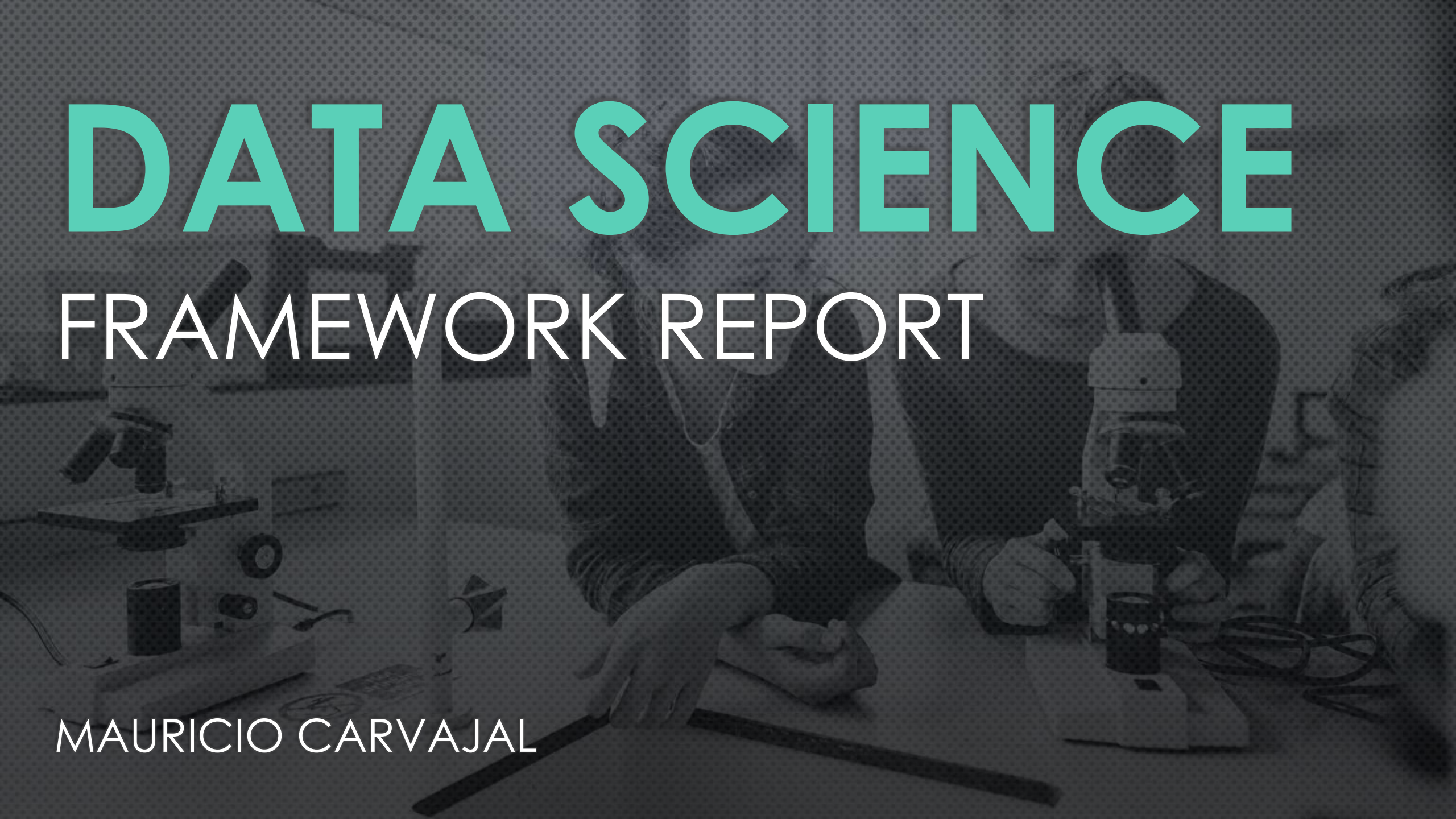


DATA SCIENCE

FRAMEWORK REPORT

MAURICIO CARVAJAL

AGENDA

- Problem statement
- Proposal data science process framework
- Data management
- Risk management
- Insights

PROBLEM STATEMENT

- Over the past year or so Credit One has seen an increase in the number of customers who have defaulted on loans they have secured from various partners, and Credit One, as their credit scoring service, could risk losing business if the problem is not solved right away.

