# Machine Learning in Production
# Midterm 1, Fall 2024

Christian Kaestner and Sherry Wu

**Name:** _____

**Andrew ID:** _____

Instructions:

- Not including this cover sheet and the scenario, your exam should have **9** pages. Make sure you are not missing any pages. *You may detach the last page and recycle it after the exam.*
- All questions in this midterm refer to the scenario on Page 9. Answers are graded in the context of the scenario; **generic answers that do not relate to the scenario will not receive full credit.**
- The exam has a maximum score of **55** points. The point value of each problem is indicated. We designed the exam anticipating approximately one minute per point.
- **Please write legibly**. We are unlikely to be able to grade your solution if we can't read it.
- We give an amount of space commensurate with what we expect you to need for each question. We use horizontal lines to suggest where to not use the full page. You may exceed those limits if it is clear where to find the rest of your answer. However, we strongly recommend writing concise, careful answers; short and specific is much better than long, vague, or rambling. However, **do NOT write anything you want us to grade on the back of pages.** We will scan the exam and will not look at the back sides.
- This is a **closed book exam**; no books or electronics allowed. You may refer to 6 sheets of notes (handwritten or typed, both sides).

# Question 1: Goals and Telemetry [12 points]

*All questions in this exam relate to the scenario on the last page.* You may detach the last page if you like. Your first task is to explore and document goals for the FoodFlux project supporting the nonprofit organization and identify how you can measure success.

**(a)** [3 points] State an *organizational goal* for the nonprofit 412 Food Rescue that you are trying to support. (no measure required)

**(b)** [3 points] State a user goal from the perspective of a *412 Food Rescue volunteer* receiving messages from this system. (no measure required)

**(c)** [6 points] You plan to evaluate how the **news analysis model** does *in production*. In particular, you would like to see how often the model predicts a large donation event that does not happen. Design a measure and suggest what data to collect and how to operationalize the measure with telemetry. The measure can be an approximation, but must be plausible within the realism of the scenario.

**Measure**:

**Data to collect (what and how)**:

**Operationalization**:

## Question 2: Model and Data Quality [14 points]

**(a)** [4 points] In the **forecasting model** (see scenario on the last page) initial results when randomly splitting training and testing from past data were surprisingly good until you realized that your training and test data violated the i.i.d. assumption ("independent and identically distributed"). What was the common problem violating the i.i.d. assumption with regard to *time series data* that has gone wrong here?

**(b)** [6 points] You consider whether the evaluation of the **news analysis model** can be improved by combining the idea of *slicing* with *synthetic data generation.*

(1) Provide a concrete example of something you could test with this approach, demonstrating an understanding of the concepts of slicing and synthetic data generation.

(2) Briefly argue whether you think this approach would be useful in this scenario.

**(c)** [4 points] Provide a plausible concrete example of *concept drift* in the scenario (i.e., changes in decision boundaries, not in data distributions) that may degrade the accuracy of your **news analysis model** in production over time.

---

(writing below this line is allowed but discouraged)

## Question 3: Trade-offs [11 points]

You can easily run the forecasting model locally, but you are not so sure about the LLM for the news analysis. You consider three options: (a) Using OpenAI's GPT4o model which charges based on token usage and API calls, (b) buying a high-end GPU and running Meta's "open source" *Llama3.2* model from HuggingFace, or (c) hosting the same open source model on Amazon's AWS cloud infrastructure (AWS Bedrock), which then charges per inference.

**(a)** [5 points] Identify and roughly rank two qualities that are important for the decision in this scenario and one quality of little importance (no measure required for any of them). Provide a brief justification of why they are important or not important:

**Quality 1** (most important):

**Quality 2** (second most important)**:**

**Quality 3** (low importance):

**Justification:**

---

(writing below this line is allowed but discouraged)

**(b)** [6 points] Make a recommendation with a brief justification of how/where to deploy the model for news analysis, considering the tradeoffs between the qualities. Refer explicitly to the important qualities identified previously and underline them in your text. Your answer must relate to the scenario. If you are missing information to make that decision, describe what information you would need and how you would make a recommendation with it.

