Credit One Default Risk Analysis

- A DATA SCIENCE PROCESS FRAMEWORK -

Overview

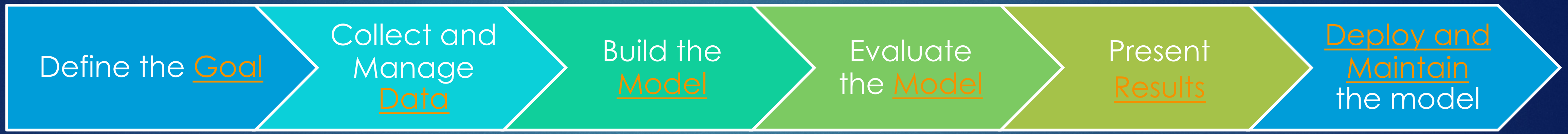
- **Credit One Main facts:**
  - A third-party credit rating authority that provides retail customer credit approval services to Blackwell Electronics and other partners
  - The number of customers who have defaulted on their loans have increased over the past year
  - As a credit scoring service provider, Credit One could risk losing business

- **Desired outcome:** Minimize Credit One's partners risk exposure

- **Questions to Answer**

Which customer attributes might relate to whether or not a customer is likely to default on their current credit obligations?

Data Science Process Framework



Define the Goal

Collect and Manage Data

Build the Model

Evaluate the Model

Present Results

Deploy and Maintain the model

- This framework is aligned to the data science process followed in the previous task
- Potential pitfalls:
  - Goals misaligned to the business
  - Poor quality data
  - Not being able to get good predictions after modeling
  - Not buying in from stakeholders
  - Poor deployment and maintenance

(*) Based on Zumel and Mount, Practical Data Science with R, chapter 1

Goals

- Define a Data Science Process to understand how much credit should CREDIT ONE allow someone to use or, if someone should not be approved
- Identify which customer attributes might relate to whether or not a customer is likely to default on their current credit obligations

Collect and Manage Data

- Data Available: Credit.csv file saved to local computer

  Owner: Credit One. Access limited to Data Science team
- Data shape: 30.000 + observations

  25 attributes. See data dictionary
- Data Types: All attributes are objects. Data types Conversions are necessary. Categorical variables need to be represented as numbers and discretization for some variables are needed too
- Other Preprocessing tasks:

  Will drop these from the data set:

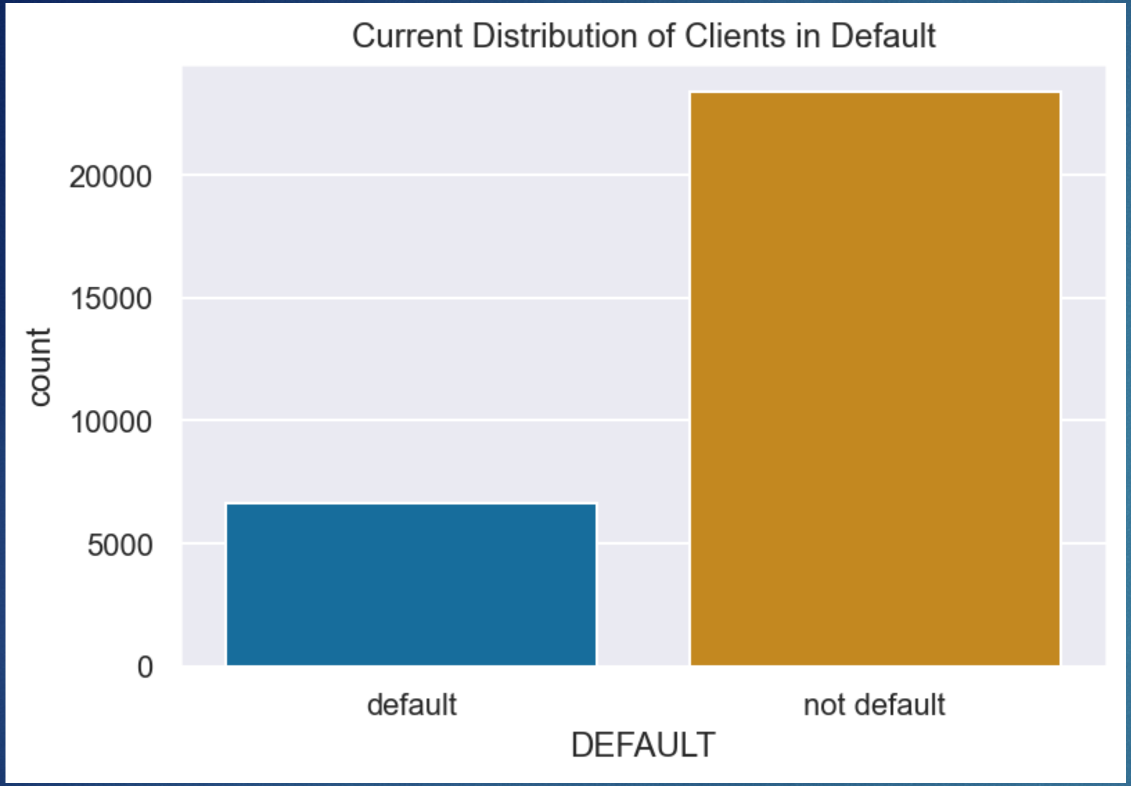
  - First observation corresponds to the attribute description

