- Support Vector Machine
- Decision Tree
- Gradient Boosting Model

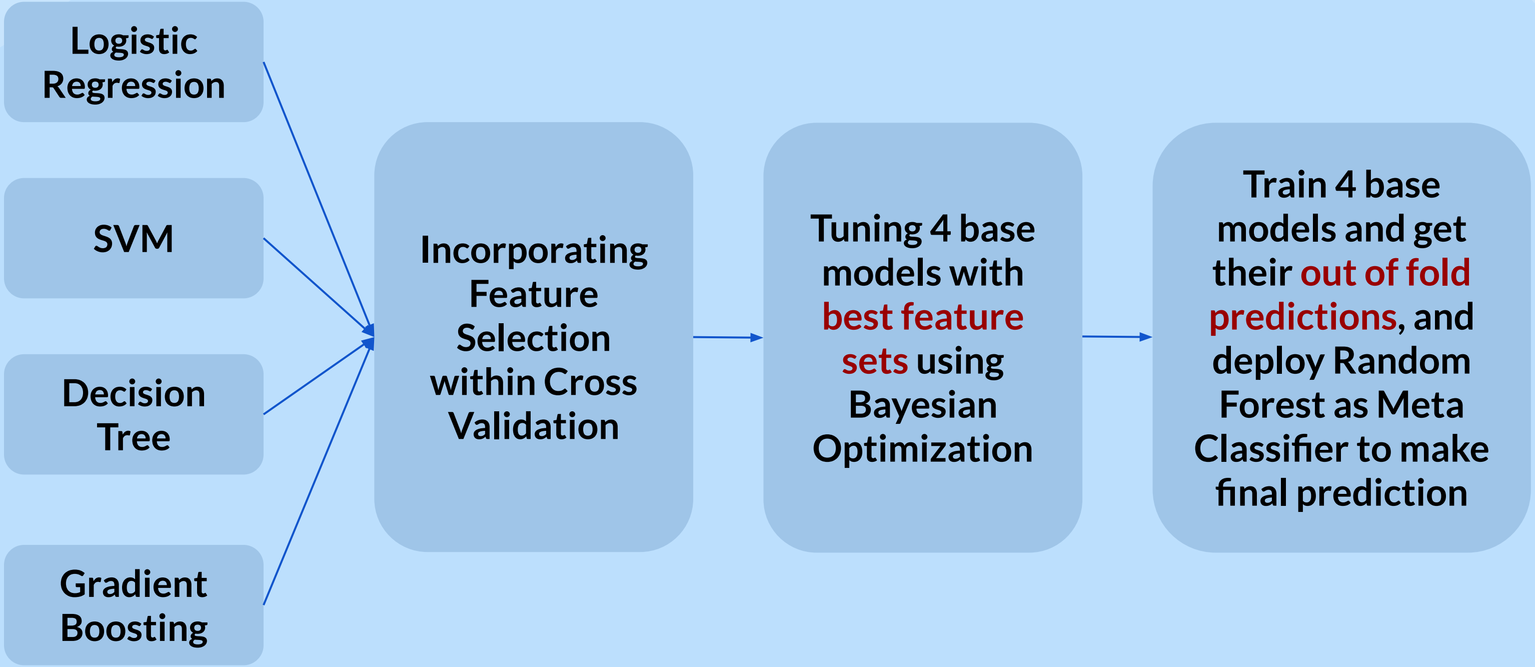
Meta Classifier: Random Forest

VS

LightGBM

XGBoost

Stacking System Pipeline



Logistic Regression

SVM

Decision Tree

Gradient Boosting

Incorporating Feature Selection within Cross Validation

Tuning 4 base models with best feature sets using Bayesian Optimization

Train 4 base models and get their out of fold predictions, and deploy Random Forest as Meta Classifier to make final prediction

Average CV Training Accuracy: 0.906

Feature Selection for LightGBM

Strategy:
Backward Stepwise Selection
Best Train Accuracy Score:
0.879

Numerical Features: **7**
age, balance, duration, campaign, pcontact, previous, day_of_month

