Improving marketing strategy using Machine Learning : Tele-marketing for Fixed Deposit (FD)

Brought to you by : Emma Tan

Capstone Project – Data Science and AI (IOD, Singapore)

Lead Trainer: Sifat Khan
Assistant Trainer: Eric Chan

AGENDA

- Study benefits
- Data exploration & processing
- Insights gained
- Model training & performance (supervised classification)

Basic Bank Operation:

Used in

Bank's Capital Source

Deposits

Fixed Deposit is **preferable**

1. Allows banks to hold onto the capital for a specific time frame

Loans

Revenue Generation

Interest on loans

Return on Investment

Trading & Sales

Deposit: A 1 trillion dollars market!

BOA: World's Top Banks

$1.9T (65% of BOA capital)

	YE 2022
Loans	$1.0 T
Deposits	$1.9 T
Net interest income	$52.5 B
Expenses	$61.4 B
Net charge-offs	$2.2 B
Earnings	$27.5 B
Headcount	217 K
Number of shares	8.0 B
Book value per share	$30.61
Active digital customers	44 MM
Customer satisfaction	87%
Employee satisfaction	85%

Extract from Bank of America Y2022 Annual Report

DBS: Singapore's Top Banks

DBS

$203bil (~40% DBS capital)

29. Deposits and Balances from Customers

Analysed by product	
Savings accounts	186,727
Current accounts	130,855
Fixed deposits	203,545
Other deposits	5,873
Total	527,000

Extract from DBS Y2022 Annual Report

BENEFITS OF STUDY

Stakeholders:

Managers and upper-level executives involved in business planning, strategic marketing, and communications.

Client patterns identification = improved marketing efficiency

- ✓**Time Saved**
- ✓**Dollar Saved**
- ✓**Improved Staff morale**

Prepared by Emma T. (29 Sep 2023)

DATA EXPLORATION & PROCESSING

Dataset overview

Marketing Outcome:
Fixed Deposit Subscription
Yes / No

Client Basic Information

- Age
- Job
- Marital Status
- Education Level
- Credit Default
- Housing Loan
- Personal Loan

Historical Marketing Campaign Info

- Contact type
- Month Contacted
- Day Contacted
- Contact Duration
- Contact Frequency

