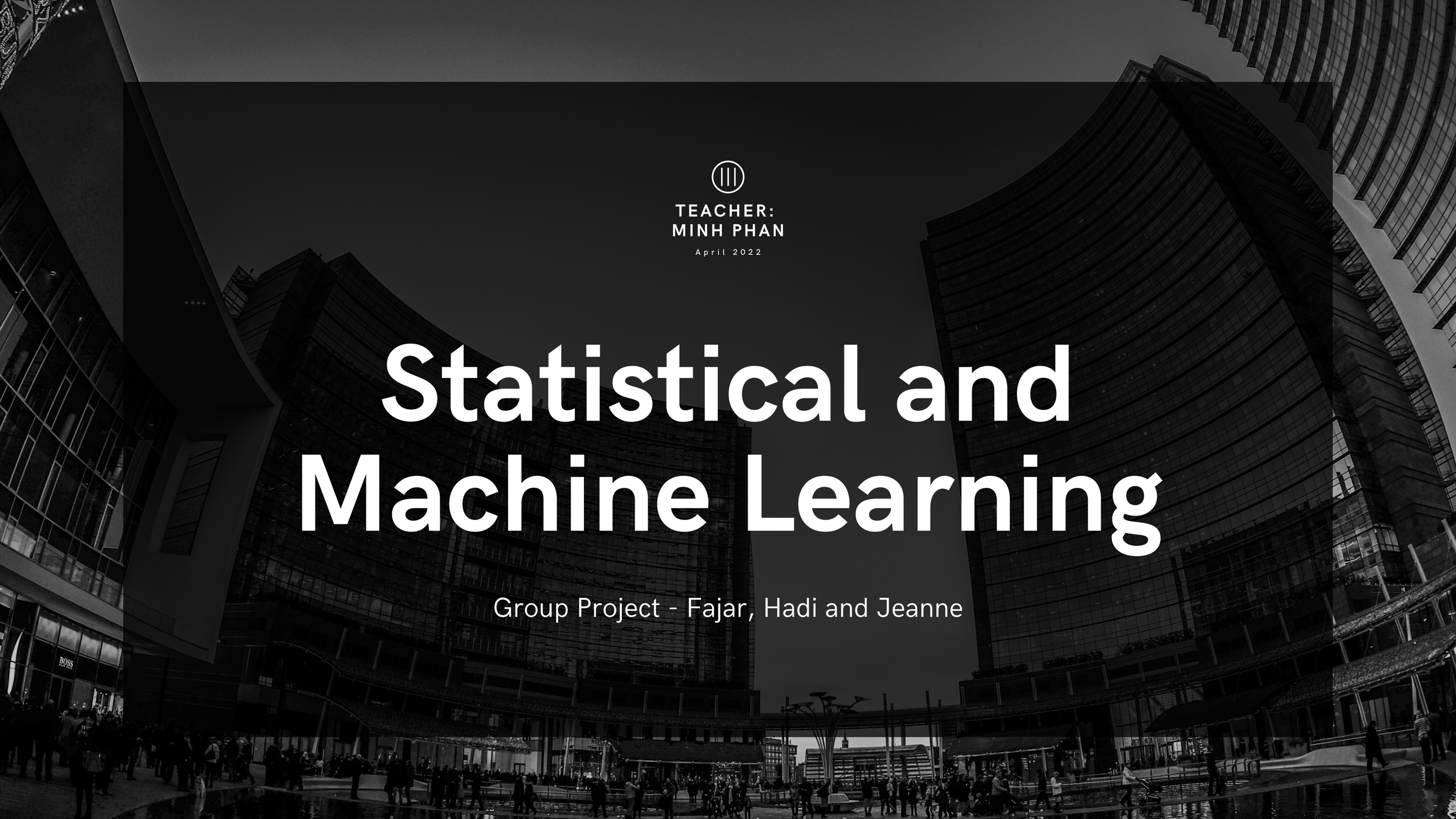


TEACHER:
MINH PHAN

April 2022

Statistical and Machine Learning

Group Project - Fajar, Hadi and Jeanne

STATISTICAL
GROUP PROJECT

Summary

What this report covers

01 Problem Definition

02 Data Summary & Processing

03 Methodology

04 Result

05 Conclusion

Problem Definition

Business Problem

A Portuguese retail bank is trying to sell long-term deposits to its clients through telemarketing campaign

