Predicting Bank Term Deposit Subscriptions

Advanced Machine Learning Model Development in SAS Viya for Targeted Marketing

SAS® Viya™

Created By :
Akant Bhola

Agenda

01 Business Context, Problem & Objective

02 Dataset Overview & Codebook

03 Exploration and Visualization

04 Model Development & Comparison

05 Conclusion

01

Business Context, Problem & Objective

Business Context, Problem & Objective

Business Context

Direct marketing campaigns by banks are costly and inefficient when targeting non-responsive customers.

Problem Statement

How can we identify customers most likely to subscribe to a term deposit?

Objective

- Predict term deposit subscription using ML
- Compare multiple models for performance
- Recommend actionable targets for campaigns

02

Dataset Overview & Codebook

Dataset Overview & Codebook

Source: UCI Machine Learning Repository — Bank Marketing Dataset

Records: 4,521
Variables: 16 + target (y)

Goal: Predict customer subscription (y = yes/no)

Variable Name	Role	Type	Description	Missing Values
age	Input	Integer	age (years)	no
job	Input	Categorical	type of job (‘admin.’, ‘blue-collar’, ‘entrepreneur’, ‘housemaid’, ‘management’, ‘retired’, ‘self-employed’, ‘services’, ‘student’, ‘technician’, ‘unemployed’, ‘unknown’)	no
marital	Input	Categorical	marital status (‘divorced’, ‘married’, ‘single’)	no
education	Input	Categorical	level of education (‘primary’, ‘secondary’, ‘tertiary’, ‘unknown’)	no
default	Input	Binary	has credit in default?	no
balance	Input	Integer	average yearly balance (euros)	no
housing	Input	Binary	has housing loan?	no
loan	Input	Binary	has personal loan?	no
contact	Input	Categorical	contact communication type (‘cellular’, ‘telephone’, ‘unknown’)	no
day	Input	Integer	last contact day of the month	no
month	Input	Categorical	last contact month of year ('jan', 'feb', 'mar', ..., 'nov', 'dec')	no
duration	Input	Integer	last contact duration (seconds).	no
campaign	Input	Integer	number of contacts performed during this campaign and for this client (includes last contact)	no
pdays	Input	Integer	days since client last contacted (-1 means client was not previously contacted)	no
previous	Input	Integer	number of contacts performed before this campaign and for this client	no
poutcome	Input	Categorical	outcome of the previous campaign (‘failure’, ‘success’, ‘unknown’, ‘other’)	no
y	Target	Binary	has the client subscribed a term deposit? (yes/no)	no

03

Exploration & Visualization

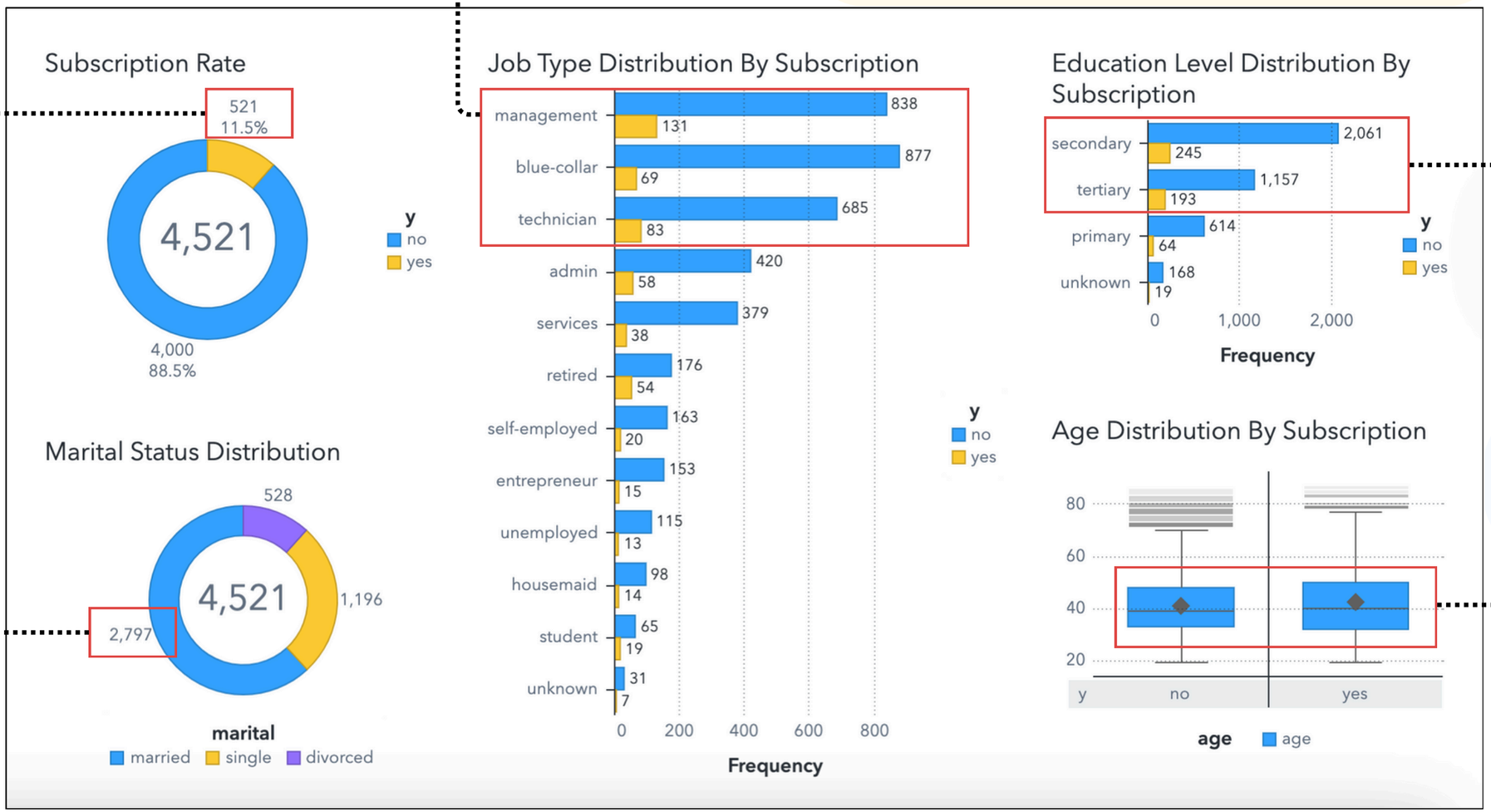
Customer Profile

Only 11.5% of customers subscribed to term deposits.

Management/technician and tertiary-educated clients show higher conversion rates than blue-collar and secondary clients.

Most clients are married

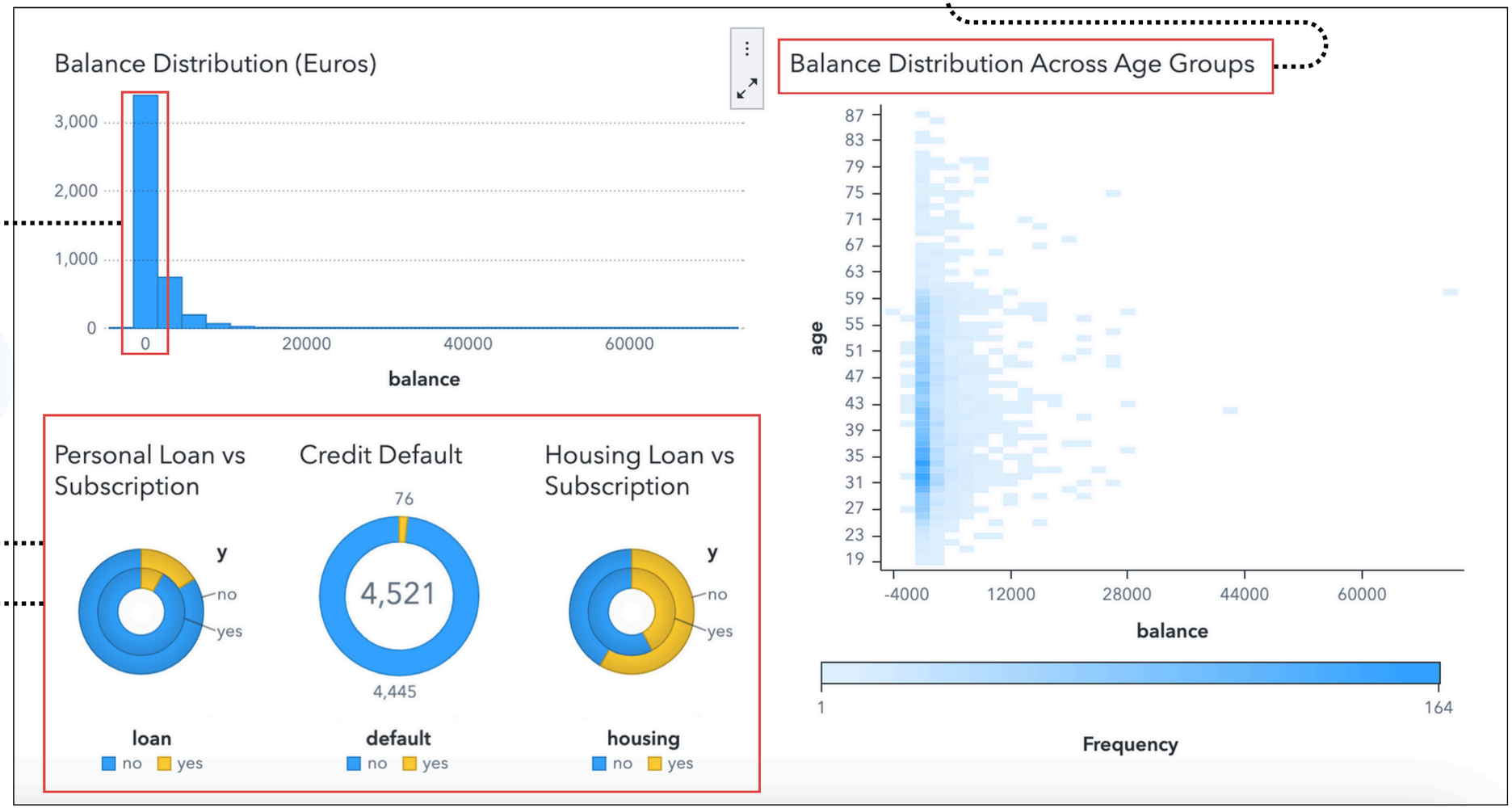
Age is right-skewed with outliers. Median age: ~40 years.