---

(writing below this line is allowed but discouraged)

## Question 4: Risks and Mitigation [14 points]

To plan for mistakes you try to better understand the requirements and risks of the product (see the scenario on the last page). For this question, you focus on the following requirement:

> *"When an unusual surplus of donations is available on short notice, the FoodFlux system should promptly alert the on-call volunteer of 412 Food Rescue via text message."*

**(a)** [3 points] Classify the following parts of the FoodFlux system into world and machine entities (in the world vs machine sense from the reading and lecture):

- Text of news articles to be analyzed ☐ *world entity* ☐ *machine entity*
- The volunteer waiting for text messages ☐ *world entity* ☐ *machine entity*
- The network connection to the news provider ☐ *world entity* ☐ *machine entity*
- The forecasting model trained on past data ☐ *world entity* ☐ *machine entity*
- The Picklesburgh food festival ☐ *world entity* ☐ *machine entity*
- The GPU executing LLM prompts ☐ *world entity* ☐ *machine entity*

**(b)** [2 points] State one **software specification** that is necessary for the system to satisfy the above requirement.

**(c)** [3 points] State one **environmental assumption** that might be assumed in the system design to satisfy the above requirement, *but that is likely not actually true* and may hence lead to failures to meet the requirement in the running system.

---

(writing below this line is allowed but discouraged)

**(d)** [6 points] If the model incorrectly misses a surprising food surplus, the system may miss opportunities for collecting food donations. Describe a *mitigation* to make it less likely that this problem happens. *The mitigation should be at the system level, outside of the ML component* (i.e., not just "train a more accurate model" or “use an ensemble model”). In addition, think about how you would update a fault tree (not provided) with this mitigation and select one of the options below.

**Mitigation description:**

**Update to a fault tree** (check the option corresponding to your mitigation below, no further explanation needed):

- ☐ *eliminate basic event*
- ☐ *add basic event with an AND connection*
- ☐ *add basic event with an OR connection*

# Question 5: Teamwork [4 points]

In a CMU collaboration you are working with a team of students who plan to extend the news analysis model to social media (facebook, reddit, nextdoor, twitter). The students are motivated, but you fear that they will be less effective than they could be due to *groupthink.*

(1) Briefly give a concrete example of a problem that can occur due to groupthink in this scenario.

(2) Suggest a mitigation strategy that the team could adopt to reduce groupthink.

# Scenario: Forecasting Food Donations


(The scenario is grounded in real organizations in Pittsburgh but otherwise fictional. You may detach this page from the exam.)

You have been volunteering for several years for the local nonprofit **412 Food Rescue**, a Pittsburgh nonprofit which describes itself as *"We work with food retailers to prevent surplus food from going to waste. Transported by a growing network of volunteers, 412 Food Rescue directly transfers food to nonprofit partners that serve those who are food insecure."*

After observing often unexpected big variations in the amount of food donations you consider whether you can use your machine-learning skills to help the organization, and possibly others like it, by better predicting surpluses and needs. Some of the variation is seasonal or weather related, others relate to specific events, such as big street festivals (e.g., Picklesburgh, Little Italy Days).

You plan to combine two models in what you coined project **FoodFlux**:

- A traditional **forecasting model** (linear regression) for timeseries data, based on past donations and needs, seasons, holidays, and weather reports.
- A **news analysis model** built with an LLM and prompts that can analyze local news articles, social media, and official announcements from local governments for reports that mention restaurants, cancellations, large events, and particularly food-related events. It will also look for events that might indicate a particularly high need of donations, such as regional flooding or long power outages. The LLM will extract whether a relevant event will happen or has happened and where, and it will classify the event in one of multiple categories. This information is used as additional features in the forecasting model.

Beyond the model, the FoodFlux system will maintain a list of volunteers who are *on call* at any time, able to follow up on predictions by calling the event or driving there. The on-call list can be maintained through a web application. Volunteers are informed by email and text messages when the model predicts unusual surplus foods or unusual needs.

You have recruited two friends to work on the *FoodFlux* project with you. One has experience with statistics and forecasting models but no system-building experience and the other is an organizer and long-time volunteer for 412 Food Rescue with lots of contacts into the community but little technical background. Other volunteers are supportive of the project and are likely willing to help if asked. 412 Food Rescue has contacts to professors at CMU who might support you by recruiting students part time for relevant engineering or research projects, usually in semester-long team projects. You plan to get the work into a state that it can be easily used by volunteers and where it runs automatically over extended periods of time. Ideally it should be extensible for multiple locations, so that it can be used by other nonprofits in other cities too.

You received a local grant from a charitable foundation in Pittsburgh for this project that will allow you to work full time on this project for 6 months while you are in between other projects. Failure to deliver a working prototype in that time might make 412 Food Rescue look bad and could make it more challenging to raise funds in the future. Beyond funds for your time, resources are limited, especially when it comes to computational power for training and deploying models.