Goal

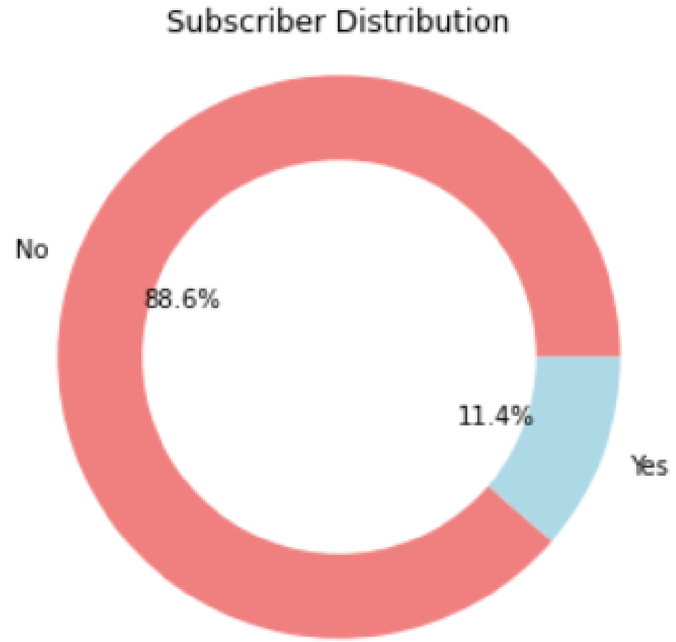
Develop a high performance & reliable model that can be used to improve decision making

Data Science Approach

We aim to develop a binary classification model that helps the bank to focus towards potential clients

Data Summary

- 20,000 bank clients data recorded from 2008 - 2013
- Covers demographic, social & economic, and last campaign information
- More than 60% clients in their 30s - 40s
- Almost 25% works as an admin
- Almost 30% has university degree



Data Processing

Numerical Variables

- Imputation of Missing values (mean & constant)
- Missing value indicator variables

Categorical Variables

- Imputation of Missing values (constant)
- Missing value indicator variables
- Value representation (label encoding + factorization)