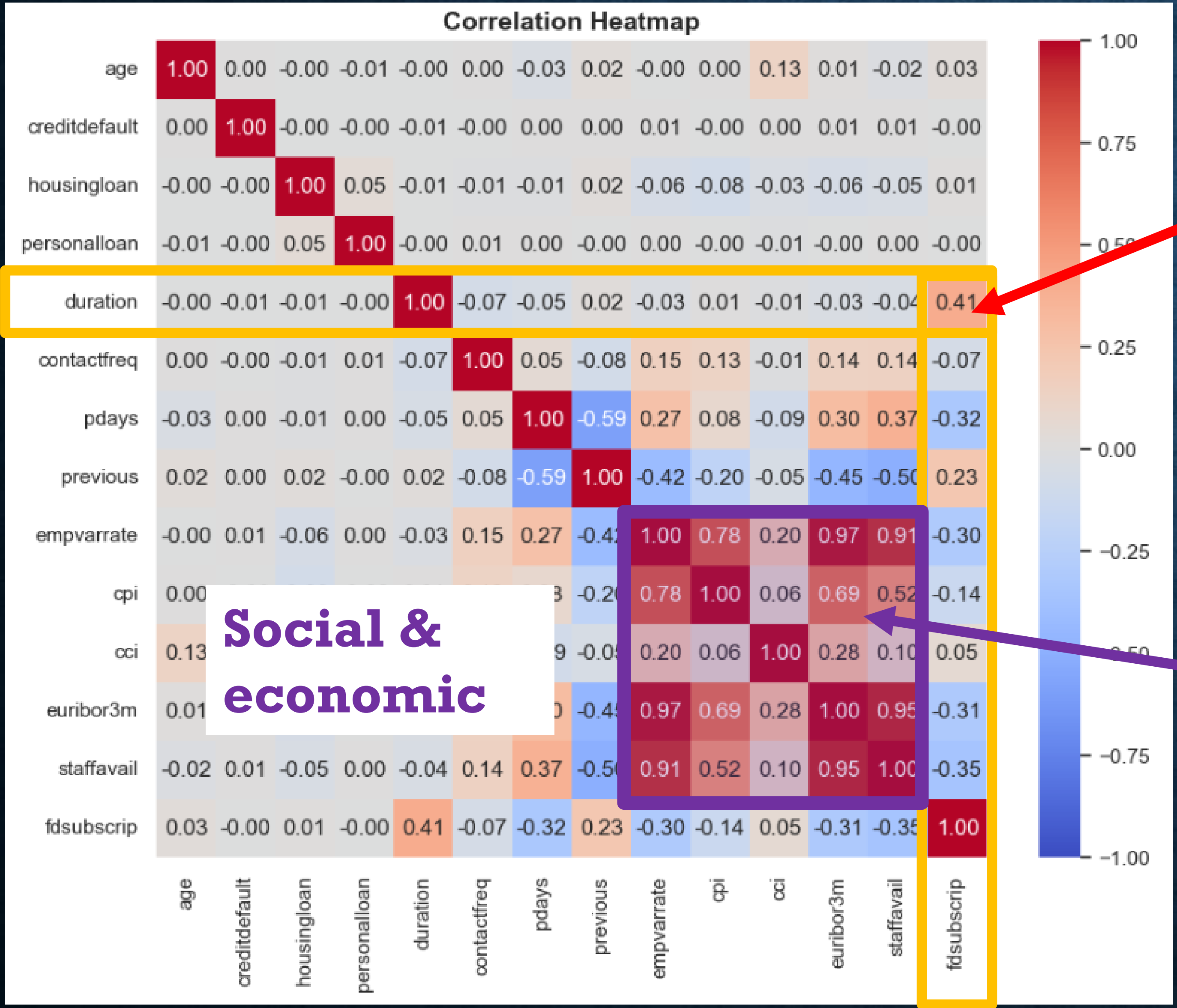
Social & Economic Context

- Employment variation rate
- Consumer confidence Index
- Consumer Price Index
- Euribor 3 months rate
- Number of employee

Prepared by Emma T. (29 Sep 2023)

DATA EXPLORATION & PROCESSING

ATTRIBUTES OVERVIEW



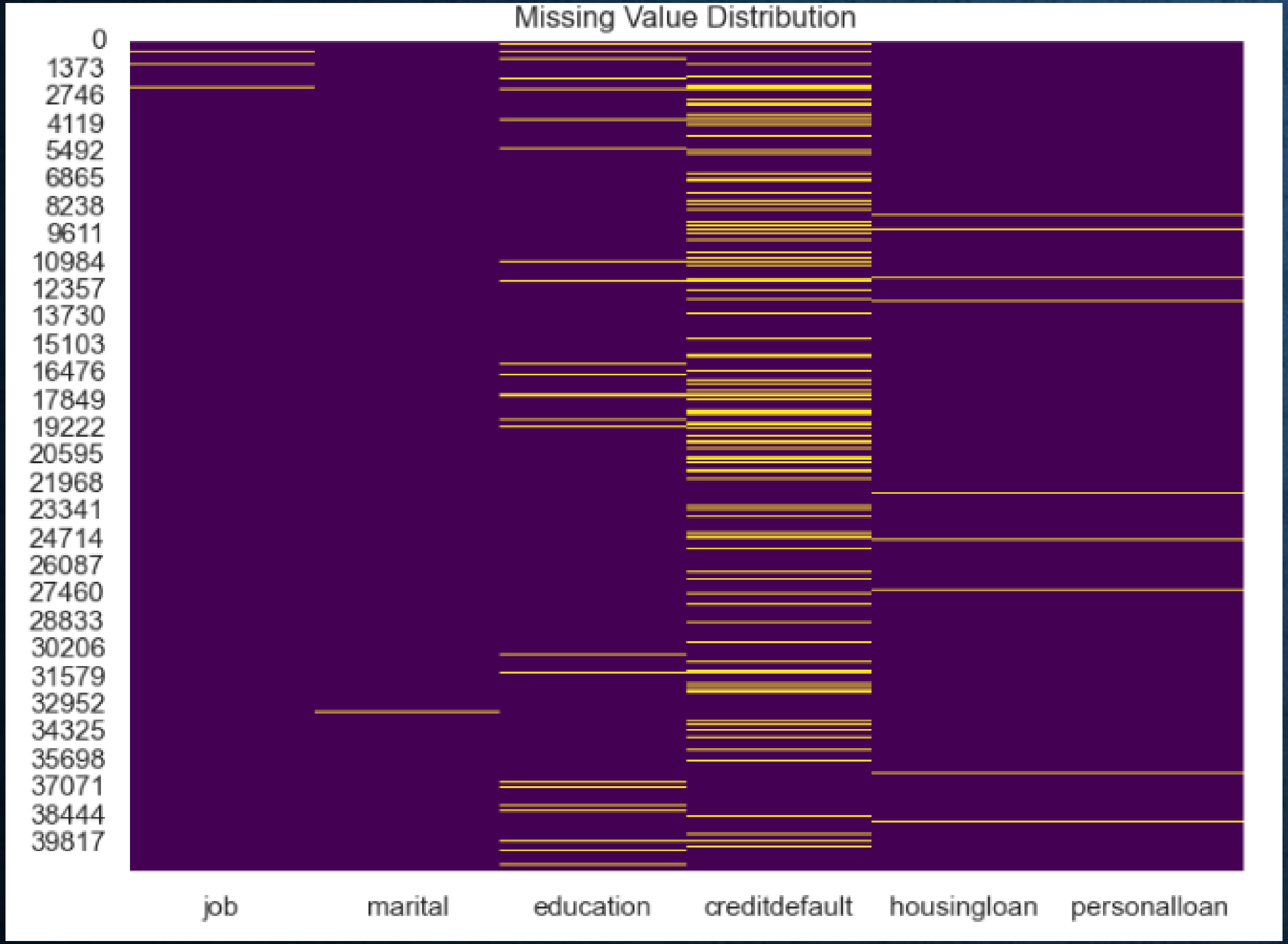
1. Highest correlation to FD Subscription: contact duration (score 0.41)

