Online activity prediction model

Predicting User Behavior for Enhanced Engagement and Resource Allocation

Introduction

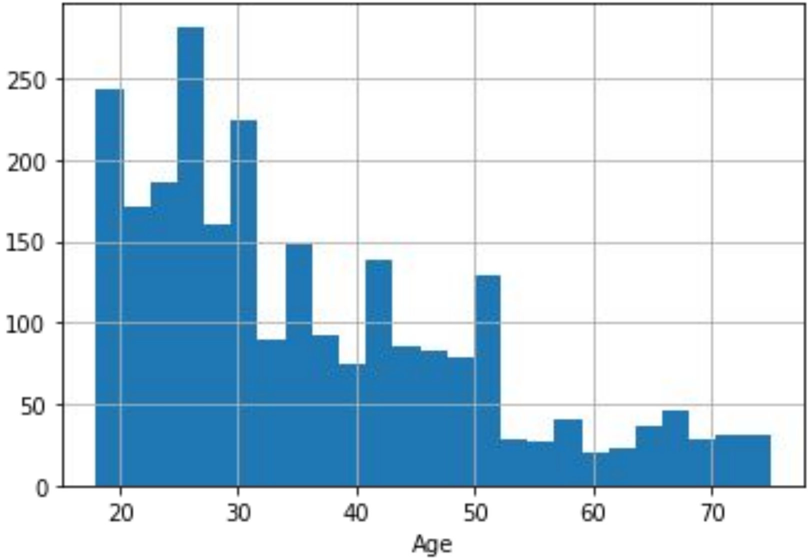
Objective: Predicting user online activity to help companies respond proactively to user demands and offer personalized experiences.

By accurately forecasting user online hours, businesses can allocate resources efficiently and enhance customer satisfaction.

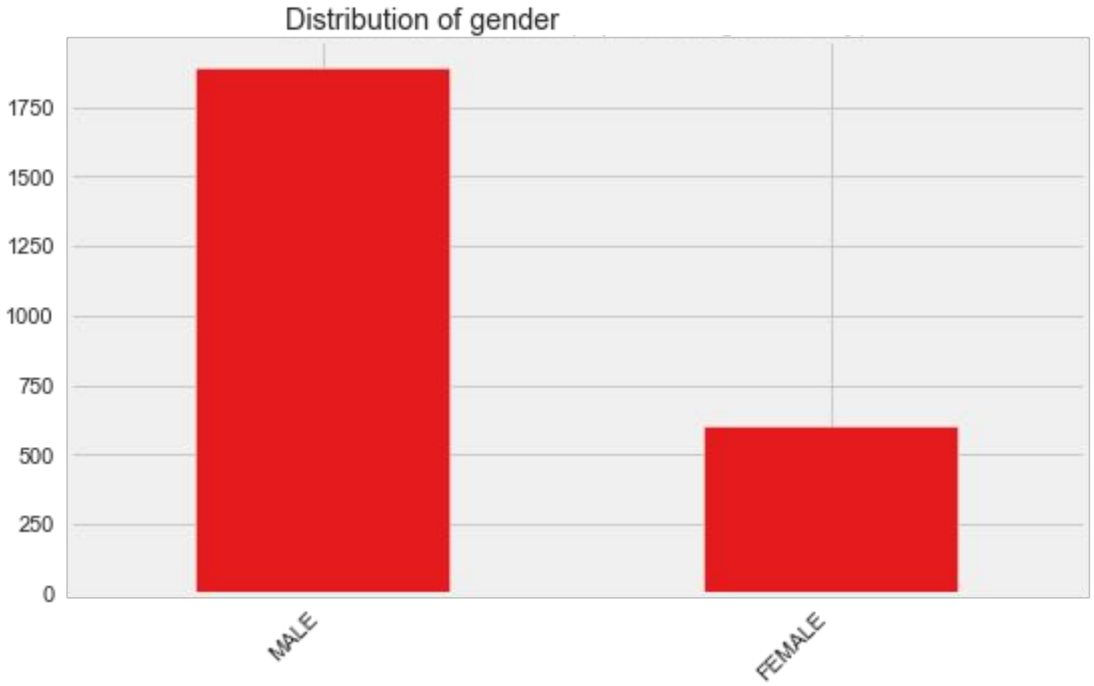
Data Exploration

Customers Dataset:

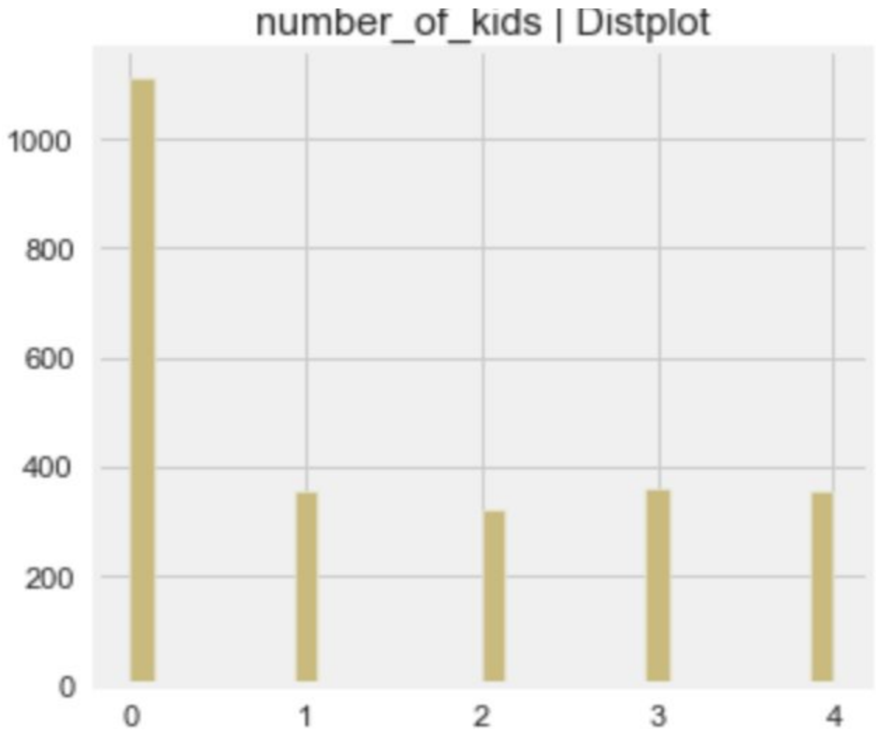
- **Age**: provides insights into their age demographics.



- **Gender**: Allows us to analyze the distribution of male and female customers.



- **Number of Kids**: We have data on the number of kids for each customer, which helps understand the family size of our user base.