Financial and Loan Behaviour

Most clients have low balances (<€2,000)

Dense cluster of clients aged 30–40 with balances below €2,000. High balances are spread across age groups, with no strong pattern.



Clients without personal or housing loans are slightly more likely to subscribe.

Majority have no credit default

Campaign Engagement

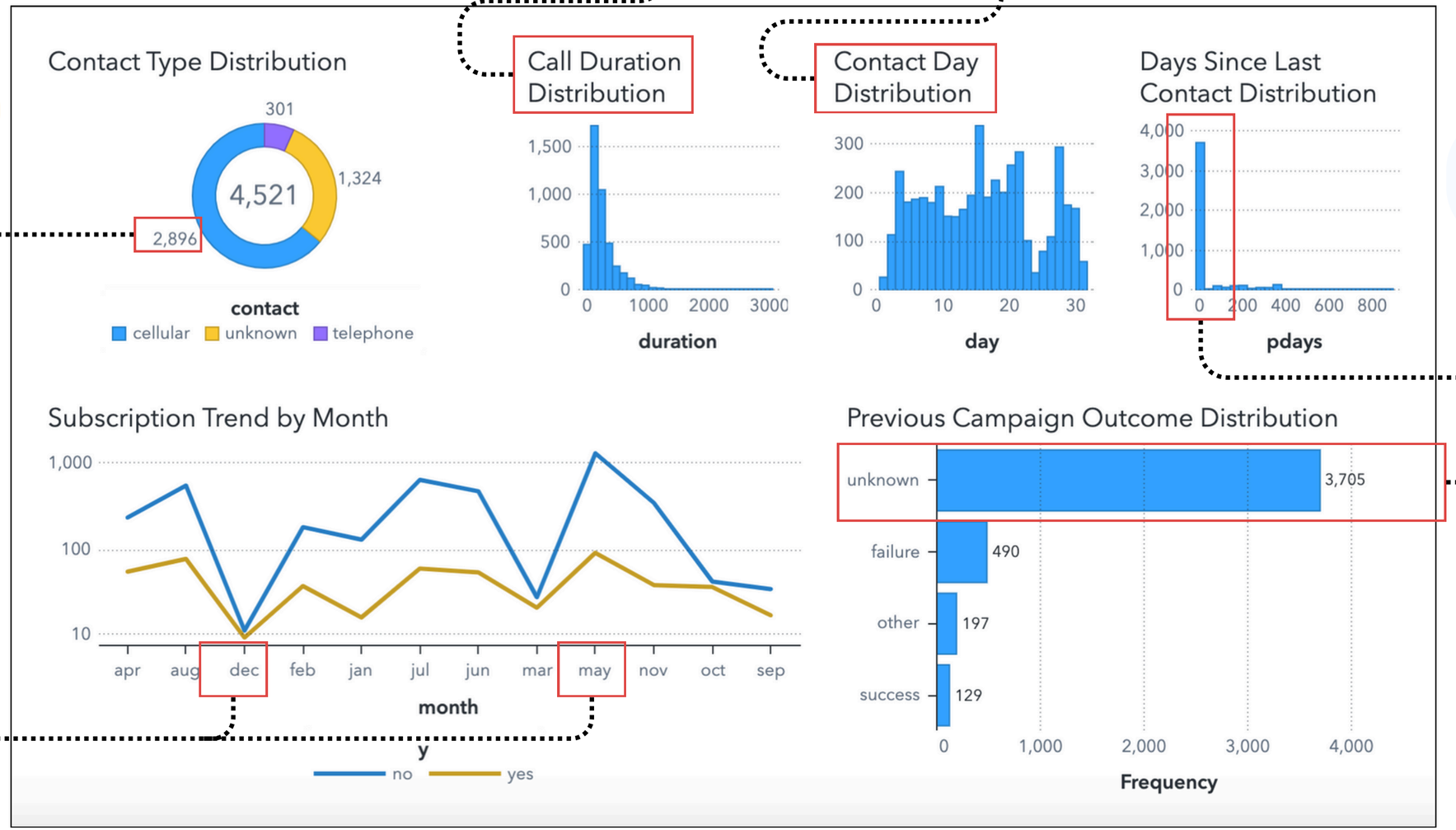
Most calls are short (<180 seconds)

Calls are more frequent around mid and end of the month.

Cellular contact is the most used channel

Most clients were never previously contacted (3,700+).

Subscriptions peak in May and dip in December



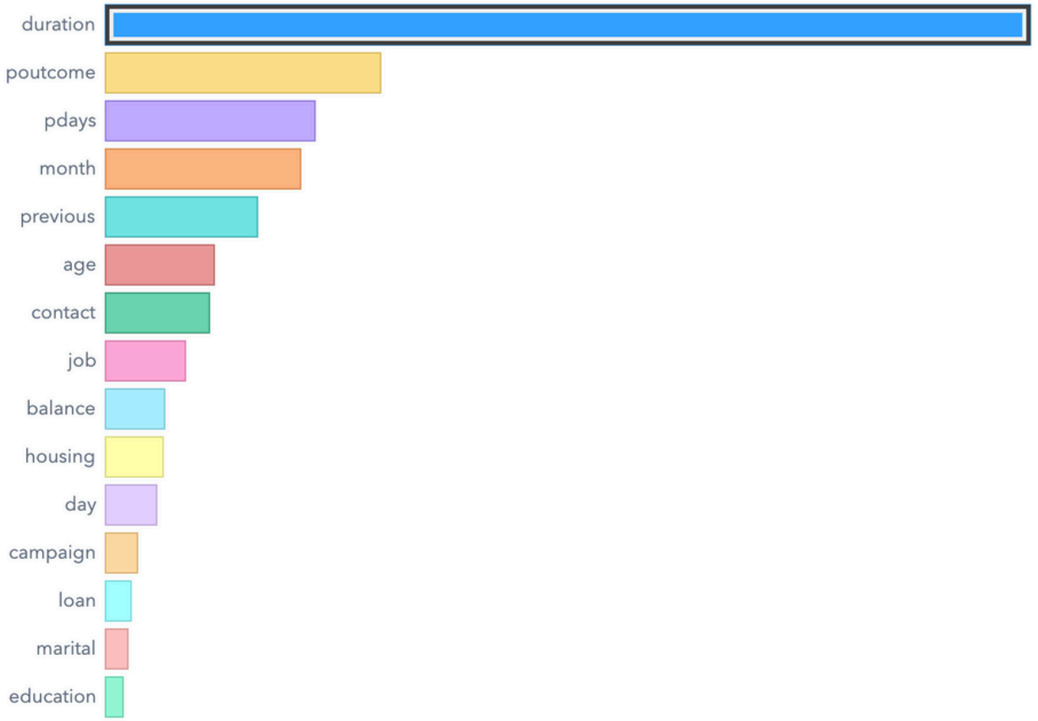
Factors Most Related to Subscription

What are the characteristics of y?

