Project Documentation

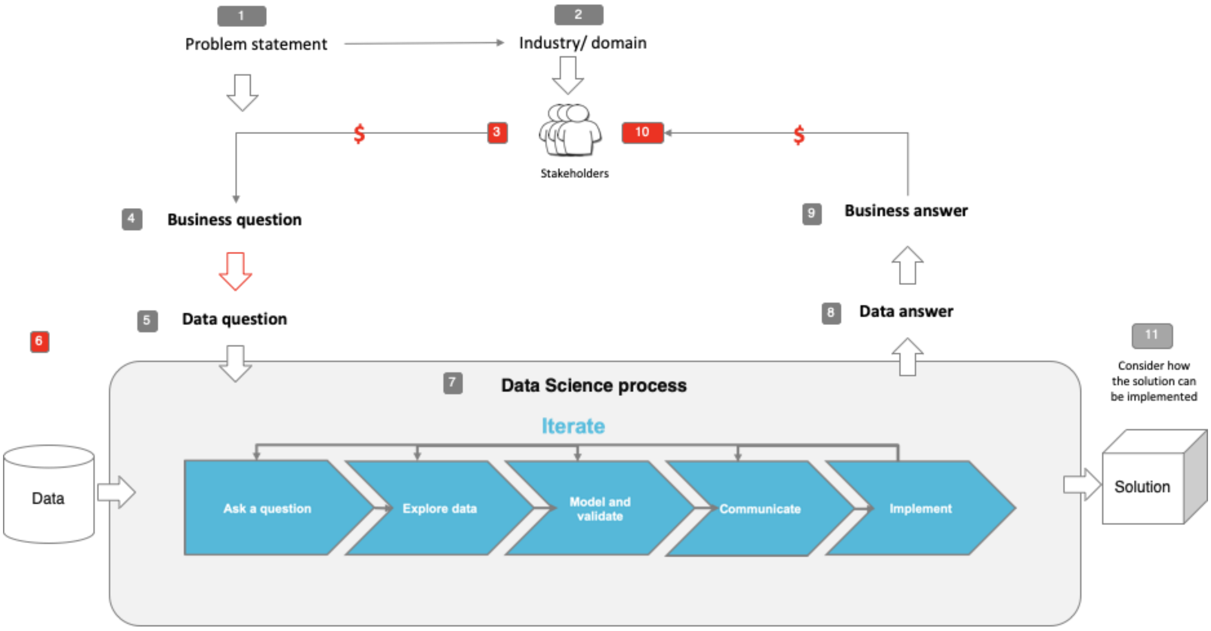
Project Author: Emma Tan
Document Date: 29 Sept 2023

Capstone Project Title:

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

Process overview

The following diagram shows the overall end-to-end process for defining, designing and delivering the Capstone project.



Note: The following are the candidate sections of the document. They are presented here for guidance.

Problem statement

- What is the problem or the opportunity that the project is investigating?
  - Deposit is a 1 trillion dollar market (Source: Bank of America Y2022 Annual Report).
  - Among top banks in Singapore, fixed deposit account for 40% of DBS capital (Source: DBS Y2022 Annual Report)
  - Getting more fixed deposit subscriber is essential for bank's overall funding and financial stability
- What is the desired state?
  - An efficient marketing strategy will result to time saved, dollar saved and higher employee morale.

Industry/ domain

- What is the industry/ domain?
  - Banking sector (specific to study dataset)
- What is the current state of this industry?
  - Absence of machine learning model
  - Marketing approach through guessing (50% success rate)
- Is the project relevant to other industries?
  - Approach used in this study can be adapted to sales and marketing department in other industry.

Stakeholders

- Who are the stakeholders?
  - Marketing managers
- Why do they care about this problem?
  - 1 trillion dollar market is not to be underestimated
- What are the stakeholders' expectations?
  - An improved marketing strategy which can provide substantial cost and time savings

Business question

- What is the main business question that needs to be answered?
  - How to improve marketing strategies (getting more Fixed deposit subscriber) for high capital generation at lowest cost and time spent?
- What is the business value of answering this question?
  - Up to 90% time and cost saving is possible if choose to adopt the machine learning model
- What is the required accuracy? What are the implications of false positives or false negatives?
  - Due to imbalance dataset and business context, recall is used instead of accuracy to give a more accurate assessment to model performance.
  - Recall measure model's ability to capture all possible FD subscriber.
  - High False Negative (FN) will lead to significant loss in potential capital gain (missed out to call client who will actually subscribe to Fixed Deposit). Hence, model strive to minimise FN.
  - False Positive will lead to time wasted to call up clients who is not likely to subscribe.