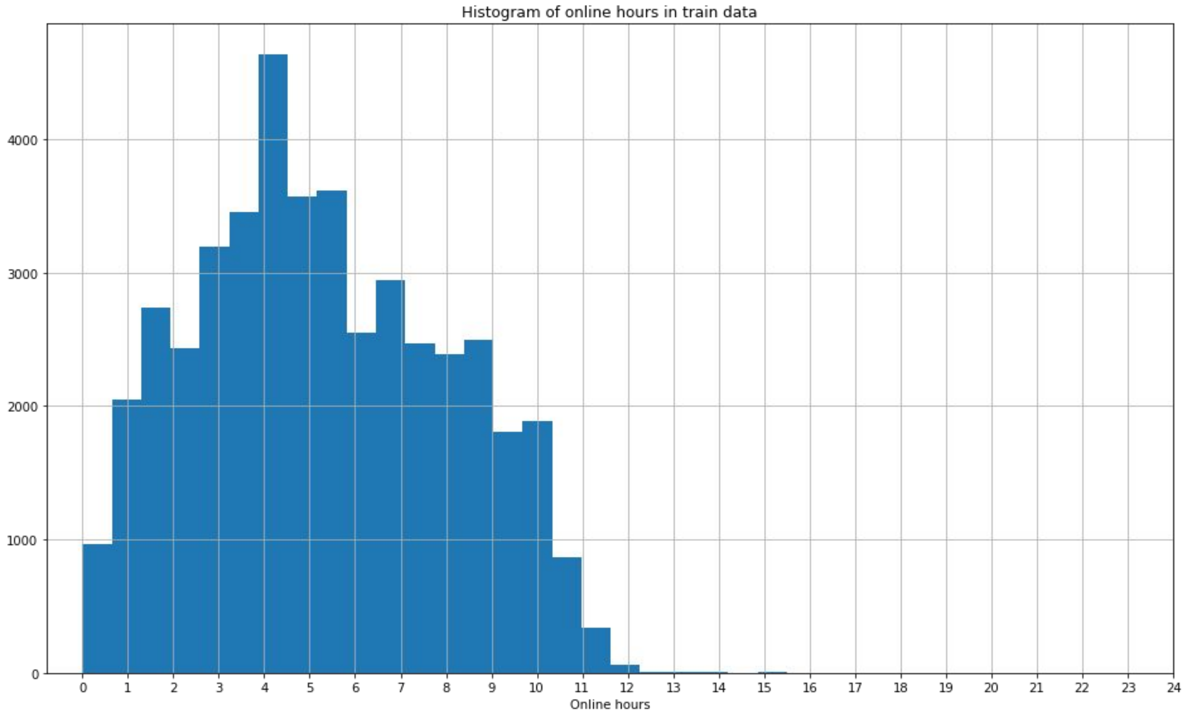
Data Preprocessing

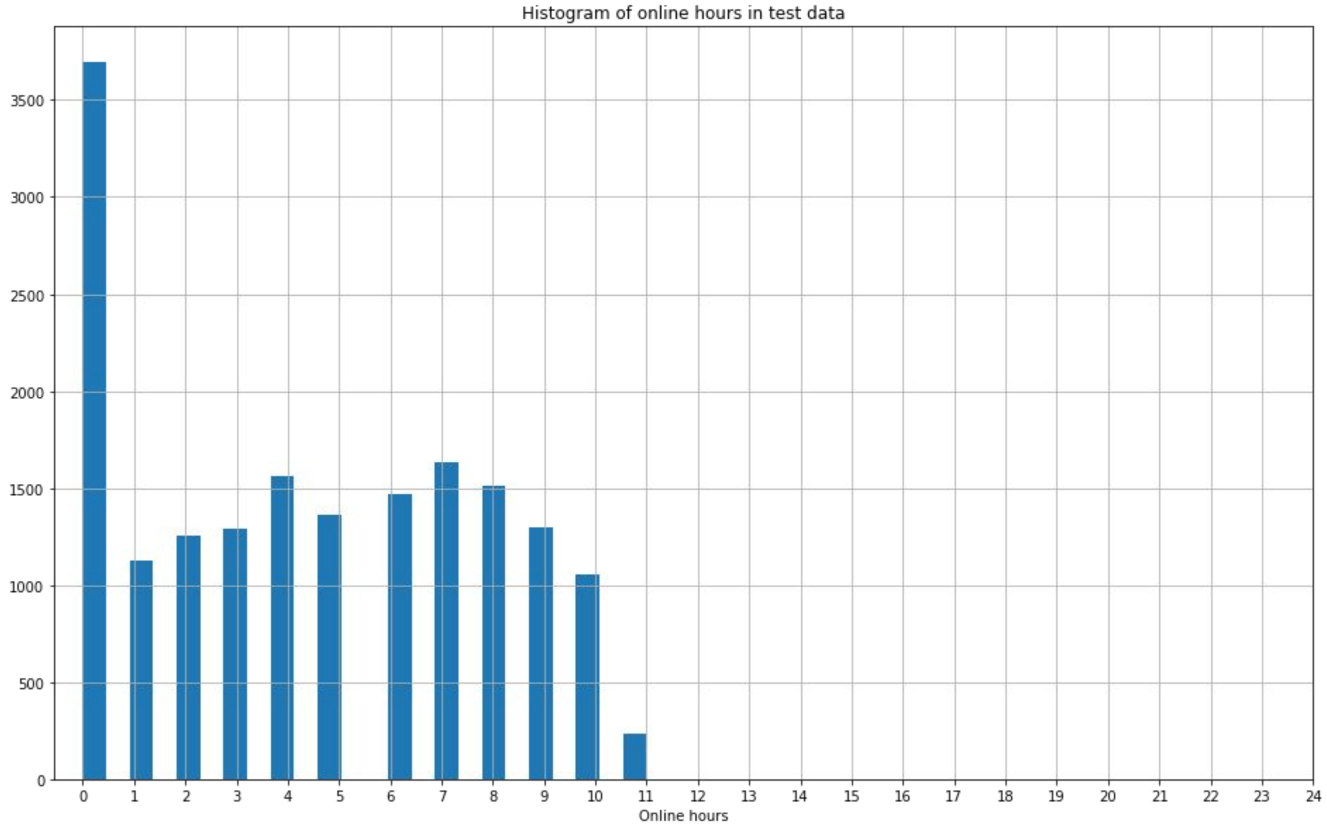
- Converting Timestamps to Datetime
- Removing Duplicates
- Sorting by Timestamp
- Calculating Online Hours
- Handling Missing Values

