Enhance Marketing Campaign: Optimizing Certificate of Deposit (CD) Subscriptions with 2-Fold Strategies of Segmentation and Classification

Nawaraj Paudel, PhD (Quantitative Modeling of Materials)
Data Scientist and ML Engineer

CDs offer high return with no capital risk

Certificate of Deposits (CDs)

- Lock-up period: 1 month to several years
- Higher interest rates than traditional savings accounts
- FDIC insured up to $250,000
- Less popular despite better returns (over 6% in 2023)

Market Comparison

- CDs: Secure returns, appealing when federal interest rates are high
- S&P 500: Higher returns (over 7%) but with market volatility and risk of capital loss
- Inflation: CD yield of 6% helps stay ahead of average inflation (over 2%)

CDs market potential and challenges

- 2023 Forbes Advisor survey: 3% never opened a savings account vs. 41% never opened a CD
- High-yield savings accounts preferred due to recent interest rate hikes
- Understand customer motivations: Lack of product knowledge, ineffective marketing, or preference for flexibility
- Target potential CD customers: Analyze data to identify characteristics and barriers

The Portugese bank data (2008-2012) contains 20 features

Demographic Features

- Age
- Job
- Marital
- Education

Behavioral Features

- Housing
- Loan
- Default

Social and Economic Features

- Emp.var.rate (Employment variation rate)
- Cons.price.idx (Consumer price index)
- Cons.conf.idx (Consumer confidence index)
- Euribor3m (Euro interbank offered rate)
- Nr.employed (Number of employees)

Other Features

- Duration
- Poutcome (Previous outcome)
- Pdays (Number of days after last contact)
- Campaign (Number of calls made)
- Previous (Number of calls before this campaign)
- Contact
- Month
- Day_of_week

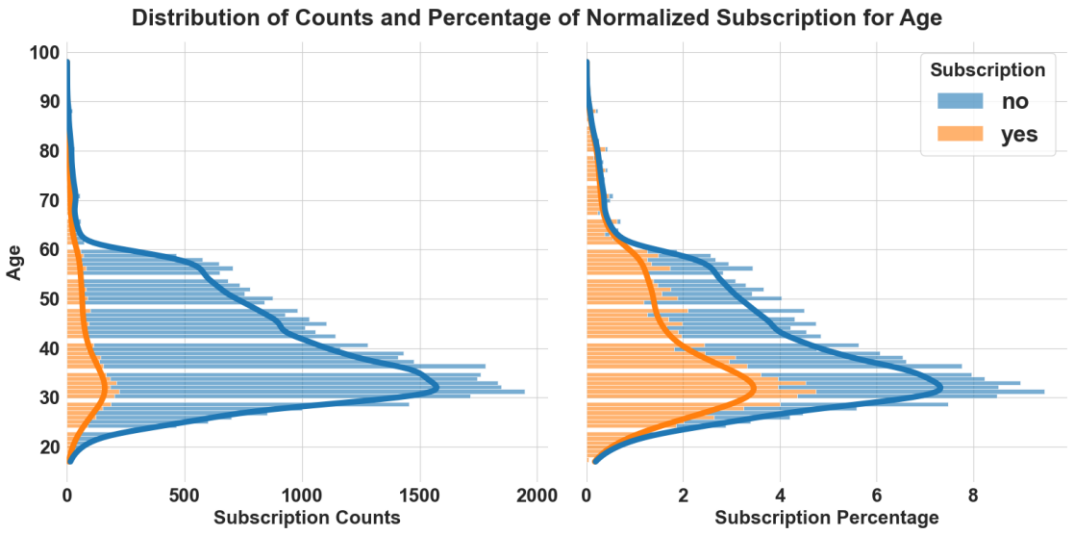
→ Subscribe CD? (Yes/No)

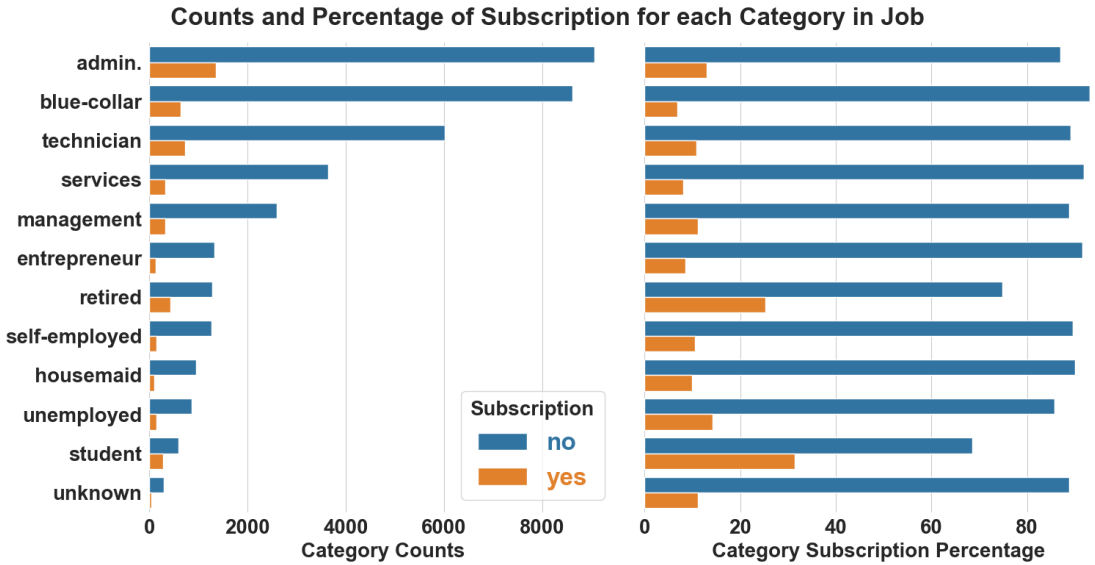
Missing data were labeled 'unknown' for categorical features

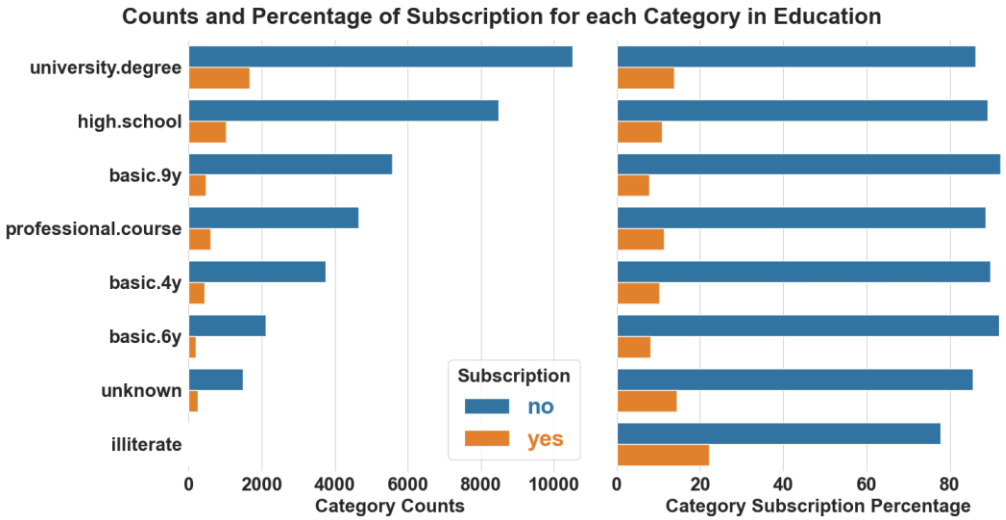
Features	Sub-categories	Count	Percentage
job	admin.	10422	25.30
	blue-collar	9254	22.47
	technician	6743	16.37
	services	3969	9.64
	management	2924	7.10
	retired	1720	4.18
	entrepreneur	1456	3.54
	self-employed	1421	3.45
	housemaid	1060	2.57
	unemployed	1014	2.46
	student	875	2.12
	unknown	330	0.80
marital	unknown	80	0.19
	married	24928	60.52
	single	11568	28.09
	divorced	4612	11.20
education	unknown	1731	4.20
	university.degree	12168	29.54
	high.school	9515	23.10
	basic.9y	6045	14.68
	professional.course	5243	12.73
	basic.4y	4176	10.14
	basic.6y	2292	5.56
	illiterate	18	0.04
default	unknown	8597	20.87
	no	32588	79.12
	yes	3	0.01

