EPFL CS-421 : Machine learning for behavioral data - BRAM

Antoine Munier, Romain Berquet
School of Computer and Communication Science, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

***Abstract*—Social media platforms such as Snapchat, Instagram, and TikTok share a common characteristic – their immense appeal to users. Once individuals experience these platforms, they are instantly drawn to use them repeatedly. This phenomenon is known as user attraction. Mobile application developers strive to create captivating applications by offering engaging content, user-friendly interfaces, and other enticing features. However, can we quantify and measure the extent to which an application attracts its users? Moreover, can we predict the future behaviour of early users? These are the fundamental questions we aim to address in our research endeavour.**

I. Introduction

In assessing user attraction towards an application, various criteria can be considered. Following exploratory analysis of the data, our focus has shifted towards predicting the probability of user retention in the application after their initial 10 sessions. Specifically, we aim to determine whether a user will continue to engage with the application by opening subsequent sessions beyond this threshold. In addressing this objective, we also endeavour to analyze the diverse features that influence the application's appeal to users, offering valuable insights into its strengths and weaknesses.

The application under investigation is Lernanvi, a platform that facilitates interactive and immersive learning experiences in mathematics and German language learning for students. However, despite its promising characteristics, the application has exhibited a noticeable pattern of user attrition, with a significant proportion of users discontinuing their engagement after only a few sessions. Understanding the underlying factors contributing to this phenomenon constitutes a crucial aspect of our research endeavour.

II. Data Preparation and Exploratory Analysis

In order to ensure data readiness for our model, we conducted a series of data-cleaning steps.

A. Data Structure

The dataset comprises seven distinct tables, each serving a specific purpose. A concise description of each table is provided below:

- *Users*: This table contains demographic information pertaining to individual users.
- *Events*: This table captures the user interactions and activities within the platform.
- *Transactions*: This table encompasses details regarding user actions and questions.
- *Documents*: This table encompasses the various questions and problems available for user solving within the application.
- *Topics_Translated*: This table represents the taxonomy of categories featured in the Deutsch and Math dashboards.
- *Topic_Trees*: This table illustrates the hierarchical structure of topics.
- *Feedback_History*: This table contains a historical record of hints or responses provided to students following their task completion.

For a more comprehensive understanding of the tables and their corresponding structures, refer to the documentation available here. The dataset encompasses a total of 30,929 distinct users.

B. Data Cleaning and Preprocessing

To address this specific aspect, we executed the subsequent procedural stages

1) Elimination of users lacking events, resulting in the removal of 8459 users.
2) Removal of events and transactions from their respective tables when no corresponding key was found in the complementary table. Consequently, 44801 transactions and 5871090 events were expunged.
3) Application of the classical quantile method to eliminate outliers present in the tables.

C. Summary Statistics

In our quest for a deeper understanding, our study aims to explore the diverse distributions manifested by the features that are relevant to our research problem, already present within the dataset. As we delved into the examination of various features, we drew inspiration from a notable paper that focuses on specific characteristics within the dataset to capture user behaviour. This paper served as a valuable reference, informing our approach to feature analysis and selection, and guiding us towards relevant insights and methodologies in understanding user behaviour patterns. [1]:

1) Number of sessions per user: Initially, our investigation focused on analyzing the distribution of user sessions to establish a threshold indicative of the level of engagement

with the application. This analysis aided in determining the degree of user attraction towards the application.

The statistical summary for the number of sessions per user is presented below:

- Count: 19,570
- Mean: 12.073480
- Standard Deviation: 22.262449
- Minimum: 1
- 25th Percentile: 2
- Median: 5
- 75th Percentile: 13
- Maximum: 638

These statistics provide an overview of the distribution of user sessions, illustrating key measures such as the average number of sessions per user (mean), the degree of variability in session counts (standard deviation), as well as the minimum, maximum, and quartile values. Upon examining the plot depicting the distribution of the number of sessions per user, we observe that approximately 50% of the users have five or fewer sessions.1 This observation holds significant implications for our feature selection process, as it highlights the importance of considering this characteristic when choosing relevant features for our analysis.