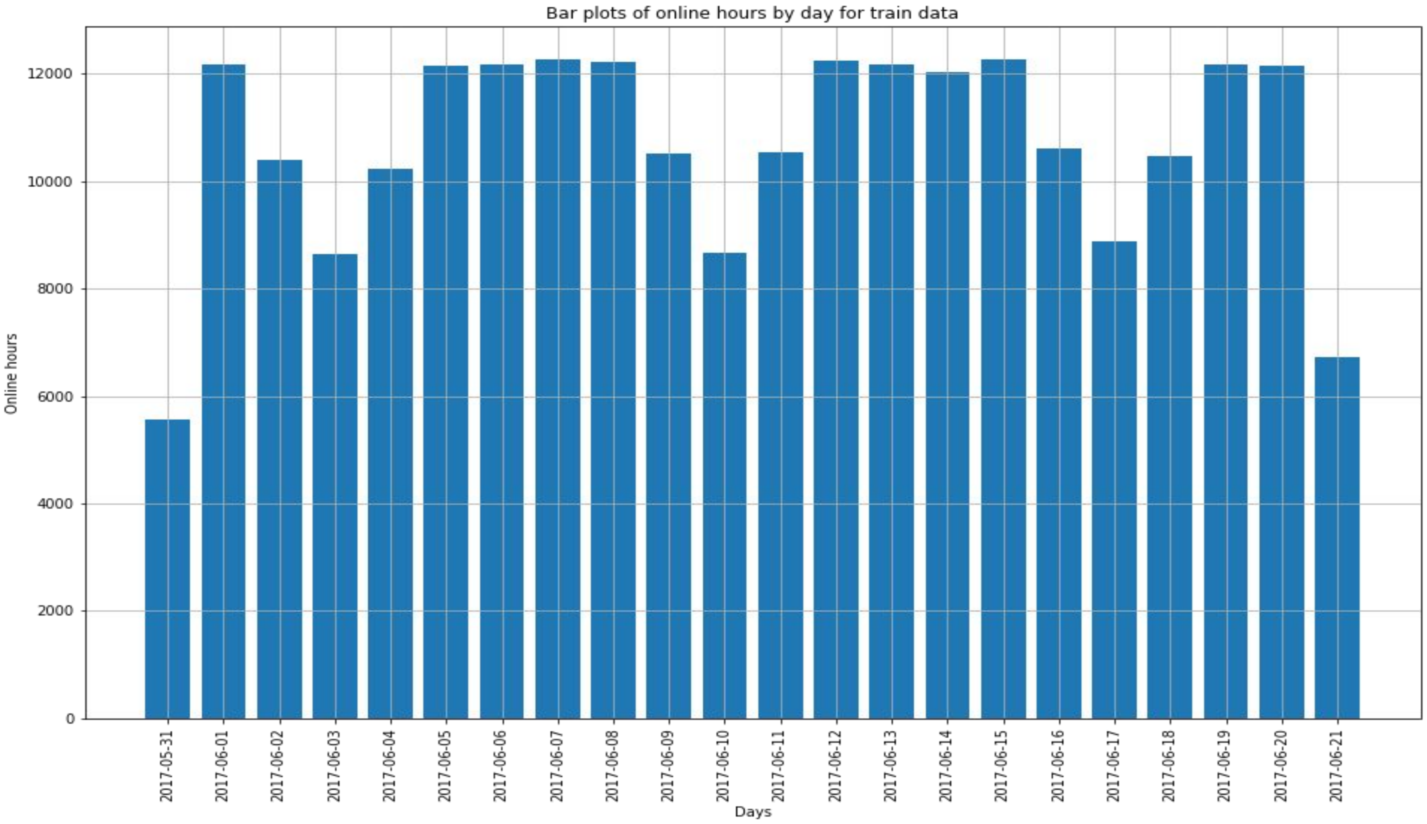
Calculation of Online Hours

- Grouping by User and Date
- Calculating Time Differences
- Converting Timestamp Differences to Hours
- Handling Short Sessions

Data Analysis and Visualization




Histogram of online hours in test data
0
500
1000
1500
2000
2500
3000
3500
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Online hours


Bar plots of online hours by day for train data
Online hours
0
2000
4000
6000
8000
10000
12000
2017-05-31
2017-06-01
2017-06-02
2017-06-03
2017-06-04
2017-06-05
2017-06-06
2017-06-07
2017-06-08
2017-06-09
2017-06-10
2017-06-11
2017-06-12
2017-06-13
2017-06-14
2017-06-15
2017-06-16
2017-06-17
2017-06-18
2017-06-19
2017-06-20
2017-06-21
Days

Baseline Model

- **Data Preparation**: Splitting the dataset into training and validation sets based on chronological order. The training set includes data up to a certain day, and the validation set contains data for the next day.
- **Feature Selection**: For the baseline, I use only one feature - the online hours of the previous day (lag 0).
- **Model Training**: I train a basic regression model (e.g., LGBMRegressor) using the training set.
- **Prediction**: Using the trained model, I predict the online hours for the next day (day + 1) based solely on the online hours of the previous day (day).
- **Evaluation**: I evaluate the model's performance by comparing the predicted online hours with the actual online hours in the validation set.
- **Root Mean Squared Error (RMSE):** To measure the model's accuracy, I calculate the RMSE between the predicted and actual online hours. A lower RMSE indicates a better-performing model.