Categorical Features: **29**

job:
job_blue-collar,
job_entrepreneur,
job_housemaid,
job_management,
job_self-employed,
job_services,
job_student,
job_technician,
job_unemployed

marital:
marital_married,
marital_single

education:
education_secondary,
education_tertiary

default:
default_yes

housing:
housing_yes

month:
month_aug,
month_dec
month_feb,
month_jan,
month_jul,
month_jun,
month_mar,
month_may,
month_nov,
month_oct,
month_sep

loan:
loan_yes

contact:
contact_telephone

poutcome:
poutcome_success

Feature Selection for XGBoost

Strategy:
Backward Stepwise Selection
Best Train Accuracy Score:
0.890

Numerical Features: 6
age, balance, duration,
campaign, pcontact,
day_of_month

Categorical Features: 28

job:
job_blue-collar,
job_entrepreneur,
job_housemaid,
job_management,
job_self-employed,
job_services,
job_student,
job_technician,
job_unemployed

marital:
marital_married,
marital_single

education:
education_secondary,
education_tertiary

housing:
housing_yes

month:
month_aug,
month_dec
month_feb,
month_jan,
month_jul,
month_jun,
month_mar,
month_may,
month_nov,
month_oct,
month_sep

loan:
loan_yes

contact:
contact_telephone

poutcome:
poutcome_success

Hyper-parameter Tuning for LightGBM and XGBoost

Strategy:
Bayesian Optimization
Best Train Accuracy Score:
LightGBM: **0.934**; XGBoost: **0.931**

LightGBM

num_leaves: 104,
learning_rate: 0.14,
max_depth: 15,
min_child_samples: 13,
subsample: 0.58,
colsample_bytree: 0.39,
reg_alpha: 0.95, reg_lambda: 0.55,
min_child_weight: 5.30,
feature_fraction: 0.47,
bagging_fraction: 0.99, bagging_freq: 6,
max_bin: 803, min_data_in_leaf: 62

XGBoost

learning_rate: 0.20, max_depth: 10, min_child_weight: 7,
subsample: 0.57,
colsample_bytree: 0.55,
gamma: 0.69,
reg_alpha: 0.78,
reg_lambda: 0.31,
num_leaves: 56,
min_child_samples: 16,
feature_fraction: 0.58,
bagging_fraction: 0.57,
bagging_freq: 2,
max_bin: 238,
min_data_in_leaf: 17

Comparison of Three Classifiers

Stacking

Base models:
Gradient Boosting,
Decision Tree,
Logistic Regression,
Support Vector Machine

