Credit One: Data Science Framework Report

Module 2 Task 1

Isabel Lantero Hernández

1. Goals

- Examine current customer demographics at Credit One.
- Understand whch attributes significantly relate to a customer’s probability of defaulting on its credit obligations.
- Build a predictive model that Customer One can use to anticipate customers likely to default.

2. Framework One

Framework One - Zumel and Mount, Practical Data Science with R, chapter 1:

1. Define the goal:
   - Stakeholders want the model to understand which customers will default.
   - They need the project to know how much to allow someone to use.
   - The resources needed are: a dataset including customer demographic data and a Jupyter Notebook.
   - How will the result be deployed: given a high probability of default, Credit One might chose to lend less or no money, demand a higher interest rate…
2. Collect and manage data
   - Data available is directy from Credit One, which stores all of it in a MySQL database.
   - Use SQL to query to the database table and retrieve the data.
   - Quality of data: 30000 registers without duplicates (big enough), no missing values, 201 duplicates, review for outliers (see distributions).
3. Build the model
   - Extract useful insights (see last slide)
   - Machine learning regression methods will be used
   - Dependent variable: Y: client's beahaviour (default / not default)
   - Divide data: 75% training, 25% testing
   - Use cross validation to optimize the hyperparameters
   - Create different regression models and compare them to chose the best one: Linear regression, random forest regressor, SVR…

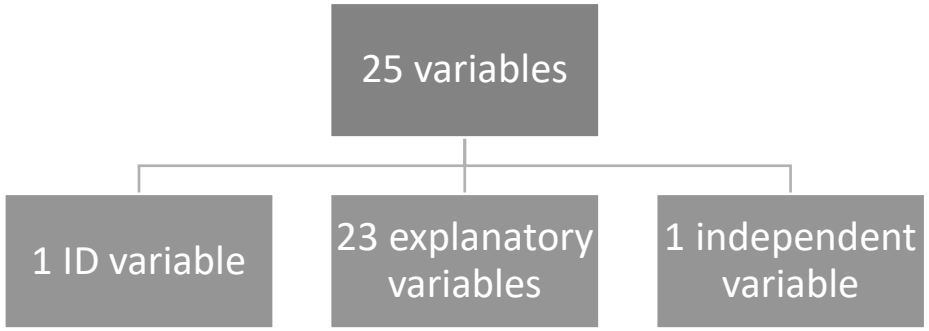
2. Framework One (II)

4. Evaluate and critique the model:
   - Assess if the model is accurate enough to meet stakeholders' needs using the following metrics:
     - R Squared Score
     - Root Mean Square Error (RMSE)
   - Compare the performance with other models and choose thebest one
   - Analyse whether the results of the model make sense in the context of the real-world problem domain
5. Present results and document them
   - Stakeholders should interpret the model taking into account that these are just predictions based on a limited amount of data.
   - The confidence of stakeholders on the predictions should be based on the metrics shown above. The better, the more confidence.
   - However, they should overrule the model's predictions in cases when external factors that are not shown in tha dataset might affect the client's probability of default (e.g: he will son receive a large inheritance, he has defaulted many times with other credit institutions, his business has recently gone bankrupt).
6. Deploy and mantain the model
   - The model should be easly accessed and used once it is handed to "production"
   - Take into account feedback and check for bugs in case anythong has to be fixed or re-enhanced. Initially, the model shoud be in constant revisión as any errors should be fixed as son as possible. However, once it is working correctly, regular revisions should be made in case any tren has changed (eg: older people now have a higher probability of default tan before…)