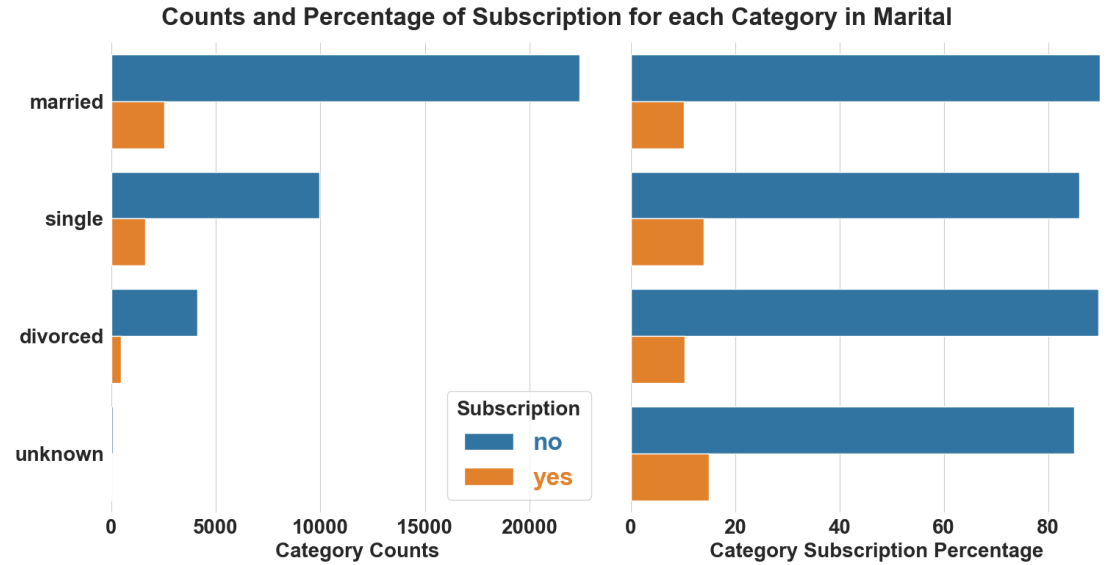
Features	Sub-categories	Count	Percentage
housing	unknown	990	2.40
	no	18622	45.21
	yes	21576	52.38
loan	unknown	990	2.40
	no	33950	82.43
	yes	6248	15.17
Subscription	no	36548	88.73
	yes	4640	11.27
contact	cellular	26144	63.47
	telephone	15044	36.53
month	may	13769	33.43
	jul	7174	17.42
	aug	6178	15.00
	jun	5318	12.91
	nov	4101	9.96
	apr	2632	6.39
	oct	718	1.74
	sep	570	1.38
	mar	546	1.33
	dec	182	0.44
day_of_week	thu	8623	20.94
	mon	8514	20.67
	wed	8134	19.75
	tue	8090	19.64
	fri	7827	19.00
poutcome	nonexistent	35563	86.34
	failure	4252	10.32
	success	1373	3.33

People over age of 60, retiree, single and with university degree are highly likely to subscribe to CDs