Meta model:
Random Forest

Train Accuracy: 0.906

LightGBM

Train Accuracy: 0.934

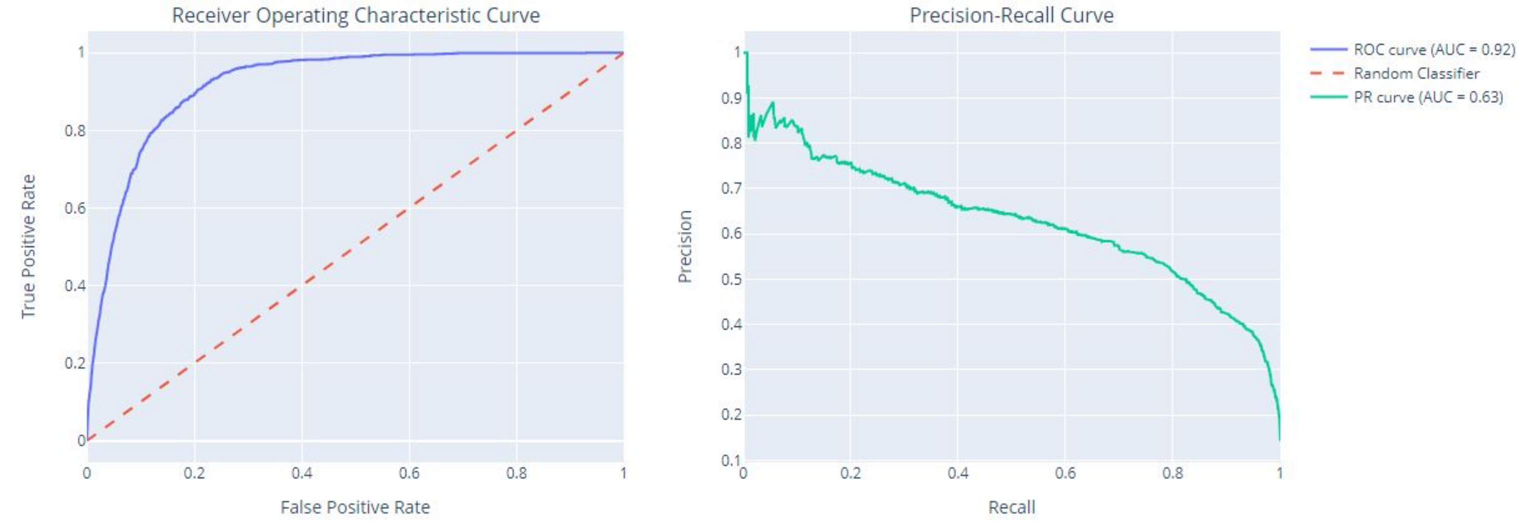
↓

Test Accuracy: 0.888

XGBoost

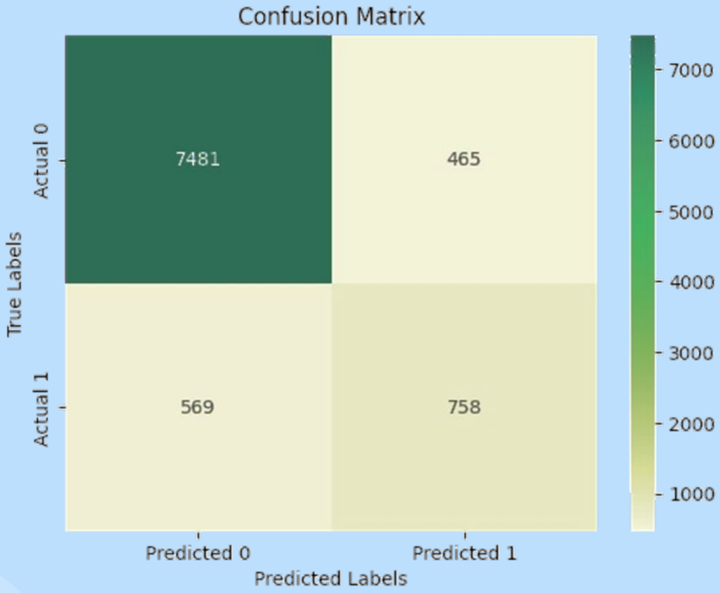
Train Accuracy: 0.931

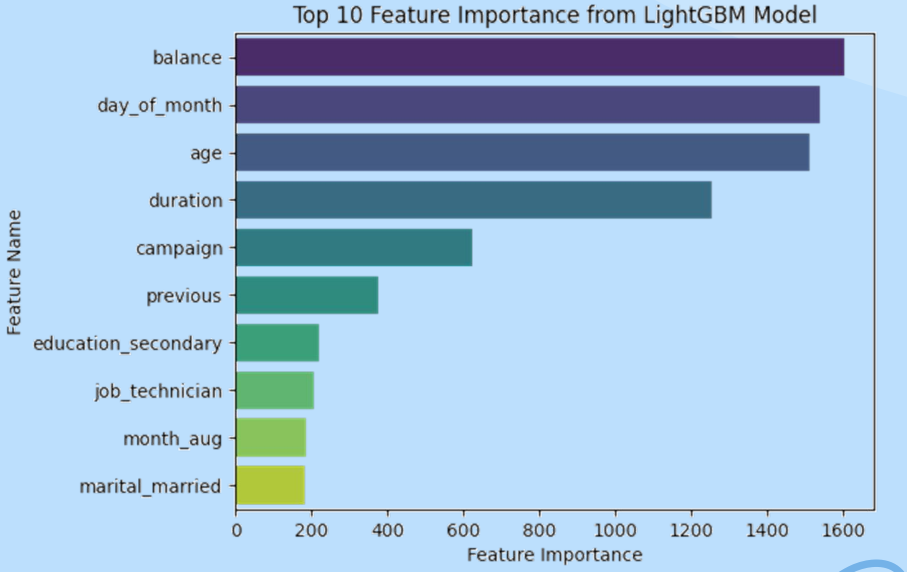
Performance Summary