Feature engineering & Selection

- Univariate Selection
- Smoothing Parameters & Polynomial term

Methodology

Data Processing

Validation Data processing is based on train Data

Model Evaluation & Selection

Scoring criteria: AUC on validation set

Train & Validation Split

Data splitting through stratified sampling

Model Development

Different models are built and compared

Final Prediction

Final Prediction on unlabeled test set

Experimental Setup

Data & Features

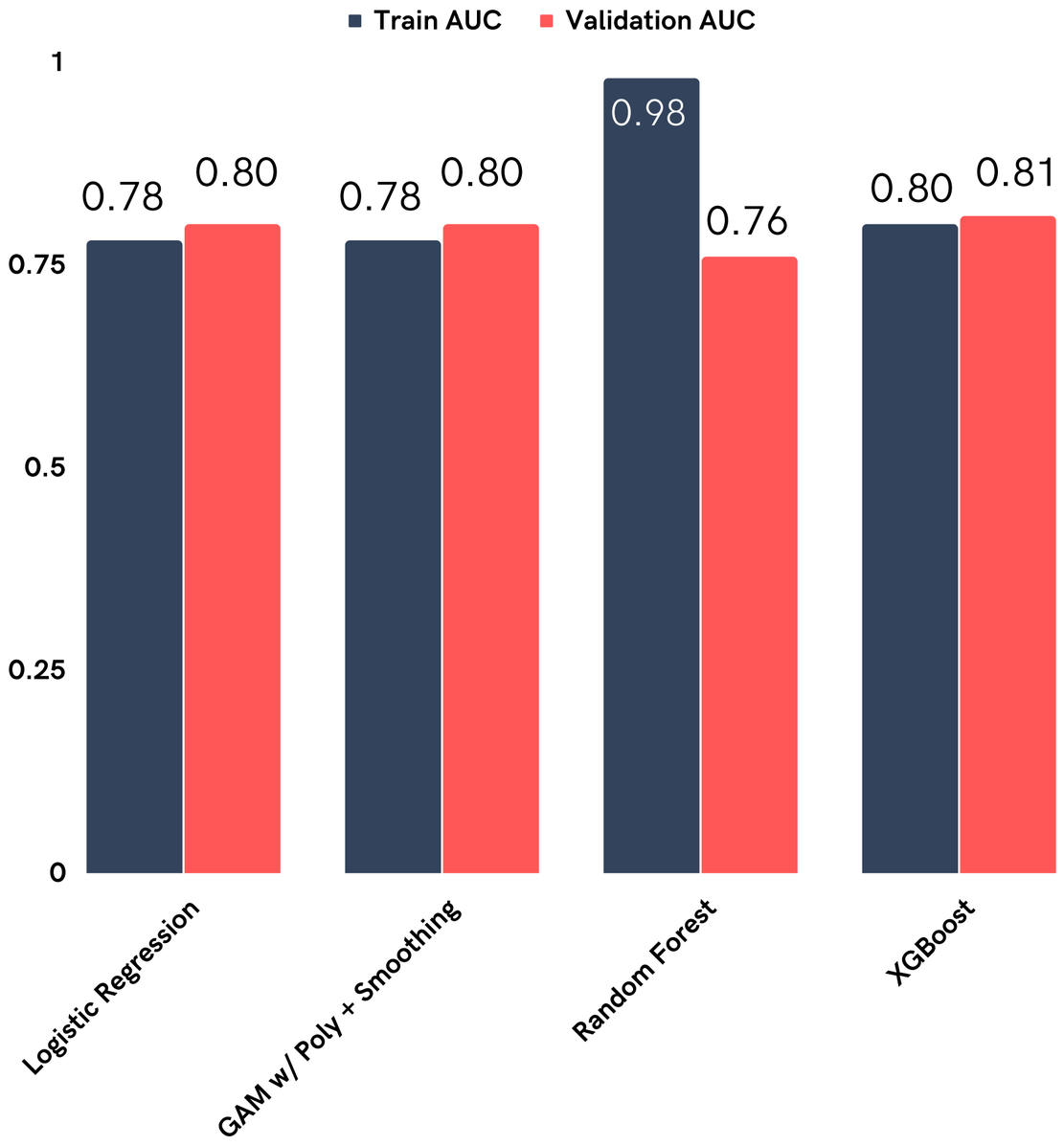
- Stratified Sampling & SMOTE are tried to handle imbalance class issue
- Univariate approach for feature selection

Modelling

- Different Models are developed and compared one another
- Trained using the same training set
- Evaluated using the same validation set

Model Evaluation & Selection

- AUC as scoring criteria
- Validation set as the main set criteria
- Score on the training set is also evaluated to check overfitting potential issues



Result

XGBoost as our best performing Model

- Multiple approaches are tried other than these 4 models
- Most model have somewhat similar performance
- Overfitting issue only occurs on Random Forest
- XGBoost is the best followed by GAM and LR

Result (cont'd)

- A hybrid XGBoost + GAM was chosen as the final Model
- Age, euribor 3 month rate, and job are some of the most important criteria
- Model has a tendency to predict "0" due to class imbalance
- AUC score on 50% unlabeled test set yields a score of 0.797

Conclusion

- A good performance Model has been succesfully developed to overcome the business problem
- Hybrid XGBoost + GAM as the final model
- Bank telemarketers should follow probability-based approach towards contacting customers

Thank you !

Group project - Fajar, Hadi and Jeanne

April 2022

Statistical and Machine Learning

Professor : Minh Phan