“NORDEUS DATA SCIENCE CHALLENGE 2024: PREDICTING USER ACTIVITY FOR ENHANCED RETENTION“

A Machine Learning Approach to Forecasting User Engagement in Top Eleven

Presenter: Konstantinos Soufleros

Date: November 2024

Project Overview

- **Objective:** To predict user activity in the first 28 days after re-registration to Top Eleven, using historical and re-registration data.
- **Why It Matters:** Accurate predictions help Nordeus tailor personalized experiences, leading to increased user retention and engagement.
- **Key Challenge:** Re-engaging users who have uninstalled and later re-registered.

Photo by Lukas ter Poorten on Unsplash

Problem Statement

Core Problem
User retention is crucial in mobile gaming. Re-registered users represent a valuable opportunity for re-engagement.

Target Variable
Number of days a user is active during the first 28 days post re-registration (integer value between 0 and 28).

Business Impact
Predicting user activity enables targeted re-engagement strategies, improving user satisfaction and retention rates.

Project Objectives

- **Predict User Activity:** Develop a robust model to forecast user engagement for the first 28 days after re-registration.
- **Enhance User Retention:** Provide actionable insights to help Nordeus optimize marketing and content strategies for returning players.
- **Deliver Insights:** Identify key features influencing user activity and offer data-driven recommendations for product teams.

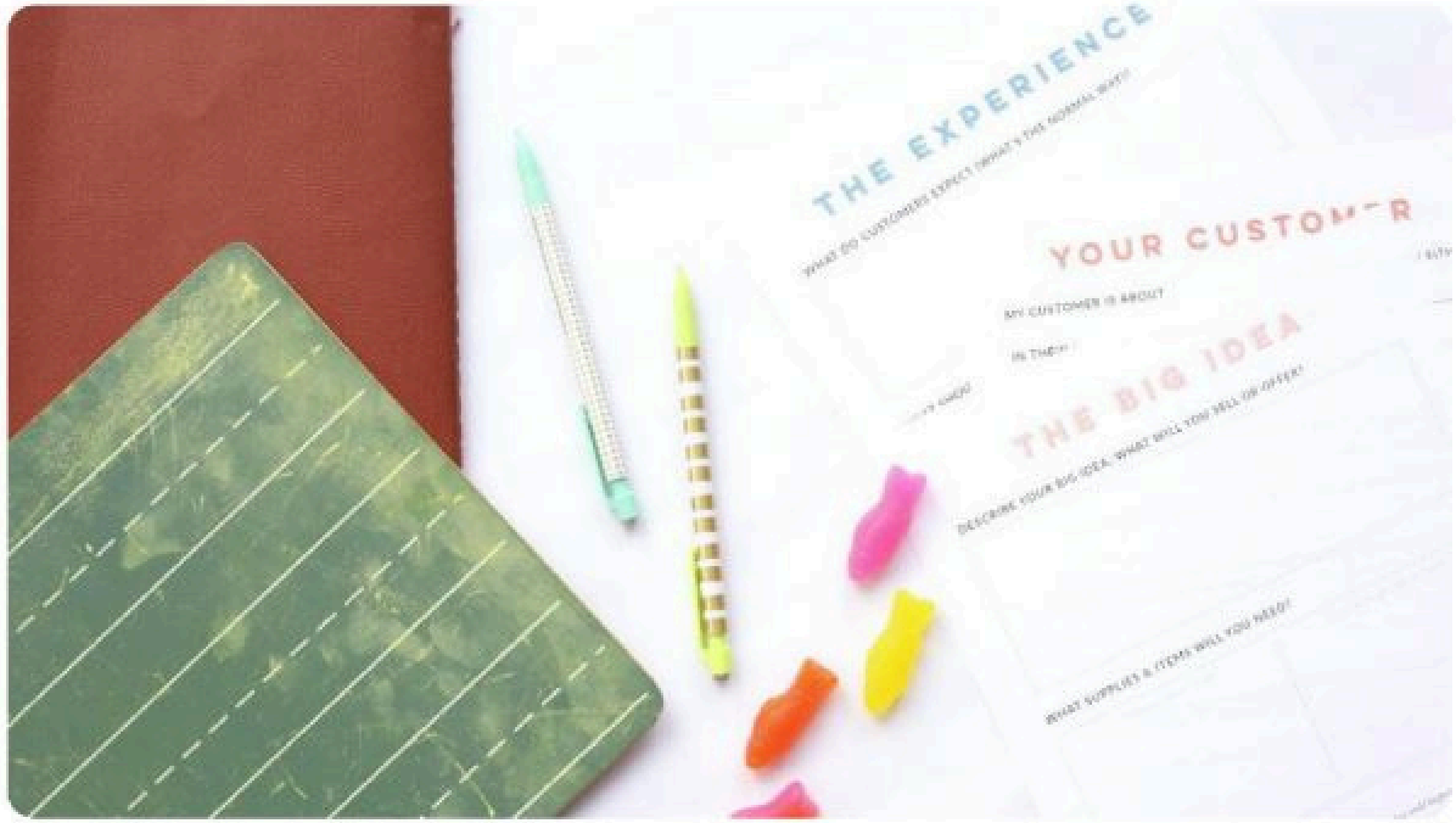


Photo by The CEO Kid on Unsplash

Data Description

- **Datasets Used:** Historical gameplay data and re-registration data: previous_lives_training_data.csv, registration_data_training.csv, previous_lives_test_data.csv, registration_data_test.csv.

