MLHB 23: Final Project Questionnaire

The questions below are intended to guide you in putting together the different pieces of your project. When the questionnaire is complete, you should have in hand what you need to form a concrete and well-thought-out plan, ready for execution.

Please fill in these questions and submit them (see link on course website).
This will account for 10% of your project's grade, so invest time and effort appropriately.
(Note: these questions may not be a perfect fit for all topics and projects; if you feel yours does not fit well within these guidelines - talk to us about it)

1. Chosen topic (number and title):

לפעמים אנחנו כותבים גם הסברים בעברית מתחת לתשובות באנגלית כדי שיהיה יותר מובן אבל התשובות המלאות שאנחנו מגישים הן התשובות באנגלית

5 – Strategic Classification (with some dynamics): The point of dirrt train data, as we will see users will make X dirty by strategically improve their model score by walking with gradient.
Also the 11 features have 2 binary features (isEducated, isSelfEmployed).

---

2. In your own words - what is your project about? Try to convey why you think or feel this is a good choice for you as a final project in our course.

Our project is about examining the dynamics that arise from strategic users who identify which features the model heavily weights, and the model's desire to retrain in order to overcome these strategic players with minimal harm to genuine users who truly should receive a loan and are not engaging in strategizing. We believe this is a good choice for us because it incorporates various aspects of the course such as dynamics, user strategy, and it's also interesting for us to see where this dynamic leads and whether it converges, for example.

---

3. What is the main phenomenon you seek to capture? What are the key mechanisms that are at play (i.e., that generate the phenomenon, or that are affected by it).

We expect to find a phenomenon where, as users strategically modify a feature,
the model gradually gives less importance to this feature when calculating the user's score after retraining.

Another phenomenon is that we want to see to what importance to features the system will converge after many timestamps, while in each timestamp: users below threshold change the features that are currently important to the model (high absolute gradient),

then, after some time, the model "notice the strategy" and retrains on the original labels and the features with strategy.
We want to see the difference in the feature importance vector over time and what it will converge to under different assumptions.
For example, under the assumption that over time, users can only increase each of their features by a total of thirty percent, say, from its original value, we expect good performance after a long time because it's like a bias for a linear model that needs to be adjusted.
Under the assumption that there is no limit on changing the features, as time passes, the training data will become more noisy due to the strategies, and the model's performance will decrease.

אנחנו מצפים למצוא תופעה שבה כשמשתמשים משנים פיצר בצורה אסטרטגית - המודל יתן לאט לאט פחות חשיבות לפיצר זה כשמחשב את הציון של המשתמש.
תופעה נוספת - היא שאנחנו רוצים לראות לאיזה חשיבות לפיצרים המערכת תתכנס אחרי הרבה Timestamps
שבכל אחד הם משנים את הפיצרים שחשובים כרגע למודל, ואז כביכול אחרי כמה זמן המודל "עולה עליהם"
ומתאמן מחדש על התיוגים המקוריים והמידע הרועש,
ורוצים לראות מה ההבדל בוקטור חשיבות הפיצרים במהלך הזמן שעובר ולמה זה יתכנס תחת הנחות שונות.

למשל, תחת הנחה שבמהלך כל הזמן משתמשים יכולים בסך הכל להעלות כל פיצר שלהם בשלושים אחוז נגיד מערכו המקורי - נצפה לביצועים טובים אחרי הרבה זמן כי זה כמו ביאס למודל לינארי שצריך לכוון.

תחת הנחה שאין מגבלה על שינוי הפיצרים, ככל שיעבור הזמן המידע אימון יותר יהרס ויהיו ביצועים נמוכים למודל.

---

4. What are the key metrics you intend to measure? How do they relate? Explain their relation to the phenomena and/or mechanisms described above.

We will examine the explainability vectors under different strategies and dynamic simulations.
This essentially describes the phenomenon we expect to see of a decrease in the importance of strategic features.
We will also examine metrics such as F1, MCC, and ACC, which are related to the phenomenon of decreasing importance of features that have been strategized in such a way that maximizing them requires this phenomenon of decreasing importance of strategic features.
That’s because, relying on features that have been strategized is not a good method because it falls into the hands of strategic users, and relying on strategic features will likely hurt model performance because these are noisy features.

