DS Intermediate Level Capstone Project

DS-XL

Jan. 2023

—

Data

The dataset **data_train.csv** will be provided and used for the entire project. Data is collected Mar 2019 for millions of bank loans with different attributes and loan statuses. This is the dataset you should use for exploratory analysis and model building.

A data dictionary **data_dictionary.xlsx** is provided to you with the definition of the variables to help you better understand the data.

One week prior to the project submission deadline (end of week 3), we will release the test dataset for model performance scoring **data_test.csv**. It will follow the exact same structure and have the exact same variable definition as the training set. Please process this data as needed to score your model to generate final predictions for submission (more details in tasks below). You should not use **data_test.csv** for any other purpose.

Training data will be sent out via OneCareer wechat group, please reach out if you have trouble accessing the datasets.

Tasks:

1. Data Exploration and Insight Finding

Perform data analysis on the given dataset, find interesting questions or insights from the dataset, and write up a data exploration report of a few pages to summarize your findings.

The target audience of this report will be your manager or business customer who would expect to understand the dataset by reading your report.

- ○ Sample structure (*you don't need to follow this, feel free to write it however you like, but make sure it's a professional style report, not a Jupyter notebook with code snippets*)

        - Overall population exploration
        - Interesting signals you observed, questions you'd like to ask
        - Data analysis to answer your questions or support your insights
        - Summary of your work and recommendations based on your findings
    - Feel free to use any open-source packages and code, but make sure to reference any external code you borrowed from others

Submission:

- A **brief (up to 2 pages) professional style report** (.pdf or .doc file, feel free to use markdown or LaTex if you'd like to) with your write-up summarizing findings, plus data visualization supporting your findings if any.
- A **.py file or .ipynb Jupyter notebook with all the code** you used to generate results and visualization you included in your report

Grading rubrics:

- Insights/questions discussed, analysis scope and depth
- Critical thinking and analytical skills
- Data analysis and visualization
- (Bonus point) Include markdown in your Jupyter notebook to make it a report

2. Modeling

Use the given dataset to build machine learning model(s) to predict *loan status* as **target**: if a loan will be charged-off, or stay current/are paid off. Don't worry if you are not familiar with the concept yet, we will give more explanation when assigning the project as well as hold office hours later.

You can train your model on the training dataset with a target, and for final submission,, you need to use the model to generate predictions on the test dataset. We will assess your predictions' AUROC on the actual values of the target variable in the test dataset.

- Feel free to use any open-source package or algorithm that you think suits the problem, you are also encouraged to write your own algorithms and methods that help to improve your model performance. Make sure to reference any external code you borrowed from others
- A couple of things we highly encourage you to try:
    - Refine your own training dataset: using only a portion of it, downsample, upsample, etc.
    - Feature engineering: dropping columns, creating new variables based on your understanding of the data, dim reduction, etc.
    - Bring in other data sources: e.g. macro economics of different years

    - Understand your model: what variables are driving your predictions, how do you explain the relationship between the target variable and the input features?
    - Thresholding: are you using or proposing any threshold for the binary target? How do you find the threshold?
    - Think about how you can improve your model and surprise us!

Submission:

- A .csv file named **predictions_[your_name_here].csv** with only two columns: [index, prediction_score]. The .csv file should have the same number of rows in **data_test.csv** and only two columns mentioned above. This should be your model prediction for all the records in the test dataset.
- A ***short*** **(up to 2 pages) technical report / model design document** explaining your model design (build sample, features used, the algorithm(s) used, validation process, etc.) and model performance analysis if you choose to perform any. You are also free to discuss how to interpret your model and how to use your model here. This should be a .pdf or .doc file, feel free to use markdown or LaTex if you'd like to.
- A **.py or.ipynb Jupyter notebook with all the code** you used to build and score your models and generate any results and visualization you included in your report

Grading rubrics:

- Model accuracy: AUROC score of submitted predictions for test dataset
- Model design and feature engineering soundness & innovation
- Model interpretation and usage proposal
- The overall flow and professional style of the report and slides
- (Bonus point) Include markdown in your Jupyter notebook to make it a report

3. Presentation

We encourage everyone to take this great opportunity to practice project presentation skills, and treat this presentation as a mock interview or real world presentation of project.

You will have up to 10 minutes to present your slides, and we will force a hard stop after 10 minutes. We will be your interviewer and give live feedback from a professional or interviewer standpoint. At the end of the class, we will do a quick discussion for projected evaluation like how we do interview/performance calibration in the real world to give you a taste of how we assess interviews.

Submission:

- A **presentation deck/slides** ***(up to 10min's presentation)*** with your findings and results ***for both sections***. Treat it as if you are sharing and presenting the results to coworkers or leadership team at a company. It could be in any format (ptt, pdf, keynote, or google slides). We will only grade on the overall flow and findings of the deck, so you don't need to spend a lot of time beautifying your slides.

Grading rubrics:

- Contents : complete, concise, covers entire project, generally follows a good logic and flow
- Quality of analysis + modeling, how well it addresses the problem
- Presentation style

Deadline

All required files should be submitted **4 weeks after initial training data is released**. Projects submitted within 4 weeks will receive the final model performance score, as well as feedback on reports :)

You can request extension on this project *in advance* if needed (please do not ask for extension on due date). Submissions without extension request will not receive any grading or feedback.

Important dates:

- Tentative training data release date: **Jan.29th**
- Test data release date: **Feb.19th**
- Submission deadline: **Feb. 26th**

Submission method

OneDrive

Please **create a folder with your Name under `project_submission` folder** in the submission OneDrive link, and upload all required files in your folder.

OneDrive Link : you will receive the link in OneCareer wechat group.

If you are not sure about the naming convention or what files to submit, you can find an example in the github folder `example_submission_Emma`.

Github (optional)

You can also submit your report and code to our class repo on GitHub for bonus points. Please create a PR for your submission and we can merge into the submission.

Office hours

We will hold office hours to answer your questions about data and the project requirement. It will be held during our Sunday live classes. However, we will not answer any questions regarding specific analysis or modeling techniques, nor discuss any results :)