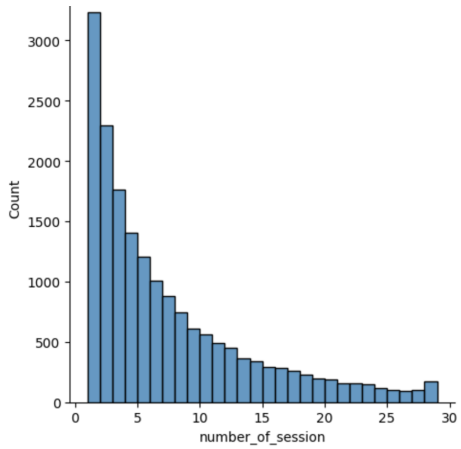

Figure 1. Distribution of the number of sessions per user

2) Type of sessions per user: Given the potentially significant influence of the type of questions answered by users on their overall engagement, we conducted an analysis to examine their distribution. Remarkably, our findings revealed that the number of German language sessions exceeded the number of math sessions by a ratio of 2:1. This discrepancy highlights the prominence of German language content within the application and suggests its potential impact on user behaviour and preferences.

3) Percentage of correct questions: Another intriguing feature available in the dataset is the distribution of correct questions answered by users, which can be easily computed. Upon examination, we discovered a significant number of users who did not answer any question correctly.3 This observation holds relevance as we proceed with feature selection, as it offers insights into user behaviour. The prevalence of users who struggle to answer questions correctly may indicate certain patterns or characteristics that warrant consideration when choosing relevant features for our analysis.

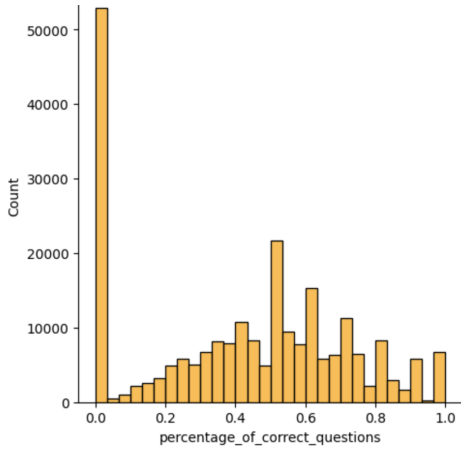

Figure 2. Distribution of the percentage of correct questions answered per user

4) Difficulty of the sessions performed by the users: In addition to analyzing user performance, we investigated the mean difficulty of questions answered by the users. Our findings indicate that, on average, questions were answered correctly less than 50% of the time. This suggests that the questions posed to users tend to be relatively challenging. Furthermore, our analysis revealed a limited number of points beyond the 70% correct answer threshold. 3 These observations will be taken into account during our feature selection process, as they shed light on the characteristics and distribution of question difficulty within the dataset.

5) Duration of the sessions performed by the users: Another intriguing feature we explored is the distribution of session durations among users. Our analysis reveals that the majority of sessions have a relatively short duration, indicating quick interactions with the application. However, the distribution exhibits a noticeable right skew, suggesting the presence of a longer-tailed portion with relatively longer session durations.4 This finding provides valuable insights into user behaviour and session patterns within the application, which will be considered in our subsequent analyses and modelling efforts.

6) Type of actions performed by the users: Lastly, upon examining the various types of actions performed by users, it becomes apparent that several actions were infrequently observed. Actions such as`'GO_TO_BUG_REPORT','GO_TO_COMMENTS', 'GO_TO_THEORY','REQUEST_HINT','SHARE, SKIP'`, and `'VIEW_QUESTION'` were rarely executed by

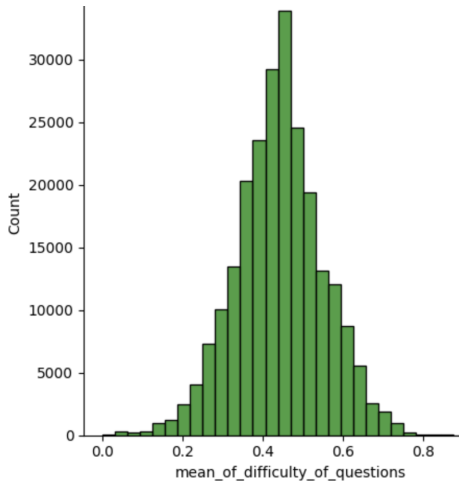

Figure 3. Distribution of the difficulty of the questions answered by the users

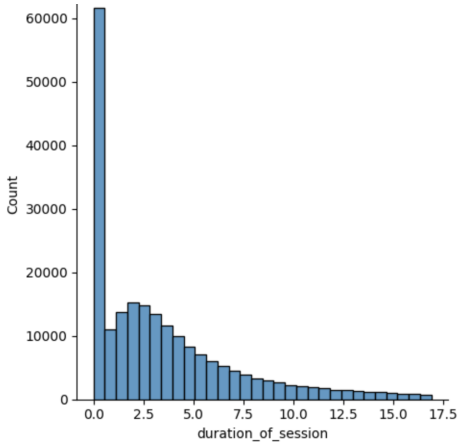

Figure 4. Distribution of the difficulty of the questions answered by the users

users. We will take note of this observation as we proceed with our feature selection process.

