Predicting Term Deposit Subscription Using Bank Marketing Data

ITEC-600 Group 3

Shu-Wen Teng, Soyoung Yoon

Yen-Jo Lee, Yen-Chun Lin

December 10, 2024

Project Objective and Business Context

Objective:

Predict customer likelihood of subscribing to term deposits and provide actionable marketing insights

Key Questions:

- Business Question: How can the bank optimize direct marketing strategies to increase subscriptions?
- Analytics Question: What factors drive a customer’s likelihood to subscribe following a campaign?

Dataset Overview

The dataset is related with direct marketing campaigns (phone calls) of a Portuguese banking institution.

- Source: UC Irvine Machine Learning Repository
- Size: **45,211** observations and **17** variables

Group	Variables
Demographic	age, job, marital, education
Financial	default, balance, housing, loan
Campaign-Related	contact, day, month, duration, campaign, pdays, previous, poutcome
Target	y: Has the client subscribed to a term deposit? (yes/no)

Hypotheses and Analytical Approach

Key Hypotheses

- Financial stability (e.g., balance) increases subscription likelihood
- Campaign effectiveness (e.g., duration, poutcome) drives success

Analytical Approach

- Balance dataset using undersampling
- Explore various machine learning models, including Logistic Regression, Decision Trees, Random Forest, and SVM, to evaluate predictive performance

Methods and Preprocessing

Data preprocessing

- One-hot encoding for categorical variables like job, marital
- Scaled numerical variables

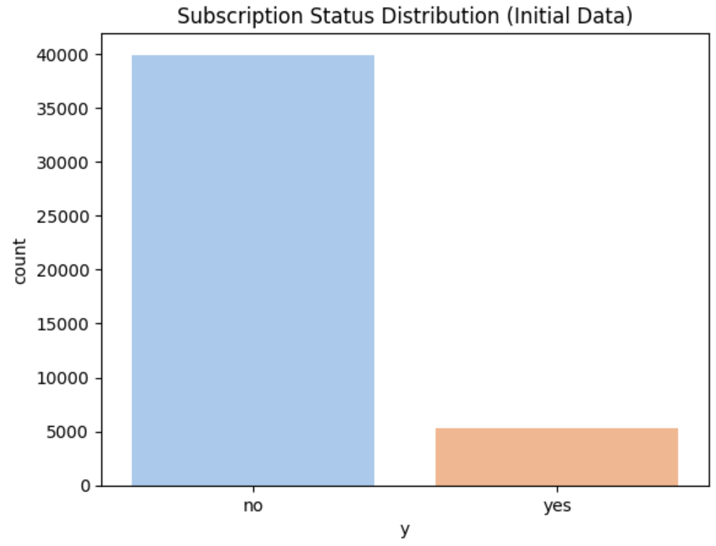
Resampling technique: Undersampling to balance y

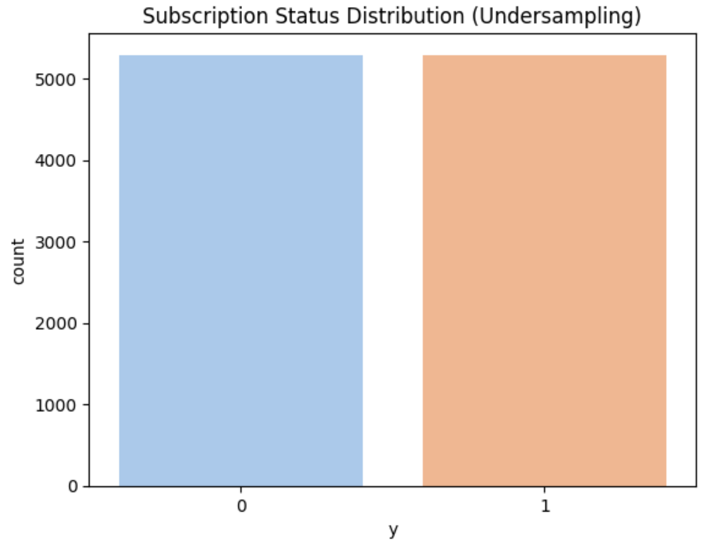
Model train-test split: 80:20

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
0	39	management	divorced	tertiary	no	517	yes	yes	unknown	14	may	1328	1	-1	0	unknown	yes
1	30	services	married	secondary	no	3929	yes	no	cellular	20	nov	593	1	-1	0	unknown	yes
2	46	management	divorced	tertiary	no	624	no	no	cellular	18	mar	420	1	276	1	other	yes
3	32	admin.	married	tertiary	no	653	no	no	cellular	2	jun	84	1	-1	0	unknown	yes
4	36	blue-collar	married	primary	no	319	yes	no	cellular	13	may	774	2	301	1	failure	yes

