Data Scientist

Case Study – 2022

Contents

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 only Swiss-based and 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: Statistical study

Tasks and questions:

The first part of the test is around a statistic modeling of the data. The requested tasks are mainly building models to describe the data. The requested task are:

1. The data is not clean, the first step will be the preprocessing of the data. Which preprocessing techniques should be used?
2. Using Python or R, write the code to clean the data.
3. Build statistical models around the data. What are the main conclusions you can see.

Part 2: Modeling

Work on the attached transaction data sample and build the credit risk model

Task:

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

Notes:

The data sample is not “clean”

Part 3: Alternative data

The aim of this part is to collect alternative data in order to have better model.

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

Thank you!

acredius