III. The Proposed Approach

In this section, we outline our proposed approach for predicting the number of sessions a user is likely to stay on the app. We begin by describing the selection of features and their organization, followed by the criteria used for dataset filtering. Subsequently, we discuss the process of feature selection based on correlation analysis, and we present the chosen set of features and provide brief explanations for each. Then we introduce how we create the Dataset with the time series and finally our model approach.

A. Feature Selection and Organization

Given the objective of predicting the number of sessions a user will spend on the app, we carefully select features and determine their appropriate representation. To maintain consistency and ensure a session-based approach, we define each line in the dataset to represent the data for one session. We initially considered using weeks or days as units of measurement; however, due to the limited availability of data, we found that a higher time value did not yield satisfactory results.

To ensure a robust distribution and enhance prediction accuracy, we focus on users who have a minimum of 10 sessions. This criterion allows us to capture a broader range of user behaviour and improve the generalization of our predictive model. Additionally, we order all sessions by their index, such that the first session for each user is assigned an index of 1, and so on.

B. Feature Selection based on Correlation

To select the most relevant features for our dataset, we conduct a correlation analysis between the week index and all available features. We employ a simple Ordinary Least Squares (OLS) regression and examine the p-values associated with each feature. Features that exhibit a p-value less than 0.05 are deemed statistically significant and are subsequently chosen for further analysis.

C. Selected Features

Based on our correlation analysis, the following features have been selected for our predictive model:

1) `'CLOSE'`: The number of times the user closes a tab from Lernavi during a session.
2) `'GO_TO_BUG_REPORT'`: The number of times the user reaches a bug report page during a session.
3) `'GO_TO_THEORY'`: The number of times the user reaches a theory page during a session.
4) `'NEXT'`: The number of times the next event occurs during a session. The next event signifies the transition to the next question after submitting an answer.
5) `'VIEW_QUESTION'`: The number of times the user engages in the view question event during a session. This event occurs when the user navigates back and forth through the list of questions at the end of a session.
6) `'mean_of_difficulty_of_questions'`: The mean difficulty level of questions within a session.
7) `'session_closed'`: The number of times the session has been marked as finished or unfinished. A session is considered finished when all questions have been answered.
8) `'duration_of_session'`: The duration of a session, calculated as the difference between the commit_time and the start_time.
9) `'math_session'`: A binary indicator (1 or 0) representing whether the session is related to math.

These selected features provide insights into user behaviour, navigation patterns, question difficulty, session

completion status, duration, and subject matter. By incorporating these features into our predictive model, we aim to capture relevant aspects that influence the number of sessions a user spends on the app.

D. Dataset Creation and Target Variable Definition

1) Dataset Creation: We first create the user's last session index, To establish it the dataset is grouped by user ID, and the maximum session index is computed. This information is merged back into the dataset, providing a reference point for subsequent data manipulation and analysis.

To ensure a manageable dataset size and capture sufficient user behaviour variability, only the first 10 sessions per user are considered. The dataset is sorted based on the user ID and session index, arranging the sessions chronologically within each user's data.

Rows containing missing values are dropped to ensure data integrity. The duration of each session is converted from the original format to minutes for ease of interpretation and analysis.

The dataset is transformed into a time series format, where the user ID serves as the index, and the session index becomes the columns. This matrix-like structure represents the sequence of sessions for each user, facilitating subsequent analysis.

2) Introduction of the Hyperparameter Omega: The hyperparameter "omega" is introduced to capture consecutive session batches. Each user generates 10 - omega + 1 batch, resulting in multiple rows of data per user. This expansion of the dataset size enables a more comprehensive analysis of user behaviour patterns.

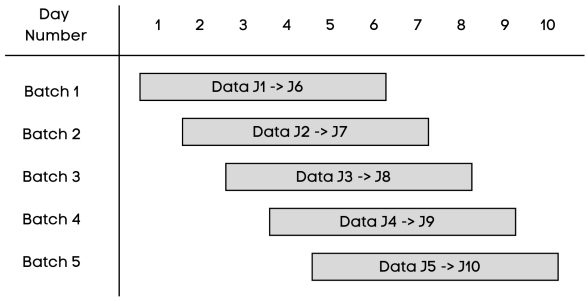

Figure 5. Distribution of the batch with an omega value of 6

3) Target Variable Definition: The target variable, denoted as "y," represents the number of sessions occurring after a specific consecutive session batch. "y" serves as a crucial metric for understanding the number of sessions occurring after a specific consecutive session batch.

