Ubiqum – Task 2

Data science with Python

Credit risk prediction

Marco Isaac Marín Granados

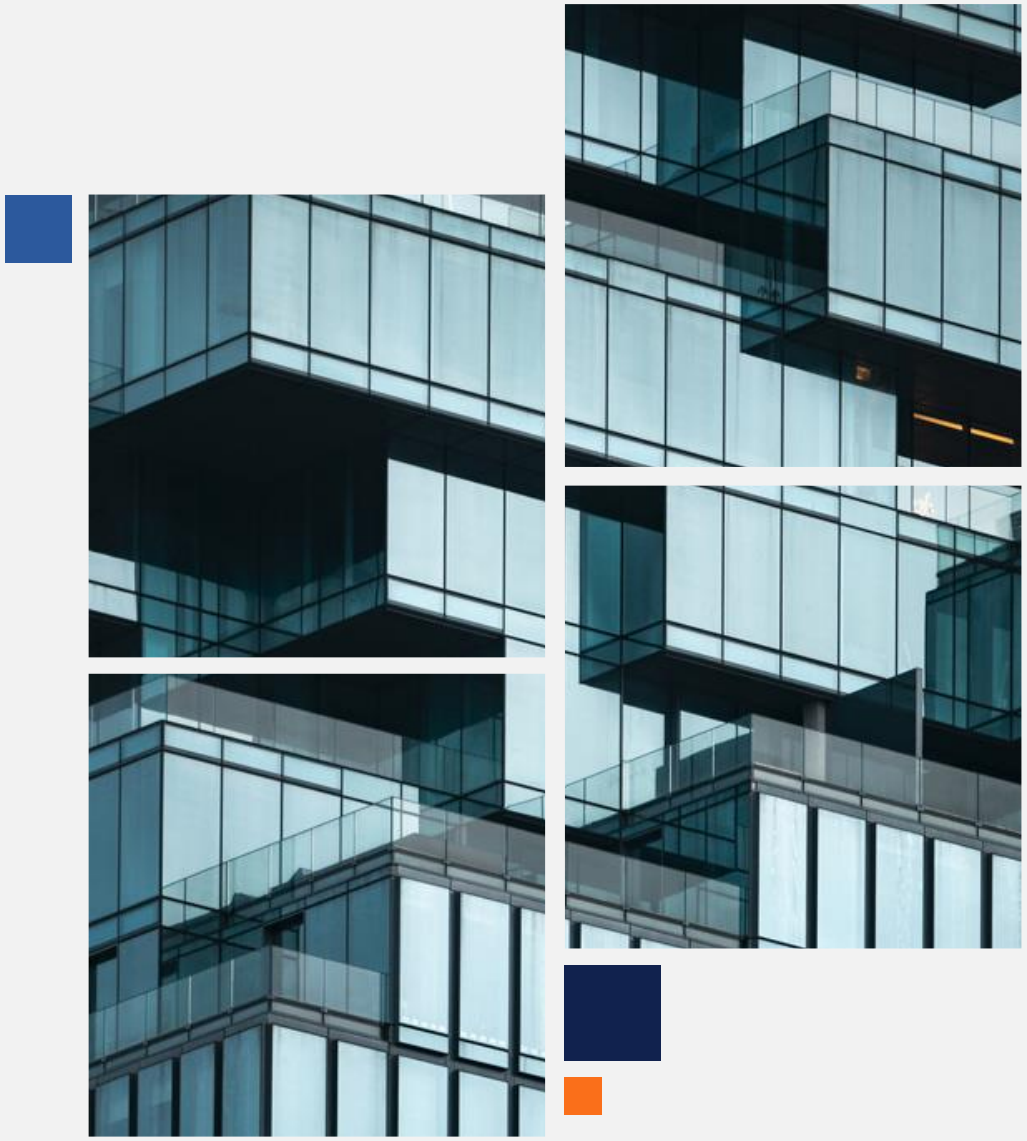
Cohort 204

Context

Capital One is a credit rating authority that provides retail customer credit approval services to Blackwell.

As a Blackwell partner, it is important that the services they provide are efficient and precise.