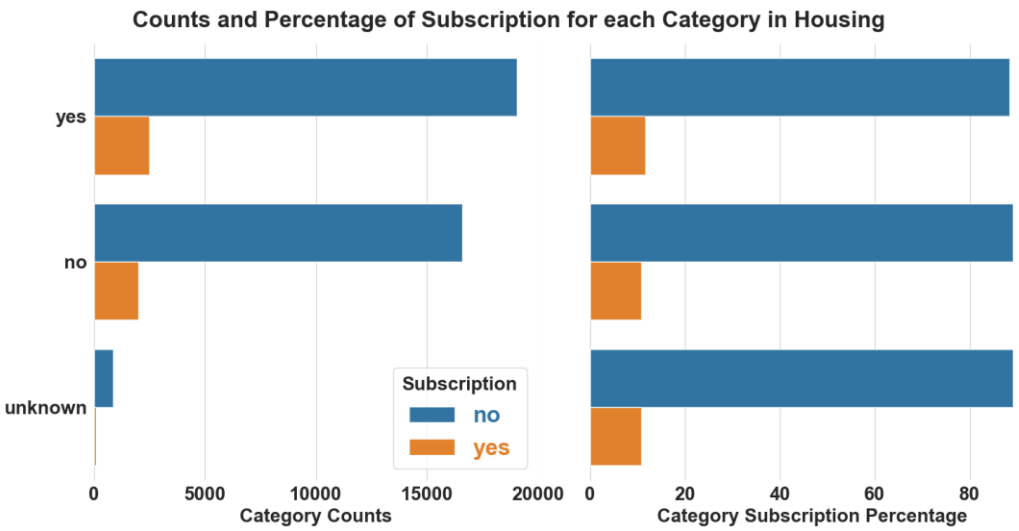


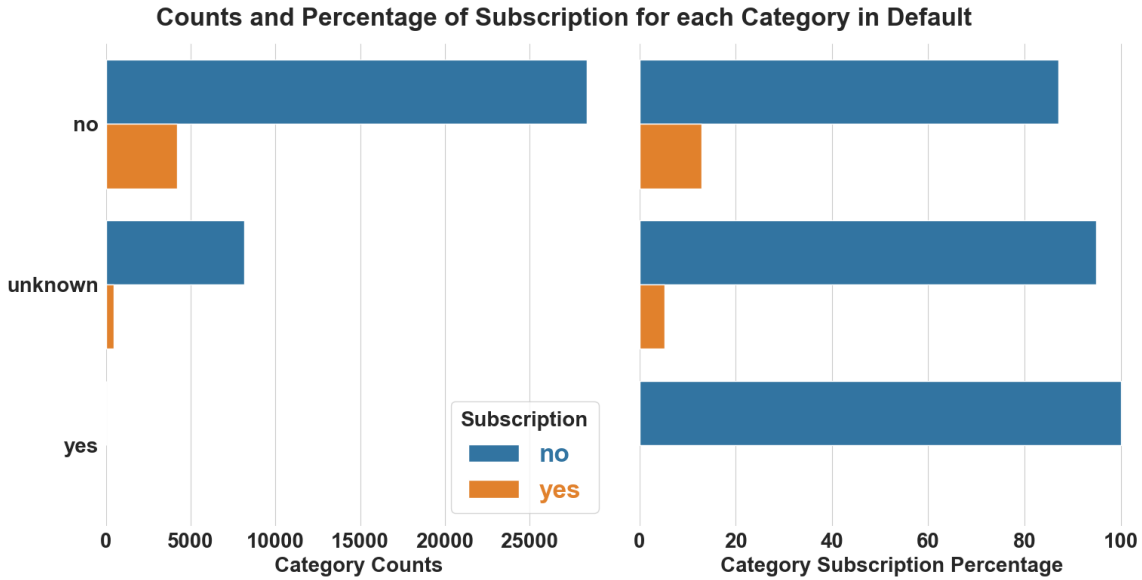


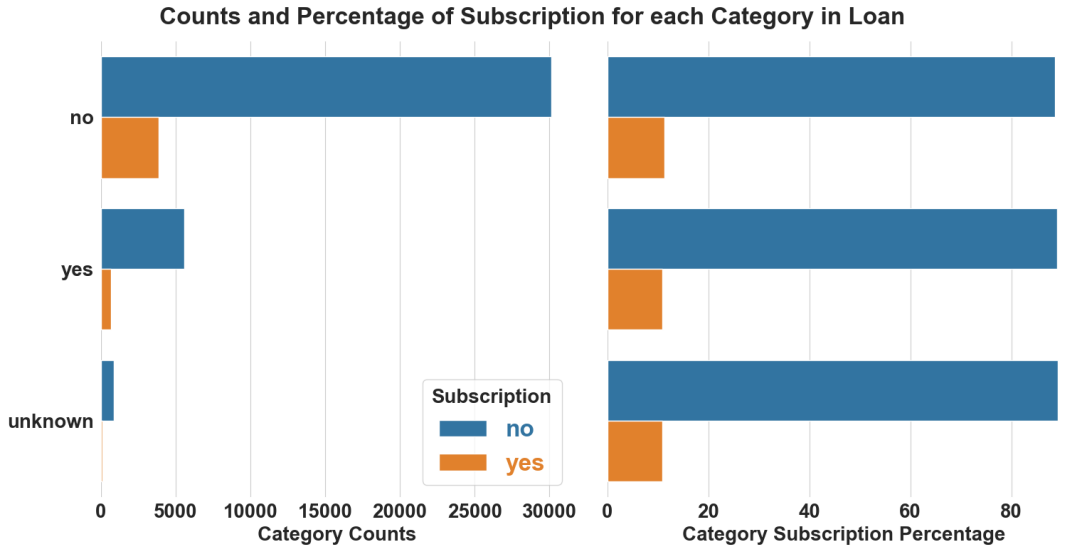




Housing and loan has no effect on CD subscription

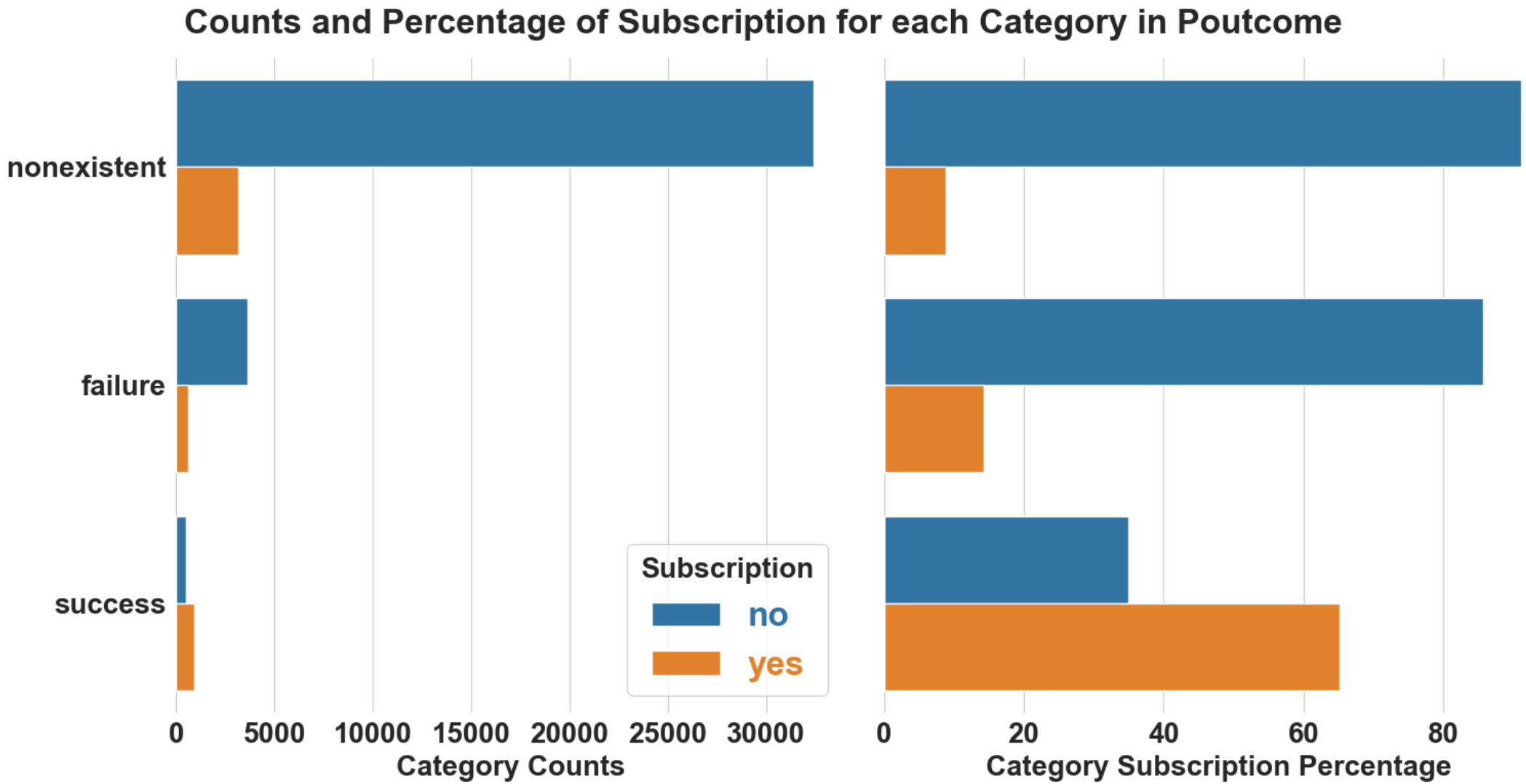






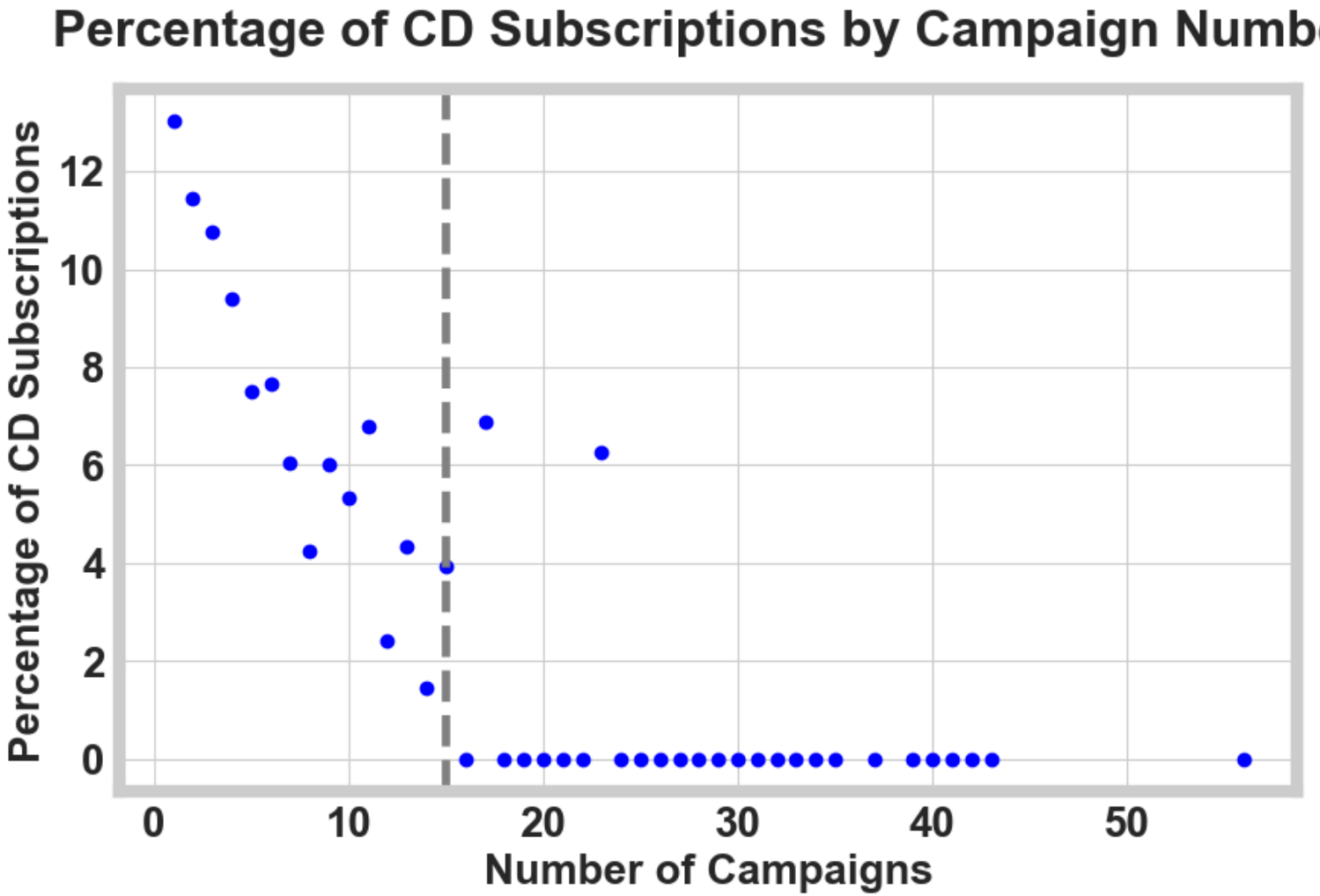
- Home ownership and loan status have no effect on CD subscription
- Customers who have never defaulted are 3 times more likely to subscribe CDs compared to those who have defaulted or did not report their default status

High retention rate: 65% of previous subscribers likely to subscribe CDs