For each batch, defined by the parameter omega, "y" is calculated by subtracting the sum of omega and the position of the consecutive sessions within the overall session batch from the last session index. This calculation reveals the number of sessions occurring after the consecutive session range, providing valuable information about user behaviour evolution.

To address potential overfitting and ensure a better distribution of the target variable, a shifted logarithmic transformation is applied. "y" is replaced with $\log(y+1)$, facilitating normalization and improving model performance.

E. Model Approach

In this study, our objective is to determine the best predictive model for analyzing user session behaviour. We compare three distinct models: Linear Regression, Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM).

1) Linear Regression: Linear Regression is a widely-used statistical model that assumes a linear relationship between input features and the target variable. In our case, we can apply Linear Regression to estimate the number of sessions occurring after a consecutive session batch. By fitting a linear function to the training data, we can determine the coefficients that best capture the relationship between the input features and the target variable.

2) Convolutional Neural Network (CNN): CNNs are powerful deep-learning models commonly used for image and sequence analysis tasks. In our context, we can leverage the temporal nature of the session data by treating it as a sequential input. By using convolutional layers, CNN can extract relevant patterns and features from the input sequences, enabling effective prediction of the number of sessions. The CNN architecture can capture local dependencies in the sequence data and learn meaningful representations, making it a promising model for session behaviour analysis.

3) Long Short-Term Memory (LSTM): LSTM is a type of recurrent neural network (RNN) that excels in modelling sequences with long-term dependencies. Unlike traditional neural networks, LSTM can retain information over long periods, making it well-suited for capturing the sequential nature of session data. By leveraging its memory cells and gates, LSTM can effectively capture temporal patterns and dependencies, enabling accurate prediction of the number of sessions occurring after a consecutive session batch. Its ability to handle varying-length sequences and retain information from past sessions makes LSTM an attractive model for session behaviour analysis.

To determine the best model among these three approaches, we will evaluate their performance on our dataset using appropriate metrics such as mean squared error (MSE), mean absolute error (MAE), and R-squared. We will also conduct cross-validation to assess the generalization capability of each model. The model that exhibits superior predictive performance and generalization ability will be selected as the optimal choice for analyzing user session behaviour.

By comparing Linear Regression, CNN, and LSTM models, we aim to identify the most suitable approach for accurately predicting the number of sessions occurring after a consecutive session batch. The selected model will enable us to gain insights into user engagement and retention patterns, ultimately contributing to the enhancement of the application's user experience.

F. Conclusion

In this section, we have outlined our proposed approach for predicting the number of sessions a user is likely to stay on the app. We carefully selected features based on their relevance and organized them in a session-based format to capture user behaviour accurately. The selected features encompass various aspects such as navigation patterns, question difficulty, session completion status, duration, and subject matter, providing a comprehensive understanding of user engagement.

By conducting a correlation analysis, we identified the features that exhibit significant relationships with the target variable. These features were selected for further analysis and incorporated into our predictive model.

To create the dataset, we grouped the data by user ID and obtained the maximum session index to establish the user's last session. We filtered the dataset to consider only the first 10 sessions per user, ensuring a manageable dataset size while capturing sufficient user behaviour variability. Missing values were removed to maintain data integrity, and the duration of each session was converted to minutes for easier interpretation.

The dataset was transformed into a time series format, where the user ID served as the index and the session index became the columns. This arrangement facilitated subsequent analysis of the sequence of sessions for each user.

We introduced the hyperparameter omega to capture consecutive session batches and expand the dataset size. Each user-generated 10 - omega + 1 batch, allowing us to analyze user behaviour patterns comprehensively.

The target variable, denoted as "y," represented the number of sessions occurring after a specific consecutive session batch. We calculated "y" by subtracting the sum of omega and the position of the consecutive sessions within the overall session batch from the last session index. To address overfitting and improve the distribution of "y," we applied a shifted logarithmic transformation.

In our model approach, we compared three distinct models: Linear Regression, Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM). Each model offered unique advantages in capturing user session behaviour, leveraging either statistical relationships or deep learning techniques. We planned to evaluate the performance of these models using appropriate metrics such as mean squared error (MSE), mean absolute error (MAE), and R-squared, along with cross-validation to assess their generalization capability. The model demonstrating superior predictive performance and generalization ability would be selected as the optimal choice for analyzing user session behaviour.

