Final Project Report: User Online Activity Prediction

1. Introduction

The goal of this project is to predict the online activity hours of users based on historical data. The data includes information about user demographics, such as age, gender, and number of kids, as well as timestamps of their online activities. The objective is to build a machine learning model that can accurately predict the online activity hours of users for a given date.

2. Problem Statement

The problem we are addressing is to predict the online activity hours of users based on their past online activity history and demographic information. This can be useful for various applications, such as resource planning, targeted advertising, and user engagement strategies.

3. Data Exploration and Preprocessing

3.1. Customers Dataset

The Customers dataset contains information about user demographics, including age, gender, and number of kids. I explored the data to gain insights into the distribution of these features.

- The majority of users are less than 30 years old, followed by the age group of 30 to 50, and the smallest group is above 50.
- Most users do not have kids, but there are users with 1, 2, 3, or 4 kids.

3.2. Pings Dataset

The pings dataset contains information about users' online activity, including timestamps of their online presence. I preprocessed the data to calculate the online hours for each user on each day.

4. Data Analysis and Visualization

I visualized the data to gain insights into the online activity patterns.

- Histograms of online hours for both train and test datasets showed that the online hours' distribution is not normal.
- The bar plots of online hours by day showed fluctuations in user activity over different days.

5. Baseline Model

I established a baseline model using a simple method of predicting the online hours for the next day based on the online hours of the previous day. I calculated the Root Mean Square Error (RMSE) for this baseline model.

6. Feature Engineering

I engineered additional features to improve the model's performance:

- I created datetime features such as month, day of the month, day of the year, and week of the year.
- I calculated rolling window mean for different time periods.
- I added lag features to capture historical online hours.

7. Model Development

I used LightGBM as our primary machine learning model for online activity prediction. I trained the model on the engineered features and evaluated its performance using RMSE.

8. Hyperparameter Tuning

I utilized Optuna to perform hyperparameter tuning for the LightGBM model. Optuna helped me find the best combination of hyperparameters that minimized the RMSE.

9. Model Evaluation

The optimized LightGBM model achieved an improved RMSE on the test dataset compared to the baseline model. The optimized model was able to better capture the online activity patterns of users.

10. Recommendations

Based on our findings and the model's performance, I recommend the following:

Utilize the optimized LightGBM model to predict online activity hours for users in the future.
Continuously collect data and retrain the model to adapt to changing user behaviors.
Consider incorporating additional features, such as user engagement metrics or external factors, to further improve the model's accuracy.
Explore other machine learning models and algorithms to compare their performance against the LightGBM model.
Conduct A/B testing to evaluate the impact of the model's predictions on business outcomes and user engagement.

11. Conclusion

In conclusion, I successfully developed a machine learning model to predict the online activity hours of users based on their historical activity and demographic information. The optimized LightGBM model outperformed the baseline model and can be used as a valuable tool for resource planning and user engagement strategies. Continuously updating and refining the model will ensure its relevance and accuracy in the dynamic online landscape.