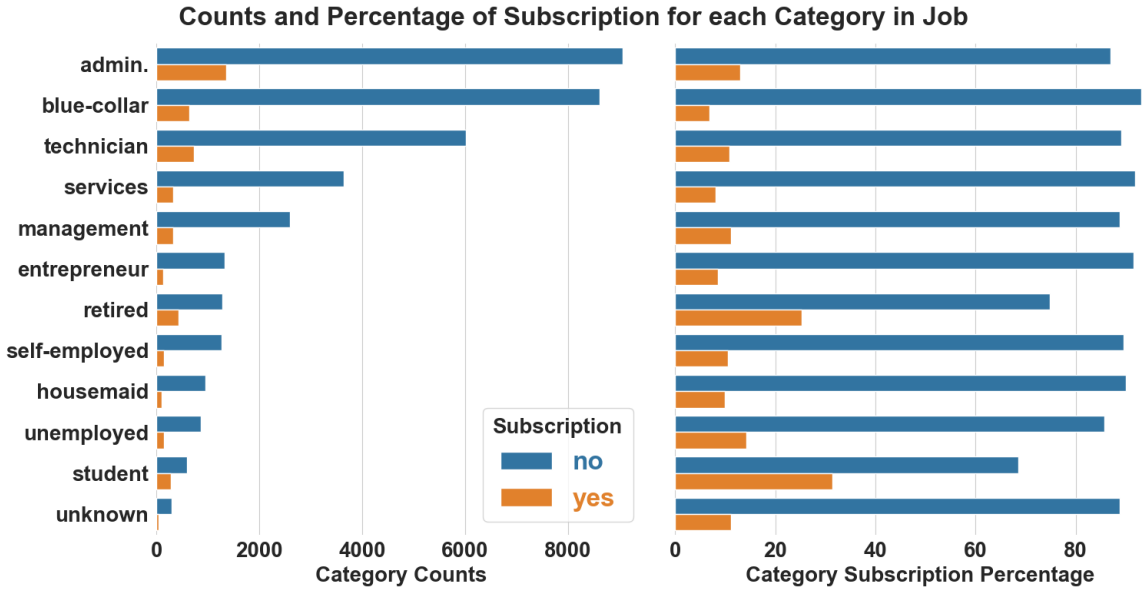
Only 3.4% were successfully contacted before this campaign and over 65% of them subscribed to CDs, nonexistent means not contacted before this campaign

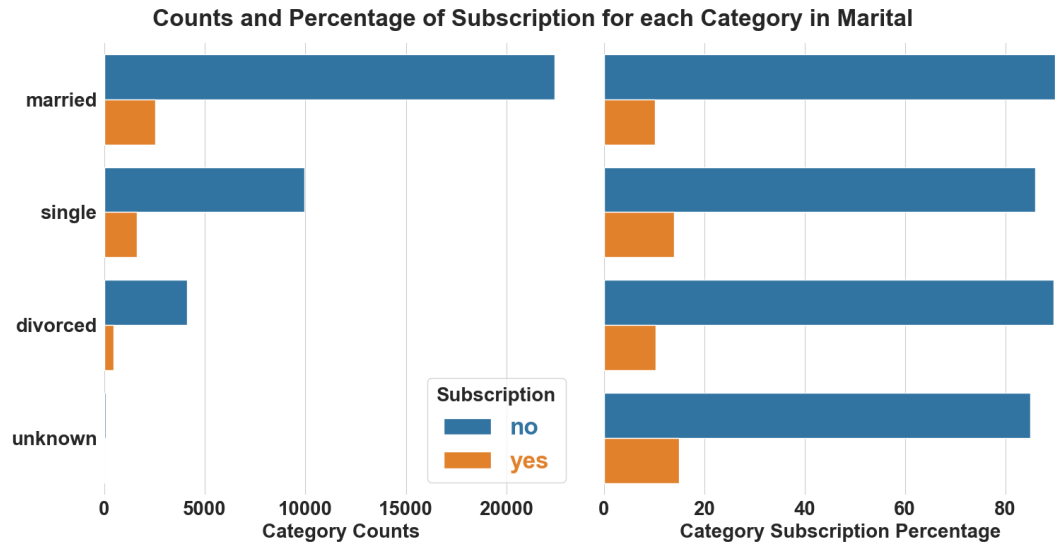
Subscription rate declines exponentially with number of calls



Subscription Campaign	No	Yes	Subscription Rate
1	15342	2300	13.04
2	9359	1211	11.46
3	4767	574	10.75
4	2402	249	9.39
5	1479	120	7.50
6	904	75	7.66
7	591	38	6.04
8	383	17	4.25
9	266	17	6.01
10	213	12	5.33
11	165	12	6.78
12	122	3	2.40
13	88	4	4.35
14	68	1	1.45
15	49	2	3.93
17	54	4	6.90
23	15	1	6.25