By identifying the best model, we aimed to gain insights into user engagement and retention patterns, contributing to the improvement of the application

IV. Experimental Evaluation

A. Linear Regression

Upon running the linear regression model with an omega value of 5, we obtained the following performance metrics averaged over 100 different episodes: a mean squared error (MSE) of 0.85, a mean absolute error (MAE) of 0.73, and an R-squared value of 0.077.

The MSE of 0.85 indicates that, on average 6, the predicted values deviate from the actual values by 0.85 units squared. Similarly, the MAE of 0.73 suggests that the average absolute difference between the predicted and actual values is 0.73 7. These metrics provide insights into the accuracy and precision of the linear regression model in predicting the number of sessions occurring after a consecutive session batch.

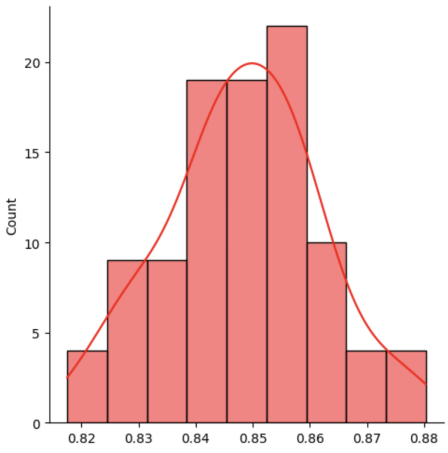

Figure 6. Distribution of the MSE over 100 episodes with an omega value of 5

The relatively low R-squared value of 0.077 implies that only 7.7% of the variance in the target variable can be

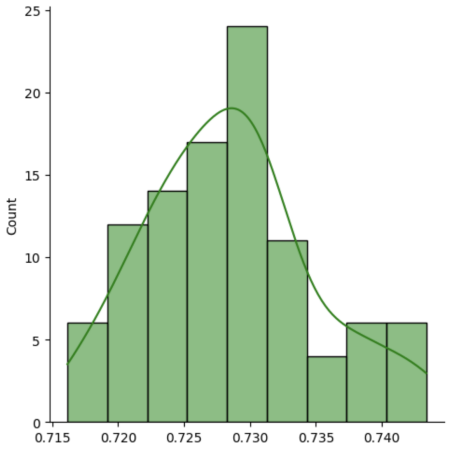

Figure 7. Distribution of the MSE over 100 episodes with an omega value of 5

explained by the linear regression model 8. This indicates that the linear relationship between the selected features and the target variable might not be very strong. It suggests that the linear regression model alone may not be sufficient for accurately predicting user session behavior based on the given features.

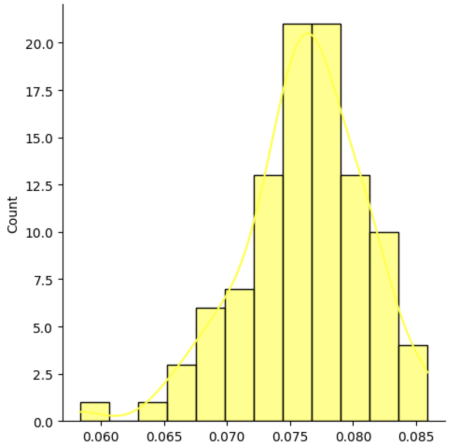

Figure 8. Distribution of the R2 over 100 episodes with an omega value of 5

The obtained results highlight the limitations of the linear regression model in capturing the complex patterns and dependencies present in the dataset. It is likely that the linear regression model oversimplifies the relationship between the input features and the target variable, resulting in suboptimal predictive performance. Therefore, it is necessary to explore more sophisticated models, such as convolutional neural networks (CNNs) and long short-term memory (LSTM) models, to potentially achieve better predictive accuracy and capture the underlying dynamics of user session behaviour.

1) Normalize vs Non-Normalize Data: In our analysis, we also compared the performance of the models using normalized and non-normalized data. Specifically, we evaluated the impact of normalizing the input features (X) on the three metrics we observed (MSE, MAE, R2).

Upon comparing the results, we observed no significant difference in the performance metrics when using normalized X versus non-normalized X. This implies that the normalization of the input features did not have a substantial impact on the predictive accuracy of the models.

The lack of difference in the performance metrics suggests that the range and scale of the input features did not heavily influence the model's ability to predict the number of sessions occurring after a consecutive session batch. Therefore, normalizing the data may not be necessary in this particular context.