Data question

- What is the data question that needs to be answered?
    - Among the given list of clients, who is likely to subscribe a Fixed Deposit and who will not?
    - Machine learning using supervised classification to predict Fixed Deposit Subscriber.
    - Any patterns in clientele and previous marketing outcome that can be used to optimise future marketing effort?
    - Exploratory data analysis to detect primary patterns.

- What is the data required to answer the question?
    - Clients info, historical marketing info and outcome
    - Relevant social and economic info (optional)
    - Outcome of FD subscription from previous marketing campaign
    - Amount subscribed in Fixed Deposit from previous marketing campaign (absent in this dataset, will be good to have to optimise marketing strategy for high return in capital)

Data

- Where was the data sourced?
    - Bank marketing dataset obtained from University of California, Irvine (UCI). Donated by a Portuguese banking institution in 2012.

- What is the volume and attributes of the data?
    - 41188 instances
    - 20 attributes
    - 1 target variable (Outcome of FD Subscription)

- How reliable is the data?
    - Reliable as it is based on actual data
- What is the quality of the raw data?
    - 30% missing values from categorical data
    - Likely due to client being reluctant or unwilling to disclose certain information
    - Severe imbalance of in positive class distribution (Fixed subscriber only 11% of dataset)
- How was this data generated?
    - Donated by a Portuguese banking institution in 2012.
- Is this data available on an ongoing basis?
    - Yes. Link: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Data science process

Data analysis

- What data pipeline was to wrangle the raw data?
    - Null values are identified & handled using 2 different approaches (keep as it is & mode imputation)
    - Categorical values are identified and converted to numerical form for machine learning
    - Presence of outliers among attributes (Future work: to be scaled to check improvement on model's performance)
    - Feature engineering on attributes that shows promising result from exploratory data analysis findings
    - Conversion of categorical data to numerical data using mapping and one hot encoder (dummies)

- What are the highlights of the Exploratory Data Analysis (EDA)?
    - Higher % of FD Subscriber among:
        1. University graduate (4% higher)
        2. Single (4% higher)
        3. Contactable by cellular (10% higher)
        4. Students and retiree (30% higher)

    - Call duration using cellular is 4-5 mins shorter than telephone (higher output with shorter time spent)
    - Contact up to 3 times is sufficient (highest % FD subscription)
    - Campaign Timings: early and end of the year potentially yield better result than mid of year

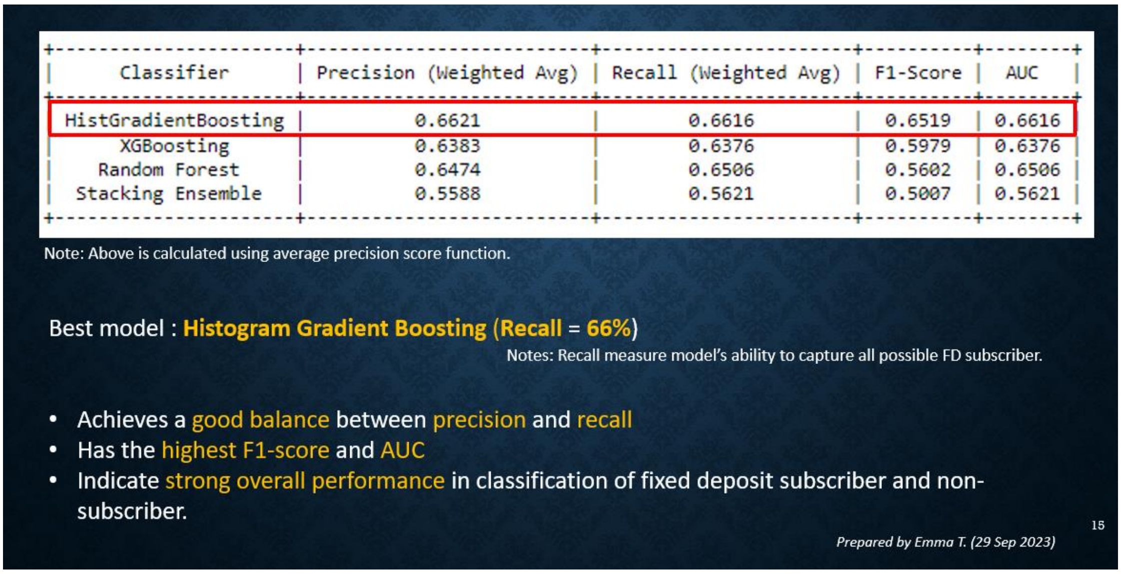
- Is the pipeline reusable? (for example, to process future data?)
    - Yes. Data however need to be cleaned and process in similar approach