How are the metrics related? F1 and MCC express the quality of the entire confusion matrix and are therefore more informative than metrics like accuracy.
Another relationship is that we expect that the lower the variance of the feature importance vector(features importance distributes uniform), the lower the performance, because it likely indicates the use of a lot of strategy that causes all features to be equally important, which is probably not correct intuitively (in our dataset for example, before using strategies – cc score is the most informative).

נבחן את הוקטורי
Explainabiliy
תחת אסטרטגיות שונות וסימולציות דינאמיות.
זה למעשה מה שמתאר את התופעה אותה אנו מצפים לראות של ירידה בחשיבות פיצרים אסטרטגיים.

נבחן גם מטריקות כמו
F1, mcc, acc
והם קשורות לתופעה של ירידת חשיבות פיצרים שעשו בהם אסטרטגיה בכך שכדי למקסם אותם צריכה לקרות התופעה הזו של ירידת חשיבות פיצרים עם אסטרטגיה, כי הסתמכות על פיצרים שנעשה בהם אסטרטגיה זו שיטה לא טובה כי נופלים לידיים של המשתמשים האסטרטגיים, והסתמכות על פיצרים אסטרטגיים כנראה יפגע בביצועי המודל כי אלו פיצרים רועשים.

איך המטריקות קשורות?
F1 mcc
מבטאות את טיב כל ה:
Confusion matrix
ולכן יותר אינפורמטיביים ממטריקות כמו
Accuracy.
עוד קשר - אנחנו מצפים שככל שהשונות
(variance)
של וקטור החשיבויות של הפיצרים יהיה יותר נמוך - ככה הביצועים יהיו יותר נמוכים, כי זה מעיד ככל הנראה על שימוש בהרבה אסטרטגיה שגורמת לכל הפיצרים להיות חשובים באותה מידה, מה שכנראה לא נכון אינטואיטיבית.

---

5. What parameters or variables will be interesting to vary or experiment with? Relate these to the metrics and measures from the previous question. Explain why you believe these variations are worthwhile to explore, and how you expect them to expose or illuminate the phenomena or effects you are after.

It will be interesting to see the effect of the number of timestamps because we expect that as time passes, the importance of features will become more uniformly distributed if there is no limitation on the percentage change of each feature.
It will also be interesting to see the effect of changing the rule that determines which users perform strategies at each timestamp.
For example, if the threshold to be classified as positive is 0.5 (we are activating sigmoid at the end of each model), and we define that only those close to the threshold try to perform a strategy (which happen in real life), for example, only users with a score from the model that is lower than the threshold by at most 0.2, then the dynamics of the simulation will likely lead to all users below the threshold minus 0.2 not trying to perform a strategy and never crossing the threshold, which will slightly weaken the phenomenon of uniformity in feature importance since the features of constant non-strategic players do not change.

It will also be interesting to change the intensity of the strategy, namely, when a user chooses to strengthen a specific feature, how they strengthen it and how much they can.
So far, we have implemented this by increasing a feature that have a positive gradient by a number uniformly sampled from a certain positive interval, and if negative gradient, sampled from a

negative interval because a strategist always wants to go in the direction of the gradient to improve its model score.
We estimate that the higher the intensity of the strategy, the quicker the decrease in model performance will be, and the phenomenon of uniform importance to features will be revealed faster because it will gradually make all strategic features less informative and advance the explainability vector towards a variance of zero.

יהיה מעניין לראות את ההשפעה של:
מספר ה:
Timestamps
כי נצפה שככל שהזמן עובר חשיבות הפיצרים תהפוך ליותר מפולגת אחיד אם אין הגבלה על אחוזי השינוי של כל פיצר.

יהיה מעניין לראות את ההשפעה של איזה משתמשים מבצעים אסטרטגיה בכל
Timestamp.
אם לדוגמא הסף כדי להיות מסווג חיובי הוא חצי, ונגדיר שרק מי שקרוב לסף מנסה לבצע אסטרטגיה,
לדוגמא רק משתמשים עם ציון מהמודל שנמוך מהסף ב:
0.2
אז הדינאמיות של הסימולציה כנראה תוביל לכך שכל המשתמשים שמתחת לסף פחות
0.2
לא ינסו לבצע אסטרטגיה ולעולם לא יעברו את הסף,
וזה יגרום קצת להחלשת התופעה של אחידות בחשיבות הפיצרים שכן הפיצרים של שחקנים לא אסטרטגיים לא משתנים.