B. CNN

Next, we proceed to evaluate the performance of our model using a Convolutional Neural Network (CNN). Our primary objective is to investigate the impact of the layer size on the model's predictive capabilities. Through experimentation, we find that the best fit is achieved with a first layer size of 8. Interestingly, we observe distinct behaviours in terms of the mean squared error (MSE), mean absolute error (MAE), and R-squared (R2) scores when comparing different second layer sizes.

For the MSE metric, we notice that a second layer size of 8 outperforms a second layer size of 32. This implies that a smaller second layer size helps minimize the overall error between the predicted and actual values. 9 However, when considering the MAE metric, the second layer size of 8 does not exhibit a notable advantage over the second layer size of 32. It suggests that the magnitude of errors between the predicted and actual values is not significantly influenced by the choice of the second layer size. 10

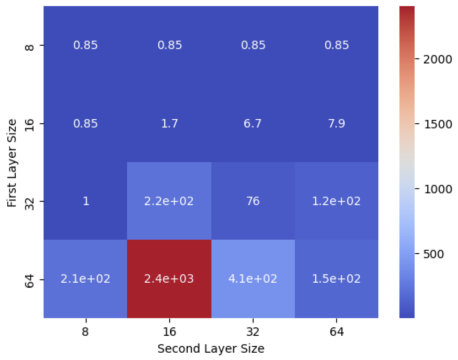

Figure 9. Distribution of the MSE for the CNN depending on the size of the two hidden layers

Initially, we observe that using an 8-unit first layer and an 8-unit second layer results in an R2 score of -77. This

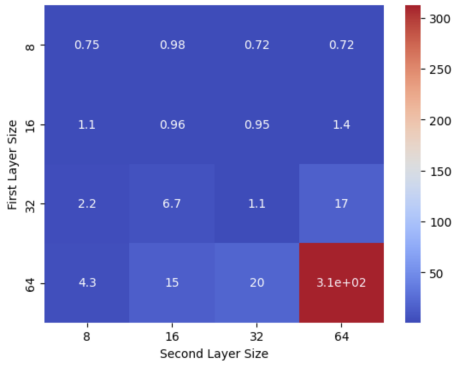

Figure 10. Distribution of the MAE for the CNN depending on the size of the two hidden layers

exceptionally low R2 score indicates a poor fit of the model to the data and suggests that the model fails to capture the underlying patterns and relationships. The negative value indicates that the model explains significantly less variation in the data compared to a constant baseline.

However, as we increase the second layer size to 16 while keeping the first layer size at 8, we observe a slight improvement in the R2 score, which now stands at -1.9. Although still negative, this indicates a better fit than the previous combination. The model shows a slightly improved ability to explain the variance in the data, although it is still far from providing a satisfactory level of prediction accuracy.

Interestingly, when we further increase the second layer size to 32 while maintaining the first layer size at 8, we achieve an R2 score of 0. This suggests that the model now explains no additional variation in the data compared to a constant baseline. In other words, the model fails to capture any meaningful patterns or relationships between the input features and the target variable. While an R2 score of 0 signifies no improvement over a baseline model, it is still preferable to the negative R2 scores observed in the previous layer size combinations. 11

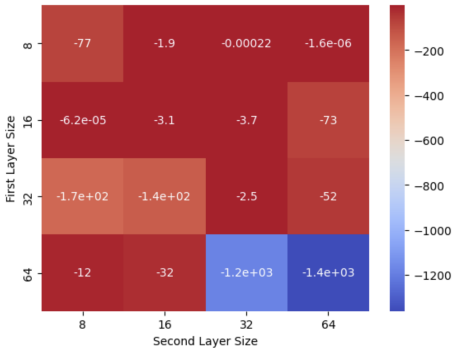

Figure 11. Distribution of the R2 for the CNN depending on the size of the two hidden layers

These findings highlight the importance of considering different evaluation metrics when assessing model performance. The performance of a model can vary depending on the chosen metric, and it is crucial to evaluate multiple metrics to gain a comprehensive understanding of its strengths and weaknesses. In our case, while the CNN model with an 8-unit first layer and a 32-unit second layer may perform poorly in terms of MSE and R2, it might still exhibit favourable performance in terms of MAE.

C. LSTM

To our surprise, the performance of the LSTM model was significantly poor across all experiments. Regardless of the variations in layer sizes and the number of episodes, the obtained results consistently indicated high error rates and low explanatory power.