During the last year it was notorious an increase in the approved credits that ended up in default status, meaning losses for Blackwell.



Data Science Framework

Zumel and Mount

From *Practical Data Science with R, chapter 1*

1. Define the goal
2. Collect and manage data
3. Build the model
4. Evaluate and critique the model
5. Present results and document
6. Deploy and maintain the model

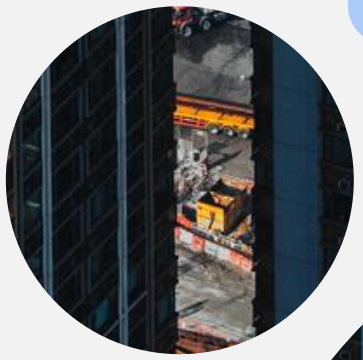
Justification

The selected framework was choosen over the BADIR one, because it provides more freedom for the team regarding the explorartion to do. BADIR seems as a good choice when there are clear hypotheis about what is expected to find from the data.

In this case, the goals are the only ones specified precissely, that's why the framework proposed by Zumel and Mount seems to fit better.

Goals

- Detect variables that can correlate with the risk of a credit to become a default.
- Develop and evaluate models capable to **classify** if a credit will become default.
- Implement and rank **regression** models for predicting the appropiate amount of credit limit.



Data source

- Credit One MySQL database with about 30K records from customers, including data from their credit, their payments log, demographics and the status regarding the default status of their credit.
- There are 23 independent variables, and a default status dependent variable.