יהיה מעניין גם לשנות את עוצמת האסטרטגיה, כלומר, כשמשתמש בוחר לחזק פיצר מסויים, איך הוא מחזק אותו וכמה הוא יכול.
בינתיים מימשנו זאת בכך שאם לפיצר יש גרדיאנט חיובי נגדיל אותו במספר שדוגמים יוניפורמית מאנטרוול מסויים, ואם גרדיאנט שלילי דוגמים מאנטרוול שלילי כי אסטרטג רוצה ללכת תמיד בכיוון הגרדיאנט.
אנחנו מעריכים שככל שעוצמת האסטרטגיה גבוהה יותר כך הירידה בביצועי המודל תהיה מהירה יותר ותופעת חשיבות אחידה לפיצרים תתגלה מהר יותר כי זה יהפוך לאט לאט את כל הפיצרים האסטרטגיים לפחות אינפורמטיביים ויקדם את וקטור ה
Explainability
לעבר שונות של אפס.

---

6. Combine your answers above to form **research question(s)** - these should be **simple, concrete**, and **testable.** State them here precisely and succinctly.
   How does the users strategic policy (who benefits from using strategy and therefore use it? How much the strategies intense?) influence across timestamps on the:
   A. performance in predictive models
   B. the features importance distribution (which is the average gradient of the score w.r.t each user's features)
   And what can be done to improve models that suffer because of these reasons?

Do such dynamic systems(loop of using strategy - retraining) converge in terms of model performance \ features importance?

---

7. Rewrite your questions as hypotheses and/or conjectures. For each, explain how you plan to test it. These will guide you throughout the entire project - put in extra effort to make sure they are **polished**.
   Important note: if your hypothesis boils down to 'we think our method will have better performance' – then this is insufficient as a research goal. Even if your project is about improving performance, try to phrase questions and hypotheses that get at the 'why' or 'how', not only at the 'if'.

   Hypotheses:
   A. The more strategy is applied to a feature by more users, the model will decrease the importance of this feature.
   B. A simulation that mimics the real world, where users close to the model's threshold perform strategies, does not significantly harm the model's performance over time because the model learns the strategies and decreases the importance of the tainted features or identifies strategic actions and accordingly lowers their score.
   C. The more intensely users can change their features, and the more users do so (including those far from the threshold, for example), the more the model's performance will deteriorate over time, and the feature importance vector will become more uniform.
   D. If a large percentage of users perform strategy that improves the score the model gives them, the model can be improved by adding a bias to the score given by the model so that the threshold effectively shifts.

   השערות:
   א. ככל שבוצע בפיצר יותר אסטרטגיה על ידי יותר משתמשים - המודל יוריד את חשיבות פיצר זה
   ב.סימולציה שמדמה את העולם האמיתי ובה משתמשים שקרובים לסף של המודל מבצעים אסטרטגיה לא פוגעת כל כך בביצועים של המודל לאורך זמן כיוון שהמודל לומד את האסטרטגיה ומוריד את החשיבות לפיצרים שהתלכלכו או מזהה אסטרטג ומוריד לו את הציון בהתאם
   ג. ככל שהמשתמשים יכולים לשנות את הפיצרים שלהם בצורה יותר אינטנסיבית, וככל שיותר עושים זאת (גם כאלו שרחוקים מהסף לדוגמא)
   ככה ביצועי המודל ידרדרו לאורך זמן ווקטור חשיבויות הפיצרים ייהפך ליותר ויותר יוניפורמי
   ד. אם המון אחוזים מהמשתמשים ביצעו אסטרטגיה כלשהיא שמשפרת את הציון שהמודל נותן להם - ניתן לשפר את המודל בהוספת
   Bias
   לציון שהמודל נותן כך שהסף למעשה יזוז

---

8. Describe your approach. Which learning algorithms will you use? Which are off-shelf tools, and which do you intend to implement yourself?