Specifically, the mean squared error (MSE) values did not improve beyond 3.7 12. Similarly, the mean absolute error (MAE) values did not demonstrate notable improvement, remaining above 1.7. 13

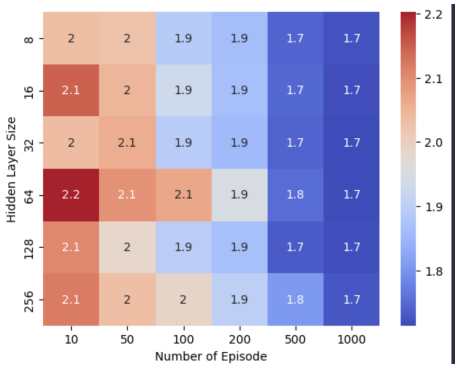

Figure 12. Distribution of the MSE for the LSTM depending on the size of the hidden layer and the number of episodes

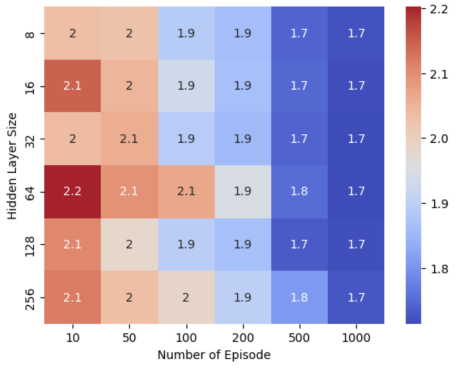

Figure 13. Distribution of the MAE for the LSTM depending on the size of the hidden layer and the number of episodes

Most strikingly, the R-squared (R2) values were consistently negative, with the best R2 value not exceeding -3 14. These negative R2 scores indicate that the LSTM model failed to capture the underlying patterns and relationships in

the data. The model's inability to explain the variation in the target variable suggests a lack of predictive power and a weak fit to the observed data.

The observed poor performance of the LSTM model is unexpected and raises important questions regarding its suitability for predicting the number of sessions each user is going to do.

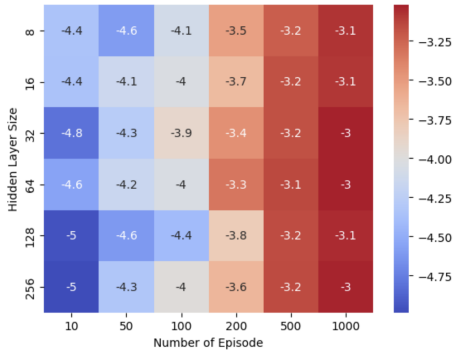

Figure 14. Distribution of the R2 for the LSTM depending on the size of the hidden layer and the number of episodes

Possible factors contributing to this suboptimal performance could include the complexity and nonlinearity of the underlying data, as well as the presence of outliers or noisy observations. It is also conceivable that the LSTM model's architecture and parameters were not appropriately configured for this particular task.

Furthermore, we noticed an interesting behaviour regarding the convergence of the LSTM model as the number of episodes increased. With a larger number of episodes, we observed that the model tended to converge to a certain level of performance. However, it is worth noting that with 1000 episodes, the influence of the hidden layer size on the model's performance diminished compared to when a smaller number of episodes was used for training.

This finding suggests that, with a limited number of episodes, the choice of the hidden layer size can significantly impact the LSTM model's predictive capabilities. However, as the number of episodes increases, the model's performance becomes less sensitive to changes in the hidden layer size. This could be attributed to the increased amount of training data available, allowing the model to learn more robust representations of the underlying patterns and relationships in the data.

V. Discussion and Implication

A. Discussion about our researches

The experimental evaluation of the linear regression, CNN, and LSTM models provided valuable insights into their predictive capabilities and limitations in predicting the number of sessions occurring after a consecutive session batch. The discussion below summarizes the key findings and their implications for understanding user session behaviour.

The linear regression model demonstrated suboptimal performance in capturing the complex patterns and dependencies present in the dataset. The relatively low R-squared value of 0.077 indicated that only 7.7 % of the variance in the target variable could be explained by the linear regression model. This suggests that the linear relationship between the selected features and the target variable might not be very strong. Therefore, relying solely on the linear regression model may not be sufficient for accurately predicting user session behaviour based on the given features. To improve predictive accuracy, it is necessary to explore more sophisticated models, such as CNNs and LSTMs, that can capture the underlying dynamics of user session behaviour.

The CNN model's performance varied depending on the choice of layer sizes. Interestingly, the model exhibited different behaviours in terms of the MSE, MAE, and R-squared scores when comparing different layer size combinations. The results indicated that the choice of layer sizes could impact the model's performance metrics. While a smaller second layer size (8 units) helped minimize the overall mean squared error, it did not show a notable advantage in terms of the mean absolute error. Furthermore, negative R-squared scores in certain layer size combinations indicated a poor fit of the model to the data. These findings highlight the importance of considering multiple evaluation metrics and conducting a thorough analysis of different model configurations to assess performance accurately.