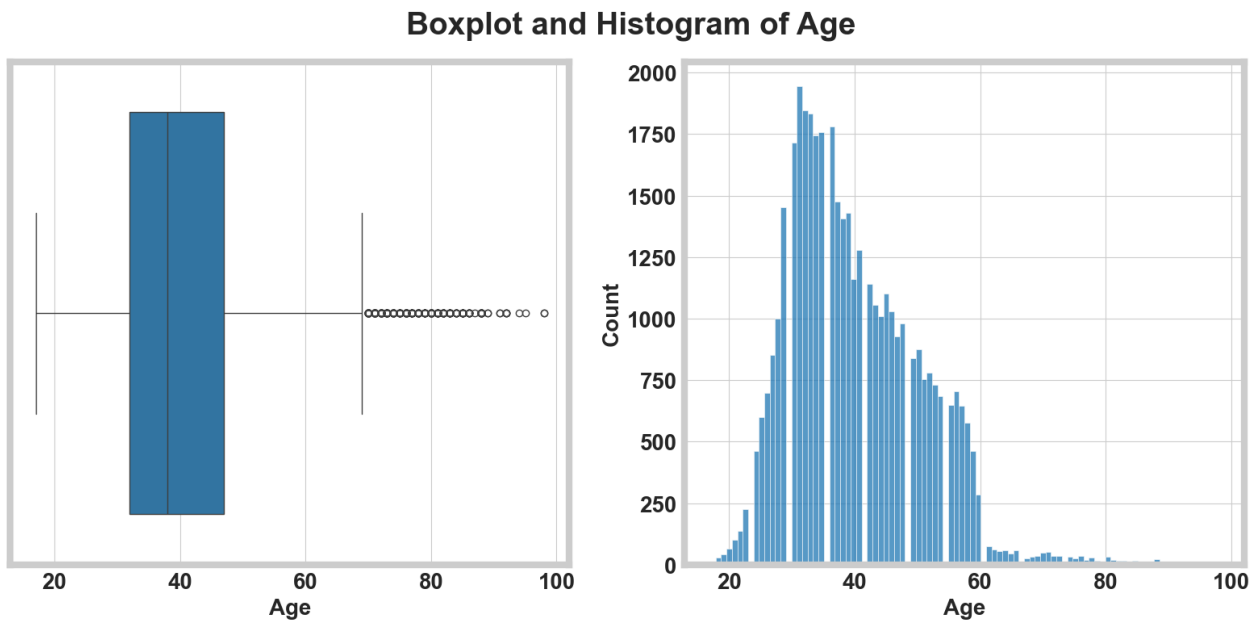
Infrequent sub-categories were merged and mutated

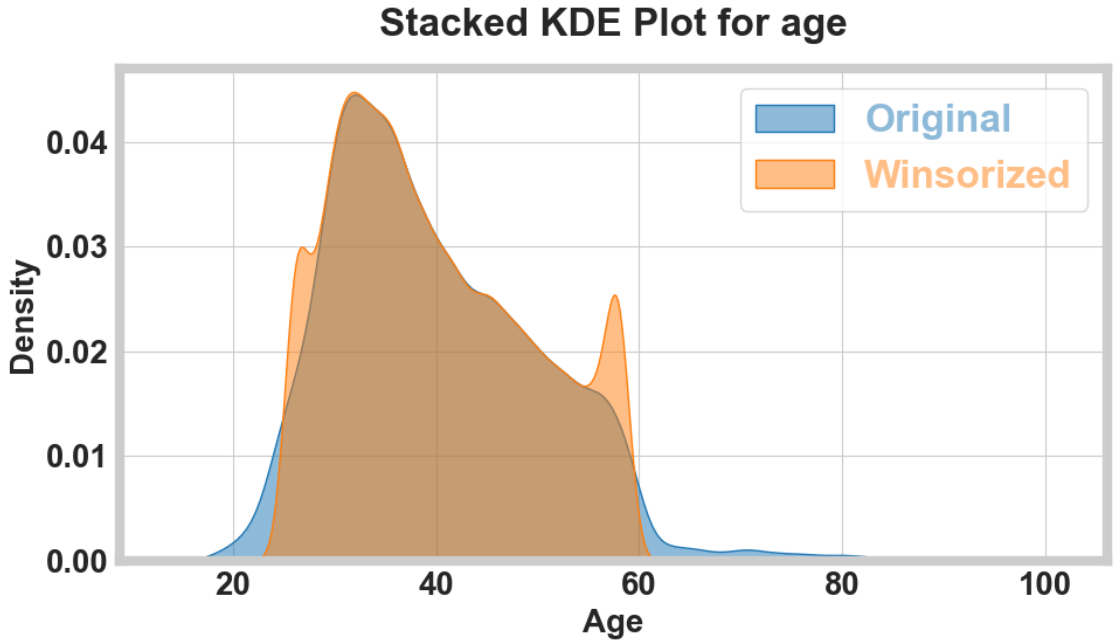




- Job sub-categories - unknown, student, unemployed and housemaid - each occurring less than 3% have been merged into a new sub-category called 'Others'
- The 'unknown' sub-category in Marital has been imputed with the most frequent sub-category 'married'
- Housing, loan and default features: Infrequent sub-categories were imputed
- Education: Sub-categories occurring less than 5% in education were merged into a new sub-category called 'Others'

The distribution of numerical features remain unchanged after handling outliers using Winsorization method





Wilcoxon rank-sum test for age:

- U Statistic: 847937868.50
- P-value: 0.9712
- Inference: The distributions of age are not significantly different (p >= 0.05)

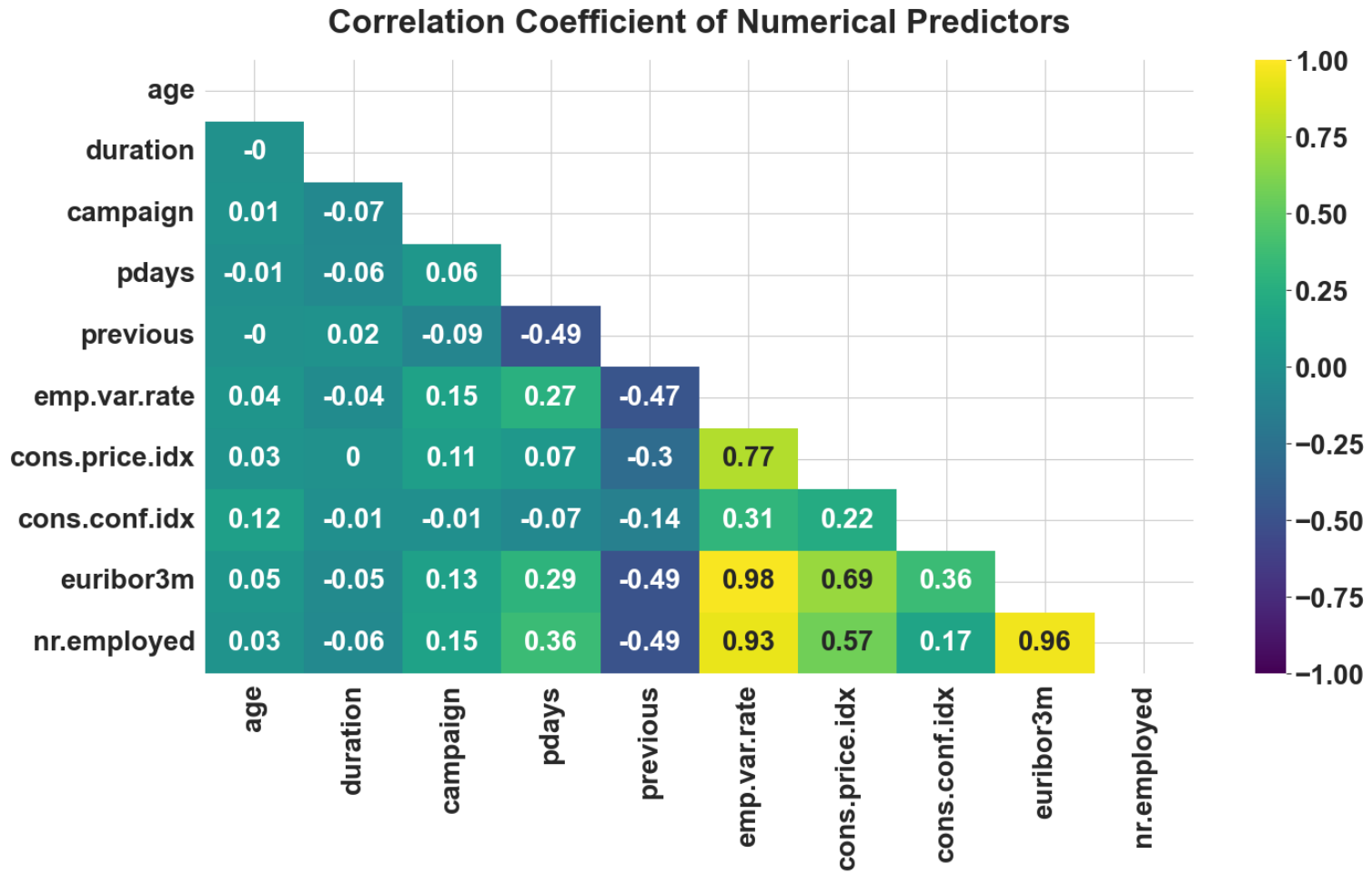
Loan, housing, and day of week features were dropped