3. Description and location of related data sources

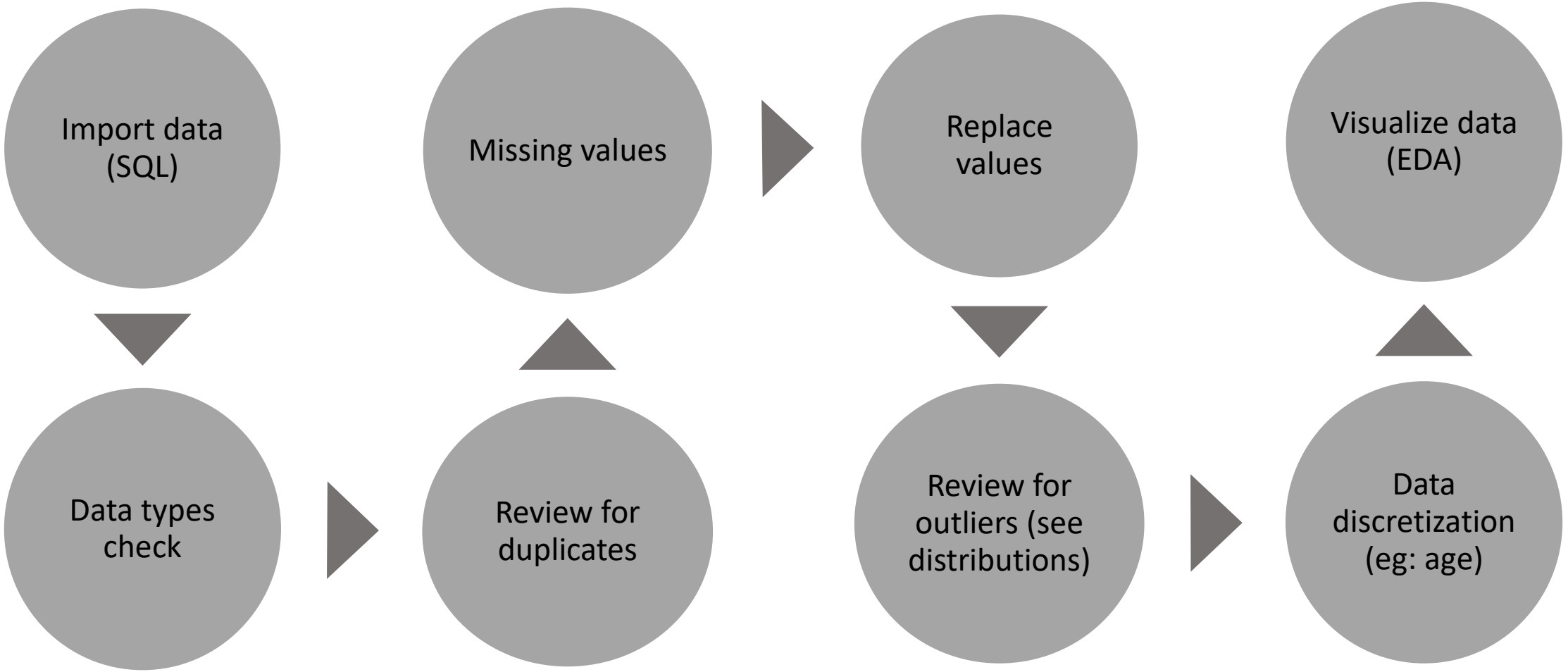
- Data source: Credit One client demographic data
- Location: Credit One stores all of their data in a MySQL database.
- Retrieving the data: Use SQL to query to the database table and retrieve the data

Description of the data:



Name of variable	Description
ID	ID of each client
LIMIT_BAL	Amount of given credit
SEX	Gender
EDUCATION	Level of education
MARRIAGE	Marrital status
AGE	Age in years
PAY_0 – PAY_6	Monthlt repayment status (from April to September, 2005)
BILL_AMT1 – BILL_AMT6	Amount of bill statement (from April to September, 2005)
PAY_AMT1 – PAY_AMT6	Amount of previous payment (from April to September, 2005)
Default payment next month	Client's behaviour (default / not default)

4. How will the data be managed



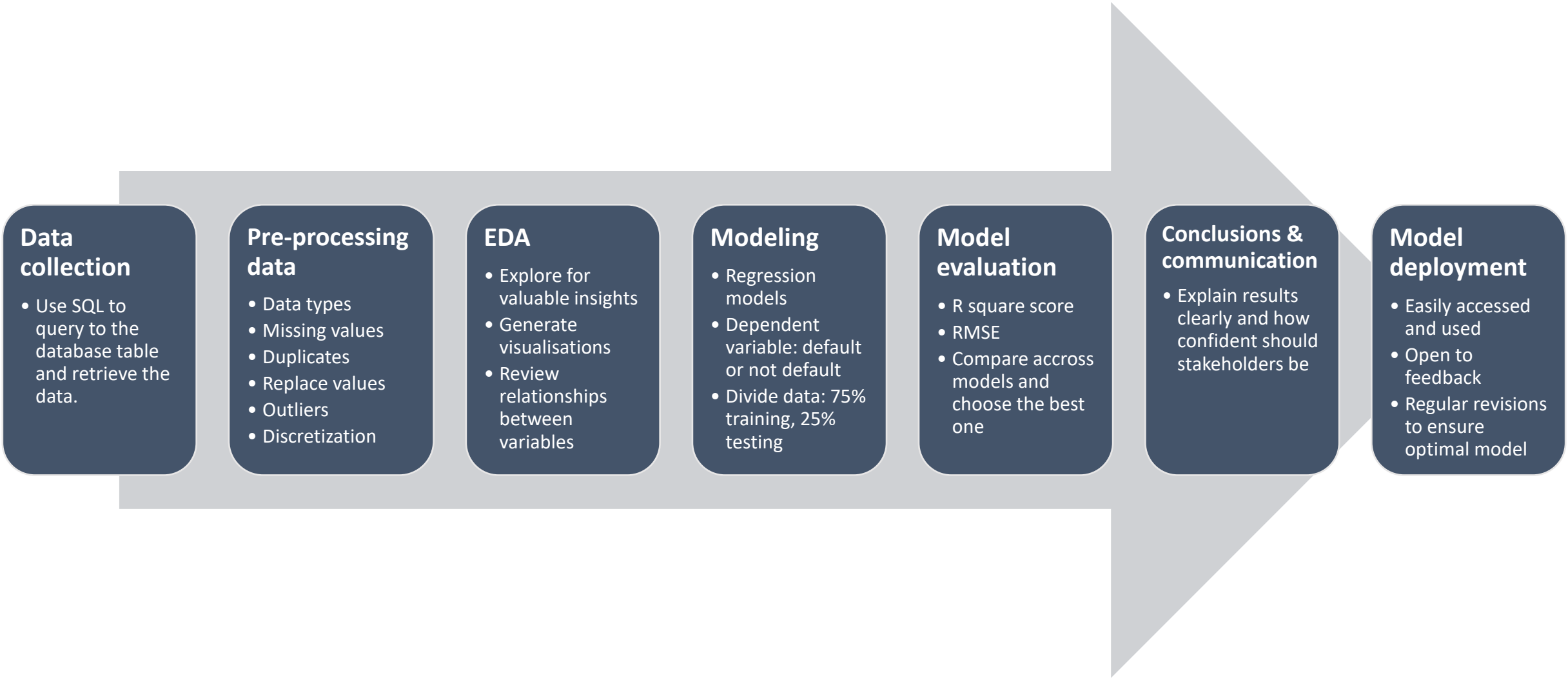
5. Issues with the data and how to address them