The LSTM model consistently performed poorly across all experiments, with high error rates and negative R-squared scores. The model's inability to capture the underlying patterns and relationships in the data suggests a lack of predictive power and a weak fit. Possible factors contributing to this suboptimal performance could include the complexity and nonlinearity of the data, the presence of outliers or noise, or inadequate configuration of the LSTM model's architecture and parameters. The poor performance of the LSTM model raises questions about its suitability for predicting the number of sessions accurately, given the specific dataset and features used in this study.

The comparison between normalized and non-normalized data revealed no significant difference in the performance metrics for all models. Normalizing the input features did not have a substantial impact on the predictive accuracy of the models. This suggests that the range and scale of the input features did not heavily influence the models' ability to predict the number of sessions occurring after a consecutive session batch. Therefore, normalizing the data may not be necessary in this particular context.

Overall, the findings from the experimental evaluation provide valuable insights into the strengths and weaknesses of the linear regression, CNN, and LSTM models for predicting user session behaviour. The limitations observed in

the linear regression model emphasize the need for more sophisticated models that can capture the complex dynamics and nonlinearity of user session data. Future research could focus on exploring advanced machine learning techniques or incorporating additional features to enhance the predictive accuracy of the models. Additionally, investigating alternative approaches, such as ensemble methods or hybrid models combining different architectures, may offer potential avenues for improving the performance in predicting user session behaviour.

Ultimately, the limitations posed by the size of our dataset and the behaviour of Lernavi application users hindered our progress. Unfortunately, a significant portion of the users merely tested the app, resulting in a mere 20-25% of the available data being truly valuable for our analysis.

B. Additional future discussions

In addition, we should also consider the limitations of our dataset and the challenges associated with predicting the exact number of days a user will stay on an application based on their behaviour. Despite testing numerous combinations, we found that our best prediction was still not sufficient. This highlights the inherent difficulty in accurately predicting user behaviour in terms of the duration of app usage.

One potential approach to address this challenge is to first classify users into different personas using a classification model. By creating personas based on various user characteristics, we can better understand and group users with similar behaviour patterns. Once personas are established, we can then employ prediction models specific to each persona to forecast their individual app usage duration. This approach acknowledges that different personas may exhibit distinct patterns of app engagement and require personalized prediction models.

Supporting this idea, it may be straightforward to demonstrate that the learning curve for a language like German is smoother compared to that of mathematics. People tend to experience less stress when learning German, and their motivation to learn the language may stem from a purer desire rather than the pressure to excel in a particular subject like mathematics. By considering these nuanced factors and tailoring prediction models to specific personas, we can potentially improve the accuracy of our predictions regarding user retention on the app.

VI. Conclusions and Future Works

A. Conclusion

We conducted a study to investigate the prediction of the number of sessions a user will engage in on the Lernavi application. Our findings led us to the challenging realization that accurately predicting the future behaviour of a user is a highly complex task. However, we did identify certain features that demonstrated a significant influence on the number of future sessions. Notably, the number of completed sessions, the number of "NEXT" actions taken, and higher mean difficulty of questions were found to be positively correlated with longer user engagement.

Regarding the model performance, both linear regression and CNN models exhibited similar levels of efficiency in predicting the number of sessions. Surprisingly, the LSTM model performed poorly on this dataset, which was unexpected as we initially anticipated it to outperform the other models in this context.

B. Future Works

In the future, we could try to set up and use metadata in our model and follow the study explored in "Metadata Matters in User Engagement Prediction" [2]. We discovered it too late but it would be interesting to make something with a model of deep learning [3] by adopting a method to generate user activity [4] before to increase the size of our dataset.

References

[1] A. A. Mubarak, H. Cao, and W. Zhang, "Prediction of students' early dropout based on their interaction logs in online learning environment," *Interactive Learning Environments*, pp. 1–20, 02 2020.

[2] X. Chen, S. Mitra, and V. Swaminathan, "Metadata matters in user engagement prediction." New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3397271.3401201

[3] M. Volkovs, F. Perez, Z. Cheng, J. Sun, S. Norouzi, A. Wong, P. Jankiewicz, and B. Rho, "User engagement modeling with deep learning and language models." New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: https://doi.org/10.1145/3487572.3487604

[4] T. W. Kim and M. Fischer, "Automated generation of user activity–space pairs in space-use analysis."