Cat. Features	Target	Cramér's V	p-value	Deg. of Freedom
job	Subscription	0.130368	1.3999e-147	8
marital	Subscription	0.053683	6.222129e-27	2
education	Subscription	0.067205	9.538740e-39	6
default	Subscription	0.099097	3.624164e-90	1
housing	Subscription	0.009860	2.529191e-02	1
loan	Subscription	0.000000	3.747513e-01	1
contact	Subscription	0.144626	1.452825e-189	1
month	Subscription	0.274006	0.000000e+00	9
day_of_week	Subscription	0.023180	2.981214e-05	4
poutcome	**Subscription**	**0.320411**	**0.000000e+00**	**2**
quarter	Subscription	0.109828	4.960959e-109	2

Cramér's V	Interpretation
V <= 0.1	Weak association
0.1 < V <= 0.3	Moderate association
0.3 < V <= 0.5	Strong association
V > 0.5	Very strong association

- We dropped the **months** feature since we introduced the quarters feature, which had a Cramér's V association of 0.99 with months. Additionally, months and **poutcome** had a Cramér's V of 0.25, while quarters and **poutcome** had a Cramér's V of 0.14
- We dropped the **loan, housing, and day_of_week features** from predictive modeling due to their weak association with the subscription variable
- For predictive modeling, categorical features were one hot encoded

Highly correlated features with variation inflation factor over 5 were dropped



Feature	VIF
duration	1.01
age	1.01
campaign	1.04
pdays	1.45
previous	1.60
cons.conf.idx	2.89
cons.price.idx	4.98
emp.var.rate	34.87
nr.employed	40.11
euribor3m	68.19

These features were dropped

Class weight, up-sampling and threshold tuning were used to handle imbalanced data set

Models Trained

- Logistic Regression
- Random Forest Classifier
- Support Vector Machine (SVM)
- XGBoost Classifier
- K-Nearest Neighbors (KNN)
- Voting Classifier
- Neural Network

Balanced class weight /scale positive weight from scikit-learn

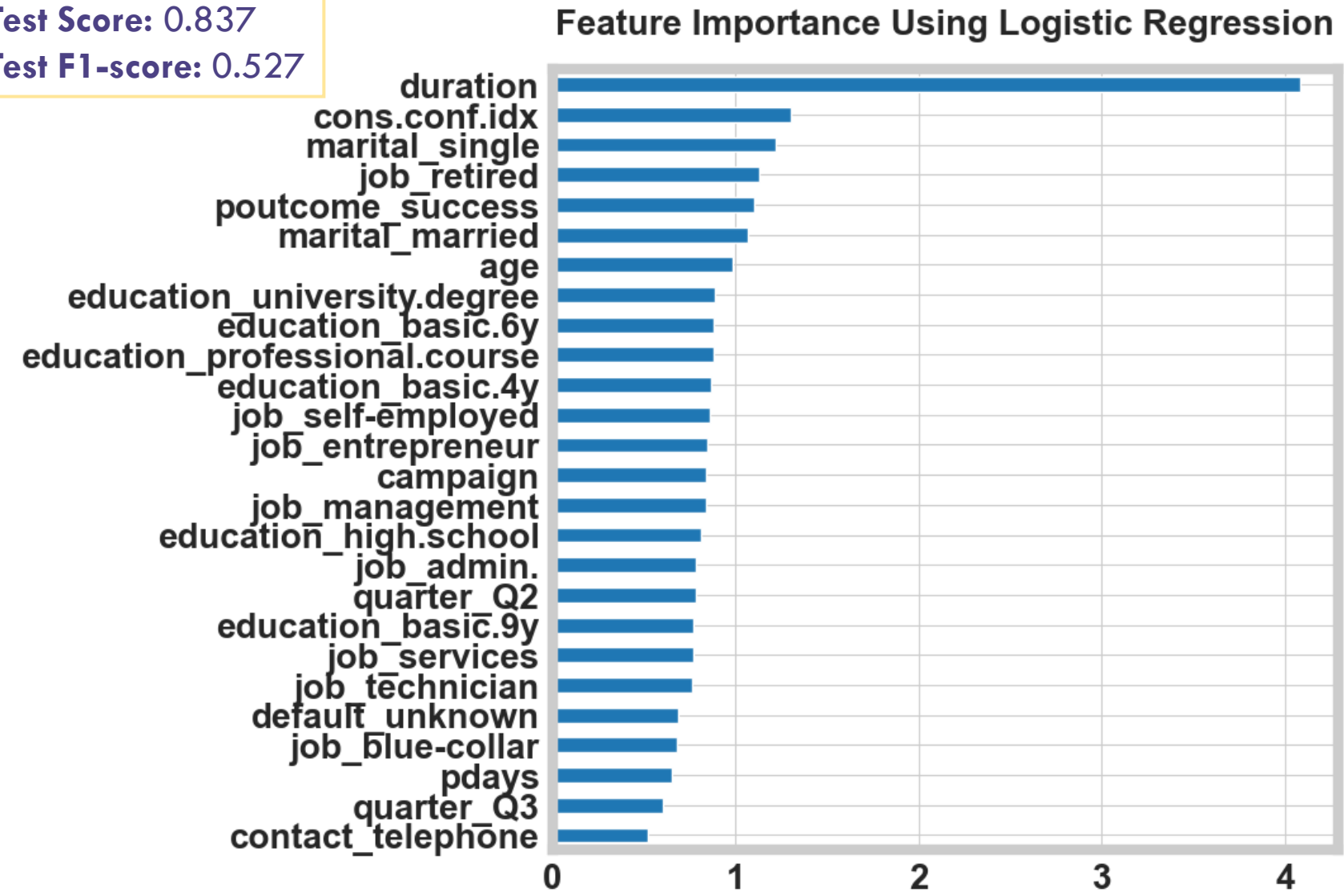
Class weight optimization and up-sampling minority class for Neural Network

Our data contains 88.7% of customers do not subscribe to CDs, while 11.3% do. To address this imbalance, we employed several techniques:

- Oversampling the Minority Class: We duplicated the minority class observations in the training dataset to balance it with the majority class
- Class Weight Adjustment: We assigned higher weights to the minority class during model training
- Threshold Tuning: The probability threshold for determining crisp labels was fine-tuned, rather than using the default threshold of 0.5