- Duplicates: In total, there were 201 duplicates. We know that this is not a coincidence of two different clients, because even the ID is the same. Therefore it has been decided to remove these duplicates from the database. Leaving 30,000 records left.

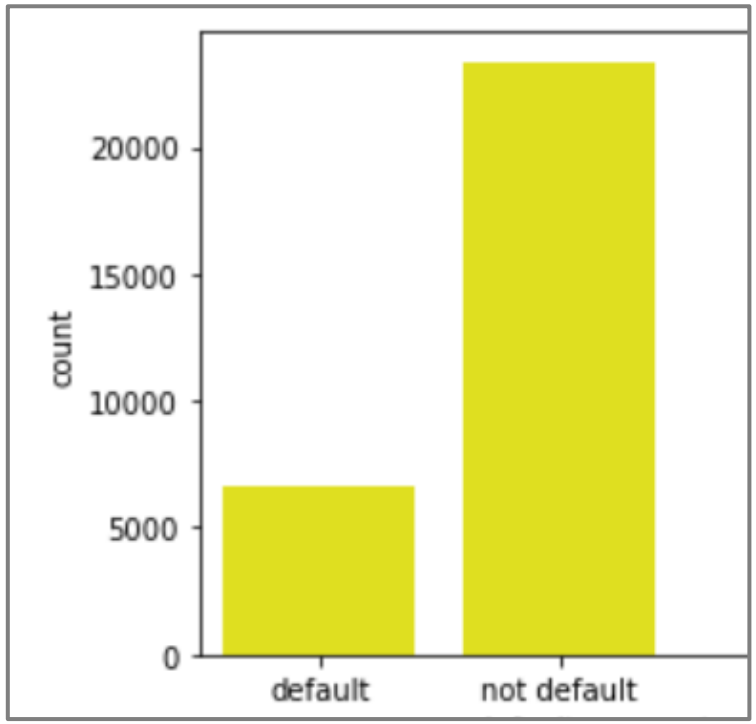
- Marriage: As the number of “0” (=Other) and “3” (=Divorce) were very small compared to the other groups (married and single), it was suggested to join these two smaller groups into the same one. However, only 26% of divorced clients default, proving that this insight is very valuable and important information would be lost if the two groups were joined.

- Sex: The variable of this variable should be changed from categorical to boolean as there are only two options

6. Flowchart of the process



7. Initial insights



There are about 4 times more “not defaults” than “defaults”, whch is still a bad result for Credit One and proves that this data science Project is necessary.

% of clients that have defaulted by category:

Gender:

- Female: 20,77%
- Male: 24,16%

Education:

- Graduate: 19,23%
- High school: 25,15%
- University: 23,73%
- Other: 7,05%

Marriage:

- Married: 23,47%
- Single: 20,92%
- Divorce: 26,01%
- Others: 9,25%

As can be seen on the left, male clients have defaulte 4% more tan female clients have.

Moreover, clients in the education level “graduate” and , especially, “other”, tend to default less tan the other two categories. High school students are the ones that default more.

Regarding marrital status, single clients and the ones in the category “other” are the ones that have defaulted less times. On the other hand, divorced people tend to default more (26%)

Additionally, to gain further insights into which categories have defaulted more times in the past, they should be combined with other variables. This Will be studied in the EDA stage when further investigations are made. For example: it might be interesting to study whether female graduate clients default less or more than male graduate clients do.