  - Erase the first observation, once column names have been changed

  - Observation with a Null values (1), and observations with non complaint values (e.g. gender = 'sex')

  Will convert categorical attributes (gender, education and default) to a numerical representation
- Exploratory Data Analysis (EDA): Learn from the data and make relevant visualizations

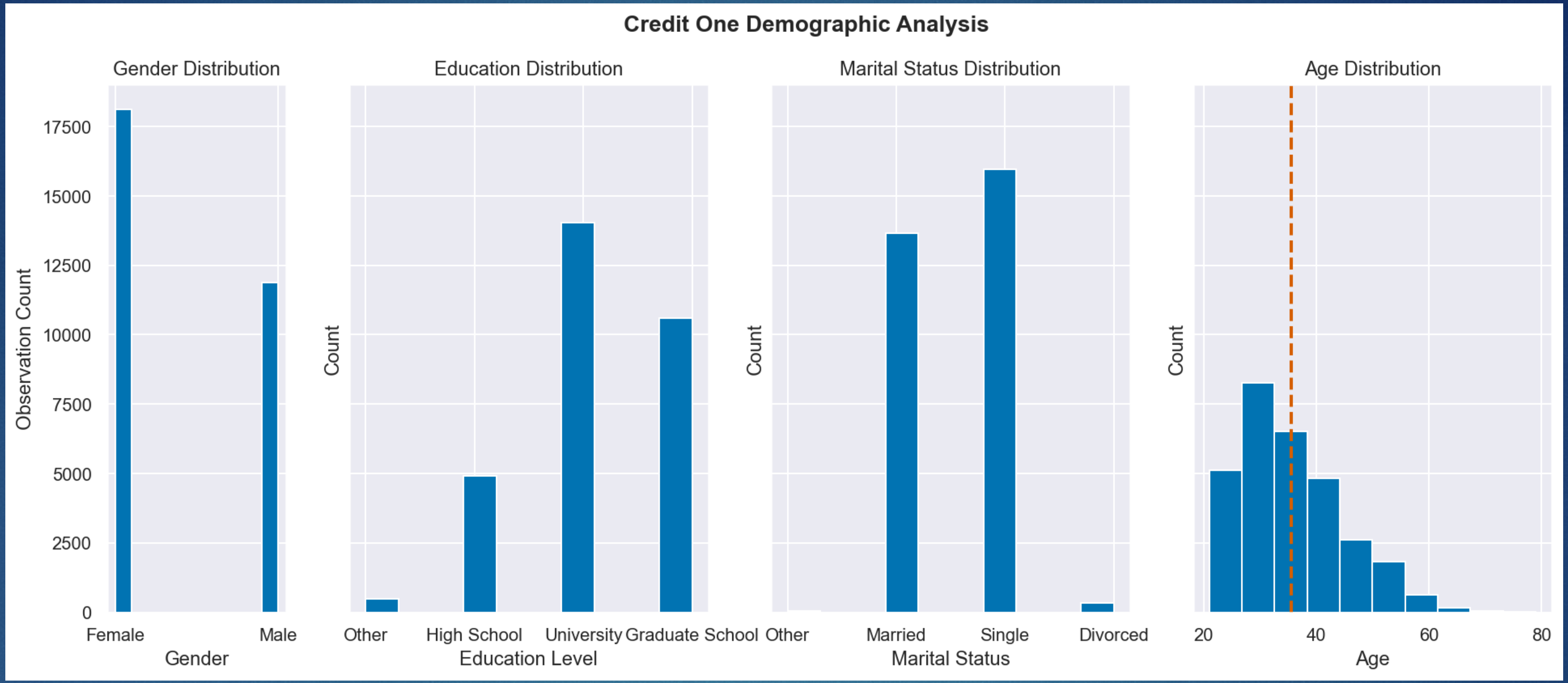
CreditOne Data Summary



- 23364 clients in good standing
- 6636 clients in default (**22%**)
- There is not information about the current balance
- See CreditProfile to see the dataset information

- It appears that the data quality is good enough after cleaning and preprocessing it
- It would have been desirable to know the current balance not just the balance limit

CreditOne Data Summary



- 0% of CreditOne clients are male, 60% are female
- 47% of CreditOne clients have a major degree, 35% a graduate school, 16% finished high school, and 2% have other kind of education
- 53% of clients are single, 46% is married, 1% is divorced and les than 1% falls under other marital status
- CreditOne clients age range is 21-79 yrs. Average age is 35.5 yrs. and 75% of customers are younger than 42 yrs.
- Actions:
- Might drop 'other's from data set depending on their default status
- Need to evaluate the presence of outliers

Build and Evaluate the model

- Target variables to predict:

  Credit Amount (LIMIT_BAL) : Amount of the given credit