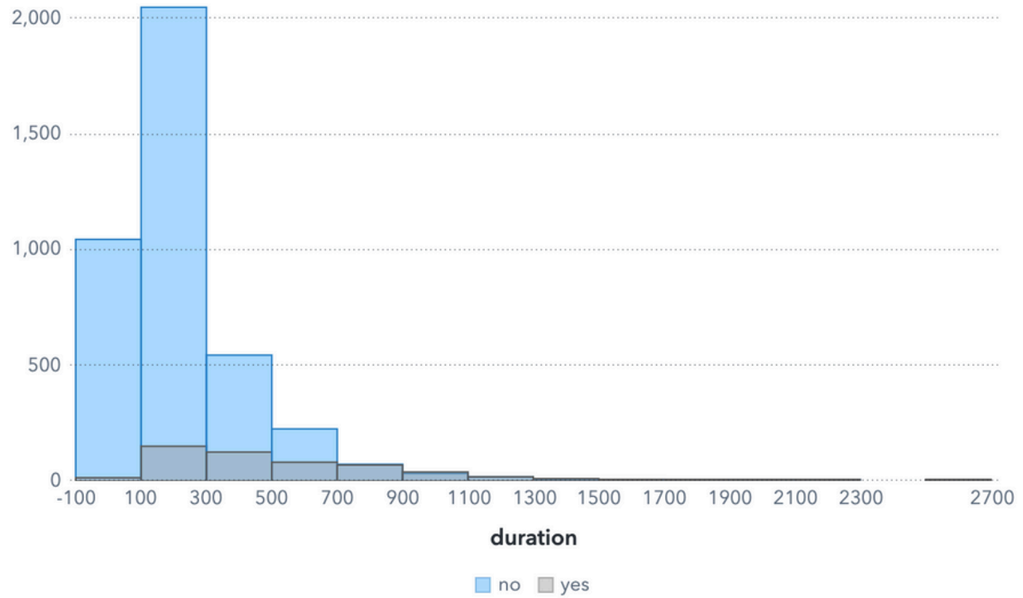
no is more common at 88.48% (4K of 4.5K). yes is less common at 11.52%. The three most related factors are duration, poutcome, and pdays.

no | yes

What factors are most related to y?

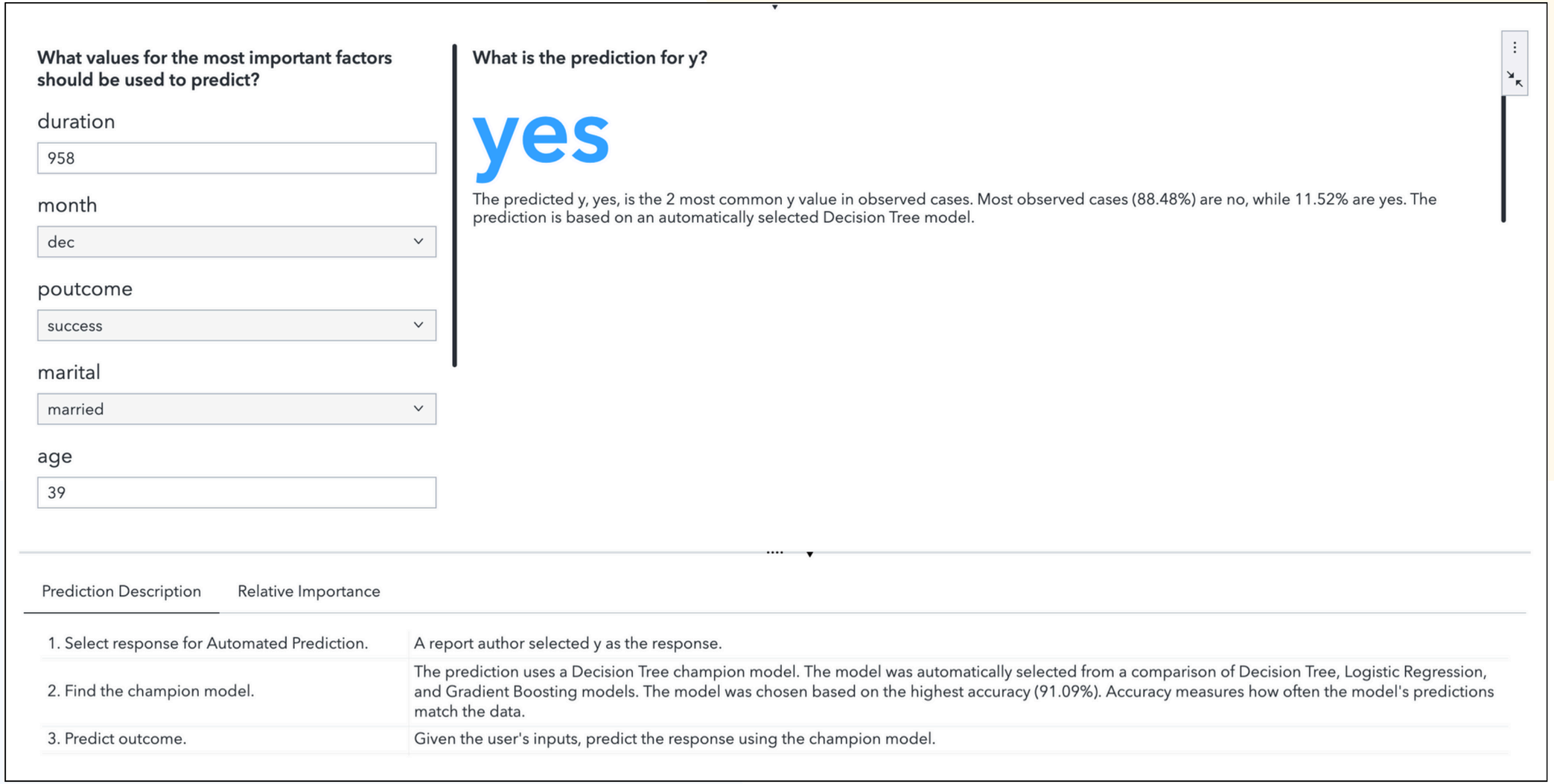


What is the relationship between y and duration?



The average duration is 226 when y is no, with a minimum of 4 and a maximum of 3K. The average duration is 553 when y is yes, with a minimum of 30 and a maximum of 2.8K. Average duration is 264, and it ranges from 4 to 3K.