GOALS

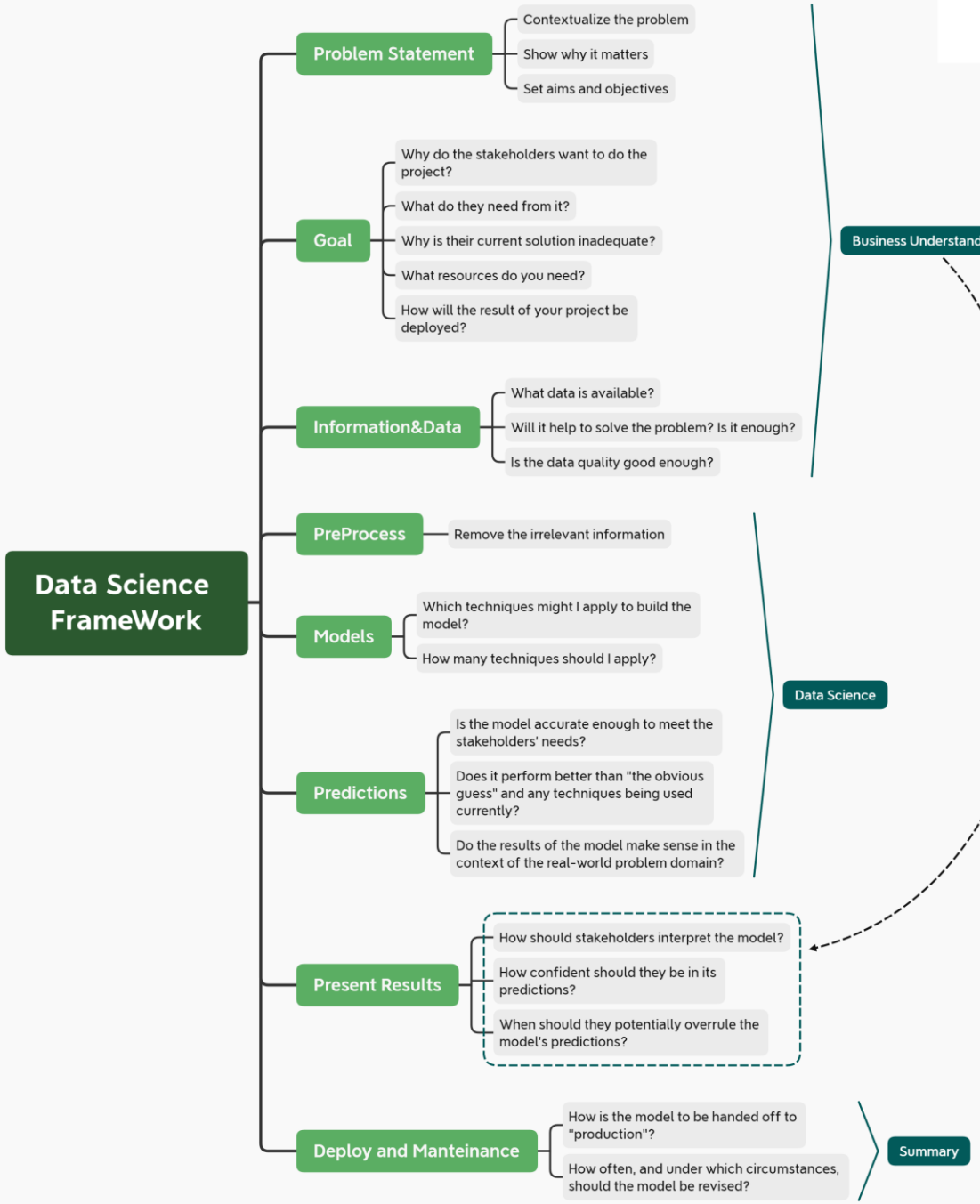
Provide a robust model to predict the default payment on clients