Demographics

- Gender
- Education
- Marital status
- Age

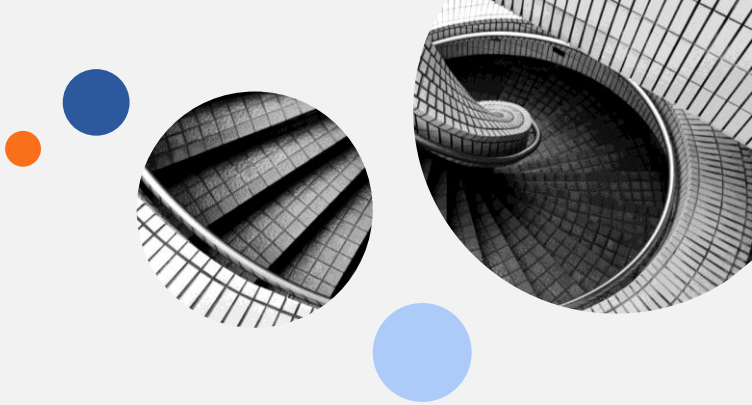
Financial

- Credit amount
- Payments log for a year
- Bill statement log for a year

Dependent variable

- Default behavior

Data management

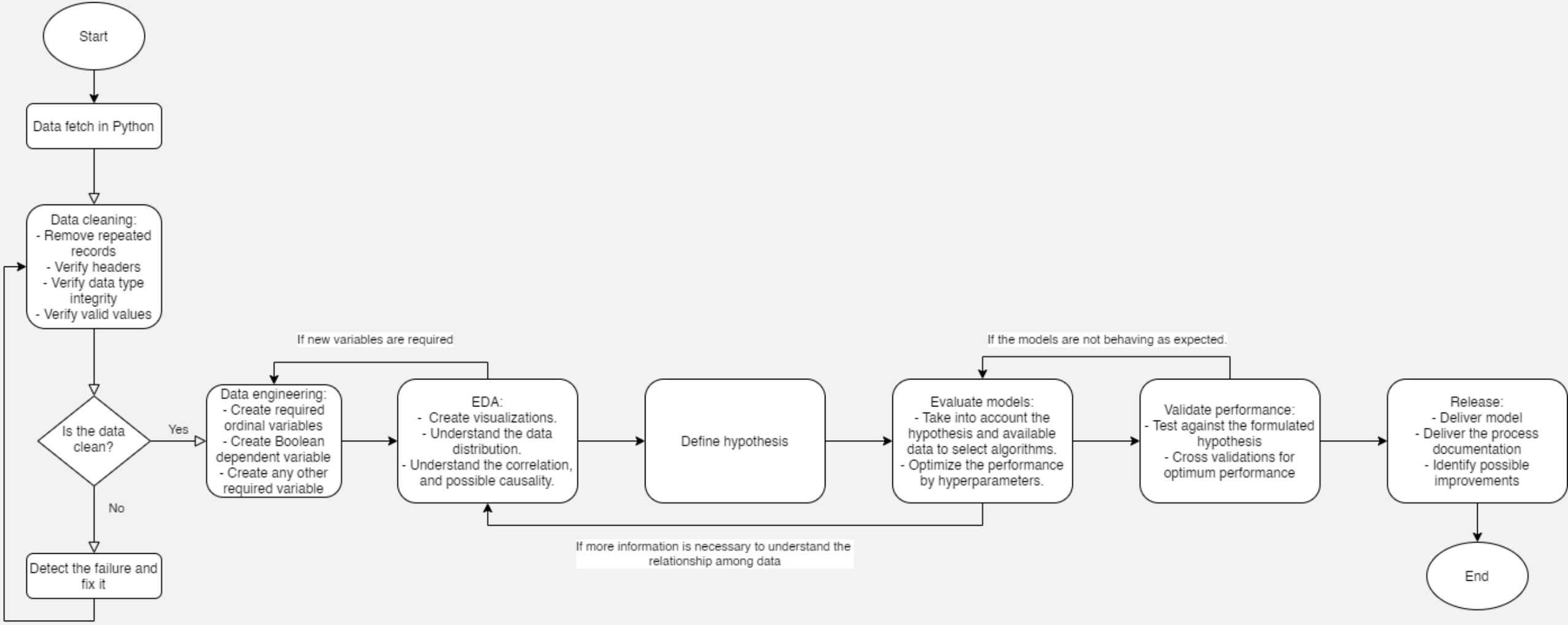


- The provided data has been previously anonymized, in order to protect the privacy of Credit One customers.
- The data will be queried by a MySQL query from the Credit One server, by using Python and some additional libraries.
- The data will be managed in Python by using a Pandas data frame.
- A cleaning process will be done to guarantee the data quality.
- An exploratory data analysis will be executed over the data.
- Key variables will be identified.
- A set of models will be proposed, explored, evaluated and optimized.

Data known issues

- More verbose header at first row, anyway they should be fixed too (PAY_0). **Replace headers.**
- Repeated records from rows 202 to 404. **Drop them.**
- The *ID* column is not consistent, it will be **dropped**.
- The values for most of the categorical variables are not numerical, as indicated in the data description document.
- The *payment* history variable contains mixed information, a **new variable** might be calculated out of it.

Workflow


Start
Data fetch in Python
Data cleaning:
- Remove repeated records
- Verify headers
- Verify data type integrity
- Verify valid values
Is the data clean?
Yes
No
Detect the failure and fix it
If new variables are required
Data engineering:
- Create required ordinal variables
- Create Boolean dependent variable
- Create any other required variable
EDA:
- Create visualizations.
- Understand the data distribution.
- Understand the correlation, and possible causality.
Define hypothesis
If more information is necessary to understand the relationship among data
If the models are not behaving as expected.
Evaluate models:
- Take into account the hypothesis and available data to select algorithms.
- Optimize the performance by hyperparameters.
Validate performance:
- Test against the formulated hypothesis
- Cross validations for optimum performance
Release:
- Deliver model
- Deliver the process documentation
- Identify possible improvements
End

Data considerations

- The repayment status variables (X6 – X11) refer to the relationship of the current payment against the previous bill. This means that if in July it has a ‘Paid in full’, it means that in July the payment done was for the whole bill of ‘June’. This can be verified by looking to the payment amounts and bill amounts.

- Are unclear the parameters used to determine the *default* status of an account, probably that process should be clarified, or additional information should be provided.

- What about the negative values in the bill amounts? There are no negative values in payment.

The Task 2 can be followed up in:

https://github.com/marinisaac1/creditRisk

Marco Marín

marinisaac1@gmail.com