Logistic regression was used as a base model

Train Score: 0.838
Test Score: 0.837
Test F1-score: 0.527

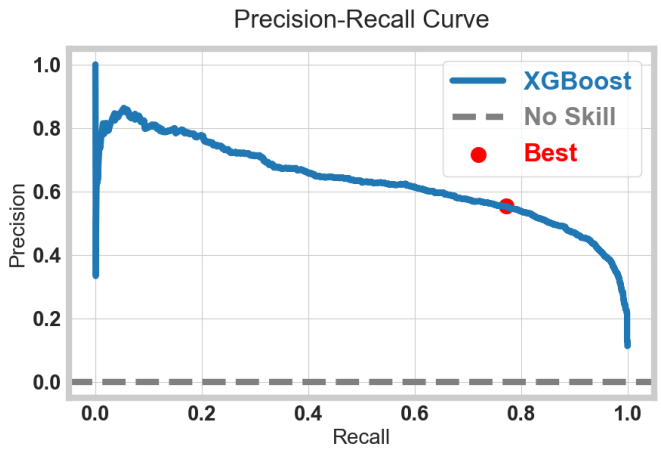


- Balanced class weight was used from scikit-learn library to handle imbalanced data
- Call duration, consumer confidence index, job retired, previous outcome success and age were the most important features. This aligns well with our observation in exploratory data analysis

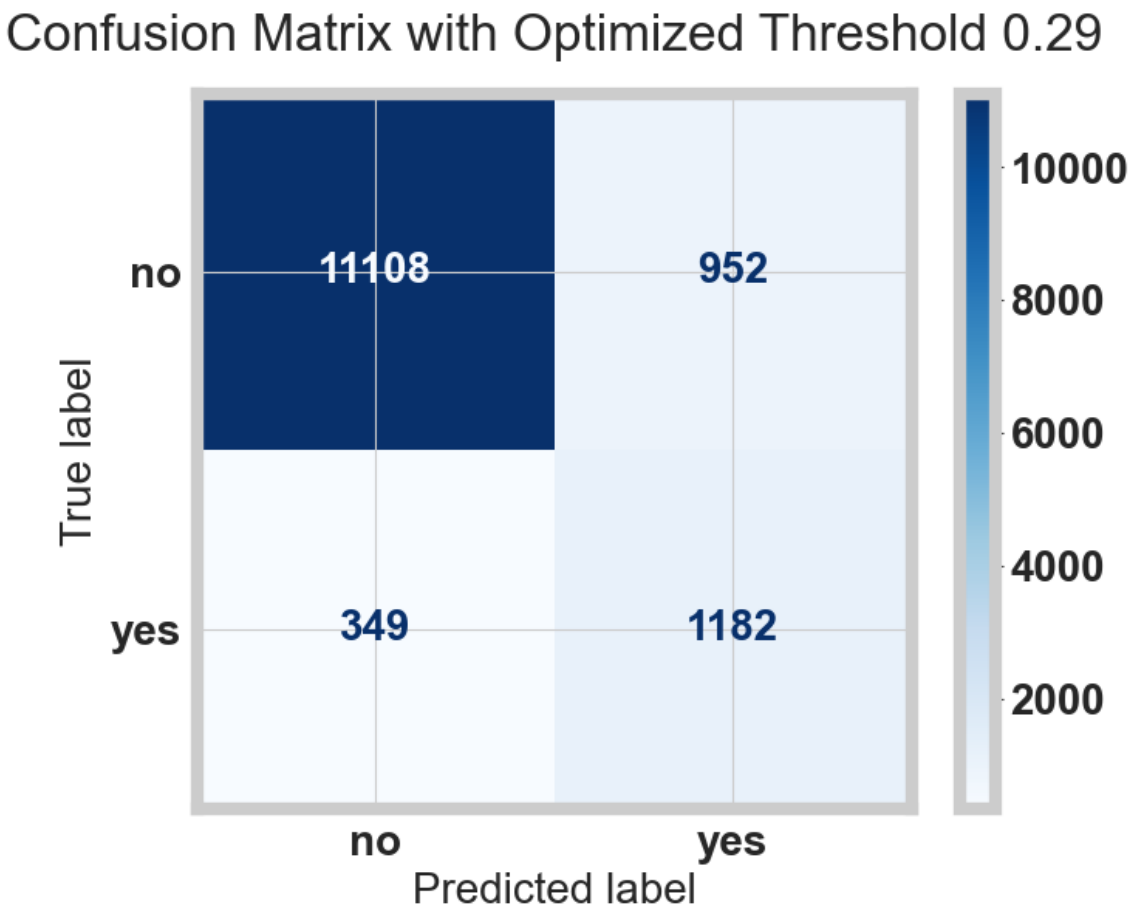
Threshold optimized XGBoost had a high recall and F1-score with better generalization to unseen data

Model	Train Score	Test Score	Train Precision	Test Precision	Train Recall	Test Recall	Test F1 Score	Test ROC AUC Score
Logistic	0.8381	0.8371	0.3934	0.3914	0.8051	0.8040	0.5265	0.9017
Random Forest	0.8505	0.8488	0.4207	0.4160	0.8681	0.8478	0.5582	0.9238
SVM	0.8453	0.8361	0.4149	0.3957	0.9099	0.8628	0.5426	0.9144
KNN	0.9249	0.8970	0.7512	0.5704	0.4992	0.3468	0.4314	0.8363
XGBoost	0.9539	0.9103	0.8501	0.6236	0.7170	0.5140	0.5636	0.9400
Hypertuned XGBoost	0.9219	0.9109	0.6956	0.6333	0.5455	0.4964	0.5565	0.9447
Threshold-Tuned XGBoost*	0.9219	0.9109	0.5710	0.5538	0.7883	0.7721	0.6623	0.9446
Hypertuned Voting Classifier	0.9296	0.9001	0.6487	0.5467	0.8181	0.7082	0.6169	0.9354
Hypertuned Neural Network	0.8157	0.7952	0.6123	0.3412	0.9045	0.8793	0.7303	0.4917

- XGBoost, Voting Classifier and Neural Network were hyper-tuned using 5-fold Grid Search Cross Validation
- The hyper-tuned XGBoost model achieved the best recall and precision, demonstrating strong generalization to unseen data. The optimized threshold was 0.29