Analyzed and understand what are the current situation of the bank

Base on data, recommendations on how improve the current situation

DATA SCIENCE FRAMEWORK

- BASE ON THE MODEL :
  - ZUMEL AND MOUNT
- DIVIDED IN 3 GROUPS:
  - BUSINESS UNDERSTANDING
  - DATA SCIENCE (DATA MANIPULATION)
  - DEPLOY AND MAINTENANCE



Data Science Framework (Business Understanding)

- Problem Statement:
  - increase in the number of customers who have defaulted on loans that will potential causes serious problem to the bank.
- Goals:
  - Analyzed what features are producing the clients defaulted
  - Create a robust model to predict possible default clients
  - Provide recommendations on how to decrease the client default paument

DATA SCIENCE FRAMEWORK (DATA SCIENCE)

- **DATA UNDERSTANDING:**
  - FIRST STEP IS TO UNDERSTAND THE DATA, FIND IF THERE ANY MISSING VALUE, WHAT REPRESENT THE VALUES, NUMERICAL, CATEGORICAL, THE INFORMATION IS RELEVANT TO THE PROBLEM STATEMENT.
- **PRE-PROCESS**
  - TRANSFORM THE DATA THAT IS NEEDED, IS THE DATA NEED TO BE ESCALATED, OR NEED TO BE REDUCED, MAYBE NEED TO BE BINNED
- **FEATURE ENGINEERING**
  - SPLIT THE DATA SETS, GROUP BY CATEGORY, PRE PROCESS IT, MAYBE APPLY PCA FOR NUMERICAL VALUES I.E AMOUNT OF BILL STATEMENT.

Data Science Framework (Data Science)

- Models:
  - Identify if is regression, or classification problem. I.e in this case is classification
  - Find a group of machine learning algorithms that preform well according with the size of the data set.
- Evaluation
  - Use Accuracy, kappa, confusion matrix to evaluate how well the is the model.
- Predictions
  - Base on the model and the results of the evaluation, make and evaluate the results of the perditions. This process can be iterative too.

DATA SCIENCE FRAMEWORK (PRESENT RESULTS)

- One the model is been proven the results is critical part for the successful of the project.
  - Use the learnings of the business understanding
  - Use a lot visualization rather to go technical
  - Try to use the best practices on "data telling"

INFORMATION OF THE DATA

Amount of the given credit

This is a numerical value, can be binned

Gender

Needs to convert to categorical value

Education

Categorical value, important to see if there any relationship between educational levels.

Marital status

Categorical

Age (year)

This is the age of the customer

INFORMATION OF THE DATA

History of past payment

Categorical value

Amount of bill statement

Numerical value, this can be reduced using PCA

Amount of previous payment

Numerical value, this can be reduced using PCA

client's behavior

This is the dependent value

DATA MANAGEMENT

1- Understanding Data

Need to convert the data, numerical to categorical

2-Proproces the information

Find a method to discard irrelevant information
- Correlation, PCA, Feature variance, recursive feature elimination

3- Find the best model

Use different machine learnings models: RF,SVM,KNN C5.0
Evaluate the best performance
- Kappa, accuracy

4-Use the best model to predict the results define in our problem statement

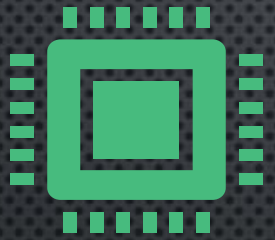
RISK DATA MANAGEMENT

- Change the customer sending habits is not posible
- Need to have clear understanding of the type of the data.
  - Conversion from numerical to categorial as needed
- Understand if the problem is classification or regression problem
- The time that is needed to build the models
- Human behavior is difficult to model

INITIAL INSIGHTS

- History of previous behavior is key to create a model that can predict the possible behavior of the customer.
- The age of the customer can another factor to take into consideration
- Gender, can be risky to analyze this feature due to any discriminations in the findings.
- The amount of given credit can be binned or grouped to have better understanding.
- The age can be grouped as: Youth, youth Adult, Adult, Seniors.

ENVIRONMENTAL CHANGES



Python

For this project the idea is to migrate from r studio to Python.

Python is more widely known

Python have similar library “sklearn” as R Caret”.

Start using git

Starting implemented the best practices of the industry using Git.

Git: Control version system

THANK YOU

LMAURICIOCARVAJAL@GMAIL.COM