Key Data Challenges

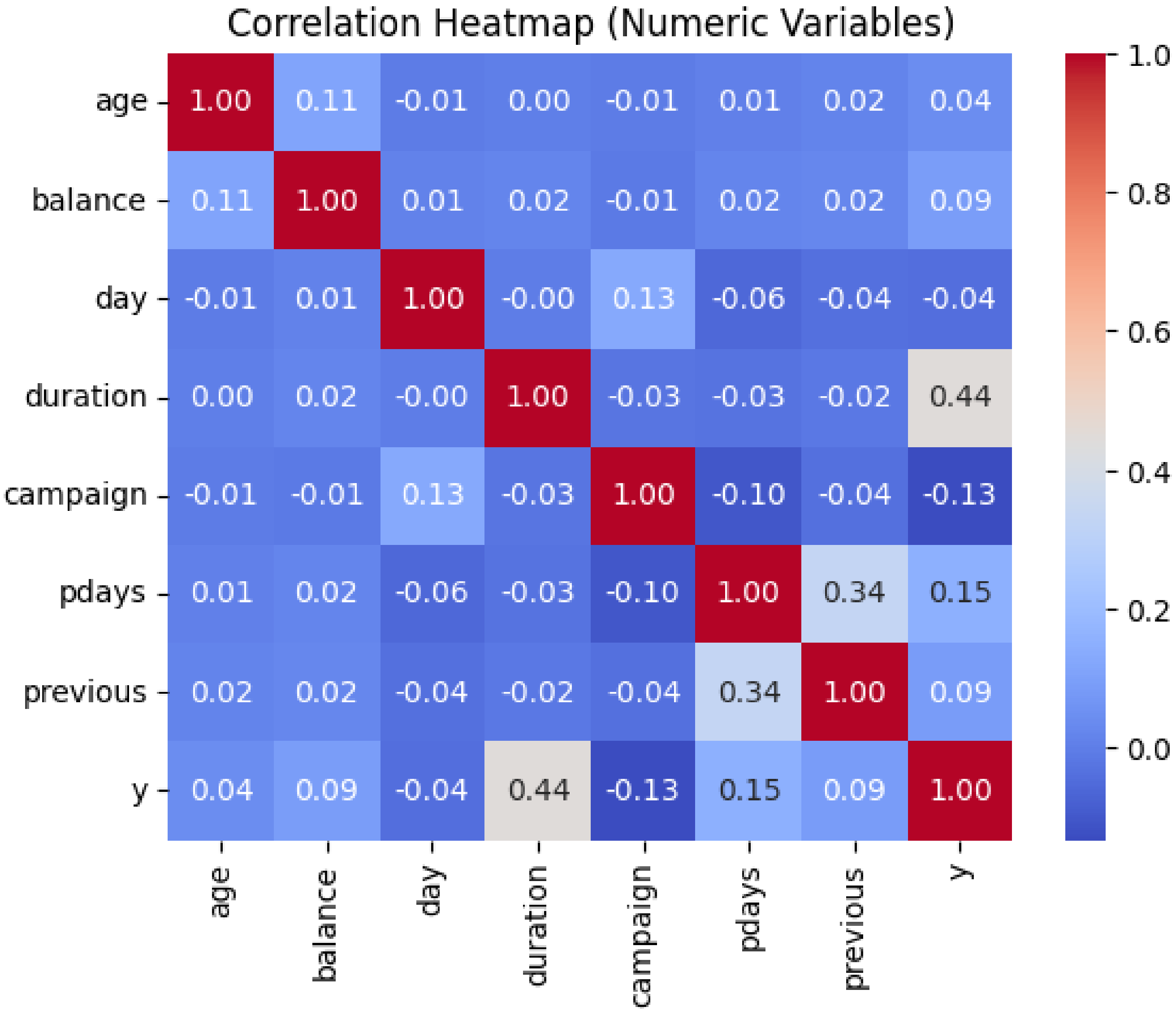
- Imbalanced target variable (y)
- Majority class (No), Minority class (yes) (39922:5289)
- Overfitting
- Undersampling (5289:5289)





Exploratory Data Analysis

- The strongest correlation: duration
- The length of the communication may be the most predictive numerical feature.
- The overall pattern of weak correlations
- Suggests that other factors, potentially categorical variables, might be more impactful.



