

Machine Learning CoE: Case Study

Assignment:

The assignment is to investigate and analyze the attached data, producing a model that can be used to predict whether a loan will be paid in full or charged off (this occurs when the loan is unpaid or delinquent for a period of time so the financial institution determines that the debt is unlikely to be collected). You will use the `loan_status` column in the dataset for the outcome, and you can use the remaining variables in the dataset to predict the loan status (although beware of variables that may contain data leakage).

We prefer that you complete this case study in R or Python, but if you are not comfortable in either of those then please feel free to use whichever tool/language you have the most fluency in.

Deliverables:

1. The code you used to perform the work (please use comments to describe your choices/steps).
2. A presentation that describes the results in 2-3 slides for a business audience. Assume the audience has minimal technical knowledge.

Expectations:

We are looking to understand your familiarity with data preparation and model building. We would also like to see your comfort level with creating business value from analytics and model results.

We are not expecting you to work on this task for more than 8 hours. If you find that you are spending too much time on the data preparation/model building section, please feel free to explain in comments any additional steps that you would like to take if this project were assigned to you in the role of a Data Scientist at BMO. For the presentation, feel free to make any assumptions that will help you put together a business case.

Please do not get too caught up in model accuracy, this exercise is not a test of building the best the model. Our goal is to understand how you structure the problem and how you connect data to business results.

If you have any questions while working on the case, please reach out to Jess Hempel at 