Real-Time Prediction Interface

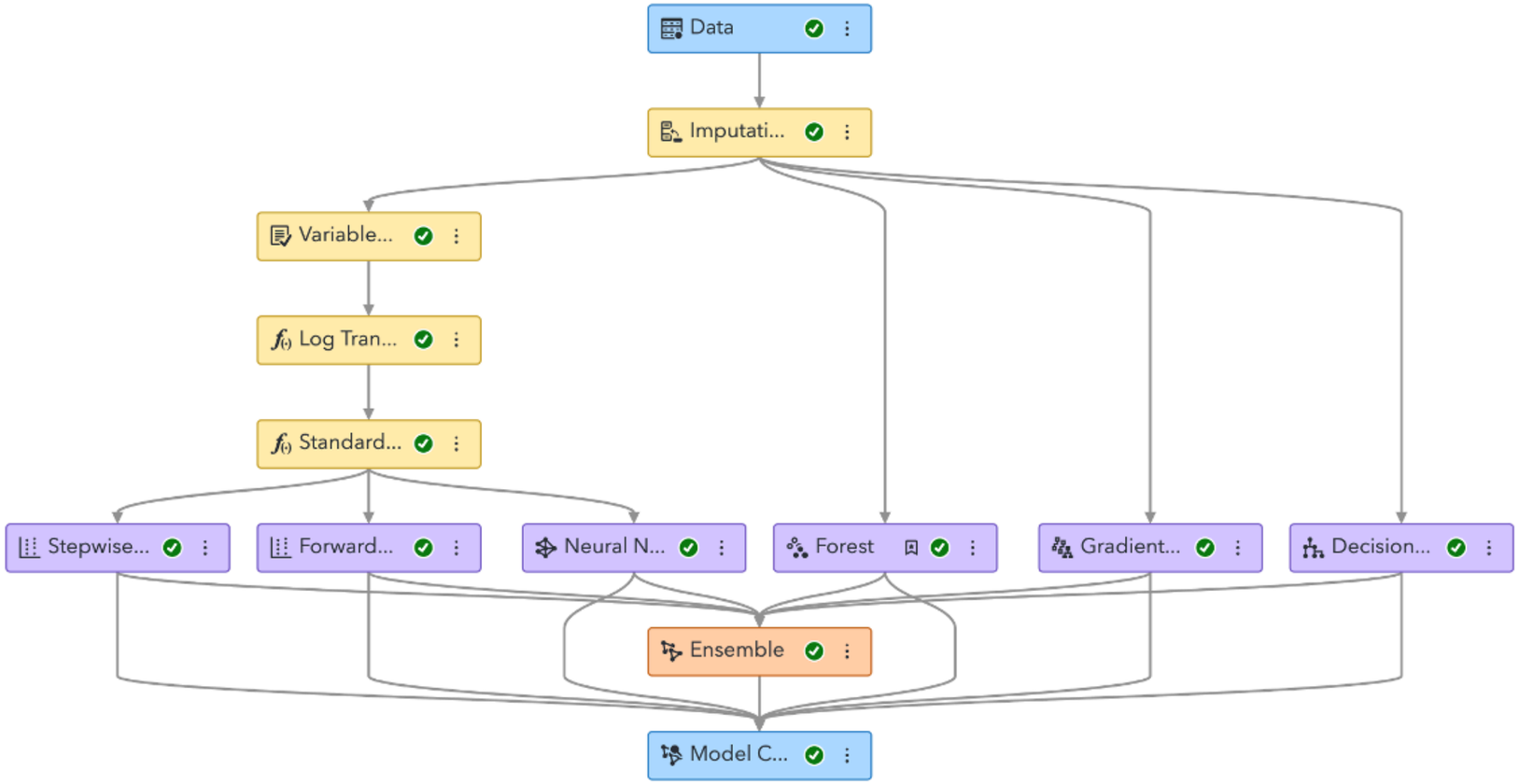


04

Model Development & Comparison

Predictive Modeling Pipeline

Seven ML models built and compared in SAS Viya



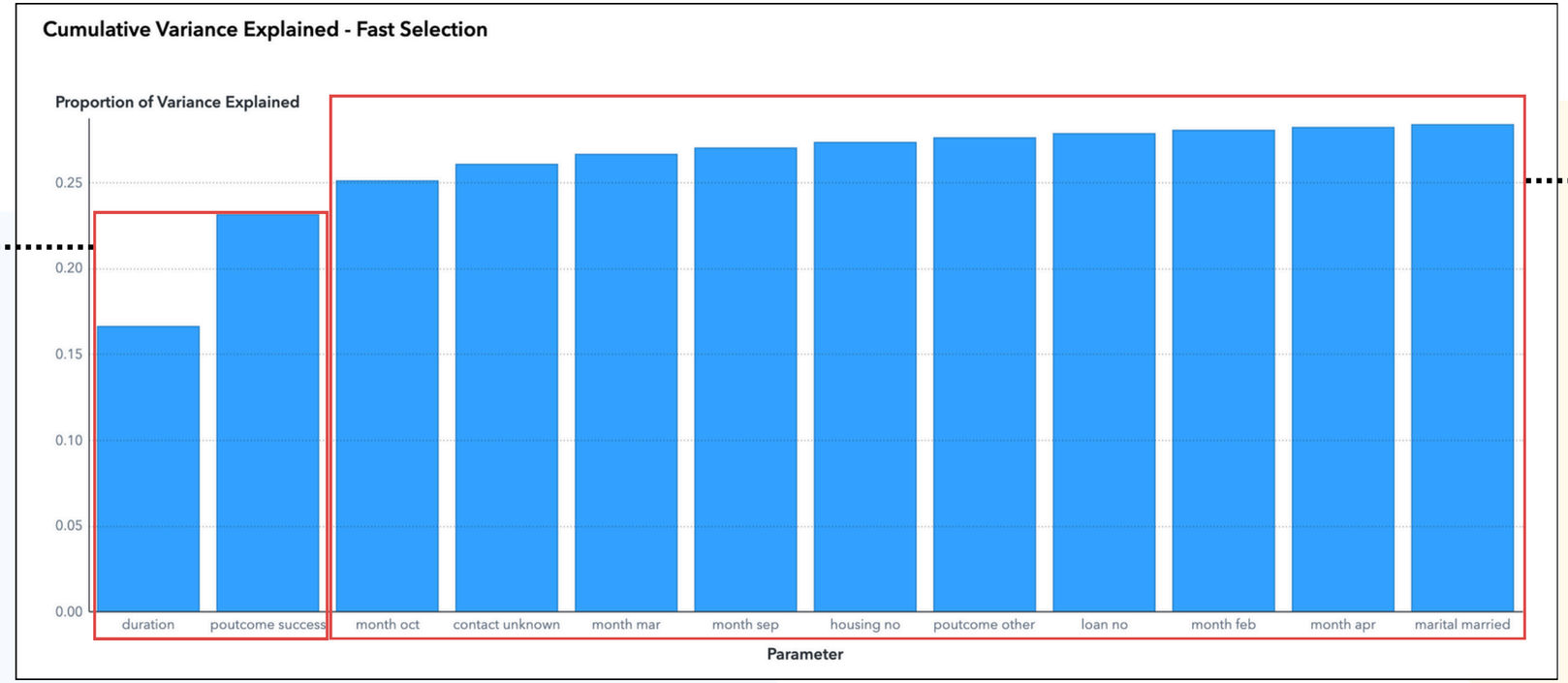
Models Used:

- Stepwise Logistic Regression
- Forward Logistic Regression
- Neural Network
- Decision Trees
- Random Forest
- Gradient Boosting
- Ensemble

Preparing Data for Modeling

Variable Selection, Log Transformation, Standardization & Data Partition

Variable Selection



duration **is the most impactful variable, explaining 16.6% of the model variation followed by** ***poutcome success*****, bringing total explained variance to 23.1%.**

Subsequent variables like ***month oct, contact unknown,*** **and** ***housing no*** **contribute incrementally (<1% each).**

Numeric features with skew were log-transformed to normalize distributions.

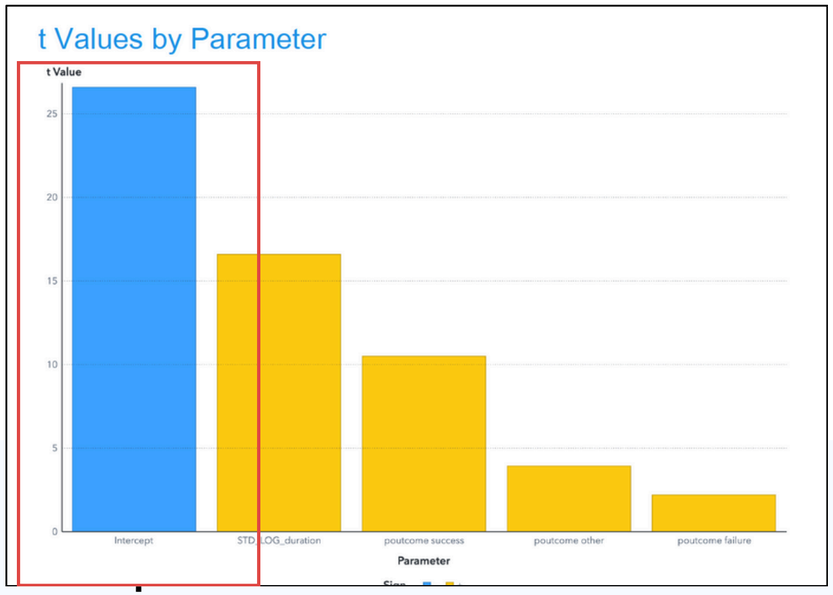
All features were standardized (mean=0, variance=1) to ensure fair comparison across models.

Dataset split into Train (60%), Validation (30%), and Test (10%)

Logistic Regression (Stepwise / Forward)

Stepwise and Forward logistic gave identical results due to predictor significance.

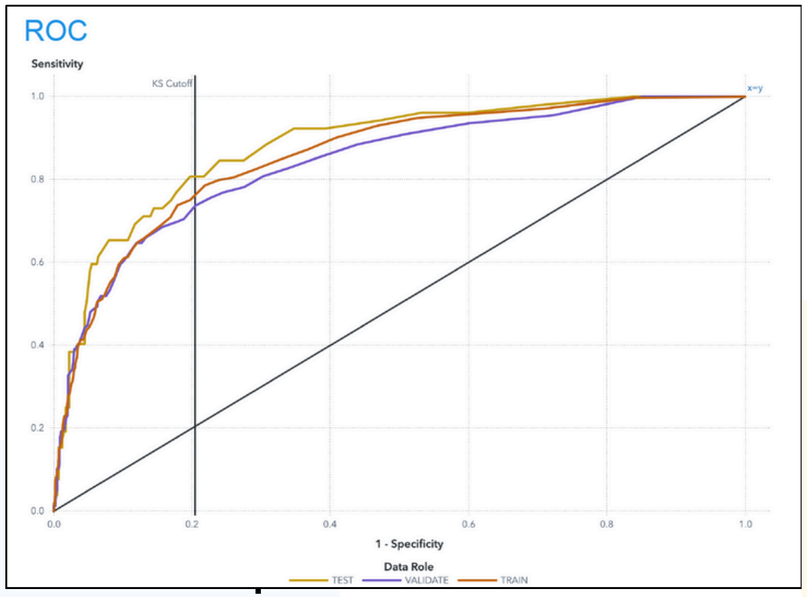
t Values by Parameter



The most significant parameter is *Intercept* with a t value of -26.587.