- The ROC curve leans towards the **upper-left corner**, and the **AUC** is close to **1**.
- The Precision-Recall curve approaches the **upper-right corner**.

Performance Summary

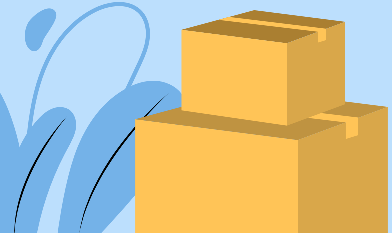




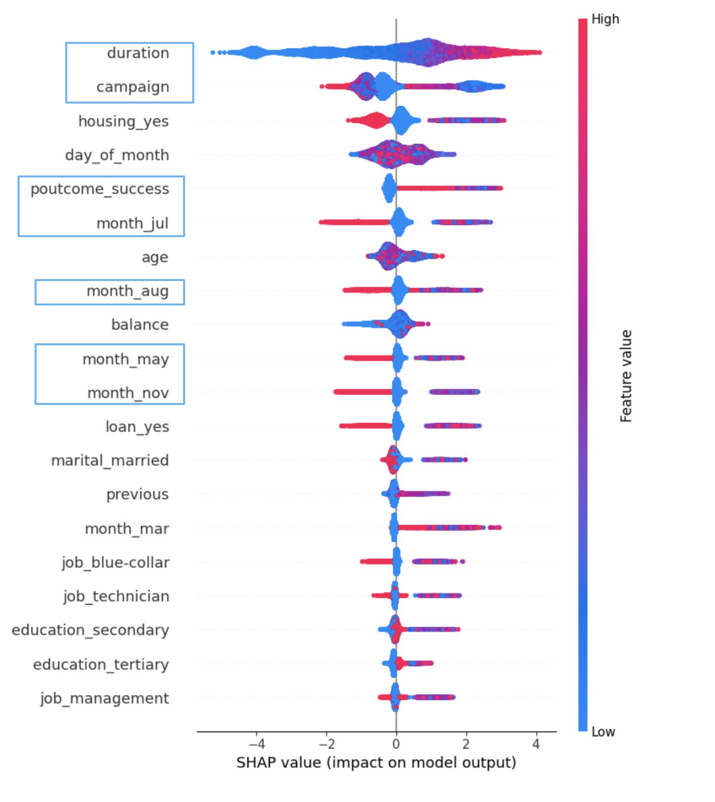
True Positive: 758; **True Negative**: 7481;
False Positive: 465; **False Negative**: 569