Feature engineering

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

- Datetime features (month, day of the month, day of the year, etc.).
- Rolling window mean for different time periods..
- Lag features to capture historical online hours.

Model Development

Overview of **LightGBM**:

- **Gradient Boosting**: LightGBM is based on the gradient boosting technique, which combines multiple weak learners (decision trees) to create a strong predictive model.
- **Lightweight and Fast**: LightGBM is designed to be memory-efficient and offers faster training speed compared to other gradient boosting libraries.
- **Leaf-Wise Splitting**: LightGBM uses a leaf-wise tree growth strategy, focusing on nodes that reduce the loss the most, resulting in better performance.
- **Histogram-Based Approach**: It employs histogram-based algorithms to bucket continuous features, further reducing computation time.

Training Process:

Splitting and Growing Trees: The algorithm iteratively creates decision trees, with each tree aiming to correct the errors of the previous one.

Gradient Descent Optimization: LightGBM uses gradient descent optimization to minimize the loss function and improve the model's predictive capabilities.

Evaluation Process:

Cross-Validation: I split the data into training and validation sets for cross-validation purposes. This allows me to evaluate the model's performance on unseen data.

Model Evaluation Metrics: I measure the model's accuracy using evaluation metrics such as Root Mean Squared Error (RMSE) for regression tasks.