2. Overall attributes are weakly correlated to FD Subscription

3. Social and economic attributes show high correlation with each others

Prepared by Emma T. (29 Sep 2023)

DATA EXPLORATION & PROCESSING



- Missing values (30% of dataset)

DATA CLEANING

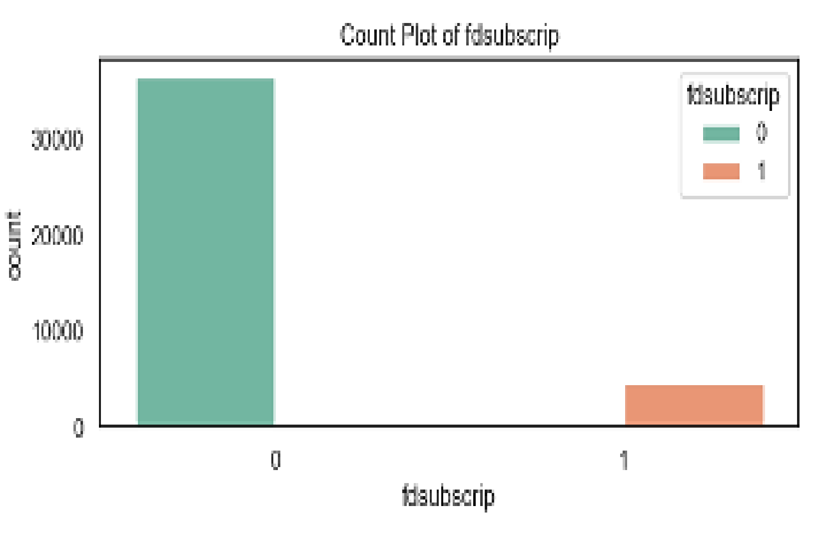
- 2 different approach used in handling null values:
  a) Leaving the null value as it is
  b) Mode imputation

- Data is otherwise cleaned and processed using similar method

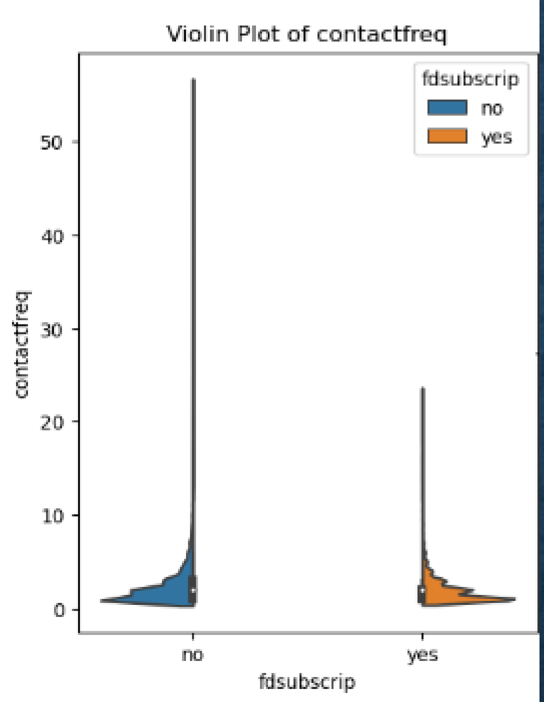
Prepared by Emma T. (29 Sep 2023)