- **Data Size:** Approximately 250,000 rows across all datasets, combining both training and test data.

- **Key Variables:** Features include user engagement metrics, re-registration details, and historical behavior data.

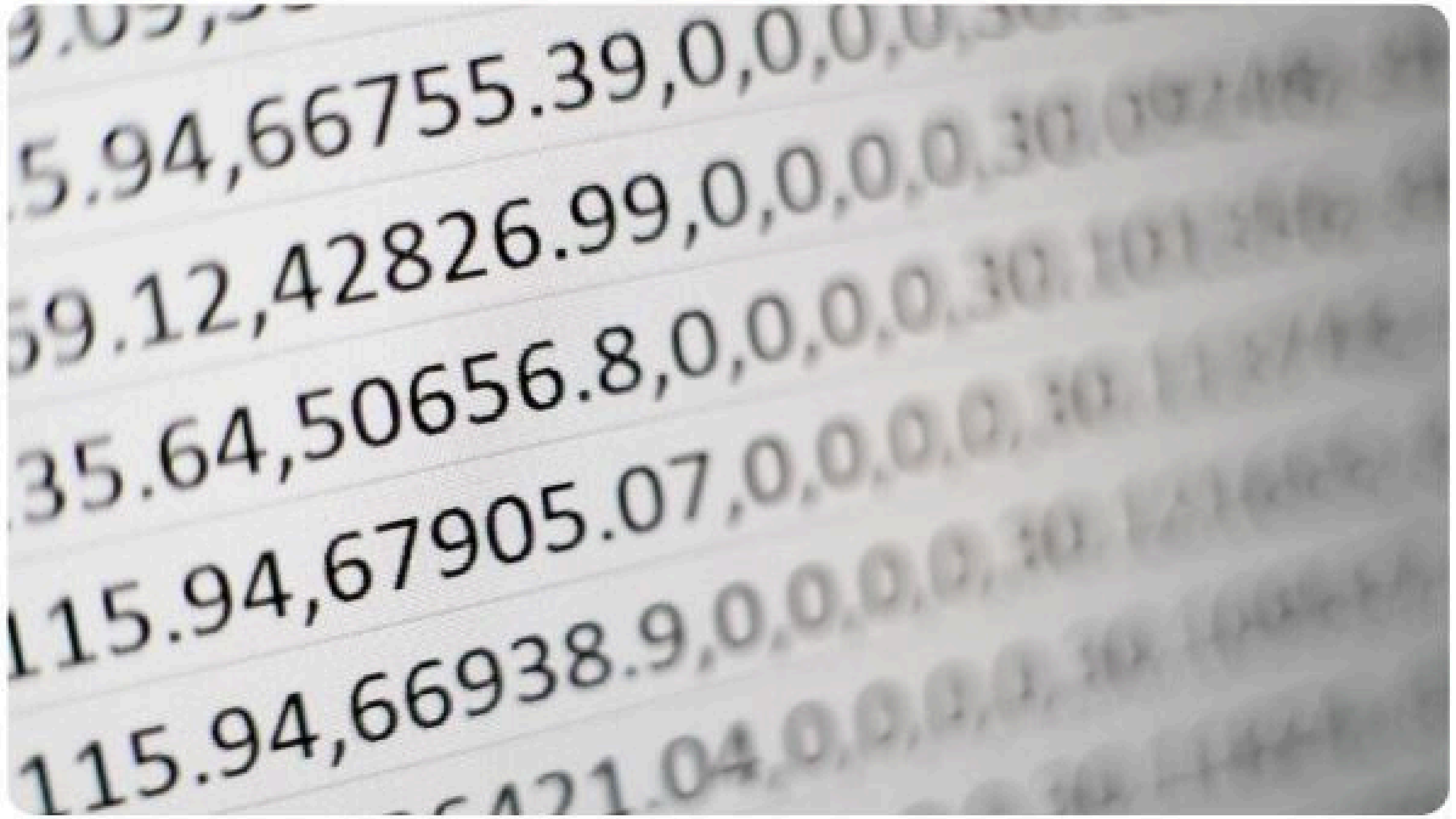


Photo by Mika Baumeister on Unsplash

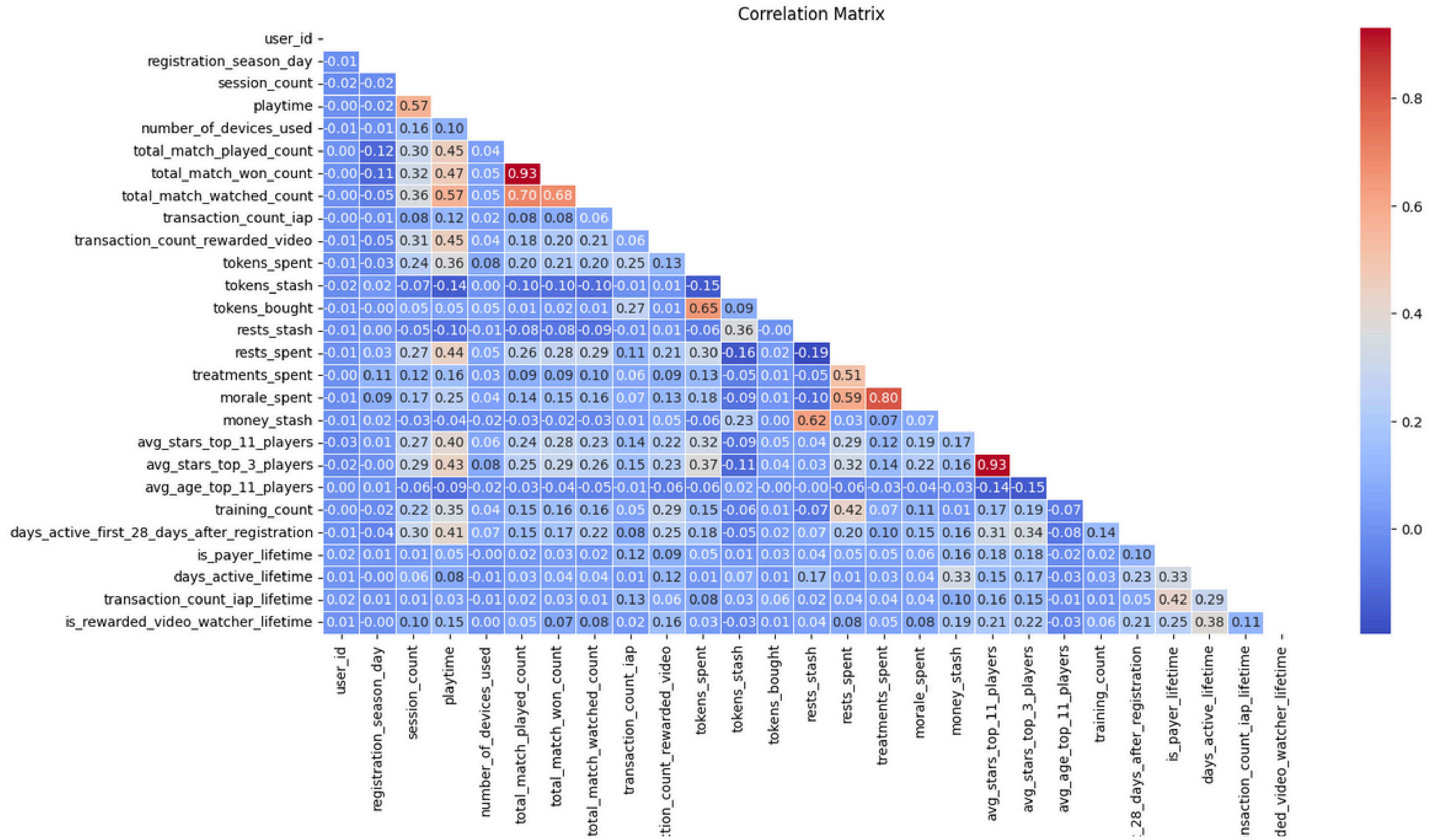
Data Preprocessing and Feature Engineering