Hyperparameter Tuning

Optuna is an automated hyperparameter optimization library that helps me fine-tune my LightGBM model for optimal performance.

Optuna for **Hyperparameter Tuning**:

- **Objective Function**: Optuna requires an objective function to optimize. In my case, the objective function is the Root Mean Squared Error (RMSE) between the predicted and actual online hours.
- **Hyperparameter Search Space**: I define a search space for hyperparameters, specifying their possible ranges and distributions.
- **Optimization Algorithm**: Optuna employs various optimization algorithms to find the best set of hyperparameters efficiently.
- **Trials**: Optuna conducts multiple trials, each representing a different combination of hyperparameters.

Optimized Hyperparameters for the Model:

The hyperparameters that yield the best performance for the LightGBM model are as follows:

- Number of Estimators: 2200
- Learning Rate: 0.0136
- Maximum Depth: 11
- Number of Leaves: 50
- Subsample for Bin: 273000
- Minimum Child Weight: 8.25e-05
- Minimum Child Samples: 30

Model Evaluation

- RMSE on **baseline** model:

  2.90213

- RMSE achieved by the **optimized** LightGBM model:

  2.50068

Recommendations and Model Deployment

Recommendations:

- **Proactive Resource Allocation**: With accurate predictions of user online hours, businesses can proactively allocate resources based on anticipated demand. This optimizes server usage, reduces costs, and ensures smooth customer experiences.
- **Personalized User Experiences**: Leveraging predictions, companies can offer personalized content and promotions during users' active hours, maximizing engagement and satisfaction.
- **Customer Retention Strategies**: Understanding user activity patterns can guide customer retention efforts. Companies can design retention strategies targeting specific user segments to improve long-term user engagement.

Model Deployment for Real-Time Predictions:

- **Real-Time Predictions**: Deploying the optimized model for real-time predictions enables businesses to deliver timely and accurate recommendations to users.
- **Scalability and Efficiency**: LightGBM's lightweight nature makes it scalable, allowing real-time predictions even in high-traffic scenarios.
- **Integration with Existing Systems**: The model can be seamlessly integrated into existing systems, making it easy to incorporate into various applications and platforms.

Insights from Model Performance and Recommendations:

- By deploying the optimized model, companies can unlock valuable insights into user behavior and provide personalized experiences, ultimately enhancing customer satisfaction and retention.

Conclusion

Key Findings and Results:

- **Accurate Online Activity Prediction**: The developed online activity prediction model based on LightGBM achieves precise forecasts of user online hours, outperforming the baseline model.
- **Influential Factors**: Feature importance analysis highlights key factors driving user online activity, enabling businesses to gain insights into user behavior patterns.
- **Proactive Resource Allocation**: Accurate predictions empower businesses to allocate resources proactively, ensuring efficient server usage and cost optimization.
- **Personalized User Experiences**: Leveraging predictions, companies can offer personalized content and promotions during users' active hours, boosting user engagement.

Model's Potential Impact on Business Outcomes:

- **Improved Customer Satisfaction**: Personalized experiences and timely content delivery enhance customer satisfaction, leading to improved loyalty and retention.
- **Enhanced User Engagement**: By offering targeted recommendations during peak online hours, businesses can increase user engagement and interaction.
- **Resource Optimization**: Proactive resource allocation based on predictions minimizes server downtime and enhances overall operational efficiency.
- **Data-Driven Decision-Making**: Insights from the model enable data-driven decision-making, guiding businesses to tailor strategies and marketing efforts.

Enhancing User Engagement:

- **Tailored Retention Strategies**: Understanding user behavior allows businesses to design personalized retention strategies, reducing churn rates and fostering long-term user engagement.
- **Competitive Advantage**: Accurate predictions and personalized experiences provide a competitive advantage, distinguishing a business from its competitors in the market.

Conclusion:

The online activity prediction model holds the potential to revolutionize how businesses interact with their users, driving improved customer satisfaction, resource optimization, and overall business success.