Exploratory Data Analysis

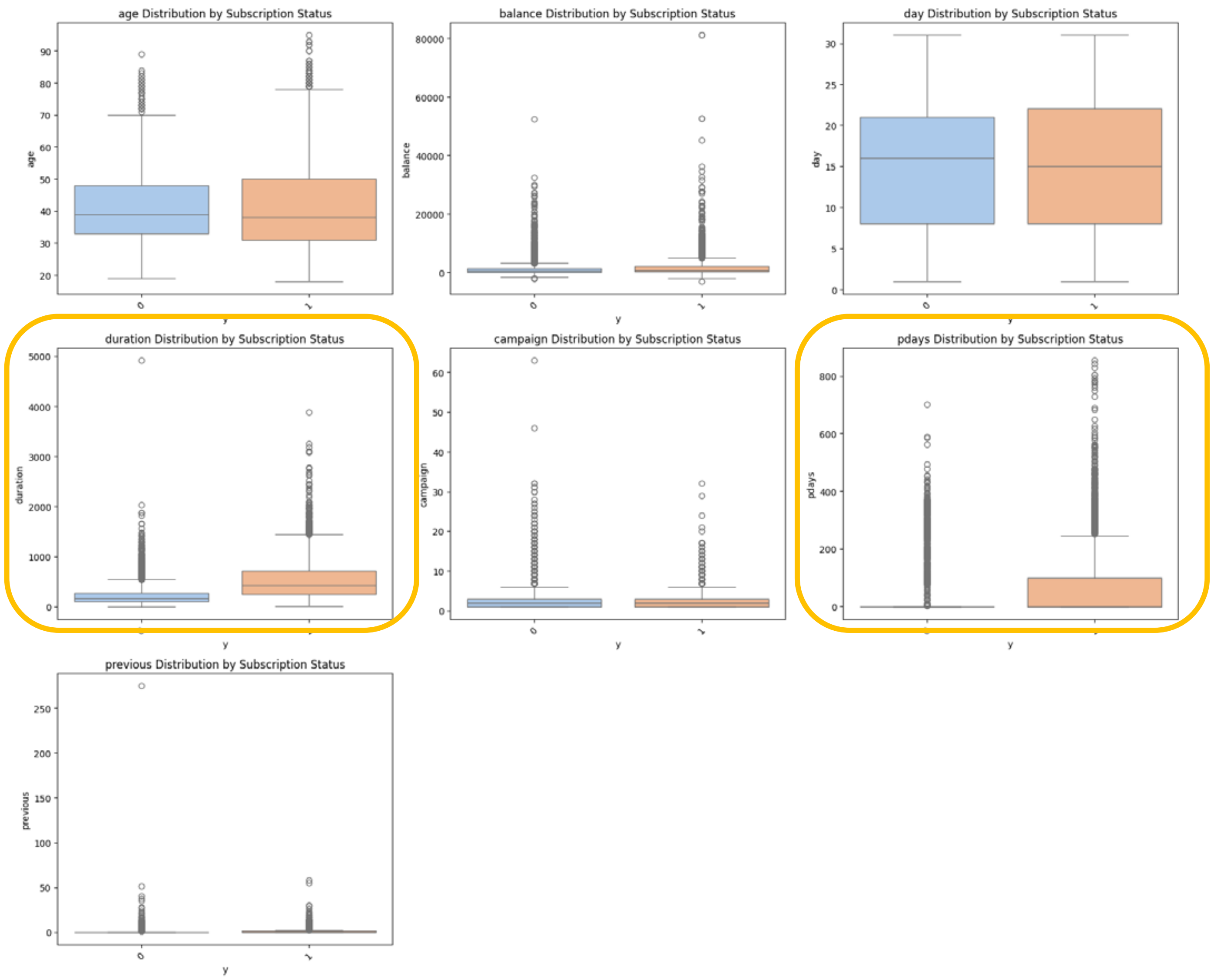
Variables related to subscription

- **Duration** and **pdays**

Variable less important in distinguish subscription

- **Age, balance, day, campaign, pervious**

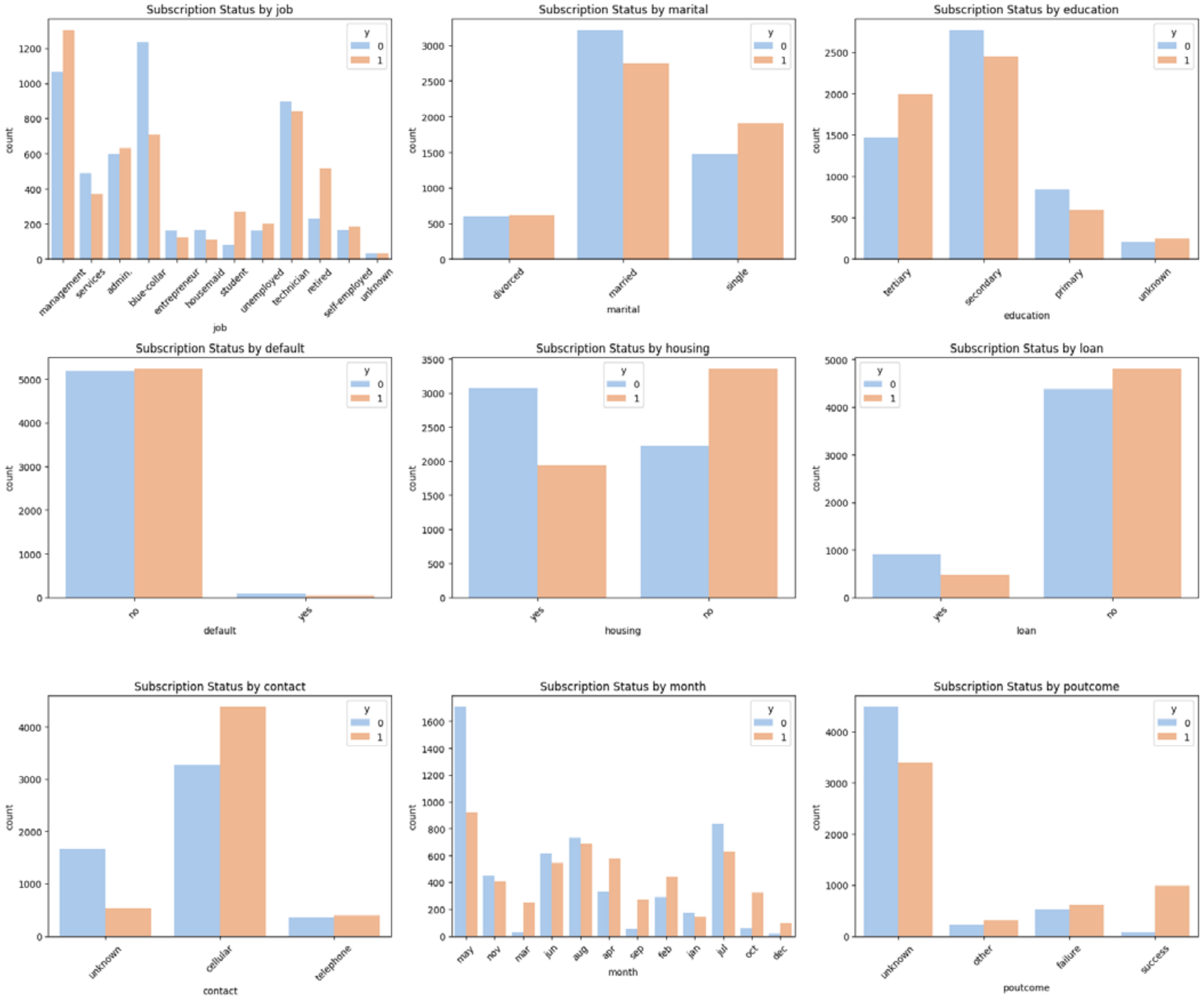
Outliers could indicate special cases or customer segments that may warrant further analysis.



Exploratory Data Analysis

Higher count of successful subscriptions (y=1)

- Job: management, admin, technician
- Marital: married
- Education: secondary, tertiary
- Default: no
- Housing: no
- Loan: no
- Contact: cellular



Logistic Regression Model

Demographic features: (age, job, marital, education)

- The p-value is 0.797 indicating that **clients' age** does not have a meaningful impact on the likelihood of subscribing to a term deposit.
- Individuals with **tertiary education** are significantly more inclined to subscribe to a term deposit.
- **Students** have a **positive slope** and with **p-value of 0.002**, indicating a significant and strong likelihood of subscription.
- Clients who are **single** have a slightly higher likelihood of subscribing to a term deposit, though this variable is **not statistically significant**.