- What are the intermediary data structures used (if any)?
    - Data frames and arrays

Modelling

- What are the main features used?
  - All dependent attributes are used
- Did you find any interesting interactions between features?
  - Overall, features shows weak correlation to Fixed Deposit subscription
  - Social and economic attributes shows high correlation with each other's
- Is there a subset of features that would get a significant portion of your final performance? Which features?
  - Contact duration (score 0.41 correlation to Fixed deposit subscription)
- How did you select features?
  - Those that have significant different in patterns(values) between subscriber and non-subscriber are identified
- What feature engineering techniques are used?
  - Reassigning values to features of interest (e.g. simplify values of attributes: single =1, divorced & married =0 in marital status)
- What are the models used?
  - Histogram Gradient Boosting, Xtreme Gradient Boosting, Random Forest
  - Models selection based on algorithm ability to handle imbalance dataset, outliers and null values.
- How long does it take to train your model?
  - Actual model training took a few days as features need to be optimise (e.g. without vs with feature engineering) to verified what actually improve model's performance.
  - Final code computation time is negligible (longest around 3-4 mins).
- What are the tools used? (cloud platform, for example)
  - Jupyter notebook

- What are the model performance metrics?



```
+----------------------+-------------------------+-----------------------+----------+--------+
|      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.

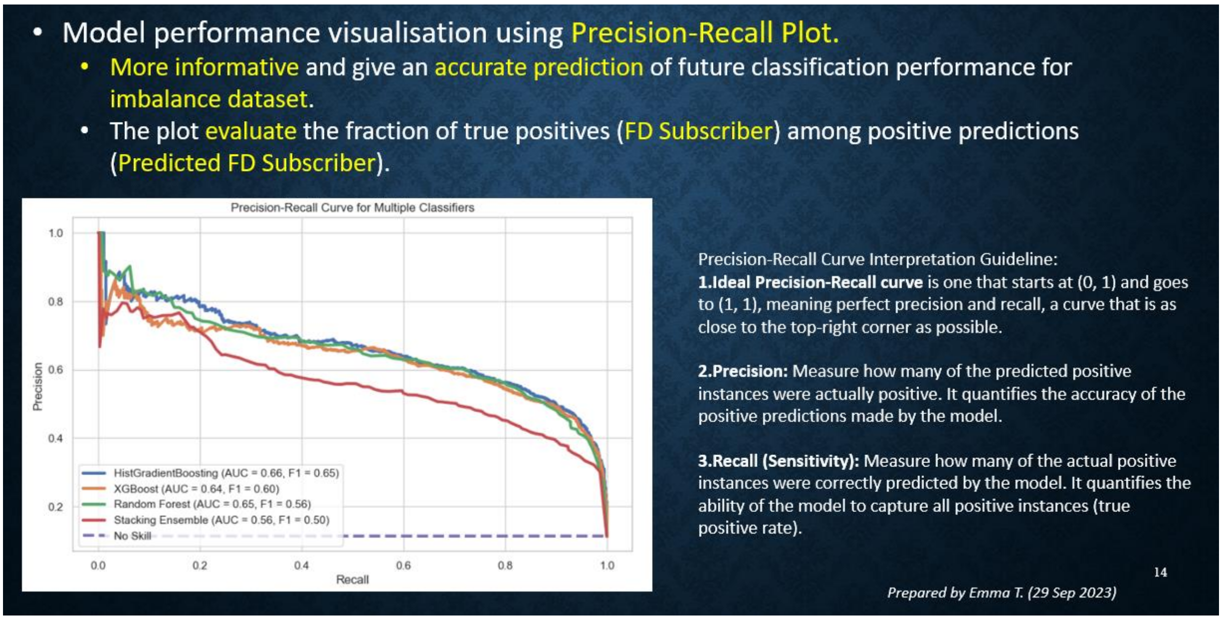
15

Prepared by Emma T. (29 Sep 2023)

- Which model was selected?
  - Histogram Gradient Boosting is recommended

Outcomes

- What are the main findings and conclusions of the data science process?
  - 66% success rate and substantial cost and time saving can be achieved (up to 90%)



- 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:

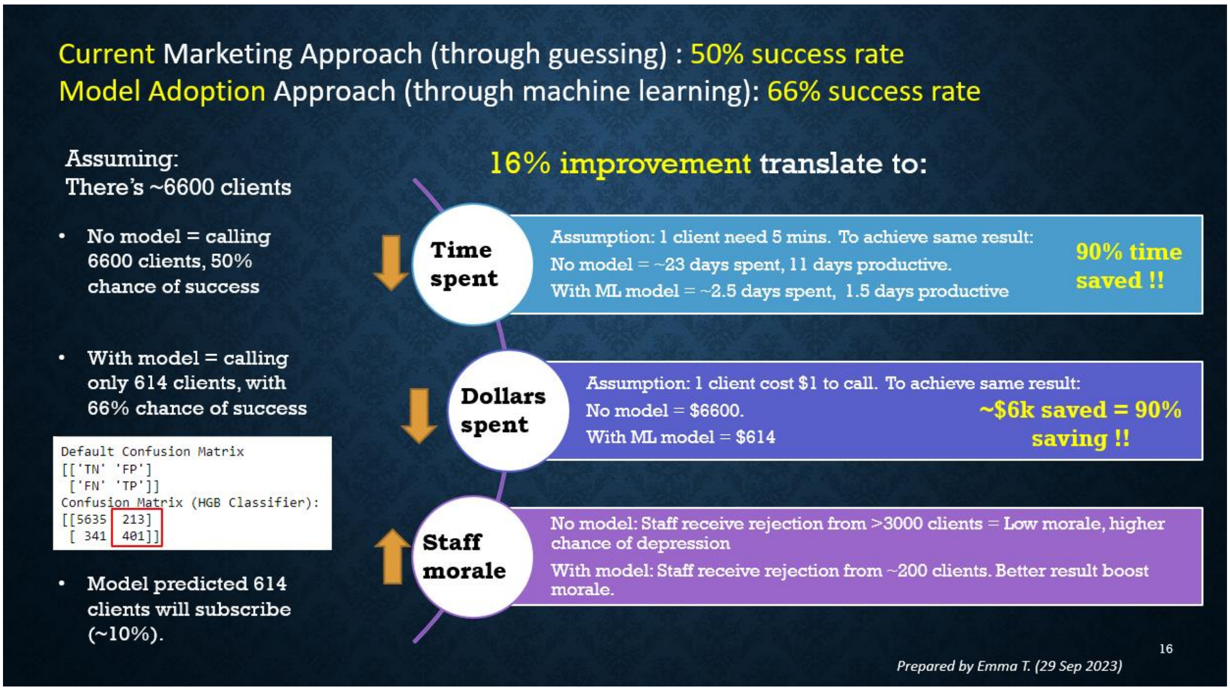
1.Ideal Precision-Recall curve is one that starts at (0, 1) and goes to (1, 1), meaning perfect precision and recall, a curve that is as close to the top-right corner as possible.

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).

