MoneyLion Machine Learning Engineer - Take Home Assessment

Build an automated machine learning pipeline to predict the risk of loan applications

Introduction

As a machine learning engineer in MoneyLion, you are responsible to architect and build machine learning pipelines. Machine learning pipelines are important because they can free up data scientists from maintaining existing models manually. Automated pipelines are also useful for enforcing machine learning governance as all newly created models are forced to adhere to required standards and best practices.

Assignment

In MoneyLion, we have a LightGBM machine learning model in production that predicts the risk of loan applications. Our machine learning model has been in production for the last 3 months and our monitoring systems have alerted us on possible feature drifts. Data scientists have been manually writing scripts to preprocess their training data and tune the model to improve the model each month. As it seems clear that this model requires retraining and fine-tuning monthly, we want to build an automated machine learning pipeline so that this loan model can get continuous updates and fine-tuning.

For this assignment, we want you to build this machine learning pipeline. This challenge is intentionally meant to be open-ended.
An input data file (loan.csv) has been provided to help you kickstart building this machine learning pipeline. This pipeline should be able to run independently when the assessment is submitted and should only require the same input file to run from end to end.

To help you get started, we break down this assignment into the following three parts:

Part 1 - Build a simple LightGBM machine learning model to predict risk

In order to help simulate the pipeline, you should start by building a simple example of a machine learning model to predict risk of loan applications. You should build this model using Python.

- The provided input data file (loan.csv) can be used to create this model and subsequently the pipeline.
- To keep it simple, you can use the `loanstatus` column as the target variable and build an equivalent LightGBM model with the dataset provided.
- However, we still expect your thought process and justification of data selection and feature engineering from the provided input data. Perform EDA or visualisations on the dataset if this will help you explain your thought process.

Part 2 - Plan and write documentation on how you would build this machine learning pipeline

Plan out how you would design and build this pipeline and write documentation on this plan.

- You should come up with a diagram and plan on how this pipeline should be built.
- Write up and discuss all possible considerations performed when coming up with this plan and for each step.
- You may refer online for technical information but **DO NOT** directly use any help from other people, sources, online forums, etc. Your submission should be solely your ideas and work.
- This documentation should be easily understood by other data scientists/machine learning engineers, with no more than 2 pages of content.

Part 3 - Build and code out the machine learning pipeline using the LightGBM model built in Part 1

With the machine learning model built in Part 1 and the plan from Part 2, build the machine learning pipeline using Python.

- You should use Python to build this machine learning pipeline.
    - An IPython notebook might be the best way to communicate your thoughts to follow along with the code.
    - The notebook can import modules and functions from other Python files.
    - You can use and modify the data provided to simulate different scenarios if necessary
    - Remember to put in some thought on how to structure your files and folders while assuming this pipeline will be placed in production!
    - You may also turn this into a tool with a UI for a demo during the interview.
- Describe in detail every step you would consider to build this pipeline while you are coding to build it.
- Your pipeline should be able to run from end-to-end from data ingestion to model deployment using the input data provided.

Tips

The Machine Learning Engineer position at Moneylion is extremely competitive and we receive many applications, so consider how you could **make yourself stand out**. Here are some skills that we're looking for:

- Detailed plans on how the design of this machine learning pipeline should look like and what considerations you would take
- Data assessment and understanding
- Clean coding style and reproducible code
- Clear communication of your thought process
- Back your considerations and assumptions with evidence

There's no expectation of the amount of time you could or should spend on the challenge. That said, do share how much time you spent on it.

Deliverables

A zip file with contents grouped into the following sub-directories (you may omit empty directories):

- Data (only if you add any new ones; **DO NOT** send back the original assessment data!)
- Notebooks (we appreciate it if you **include an HTML file of your notebook** as well as the raw file). Any code in the notebooks should be able to be reproduced and used directly.
- Other Python files and folders (if you use any custom Python files)
- Documentation about the application and pipeline. Do include workflow/architecture diagrams if applicable.