Logit Regression Results

Dep. Variable:	y	No. Observations:	8462
Model:	Logit	Df Residuals:	8419
Method:	MLE	Df Model:	42
Date:	Mon, 09 Dec 2024	Pseudo R-squ.:	0.4147
Time:	04:55:02	Log-Likelihood:	-3433.1
converged:	True	LL-Null:	-5865.4
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.2892	0.225	5.733	0.000	0.848	1.730
age	0.0113	0.044	0.258	0.797	-0.075	0.098
job_blue-collar	-0.3240	0.120	-2.693	0.007	-0.560	-0.088
job_entrepreneur	-0.1607	0.205	-0.785	0.432	-0.562	0.240
job_housemaid	-0.6481	0.217	-2.993	0.003	-1.072	-0.224
job_management	-0.1593	0.123	-1.297	0.195	-0.400	0.081
job_retired	0.2542	0.170	1.499	0.134	-0.078	0.587
job_self-employed	-0.3111	0.194	-1.601	0.109	-0.692	0.070
job_services	-0.2711	0.140	-1.943	0.052	-0.545	0.002
job_student	0.6527	0.206	3.163	0.002	0.248	1.057
job_technician	-0.1596	0.114	-1.395	0.163	-0.384	0.065
job_unemployed	-0.3135	0.188	-1.667	0.095	-0.682	0.055
job_unknown	-0.3316	0.378	-0.878	0.380	-1.072	0.409
marital_married	-0.2156	0.099	-2.175	0.030	-0.410	-0.021
marital_single	0.1618	0.113	1.429	0.153	-0.060	0.384
education_secondary	0.2258	0.105	2.155	0.031	0.020	0.431
education_tertiary	0.4399	0.123	3.569	0.000	0.198	0.682
education_unknown	0.5341	0.173	3.089	0.002	0.195	0.873

Logistic Regression Model

Financial features: (default, balance, housing, loan)

- Clients **without a housing loan** are significantly more likely to subscribe to a term deposit.
- Clients **without personal loans** are more likely to subscribe to a term deposit.
- Clients with **higher account balances** are more likely to subscribe to a term deposit due to small p-value and positive slope.

Logit Regression Results

Dep. Variable:	y	No. Observations:	8462
Model:	Logit	Df Residuals:	8419
Method:	MLE	Df Model:	42
Date:	Mon, 09 Dec 2024	Pseudo R-squ.:	0.4147
Time:	04:55:02	Log-Likelihood:	-3433.1
converged:	True	LL-Null:	-5865.4
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.2892	0.225	5.733	0.000	0.848	1.730
balance	0.1264	0.034	3.689	0.000	0.059	0.194
default_yes	0.1080	0.252	0.429	0.668	-0.385	0.602
housing_yes	-0.6708	0.071	-9.460	0.000	-0.810	-0.532
loan_yes	-0.5239	0.095	-5.490	0.000	-0.711	-0.337

Logistic Regression Model

Campaign-Related features: (contact, day, month, duration, campaign, pdays, previous, poutcome)

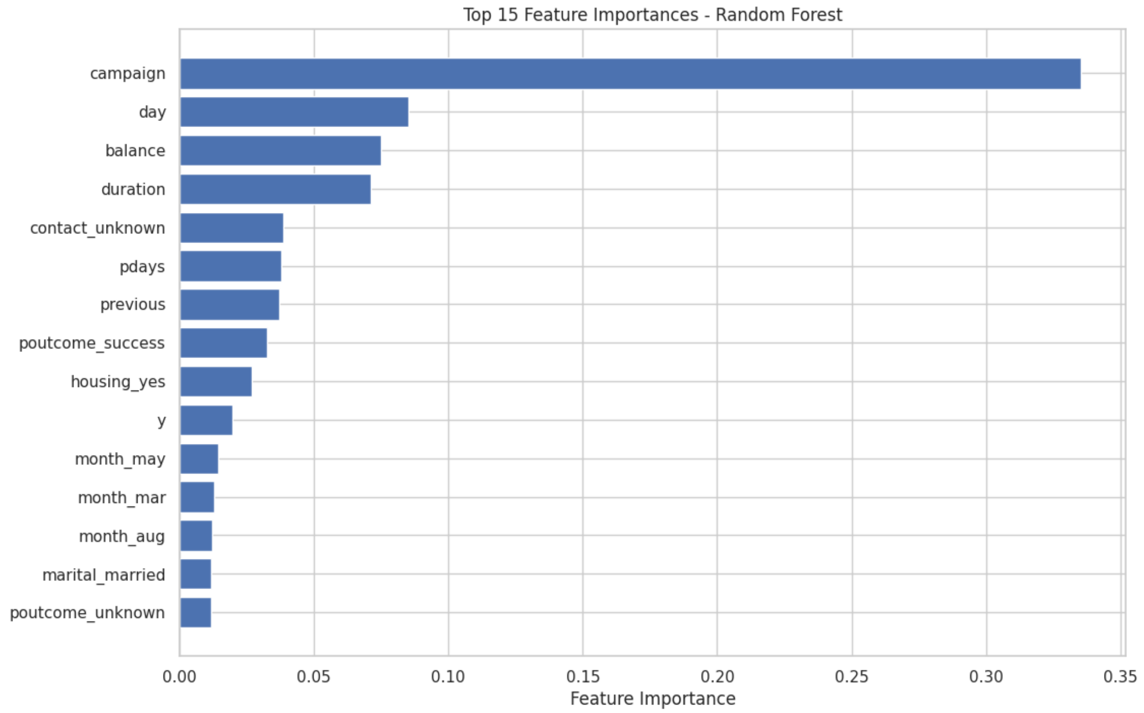
- **Longer call durations** significantly increase the likelihood of clients subscribing to a term deposit.
- **A higher number of contact** attempts during the campaign **reduces the likelihood** of clients subscribing to a term deposit.