We will get the dataset from Kaggle. We plan to use basic algorithms like Logistic Regression and Random Forest from sklearn, and also relatively simple models that we will build in PyTorch.
To calculate the gradient vectors, we will use PyTorch's autograd system.
As for the simulations, feature changes, etc., we will implement them in Python.

9. Describe the setting or environment you plan to experiment in, i.e., to test your hypotheses/conjectures. How do you intend to design and run simulations? What experiments do you plan to run? Give concrete details, and be consistent with your earlier questions - **this is the time to plan ahead**.

The environment is one where each user has 11 features indicating their ability to repay the loan, and a binary feature indicating whether the loan was repaid or approved.
In this environment, at each time unit, there are users who choose to act strategically, usually those close to the model's threshold since it is advantageous for them to make a small effort and deceive the model.
Therefore, we allow these users to modify certain features to a certain extent, which in practice will be, for example, the top 3 features with the highest absolute average gradient.

Regarding the simulations we intend to perform, we explained some of them in previous questions, but here are more details:
In one simulation, there will be 100 time units, and at each time unit, the model scores all users based on their current features, then the average gradient across all users is calculated.
From it, we take a few features with the largest absolute derivative value and allow users below the model's threshold (say from 0.3 to 0.5, which is the threshold we defined) to add(or substract if negative gradient) any value to them.
Afterwards, the model is trained on the new data with the noise, and the feature importance vector is presented to see if the model decreased the importance of the features where strategies were applied.
This continues over time units, with each time being a struggle between users close to the threshold improving their features, and the model trying to decrease the importance of strategic features.

We will run simulations that examine various parameters, such as which users decide to strategize? How much can they change their features? And see if performance \ importance converges.

We also want to examine classification methods where we give the model the original users data, and tell it in estimation how much each feature can be changed by strategy (quite logical in the real world, as the bank, for example, knows that it's hard to cheat in credit rating but easy to cheat in other things like the value of your assets).
Then compare a model that does not use this additional information to a model that does.

הסביבה היא סביבה בה לכל משתמש יש 11 פיצרים שמעידים על היכולות שלו להחזיר את ההלוואה, ופיצר בינארי שאומר האם החזיר את ההלוואה\האם אישרו לו את ההלוואה.
בסביבה זו, בכל יחידת זמן, יש משתמשים שבוחרים לפעול בצורה אסטרטגית, בדרך כלל כאלו שקרובים לסף של המודל כי כדאי להם לעשות מאמץ קטן ולהערים על המודל, ולכן אנו מאפשרים למשתמשים אלה לשנות במידה מסויימת את הפיצרים שהם רוצים, שניקח אותם בפועל להיות הטופ 3 פיצרים לדוגמא עם גרדיאנט ממוצע בערך מוחלט הכי גבוה.

לגבי הסימולציות אותן אנו מתכוונים לבצע, הסברנו על כמה מהן בשאלות קודמות, אבל נפרט עוד קצת.
בסימולציה אחת, יהיו 100 יחידות זמן, ובכל יחידת זמן המודל נותן ציון לכל המשתמשים על פי הפיצרים הנוכחיים שלהם, ואז מחושב הגרדיאנט הממוצע על פני כל המשתמשים. ממנו נלקחים כמה פיצרים שהנגזרת שלהם בערך מוחלט הכי גדולה - ומאפשרים למשתמשים שמתחת לסף של המודל (נניח מ0.3 עד 0.5 שזה הסף שהגדרנו) להוסיף להם ערך כלשהוא.
לאחר מכן המודל מתאמן על המידע החדש עם הרעש, ומציגים את וקטור החשיבויות של הפיצרים כדי לראות האם המודל הוריד חשיבות לפיצרים שבהם נעשתה אסטרטגיה.

כך ממשיכים ביחידות זמן, כשכל פעם יש מאבק בין המשתמשים שקרובים לסף ומשפרים את הפיצרים שלהם, למודל שמנסה כנראה להוריד את החשיבות לפיצרים האסטרטגיים.