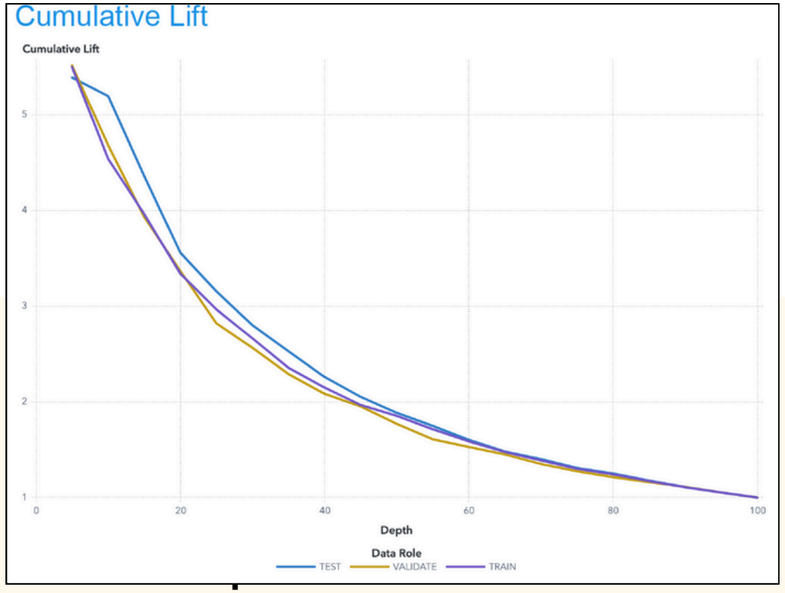
Numeric features with skew were log-transformed to normalize distributions.

ROC



The KS cutoff for the Validate partition is at 0.12, where sensitivity = 0.737 and 1-specificity = 0.205, indicating optimal class separation.

Cumulative Lift



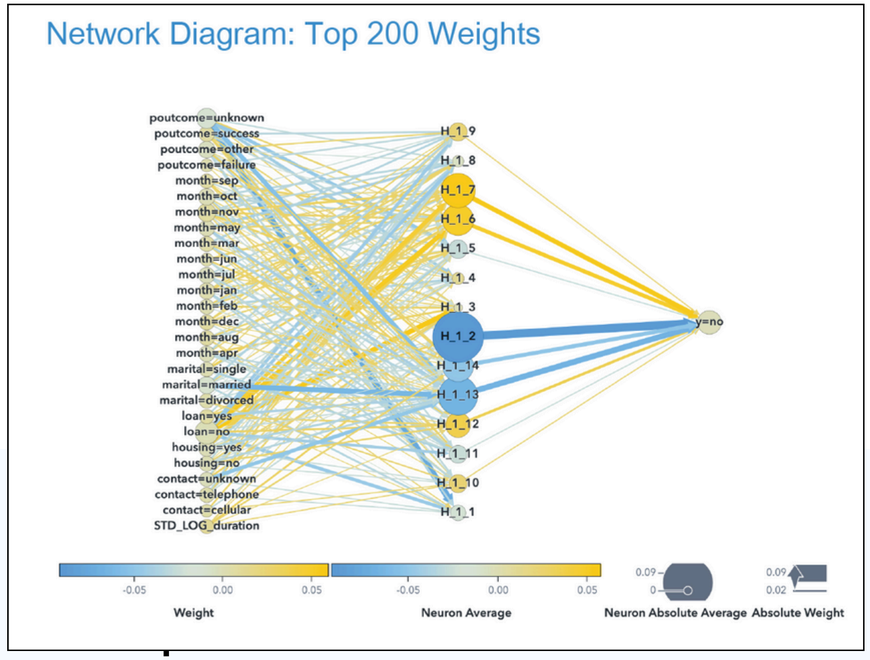
Cumulative Lift @ Top 10%:

- **Train = 4.54**
- **Validate = 4.68**
- **Test = 5.19**

→ The model is about 5× more effective than random at identifying likely subscribers.

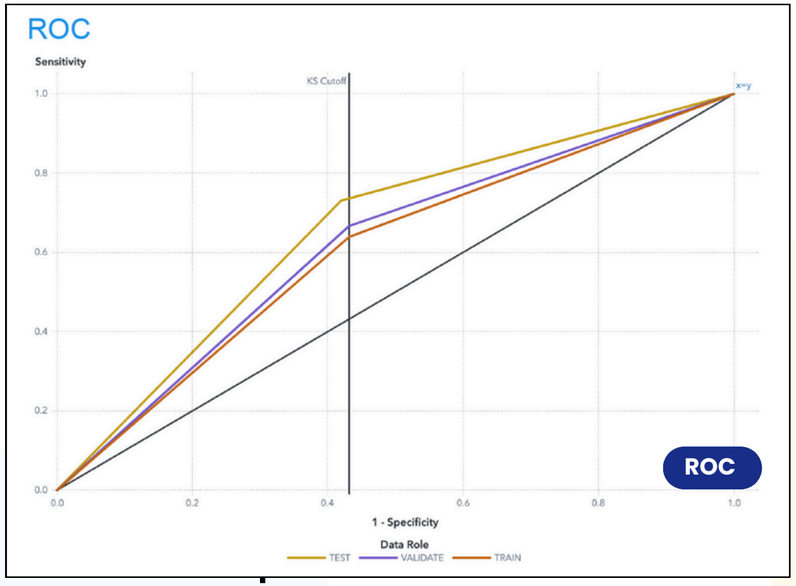
Neural Network

Network Diagram



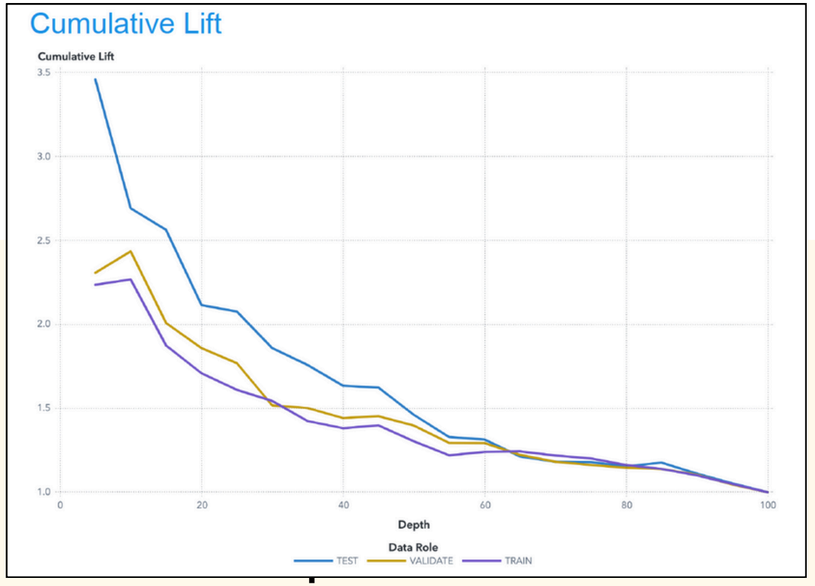
Neural Network with one hidden layer of 14 neurons using Tanh activation to model non - linear relationships.

ROC



The KS cutoff for the Validate partition is at 0.27, where sensitivity = 0.667 and 1-specificity = 0.432, indicating optimal class separation.

Cumulative Lift



Cumulative Lift @ Top 10%:

- **Train = 2.27**
- **Validate = 2.44**
- **Test = 2.69**

→ The model is about 2.5× more effective than random at identifying likely subscribers.