Dep. Variable:	y	No. Observations:	8462
Model:	Logit	Df Residuals:	8419
Method:	MLE	Df Model:	42
Date:	Mon, 09 Dec 2024	Pseudo R-squ.:	0.4147
Time:	04:55:02	Log-Likelihood:	-3433.1
converged:	True	LL-Null:	-5865.4
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.2892	0.225	5.733	0.000	0.848	1.730
day	0.0532	0.034	1.550	0.121	-0.014	0.120
duration	1.9316	0.050	38.260	0.000	1.833	2.031
campaign	-0.2502	0.041	-6.134	0.000	-0.330	-0.170
pdays	-0.0366	0.056	-0.656	0.512	-0.146	0.073
previous	-0.0020	0.024	-0.086	0.932	-0.049	0.045
education_unknown	0.5341	0.173	3.089	0.002	0.195	0.873
contact_telephone	-0.1882	0.124	-1.522	0.128	-0.431	0.054
contact_unknown	-1.6488	0.110	-14.937	0.000	-1.865	-1.432
month_aug	-0.8391	0.129	-6.523	0.000	-1.091	-0.587
month_dec	0.8328	0.352	2.364	0.018	0.142	1.523
month_feb	-0.0663	0.147	-0.452	0.651	-0.354	0.221
month_jan	-1.3559	0.190	-7.128	0.000	-1.729	-0.983
month_jul	-1.0428	0.129	-8.066	0.000	-1.296	-0.789
month_jun	0.3507	0.152	2.306	0.021	0.053	0.649
month_mar	1.9291	0.259	7.440	0.000	1.421	2.437
month_may	-0.5931	0.123	-4.823	0.000	-0.834	-0.352
month_nov	-0.8674	0.140	-6.182	0.000	-1.142	-0.592
month_oct	1.2796	0.207	6.188	0.000	0.874	1.685
month_sep	0.8111	0.225	3.601	0.000	0.370	1.253
poutcome_other	0.0479	0.152	0.314	0.753	-0.251	0.346
poutcome_success	2.3392	0.173	13.535	0.000	2.000	2.678
poutcome_unknown	-0.3930	0.154	-2.548	0.011	-0.695	-0.091

Random Forest Model

- Demographic features
  - (+): job_retired, job_student, marital_single
- Financial features
  - (+): balance, default_yes
- Campaign-related features
  - (+): duration, poutcome_success
  - (-): campaign, pdays, previous



Model Performance Summary

- Results

Model	Precision(1)	Recall (1)	F1-Score	Accuracy
Logistic Regression	0.84	0.82	0.83	0.83
Decision Tree	0.74	0.88	0.81	0.78
Random Forest	0.84	0.89	0.86	0.86
SVM	0.82	0.89	0.85	0.85

- Random Forest and SVM demonstrate superior performance, showing the best balance between precision and recall, effectively identifying clients likely to subscribe.

- Logistic Regression and Random Forest maintain the highest precision, accurately predicting positive instances of clients subscribing.

Conclusion and Recommendations

- Reduce the frequency of contacts.(campaign (-))
- Train agents to increase call duration with meaningful customer interactions (duration (+))
- Focus on customers with higher account balances and successful prior engagement.(balance(+))
- Leverage Seasonal Trends

Q&A

You have Questions

We have Answers