  Client Behavior (DEFAULT) : Whether the client is in good standing or not.

- Plan to use Machine Learning Regression methods to predict the targets
- The number of models to implement will depend on the results obtained. (Evaluation) This is an iterative process
- There is no budget constrain, but it is urgent to solve the issue

Build and Evaluate the model

- Target variables to predict:

  Credit Amount (LIMIT_BAL) : Amount of the given credit

  Client Behavior (DEFAULT) : Whether the client is in good standing or not.

- Plan to use Machine Learning Regression methods to predict the targets
- The number of models to implement will depend on the results obtained. (Evaluation) This is an iterative process
- There is no budget constrain, but it is urgent to solve the issue

Present Results and Document

- The the models built and its results will be presented to the stakeholders
  - Accuracy obtained should be > 85% to have a good level of confidence in the predictions
- The report will include recommendations for its deployment and maintenance

Data Dictionary

Attribute	Description
ID	Unique identifier
LIMIT_BAL	Credit amount
DEMOGRAPHIC DATA	
GENDER(1)	0 = female 1 = male
EDUCATION(2)	0 = other 1 = high school 2 = university 3 = graduate school
MARRIAGE	0 = other 1 = married 2 = single 3 = divorced
AGE	Year

(1) Changed order to reflect categorical encoding output
(2) Changed order to reflect educational level