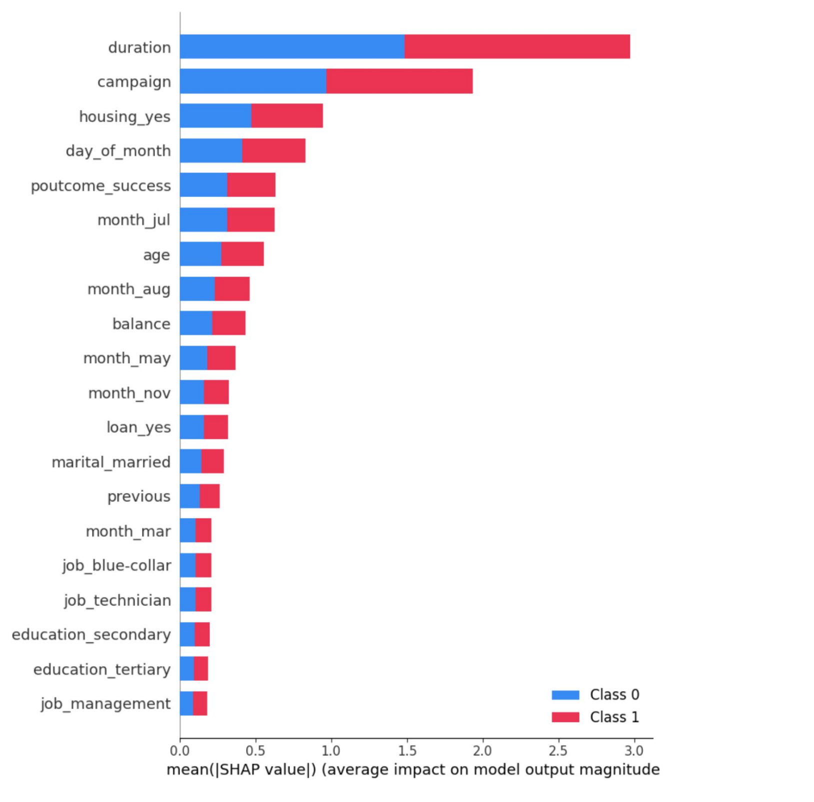
Top 4 features: above 1200;
Top 5-10 features: below 800