DATA EXPLORATION & PROCESSING

DATA PROCESSING



- Class distribution : Severe Imbalance
  - Only 11% instances is FD subscriber from 41.2k entries
  - Handling by selecting algorithm that is robust against imbalance dataset



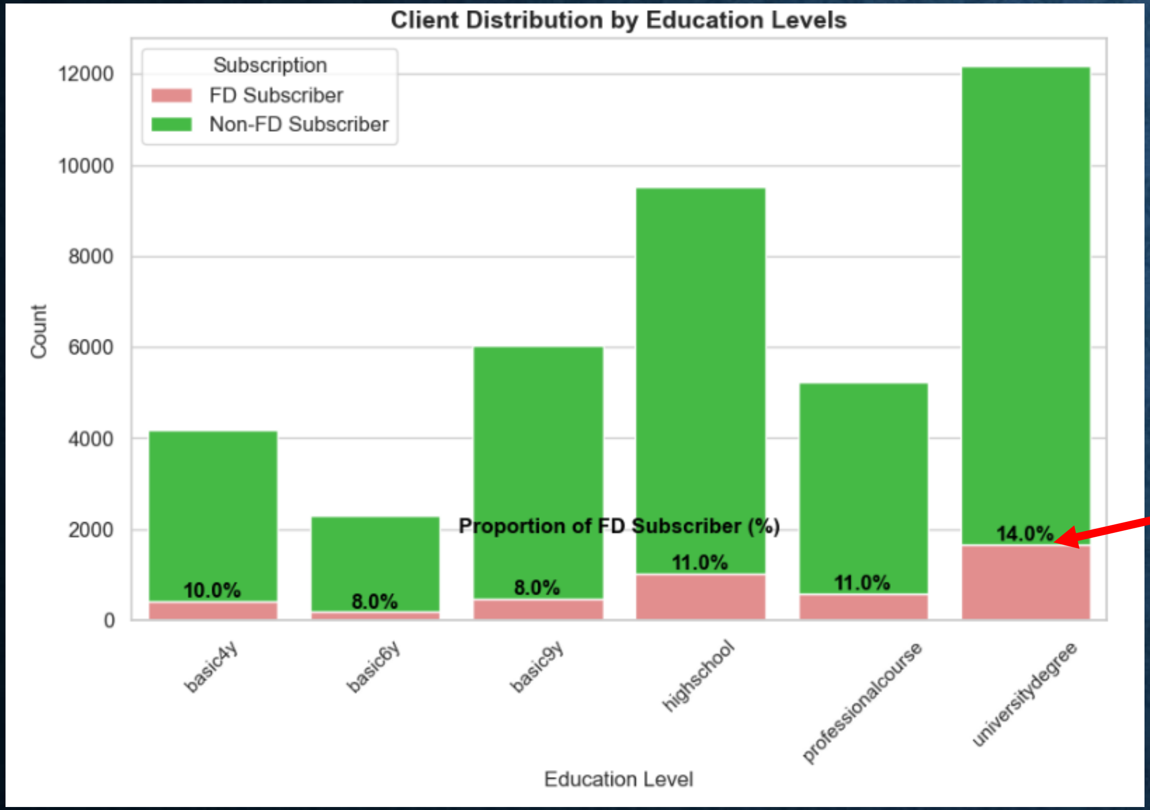
- Presence of Outliers
  - Keep as it is
  - Handling by selecting algorithm that is robust against outliers

- Feature engineering on
  - ✓ Marital status
  - ✓ Contact type
  - ✓ Education level
- Meaningful insights obtained from data exploration (next slide)

Prepared by Emma T. (29 Sep 2023)

Insights:

Education Levels



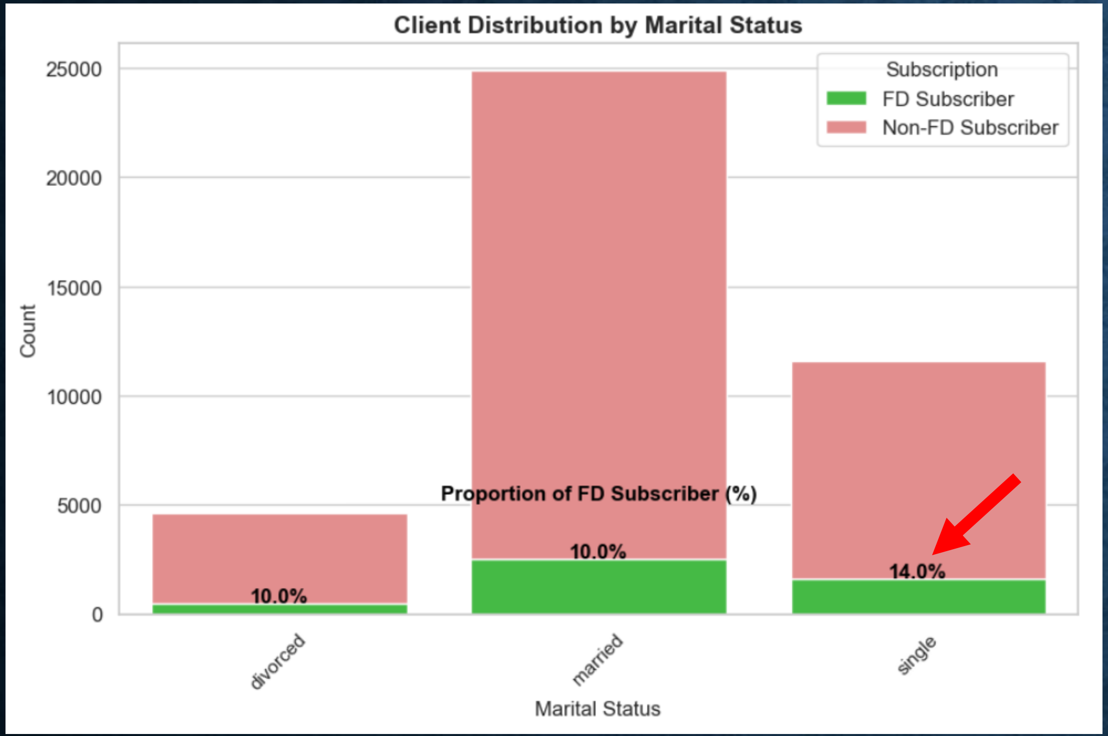
Higher % of FD Subscriber obtainable from

1. University graduate (4% higher)

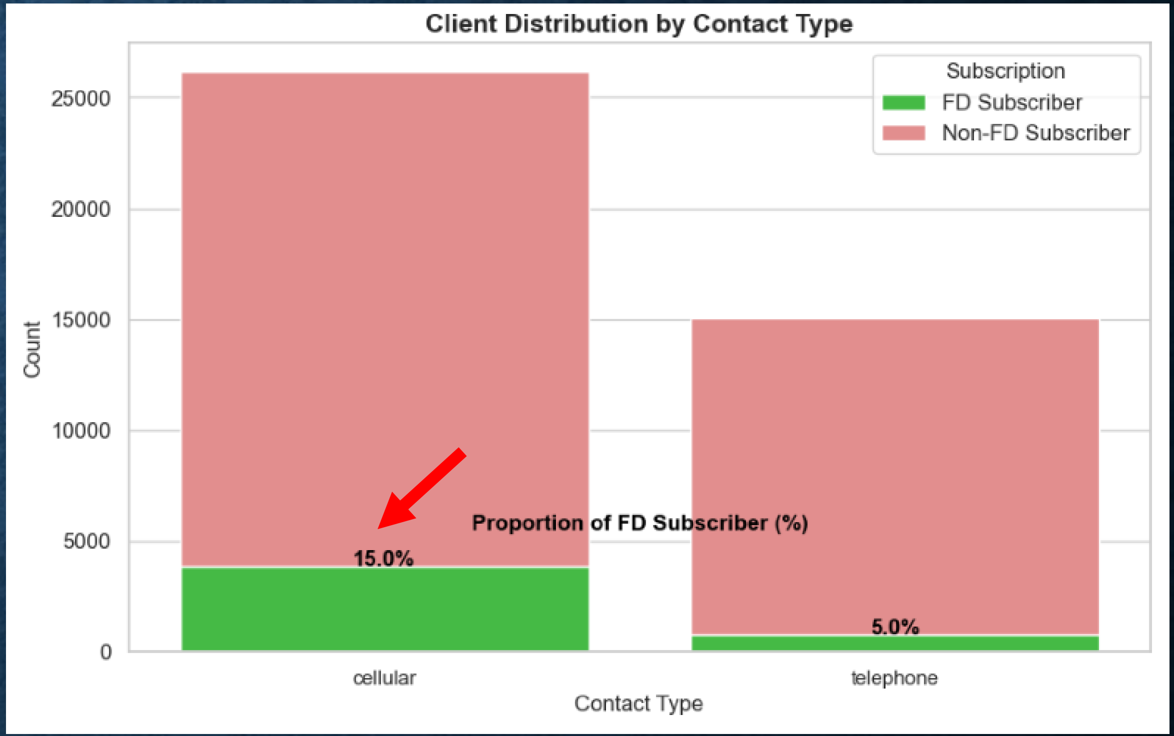
Prepared by Emma T. (29 Sep 2023)