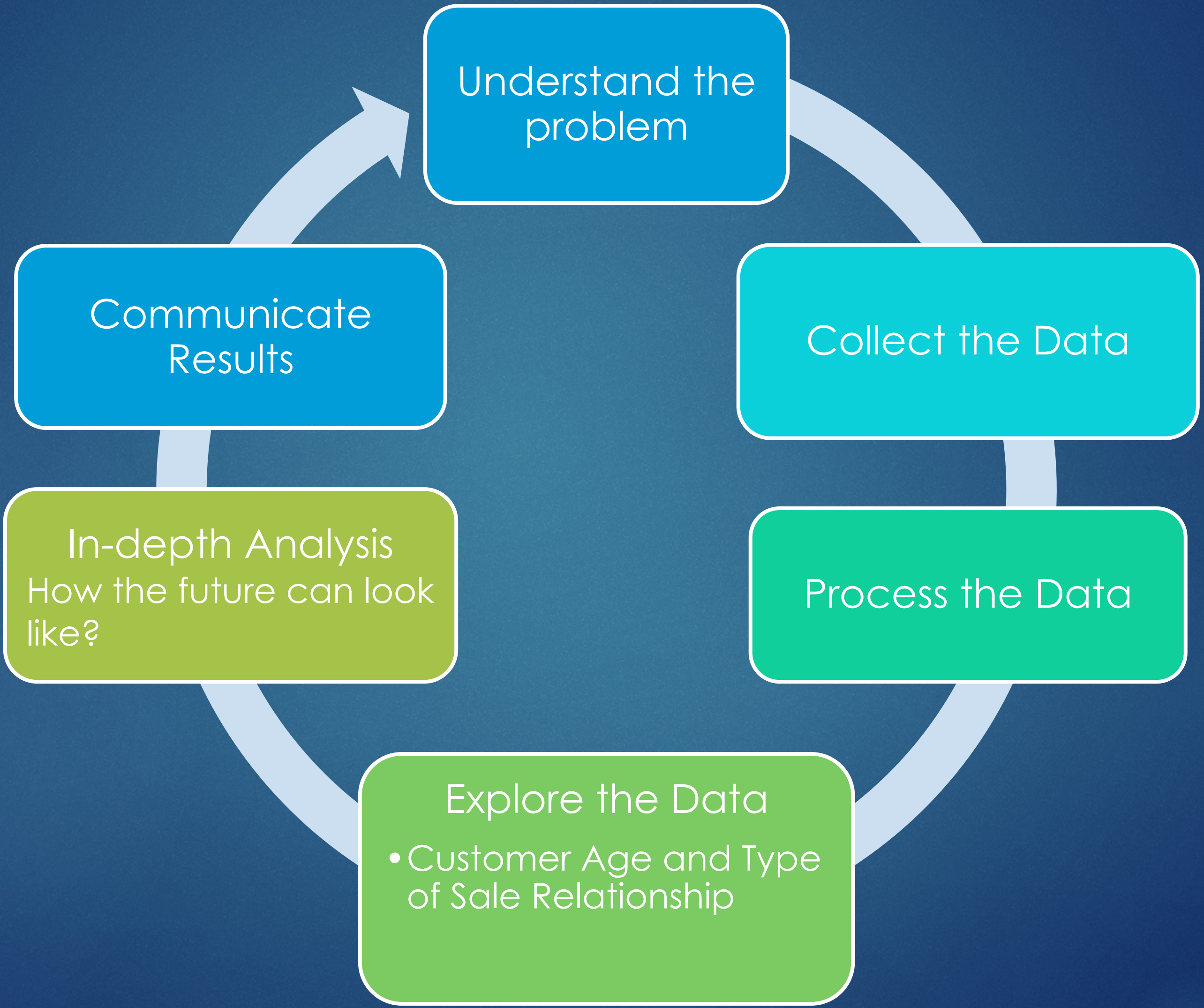
Attribute	Description
PAY_1 – PAY_6	Monthly Repayment Status: PAY_1 = September 2005 … PAY_6 = April 2005 Key: -2 = No consumption -1 = Paid in full 0 = Use of revolving credit 1 = 1 month payment delay 2 = 2 months payment delay … 8 = 8 months payment delay 9 = 9 months payment delay or more

BILL_AMT_1 -
BILL_AMT_6
	Monthly bill statement:
	BILL_AMT1 = September 2005…
	BILL_AMT5 = April 2005

PAY_AMT1 -
PAY_AMT6
	Amount previous payment
	PAY_AMT1 = September 2005…
	PAY_AMT6 = April 2005

DEFAULT	client's behavior 0 = not default, 1 = in default

Data Mining Approach



Understand the problem

Collect the Data

Process the Data

Explore the Data

- Customer Age and Type of Sale Relationship

In-depth Analysis

How the future can look like?

Communicate Results

(*) Based on AJGoldstein.com Data Science Deconstructed