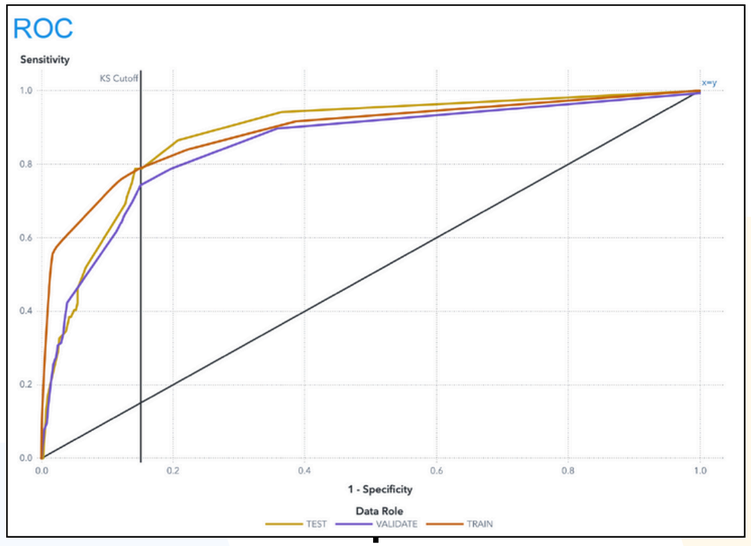
Decision Tree

Variable Importance

Variable Name	Training Importance	Training Relative Importance	Validation Relative Importance	Count	Validation Importance
duration	108.22	1	1	7	35.9
poutcome	44.58	0.41	0.62	2	22.16
month	34.89	0.32	0.23	6	8.28
day	14.18	0.13	0.13	3	4.61
marital	4.59	0.04	0.04	1	1.56
balance	9.51	0.09	0.02	3	0.76
job	5.22	0.05	0.01	2	0.26
campaign	1.49	0.01	-0.03	1	-1.14
pdays	2.92	0.03	-0.1	1	-3.62
education	5.68	0.05	-0.11	2	-3.99
loan	3.78	0.03	-0.13	1	-4.68
age	12.01	0.11	-0.17	3	-6.22

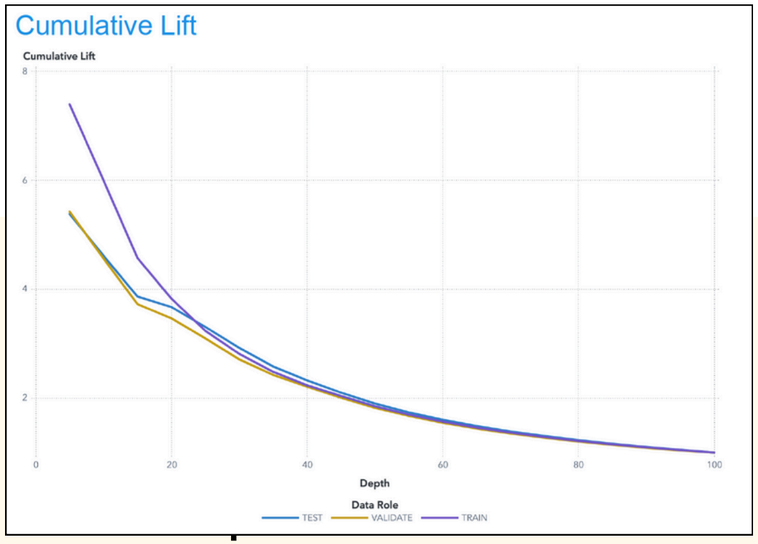
Numeric features with skew were log-transformed to normalize distributions.

ROC



The KS cutoff for the Validate partition is at 0.09, where sensitivity = 0.744 and 1-specificity = 0.151, indicating optimal class separation.

Cumulative Lift



Cumulative Lift @ Top 10%:

- **Train = 6**
- **Validate = 4.56**
- **Test = 4.62**

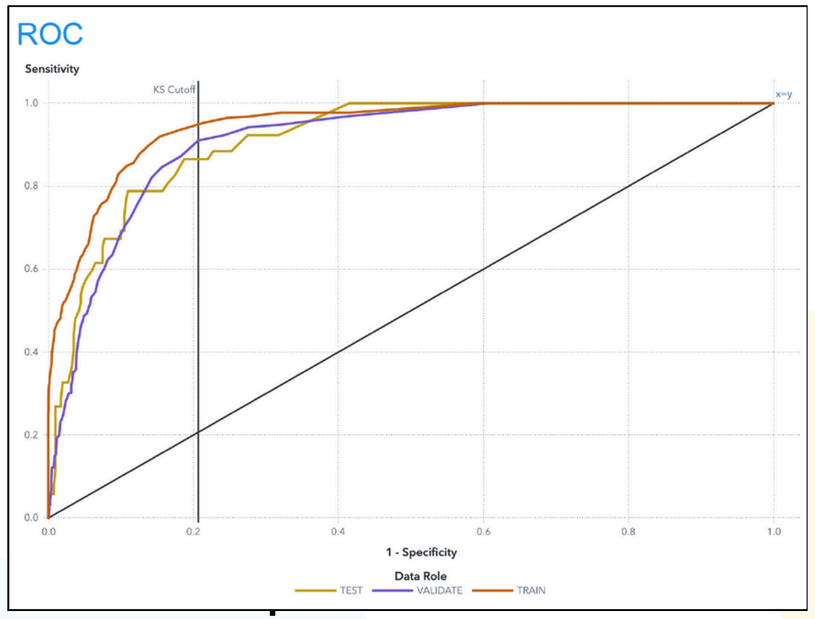
→ The model is about 5× more effective than random at identifying likely subscribers.

Random Forest

Variable Importance

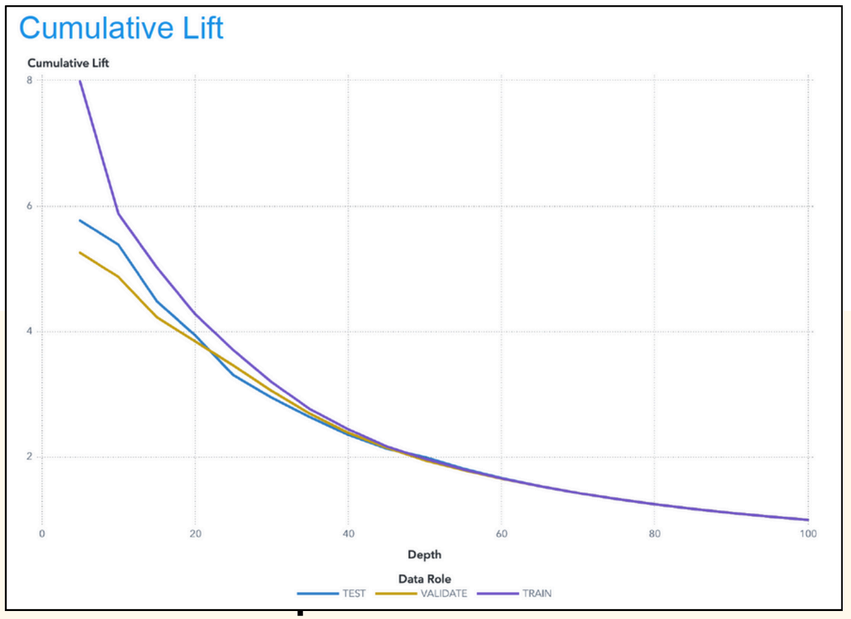
Variable	Training Importance	Importance Standard Deviation	Relative Importance
duration	59.74	16.83	1
month	22.29	7.59	0.373
poutcome	18.64	9.87	0.312
job	10.83	6.11	0.181
balance	7.46	4.53	0.125
age	7.32	4.21	0.123
day	7.3	4.65	0.122
pdays	4.6	4.36	0.077
education	3.67	2.67	0.061
contact	3.38	2.81	0.057
campaign	2.82	2.06	0.047
marital	2.54	2.28	0.042
previous	2.51	2.83	0.042
housing	1.05	1.54	0.018
loan	0.74	1.31	0.012
default	0.11	0.38	0.002

ROC



The KS cutoff for the Validate partition is at 0.9, where sensitivity = 0.91 and 1-specificity = 0.207, indicating optimal class separation.