14

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:

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 !!

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

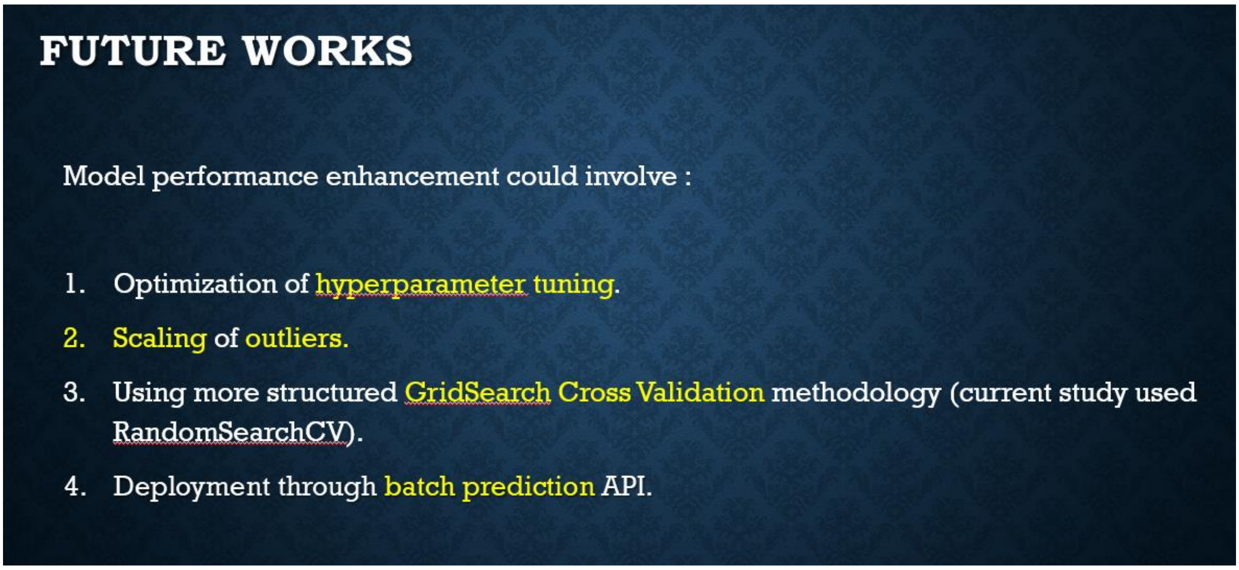
~$6k saved = 90% saving !!

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.

16

Prepared by Emma T. (29 Sep 2023)



FUTURE WORKS

Model performance enhancement could involve :

1. Optimization of hyperparameter tuning.
2. Scaling of outliers.
3. Using more structured GridSearch Cross Validation methodology (current study used RandomSearchCV).
4. Deployment through batch prediction API.

Implementation

- What are the considerations for implementing the model in production?
    - Data Quality and Privacy: Ensure high-quality data and strong data privacy measures to protect sensitive financial information.
    - Model Selection and Evaluation: Choose the right model, evaluate its performance using appropriate metrics, and validate it with robust testing.
    - Scalability and Monitoring: Plan for scalability to handle production-level workloads and set up monitoring for model performance and system health.
    - Compliance and Fairness: Adhere to relevant regulations, address bias and fairness issues, and maintain compliance with data protection laws.
    - Deployment Planning: Develop a well-documented deployment plan, consider user experience, and have a rollback strategy in case of issues to ensure a smooth production deployment.

Data answer

- Was the data question answered satisfactorily?
    - Yes
- What is the confidence level in the data answer?
    - 66% recall & precision rate

Business answer

- Was the business question answered satisfactorily?
    - Yes
- What is the confidence level in the business answer?
    - 66% success rate of Fixed deposit subscription if choose to trust and adopt the machine learning model

Response to stakeholders

- What are the overall messages and recommendations to the stakeholders?
    - Strongly recommend adoption of model trained as it give up to 90% time and cost saving, in addition of boosting sales & marketing staff's morale in getting a positive Fixed Deposit subscription

End-to-end solution

- What is the overall end-to-end solution to use the model developed in the project?

1. Data Preparation and Feature Engineering:
   Gather and pre-process relevant data, engineer meaningful features for the model.
2. Model Development and Evaluation:
   Choose and train a suitable machine learning model, assess its performance on validation and test data.
3. Model Deployment:
   Host the model in a deployment environment, create an API for interaction.
4. Monitoring and Security:
   Implement monitoring and security measures, ensuring compliance with regulations.
5. Maintenance and User Support:
   Establish a feedback loop, regularly update the model, and provide user training and support for a reliable, user-friendly solution.

References

- Where are the data and code used in the project? (show a simplified list of main items: notebooks, datasets, exported models)

1. Refer requirement.txt for libraries and environment setup needed
2. Refer bank-additional-full.csv for datasets
3. Refer code file_FTMS (final).ipynb for codes
4. Refer bankmarketing_model for model (pickle file)

- Additional documents provided:

1. Refer presentation slide_Marketing Strategies Fine-tuning Using ML (final)

- A copy of notebook, dataset, exported model and presentation slide is available at https://github.com/EmmaT0611/improving-bank-marketing-using-ML.git

- What are the resources used in the project? (libraries, algorithms, etc)

EDA, CLEANING & PROCESSING:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline

import warnings
# Suppress all warnings
warnings.filterwarnings("ignore")
```

DATA SPLITTING AND BALANCING:

```
from sklearn.model_selection import train_test_split
```

Machine Learning & Performance Visualisation (Precision-Recall Curve & Metrics calculation using Average Precision Score):

```
import xgboost as xgb
from sklearn.ensemble import HistGradientBoostingClassifier
from sklearn.ensemble import StackingClassifier, RandomForestClassifier
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score, f1_score, confusion_matrix, classification_report
from sklearn.metrics import precision_recall_curve, average_precision_score, auc
from sklearn.model_selection import RandomizedSearchCV
from sklearn.datasets import make_classification
from tabulate import tabulate
```

Saving best performing model:

```
import pickle
```

Otherwise, refer requirement.txt for the same listing.

End of documentation.
Thank you!