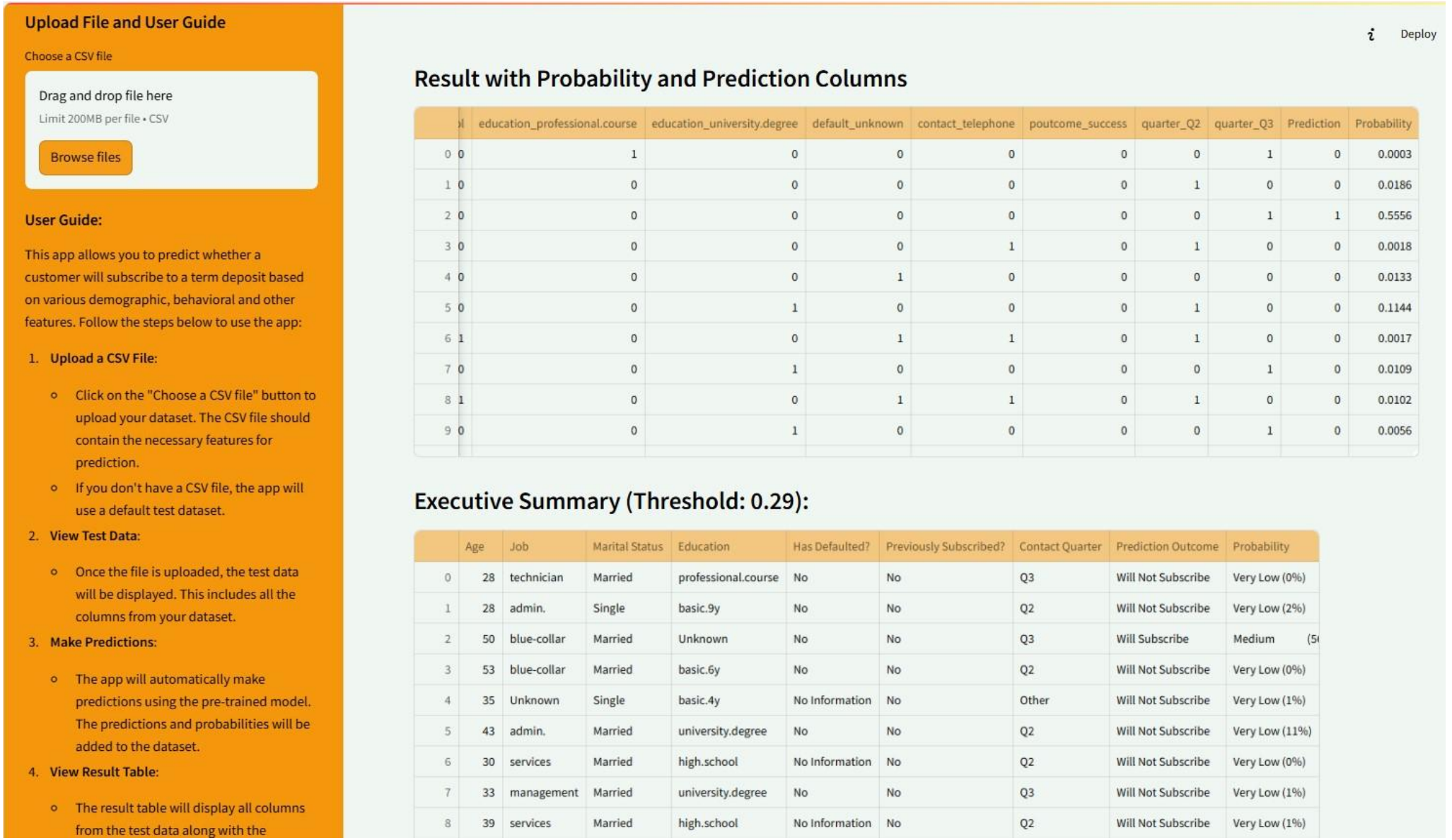
Identifying customers who will subscribe to CDs involves costs associated with targeting those who ultimately do not subscribe



The cost of missing a potential subscriber is higher, we prioritize recall to ensure we capture as many subscribers as possible

- **Optimized Precision (56.6%)**: Some marketing efforts on non-subscribers
- **Optimized Recall (76.2%)**: Reduce risk of missing potential subscribers
- **Maximized F1 Score (64.9%)**: Good balance between capturing subscribers and managing costs

Model deployment and continuous monitoring on cloud using Streamlit app



Upload File and User Guide

Choose a CSV file

Drag and drop file here

Limit 200MB per file • CSV

Browse files

User Guide:

This app allows you to predict whether a customer will subscribe to a term deposit based on various demographic, behavioral and other features. Follow the steps below to use the app:

1. **Upload a CSV File**:
   - Click on the "Choose a CSV file" button to upload your dataset. The CSV file should contain the necessary features for prediction.
   - If you don't have a CSV file, the app will use a default test dataset.
2. **View Test Data**:
   - Once the file is uploaded, the test data will be displayed. This includes all the columns from your dataset.
3. **Make Predictions**:
   - The app will automatically make predictions using the pre-trained model. The predictions and probabilities will be added to the dataset.
4. **View Result Table**:
   - The result table will display all columns from the test data along with the

Deploy

Result with Probability and Prediction Columns

	l	education_professional.course	education_university.degree	default_unknown	contact_telephone	poutcome_success	quarter_Q2	quarter_Q3	Prediction	Probability
0	0	1	0	0	0	0	0	1	0	0.0003
1	0	0	0	0	0	0	1	0	0	0.0186
2	0	0	0	0	0	0	0	1	1	0.5556
3	0	0	0	0	1	0	1	0	0	0.0018
4	0	0	0	1	0	0	0	0	0	0.0133
5	0	0	1	0	0	0	1	0	0	0.1144
6	1	0	0	1	1	0	1	0	0	0.0017
7	0	0	1	0	0	0	0	1	0	0.0109
8	1	0	0	1	1	0	1	0	0	0.0102
9	0	0	1	0	0	0	0	1	0	0.0056

Executive Summary (Threshold: 0.29):

	Age	Job	Marital Status	Education	Has Defaulted?	Previously Subscribed?	Contact Quarter	Prediction Outcome	Probability
0	28	technician	Married	professional.course	No	No	Q3	Will Not Subscribe	Very Low (0%)
1	28	admin.	Single	basic.9y	No	No	Q2	Will Not Subscribe	Very Low (2%)
2	50	blue-collar	Married	Unknown	No	No	Q3	Will Subscribe	Medium (5
3	53	blue-collar	Married	basic.6y	No	No	Q2	Will Not Subscribe	Very Low (0%)
4	35	Unknown	Single	basic.4y	No Information	No	Other	Will Not Subscribe	Very Low (1%)
5	43	admin.	Married	university.degree	No	No	Q2	Will Not Subscribe	Very Low (11%)
6	30	services	Married	high.school	No Information	No	Q2	Will Not Subscribe	Very Low (0%)
7	33	management	Married	university.degree	No	No	Q3	Will Not Subscribe	Very Low (1%)
8	39	services	Married	high.school	No Information	No	Q2	Will Not Subscribe	Very Low (1%)

- Retrain model periodically based on business needs
- Continuous monitoring of model predictive power
- Use population stability index (PSI), Z-score etc. to make sure there is no model drift

Customer segmentation and tailored marketing strategies based on subscription probability

Tier 1: High Probability (0.8 - 1.0)

- **Personalized Offers:** Tailor offers based on age, job, and education
- **Direct Communication:** Use preferred contact methods.
- **Exclusive Deals:** Highlight limited-time offers
- **Financial Advisory:** Provide personalized advice sessions

Tier 2: Medium Probability (0.5 - 0.8)

- **Targeted Campaigns:** Use digital marketing and mobile apps
- **Educational Content:** Share videos and articles on CD benefits
- **Incentives:** Offer bonuses or interest rate boosts.
- **Quarterly Promotions:** Align with financial planning cycles

Tier 3: Low Probability (0.3 - 0.5)

- **Awareness Campaigns:** Focus on CD safety and reliability
- **General Promotions:** Offer introductory rates and flexible terms
- **Cross-Selling:** Promote other banking products
- **Customer Engagement:** Use surveys to understand needs

Future enhancements: Integrating additional data sources for improved CD subscription prediction and marketing strategy

- **Enhanced Data Integration:** Incorporate client transaction history, investment activities, and savings account information to improve model accuracy
- **Behavior Forecasting:** Better predict customer behavior, opportunity costs, acquisition expenses, and potential lifetime value
- **Investment Prediction:** Analyze previous CD subscription data to determine how much a client might invest in a CD
- **Key Indicators:** Identify trends from savings patterns, transaction histories, and investment activities
- **Personalized Marketing:** Focus on customers with higher savings balances and frequent investment activities for larger CD investments
- **Lifetime Value:** Understand the long-term value customers bring and optimize acquisition costs accordingly

Acknowledgements

I would like to extend my heartfelt gratitude to everyone who provided invaluable assistance during this project.

- Vivian S. Zhang
- Cole Ingraham
- Vinod Chugani
- Khuzaima Shahid
- Jonathan Presley
- Philippe Heitzmann
- Kyle Gallatin
- All my cohorts