Cumulative Lift



Cumulative Lift @ Top 10%:

- **Train = 5.88**
- **Validate = 4.87**
- **Test = 5.38**
-

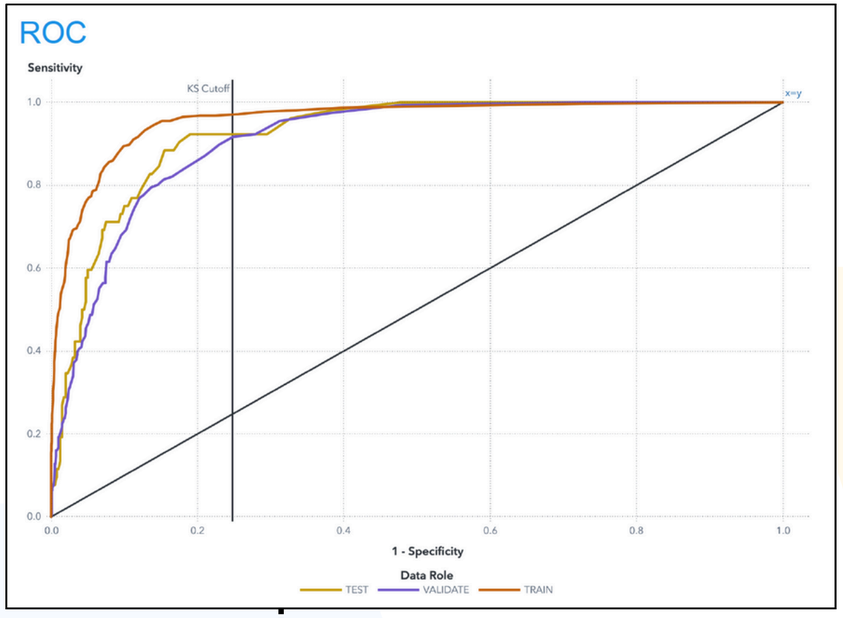
→ The model is about 5× more effective than random at identifying likely subscribers.

Gradient Boosting

Variable Importance

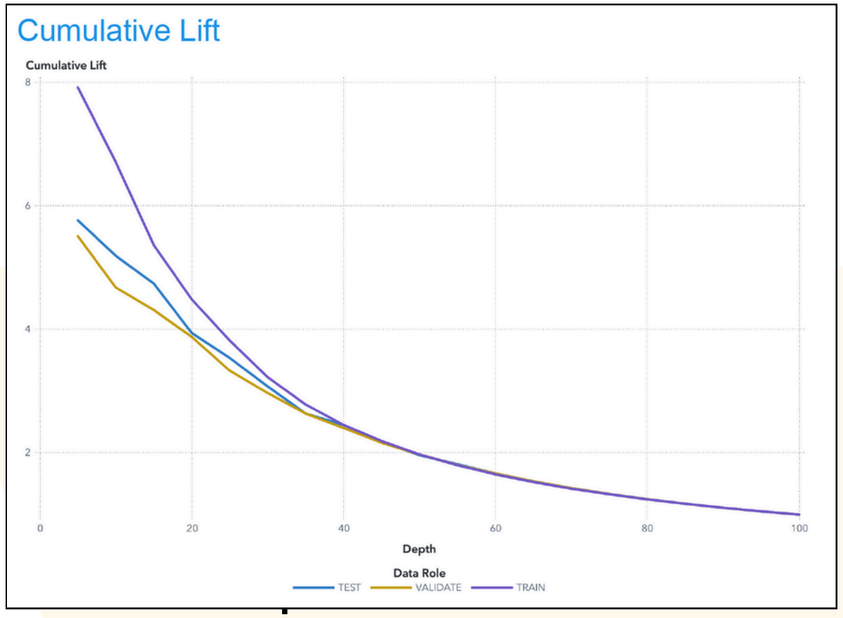
Variable	Training Importance	Importance Standard Deviation	Relative Importance
duration	14.61	23.85	1
month	7.51	4.81	0.514
job	4.3	3.13	0.294
poutcome	3.55	6.14	0.243
day	2.86	2.62	0.196
age	1.82	1.99	0.125
balance	1.4	1.8	0.096
contact	0.94	1.75	0.064
pdays	0.83	1.5	0.057
campaign	0.68	0.95	0.046
marital	0.62	1.07	0.043
education	0.59	1.37	0.041
previous	0.54	1.06	0.037
housing	0.43	1.07	0.03
loan	0.3	0.84	0.021

ROC



The KS cutoff for the Validate partition is at 0.06, where sensitivity = 0.917 and 1-specificity = 0.248, indicating optimal class separation.

Cumulative Lift



Cumulative Lift @ Top 10%:

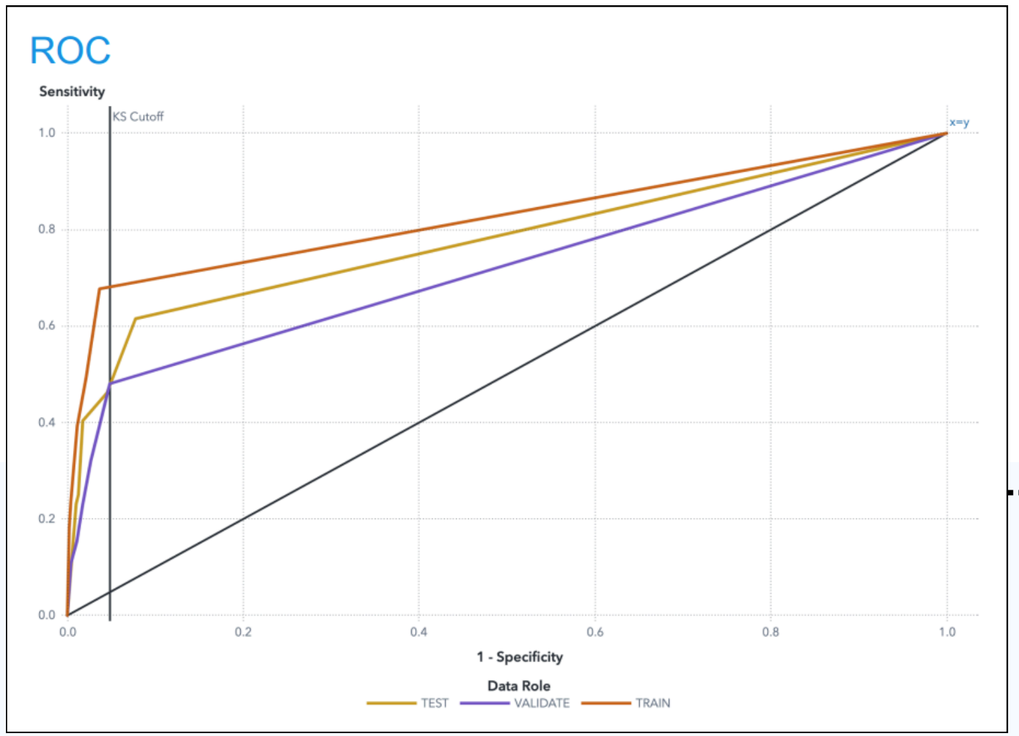
- **Train = 6.71**
- **Validate = 4.68**
- **Test = 5.19**

→ The model is about 5× more effective than random at identifying likely subscribers.

Ensemble Model

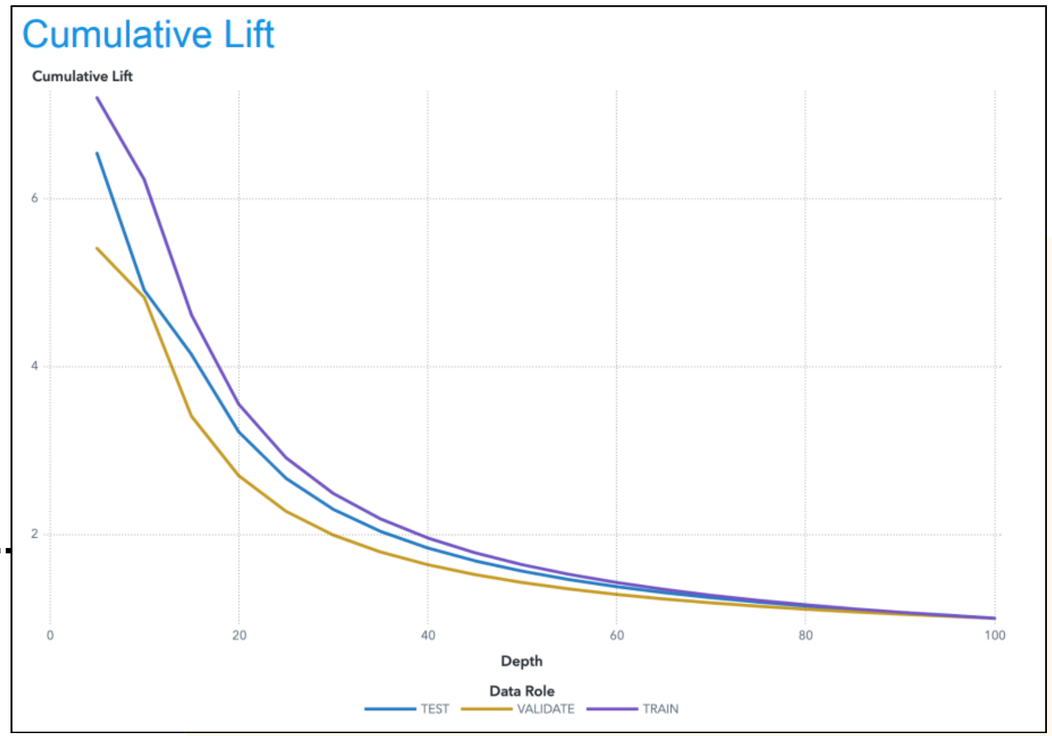
Combining multiple models for improved accuracy and robust predictions.