04. Business Insights & Conclusion

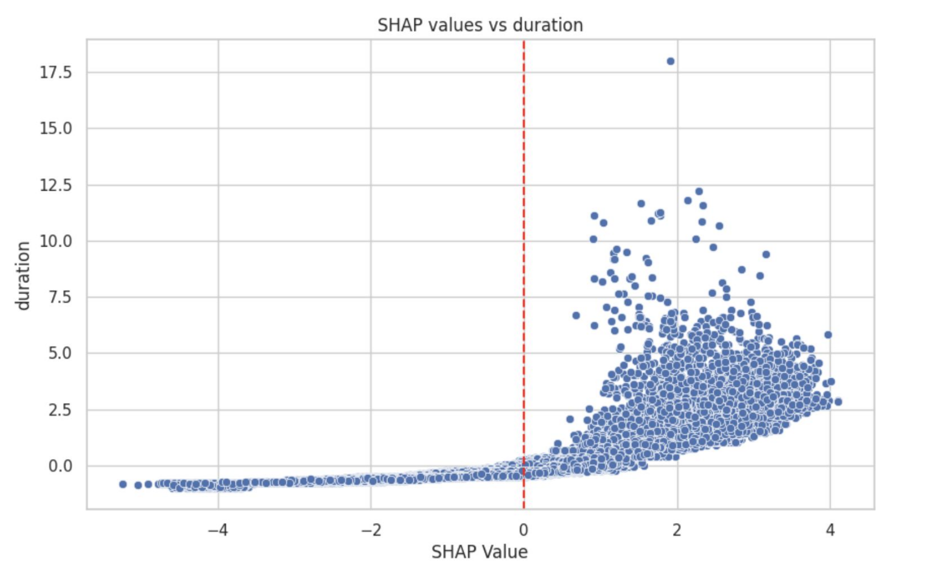


SHAP Summary Plot



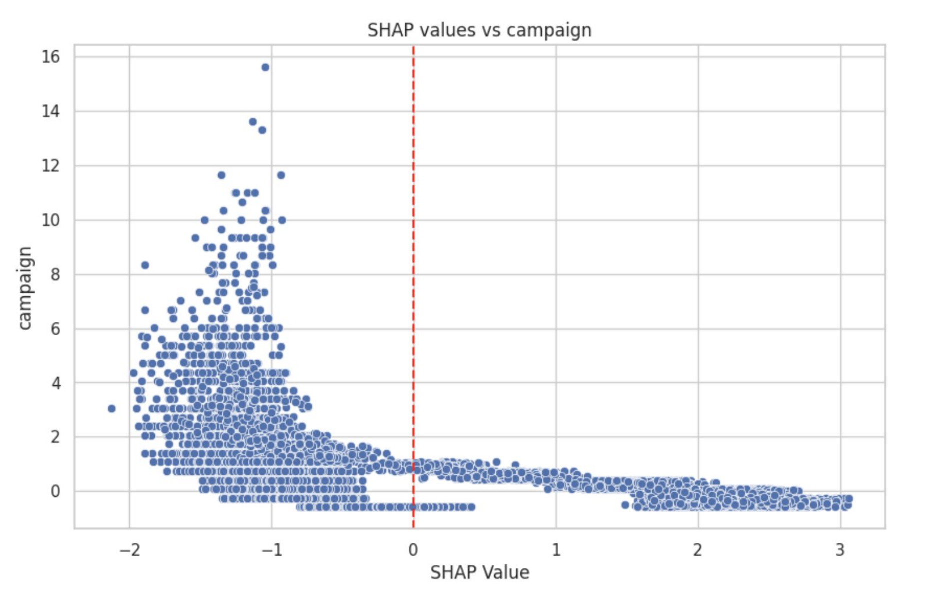


SHAP Values vs Duration



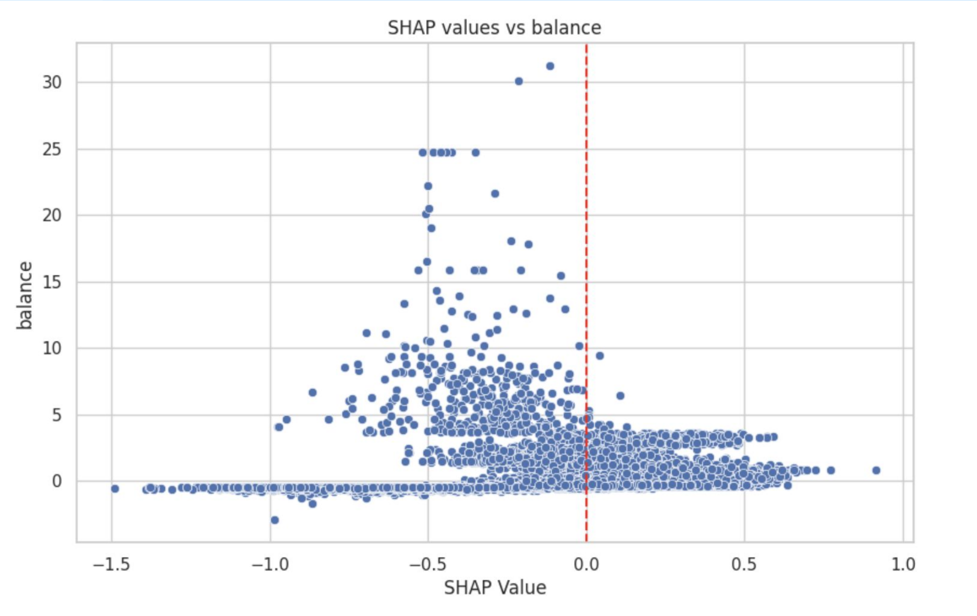
- High SHAP value
- Most impactful feature
- Duration>0, positive impact on successful outcomes

SHAP Values vs Campaign



- Number of contracts decreases the possibility of successful outcomes
- Focus more on quality rather than quantity

SHAP Values vs Balance



- Lower balance: not clear in either direction
- Higher balance: impact on prediction becomes negative

Thank You!

Q&A

Colab Link: https://drive.google.com/file/d/1BFzD6AI5aoIdC99X-aGqR2nq7aXUbOpv/view?usp=sharing