Insights on clientele

Marital Status



Contact Type



Higher % of FD Subscriber among:

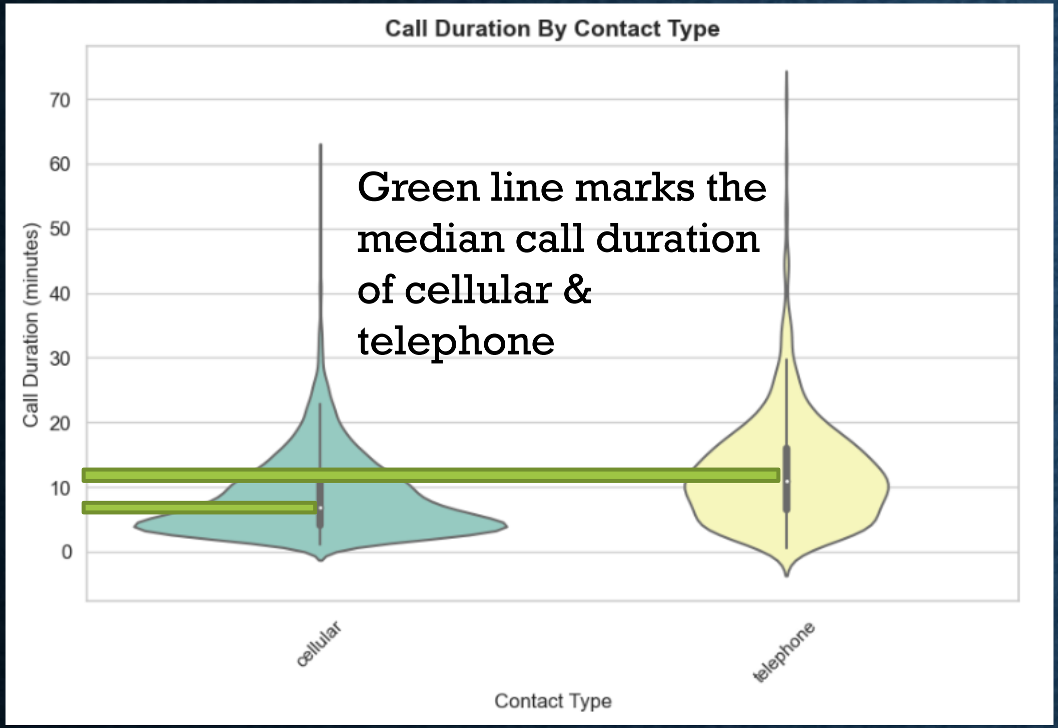
Insight Summary:

1. University graduate (4% higher)
2. Single (4% higher)
3. Contactable by cellular (10% higher)

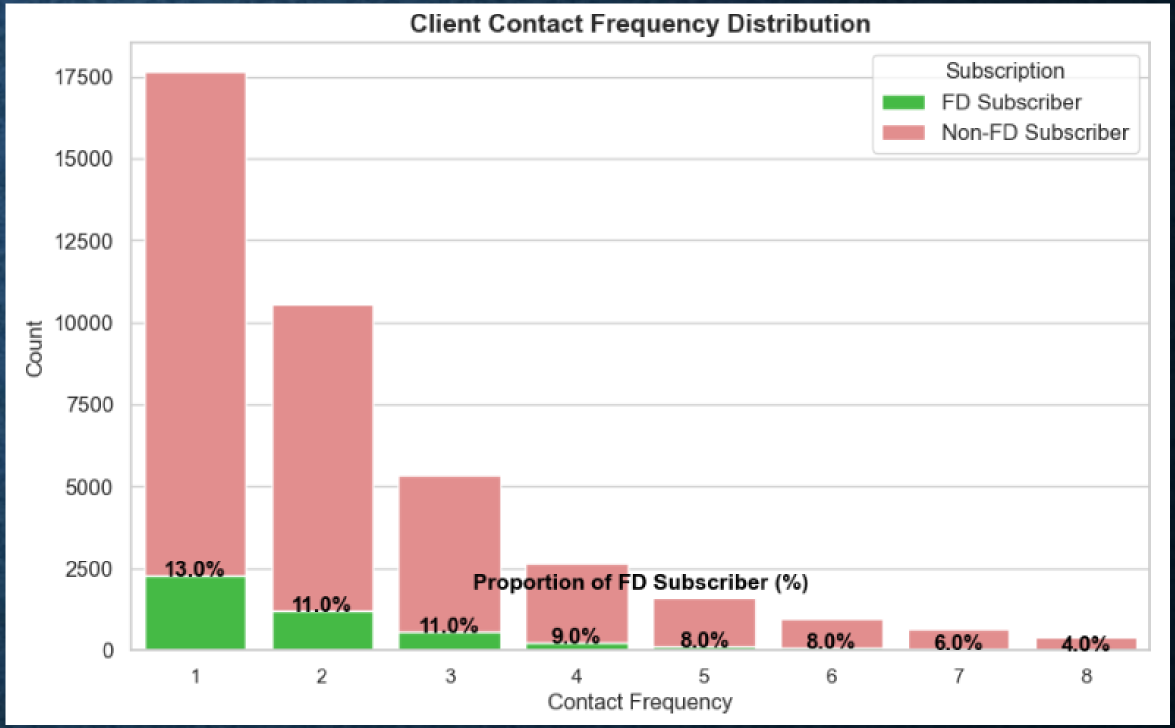
Prepared by Emma T. (29 Sep 2023)

Insights:

Call Duration by Contact Type



Contact Frequency



Unique insights:

Previous slide:

Higher % of FD Subscriber among:

1. University graduate (4% higher)
2. Single (4% higher)
3. Contactable by cellular (10% higher)

- ✓ Call duration using cellular is 4-5 mins shorter than telephone (higher output with shorter time spent)
- ✓ Contact up to 3 times (highest % FD subscription)

Prepared by Emma T. (29 Sep 2023)

MACHINE LEARNING – MODEL TRAINING & RESULT

- Approx. 26k instances are used for training (56% of dataset). Balance instances used as validation(20%) and testing(16%).

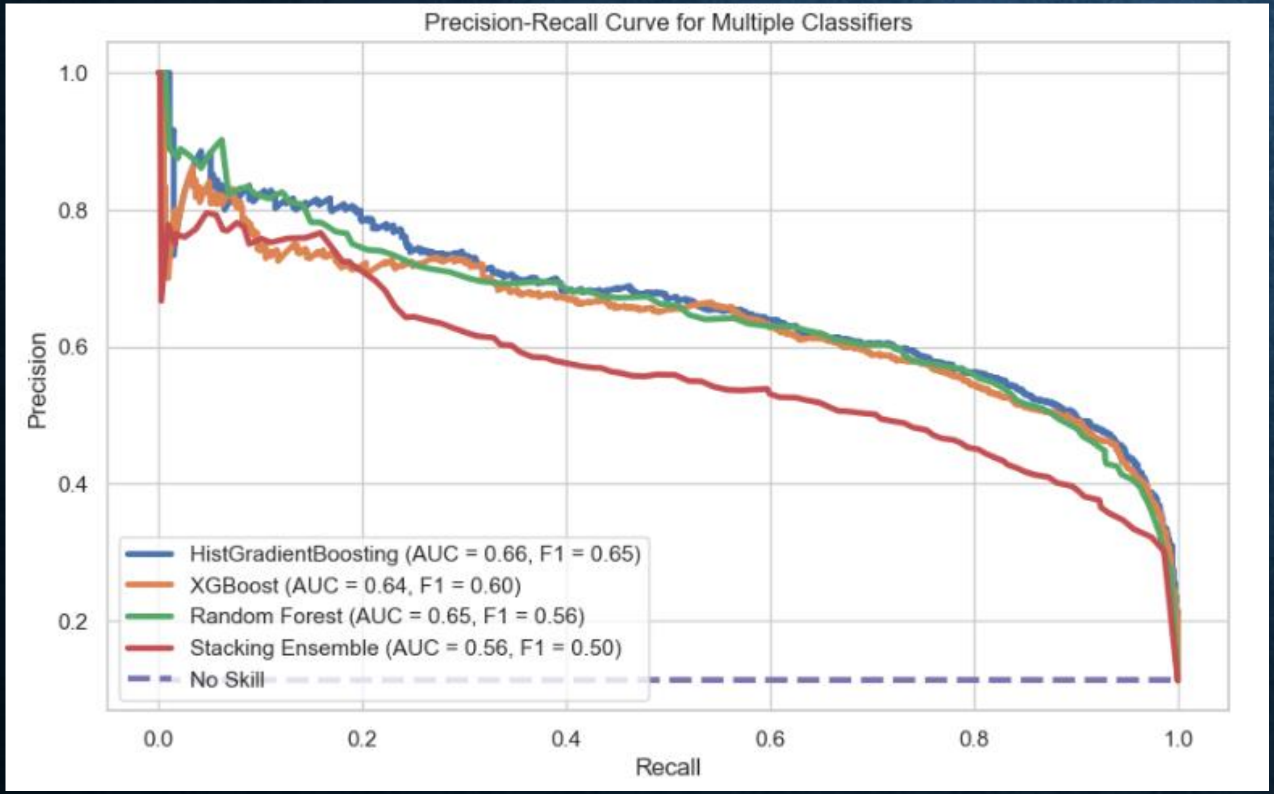
- Model selection based on algorithm ability to handle dataset challenges.

