Acredius AG Hardturmstrasse 161 I 8005 Zürich I Switzerland I **CONFIDENTIAL**

Data Scientist

Case Study – 2021/2022

Contents

Case study description

Part 1: Descriptive analysis

Part 2: Modeling

Part 3: Alternative data

Case study description

In this case study, we will be **developing a credit risk model from scratch** for the Acredius Crowdlending pillar.

As a marketplace, businesses come to Acredius to ask for loans. These are registered companies. They apply for a loan online. The **loan application is a journey** where the business precises the amount and duration of the wished loan, uploads all the documents (balance sheets, marketing documents, etc), connects its social media profiles if possible, and answers a couple of questions. Once the process is done, it receives an instant offer including the interest rate and monthly repayments. Acredius uses traditional and non-traditional data (directly collected from the applicant and/or through different APIs) to provide accurate pricing.

Some portion of the applications is rejected. The ones accepted are classified into **different risk classes**.

Part 1: Descriptive analysis

The first part of the case is understanding and describing the data sample.

1. As the data is not clean, the first step will be to pre-process it. Which preprocessing techniques should be used (for example, how to deal with missing values, etc)?
2. Using Python or R, write the code to clean the data.
3. Explore the data using different statistical techniques. What are the top 3 main highlights?

Part 2: Modeling

Build a model that predicts the interest rate of the loan, based on a selected set of variables. Please describe the step-by-step approach you used.

Expected output:

A recommendation on the best model to use.

Part 3: Alternative data

The aim of this part is to collect alternative data (like social media profiles, industry specific information, etc.) in order to improve the model you created in part 2.

1. How can we improve the previous predictive model? Is getting alternative data, a solution for that?
2. Could you explain what is the meaning of generative data? Is it useful in our case? Why?
3. What are the data sources we can add to our dataset to have better a predictive model? Can you list some sources & techniques?
4. Could you adjust the model using the new data? What are the results? What are your conclusions?

Thank you!

acredius