אנחנו נריץ סימולציות שבוחנות פרמטרים שונים כמו איזה משתמשים מחליטים לעשות אסטרטגיה? כמה הם יכולים לשנות את הפיצרים שלהם?

אנחנו גם רוצים לבחון שיטות של סיווג שבהם נותנים למודל את המידע המקורי של המשתמשים, ואומרים לו בהערכה כמה כל פיצר ניתן לשינוי בעזרת אסטרטגיה(די הגיוני בעולם האמיתי, כי הבנק לדוגמא יודע נגיד שקשה לרמות בדירוג אשראי, אבל קל לרמות בדברים אחרים כמו מה שווי הנכסין שלך).
ואז להשוות מודל שלא משתמש במידע הנוסף הזה למודל שכן משתמש

10. What code do you plan to use, and from what sources? E.g., public packages/repos (check online!), code from homework/workshops (if so, make sure your project remains distinct, and not a mild variation), new code (if so, describe it in brief).

We plan to use a dataset from Kaggle as mentioned.

For the algorithmic part, as explained, we will utilize libraries like PyTorch, pandas, and sklearn for data processing, building models, calculating gradients, and more.
Regarding the code, we are writing everything from scratch, starting with analyzing the features and their relationships, and then moving on to the simulations.

---

11. List three potential pitfalls that you anticipate may occur. Try to plan your response.

    1. In the experiment where the model receives information on how much each feature can be changed by strategy, we might not manage to use the information effectively and fail to improve compared to a baseline model that does not use this information.

    ---

    2. ---

    We may not observe the phenomena we expect, for example, the importance of features that are strategized does not decrease over time. To mitigate this, we'll need to be prepared to adjust our hypotheses and explore alternative explanations

    ---

    3.

    The gradient vector of the model's score with respect to the user's features, may vary significantly between users.
    This variation means that the average vector might not accurately reflect the importance of each feature to the model.
    To address this, we will explore new methods to obtain the explainability vector.
    There is considerable research in the area, for example, in computer vision, a method exists where gradients of the same image under different brightness levels are averaged to improve the explainability matrix.
    To adapt this to our work with 11 features, we might consider weakening a specific feature at a time and averaging the gradients.

    בעברית:

    שלושה בעיות שייתכן שיקרו לנו:
    א. בניסוי שבו המודל מקבל מידע על כמה כל פיצר ניתן לשינוי על ידי אסטרטגיה - לא נצליח להשתמש במידע בצורה אפשקטיבית ולא נצליח כל כך להשתפר לעומת מודל בסיס שלא משתמש במידע זה.
    ב. לא נחזה בתופעות אותן אנו מצפים לקבל, לדוגמא, החשיבות של פיצרים שנעשית בהם אסטרטגיה לא תרד עם הזמן.
    ג. וקטור הגרדיאנט של ציון המודל של המשתמש, כשגוזרים בכבוד לפיצרים של המשתמש - ישתנה מאוד בין משתמש למשתמש, כך שהוקטור הממוצע לא ישקף בצורה טובה את החשיבות של כל פיצר למודל.
    איך נתמודד עם זה? נחפש שיטות חדשות להשיג את וקטור ה
    Explainability,

יש המון מחקרים בתחום, נגיד בתחום של ראייה ממוחשבת יש שיטה שבה ממצעים גרדיאנטים של אותה תמונה עם בהירות שונה כל פעם ומקבלים שיפור במטריצת
Explainability.
כדי להקביל את זה לעבודה שלנו שבה יש 11 פיצרים, יכול להיות שנחליש כל פעם פיצר מסויים ונמצע גרדיאנטים

---

12. Bonus question: What is the spirit animal of your project? Use an emoji, and explain briefly.

We chose the fox.
Foxes are symbols of cunning and trickery - a reputation stemming from their known ability to evade hunters.

This reminds us of our project, where the strategic users,
who wish to obtain a loan from the bank, are the foxes trying to change the feature that is important to the model (the hunter) to deceive it into granting them a loan.

The model is the hunter, needing to catch on to the tricks of the strategic users (the foxes) and take various actions such as decreasing the importance of strategic features to classify correctly.

Good luck!
MLHB 23-24 team