Challenges	HistGradient Boosting	Xtreme Gradient Boosting	Random Forest
Missing values	/	/	X (mode imputation)
Outliers	/	/	/
Imbalance data	/	/	/

- Stacking Ensemble (combining prediction from 3 models - often improve overall performance by leveraging the strengths of different models)

MACHINE LEARNING – MODEL TRAINING & RESULT

- Model performance visualisation using Precision-Recall Plot.
  - More informative and give an accurate prediction of future classification performance for imbalance dataset.
  - The plot evaluate the fraction of true positives (FD Subscriber) among positive predictions (Predicted FD Subscriber).



Precision-Recall Curve Interpretation Guideline:

1.Ideal Precision-Recall curve is one that starts at (0, 1) and goes to (1, 1), meaning perfect precision and recall, a curve that is as close to the top-right corner as possible.

2.Precision: Measure how many of the predicted positive instances were actually positive. It quantifies the accuracy of the positive predictions made by the model.

3.Recall (Sensitivity): Measure how many of the actual positive instances were correctly predicted by the model. It quantifies the ability of the model to capture all positive instances (true positive rate).

WHICH MODEL TO USE?

Classifier	Precision (Weighted Avg)	Recall (Weighted Avg)	F1-Score	AUC
HistGradientBoosting	0.6621	0.6616	0.6519	0.6616
XGBoosting	0.6383	0.6376	0.5979	0.6376
Random Forest	0.6474	0.6506	0.5602	0.6506
Stacking Ensemble	0.5588	0.5621	0.5007	0.5621

Note: Above is calculated using average precision score function.

Best model : **Histogram Gradient Boosting (Recall = 66%)**

Notes: Recall measure model's ability to capture all possible FD subscriber.

- Achieves a good balance between precision and recall
- Has the highest F1-score and AUC
- Indicate strong overall performance in classification of fixed deposit subscriber and non-subscriber.

Prepared by Emma T. (29 Sep 2023)

Current Marketing Approach (through guessing) : 50% success rate
Model Adoption Approach (through machine learning): 66% success rate

Assuming:
There's ~6600 clients

- No model = calling 6600 clients, 50% chance of success

- With model = calling only 614 clients, with 66% chance of success

```
Default Confusion Matrix
[['TN' 'FP']
 ['FN' 'TP']]
Confusion Matrix (HGB Classifier):
[[5635  213]
 [ 341  401]]
```

- Model predicted 614 clients will subscribe (~10%).

16% improvement translate to:

Time spent

Assumption: 1 client need 5 mins. To achieve same result:
No model = ~23 days spent, 11 days productive.
With ML model = ~2.5 days spent, 1.5 days productive

90% time saved !!

Dollars spent

Assumption: 1 client cost $1 to call. To achieve same result:
No model = $6600.
With ML model = $614

~$6k saved = 90% saving !!

Staff morale

No model: Staff receive rejection from >3000 clients = Low morale, higher chance of depression

With model: Staff receive rejection from ~200 clients. Better result boost morale.

FUTURE WORKS

- Model performance enhancement could involve :

- 1. Optimization of hyperparameter tuning.
- 2. Using more structured GridSearch Cross Validation methodology (current study used RandomSearchCV).

- End of Presentation -

-Thank you!-

Prepared by Emma T. (29 Sep 2023)