ROC



The KS cutoff for the Validate partition is at 0.01, where sensitivity = 0.481 and 1-specificity = 0.048, indicating optimal class separation.

Cumulative Lift



Cumulative Lift @ Top 10%:

- **Train = 6.23**
- **Validate = 4.82**
- **Test = 4.91**

→ The model is about 5× more effective than random at identifying likely subscribers.

Model Comparison

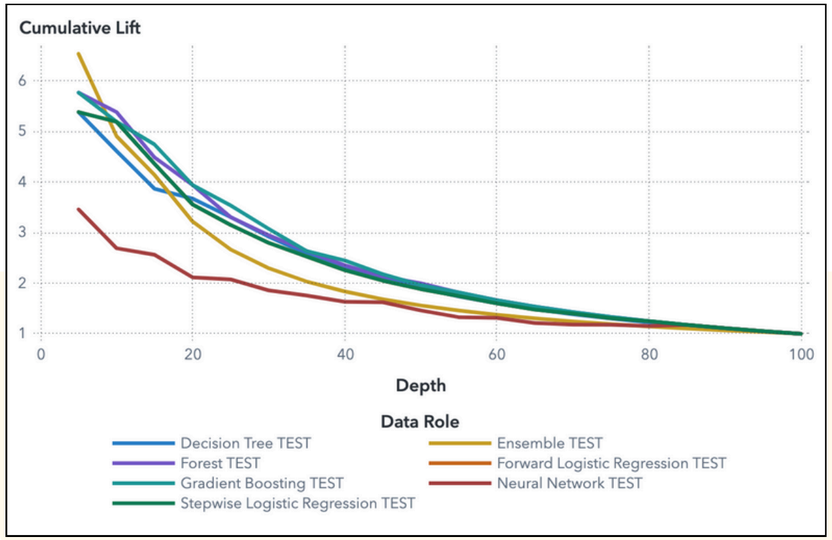
Evaluation on Test Dataset

Algorithm Name	KS (Youden)	Accuracy	Average Squared Error	Area Under ROC	Cumulative Lift
Gradient Boosting	**0.7331**	**0.9004**	**0.0661**	**0.923**	**5.1923**
Random Forest	0.6785	0.9027	0.0689	0.9127	5.3846
Decision Tree	0.6579	0.885	0.0862	0.8811	4.6154
Forward Logistic Regression	0.6102	0.9093	0.0729	0.8796	5.1923
Stepwise Logistic Regression	0.6102	0.9093	0.0729	0.8796	5.1923
Neural Network	0.3108	0.885	0.1256	0.6554	2.6923
Ensemble	0.5379	0.9159	0.0766	0.7779	4.9084

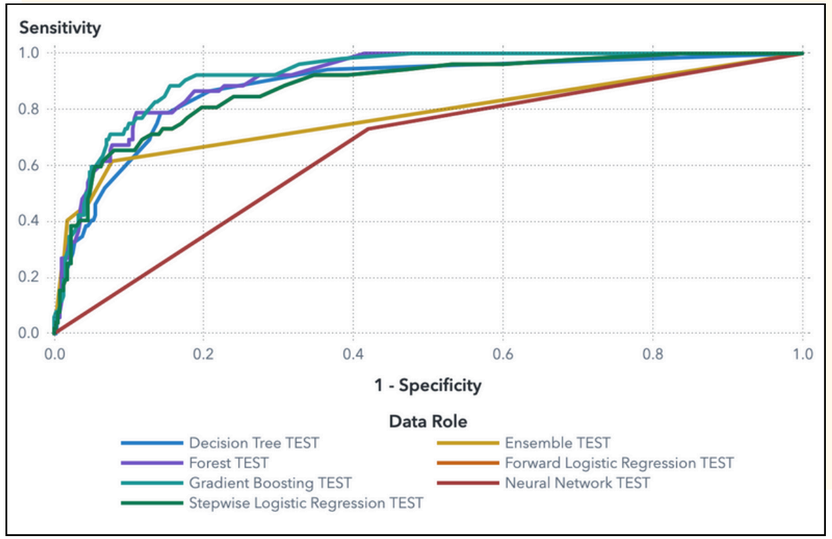
Numeric features with skew were log-transformed to normalize distributions.

Gradient Boosting emerged as the Champion Model

Cumulative Lift



ROC



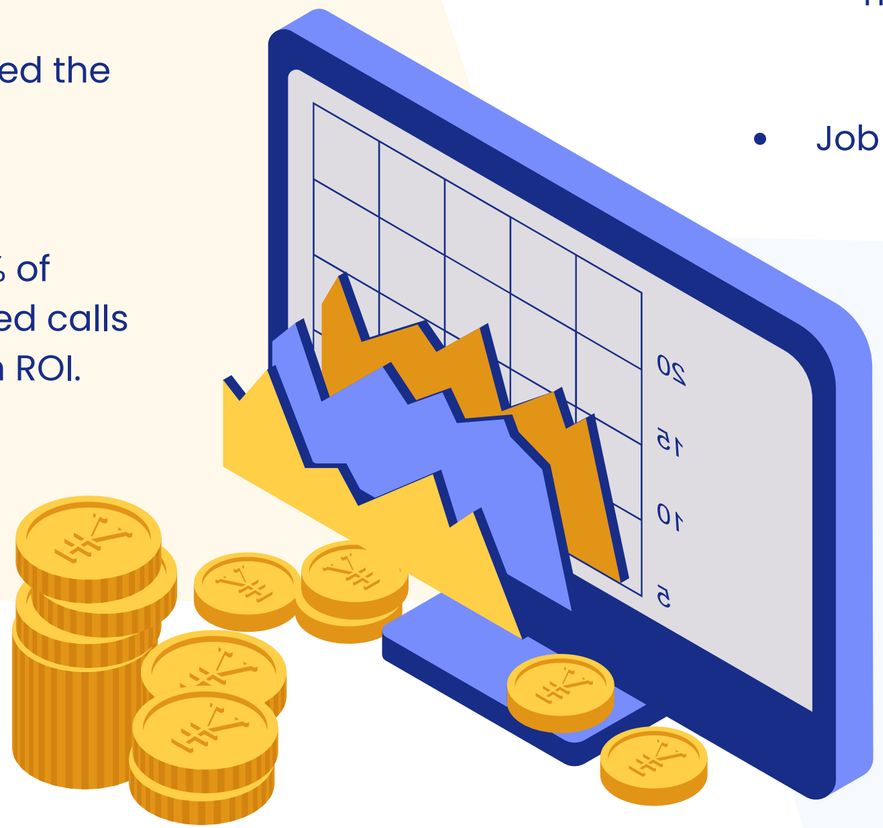
05

Conclusion

Conclusion

Champion Model

- Gradient Boosting achieved the best performance.
- Targeting the top 10% of customers captures ~50% of subscribers, cutting wasted calls and improving campaign ROI.



Key Drivers of Subscription

- Call duration was the single most important predictor.
- Timing (month, day) and past campaign outcomes influenced success.
- Job type, age, and financial balance added predictive value.

Business Value

- Reduce marketing costs by focusing on high-likelihood customers.
- Improve campaign success rates with data-driven targeting.
- Enhance customer engagement through more personalized outreach.

References

Alexandra, J., & Sinaga, K. P. (2021, October). Machine learning approaches for marketing campaign in Portuguese banks. In 2021 3rd International Conference on Cybernetics and Intelligent System (ICORIS) (pp. 1–6). IEEE. https://doi.org/10.1109/ICORIS52787.2021.9649623

Moro, S., Rita, P., & Cortez, P. (2014). Bank Marketing [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5K306.

Saxena, S. (2022). Tree-based machine learning methods in SAS® Viya®. SAS Institute Inc.

SAS Institute Inc. (2020). Machine learning using SAS® Viya® [PDF]. SAS Institute Inc. Retrieved from https://support.sas.com/content/dam/SAS/support/en/books/machine-learning-with-sas-viya/74588_excerpt.pdf

Zaki, A. M., Khodadadi, N., Hong Lim, W., & Towfek, S. K. (2024). Predictive Analytics and Machine Learning in Direct Marketing for Anticipating Bank Term Deposit Subscriptions. American Journal of Business & Operations Research, 11(1).

Thank You / Let's Connect

Akant Bhola | Data Analyst

akantbhola.AB@gmail.com

Scan to connect on LinkedIn