Milestone 5

MLBD | Milestone 5 (google.com)

What research question will you explore for Milestone 6 and 7? *

You can think of this as an extension of your work for Milestone 4, or a completely separate exploration.

M4 gave us some strong indications to support our hypothesis. Even as it encouraged our hypothesis that predicting level check improvements is possible, we were not able to identify a specific behavior as the optimal one.
We plan on keeping the same research question but trying out additional pre-processing methods, leveraging the results of milestone 4 and explorating different modeling techniques.

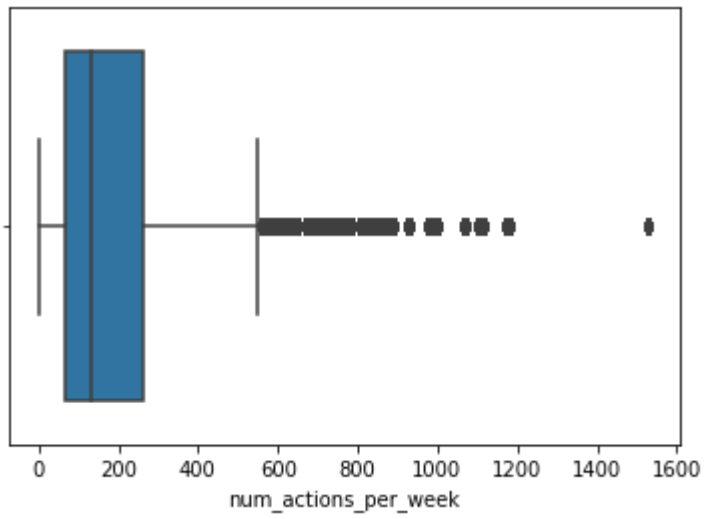
Our research question still remains:
Is it possible to identify one or more studying behaviors leading to a significant improvement in level checks results?

How will you preprocess your data?

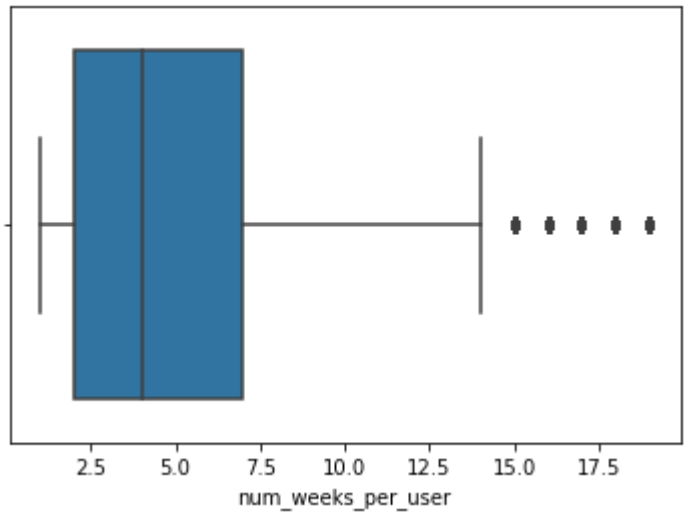
Please include which frameworks you plan to use, which visualizations you plan to create, and why you plan to preprocess the data in the way you propose. If this is the same as Milestone 4, feel free to leave this blank.

In M4 we did not remove the outliers shown in the plots below since we considered that these outliers are attributed to users with a high activity on the platform and, therefore, they could reveal interesting regularity patterns when using the platform. Nevertheless, after the clustering results obtained in M4, we have thought that maybe it is a good idea to remove those outliers in order to obtain more uniform and balanced clusters.

Num of actions per week boxplot:



Num of weeks per user boxplot:



Another thing that we are considering is splitting creating different classifiers for both german and math level checks, and perhaps focusing on one.

Describe your modeling methodology. *

Which model will you use? Why did you decide on that model for this research problem?

M4 allowed us to precisely identify the optimal classificator for level check improvements (TimeSeriesSVC). However, we are thinking of trying out regressors instead of classificators. Linear regression for example could allow us to identify which feature contributes more to the result. Another option could be to try to implement the methods studied in class with (Hidden) Markov models for predicting success in level check quizzes.

Regarding clustering, due to its lack of impact on our results, we can explore different clustering techniques such as spectral clustering (although computation times also significantly increase with it) that seem fitted to our type of data.

Further visualizations can be considered using feature dimensionality reduction such as PCA. This could also give us indications on which features are most predictive.

Do you have any questions about your plan / the data?

This is a nice opportunity to get feedback from the course TAs!

We are open to any suggestion or comment to further extend our project.