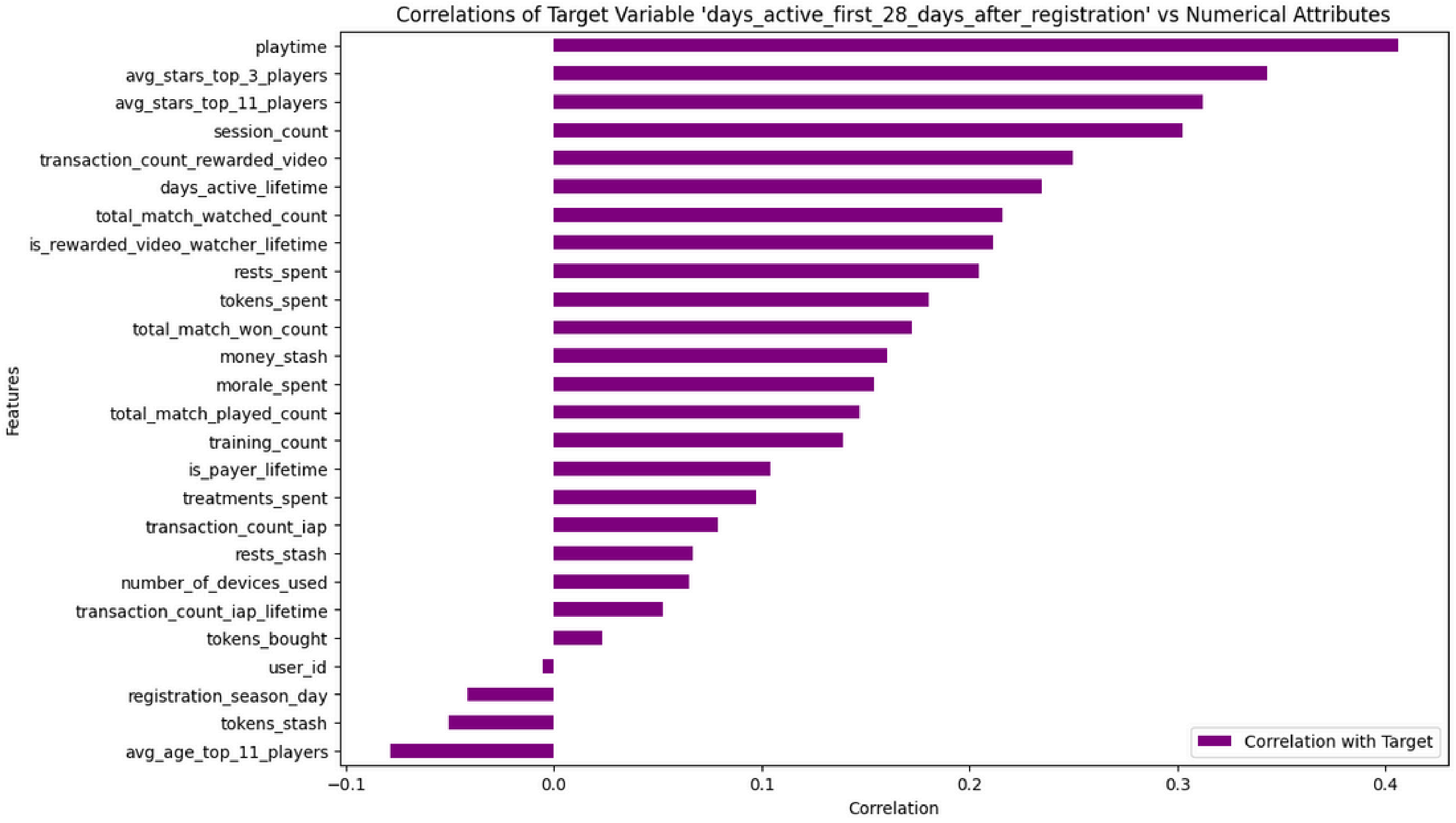
Data Cleaning
Addressed missing values, performed datetime conversion, and encoded categorical features.

Feature Engineering
Aggregated user data, merged historical and re-registration datasets, and mapped countries to continents.

Outlier Handling
Retained outliers representing valid, high-engagement user behavior.

Exploratory Data Analysis (EDA)





- **Key Insights:** High correlation observed between features indicating potential multicollinearity: total_match_played_count & total_match_won_count (0.93), avg_stars_top_11_players & avg_stars_top_3_players (0.93), treatments_spent & morale_spent (0.80).

- **Engagement Patterns:** Strong engagement observed for users with higher playtime and session counts.

- **Visualization:** Correlation heatmap and distribution plots for key features were utilized.

Modeling Approach

Baseline Models
Linear Regression, Ridge, Lasso were used as initial benchmarks.

Tree-Based Models
Random Forest, XGBoost, and LightGBM were explored for capturing non-linear relationships.

Deep Learning and Ensembles
Feedforward Neural Network (FFNN) and ensemble methods like Voting and Stacking Regressor.

Hyperparameter Tuning with Optuna

- **Optimization Strategy:** Bayesian optimization using Optuna for FFNN and LightGBM models.
- **Hyperparameters Explored:** Learning rate, number of layers, batch size, and tree depth among others.
- **Results:** FFNN achieved an MAE of 5.45; LightGBM tuning reduced MAE to 5.91.

Photo by Martin W. Kirst on Unsplash

Model Performance and Comparison

Best Model
Feedforward Neural Network (FFNN) achieved an MAE of 5.45, outperforming all other models.

Tree-Based Model Performance
LightGBM was the top-performing tree-based model with an MAE of 5.95.

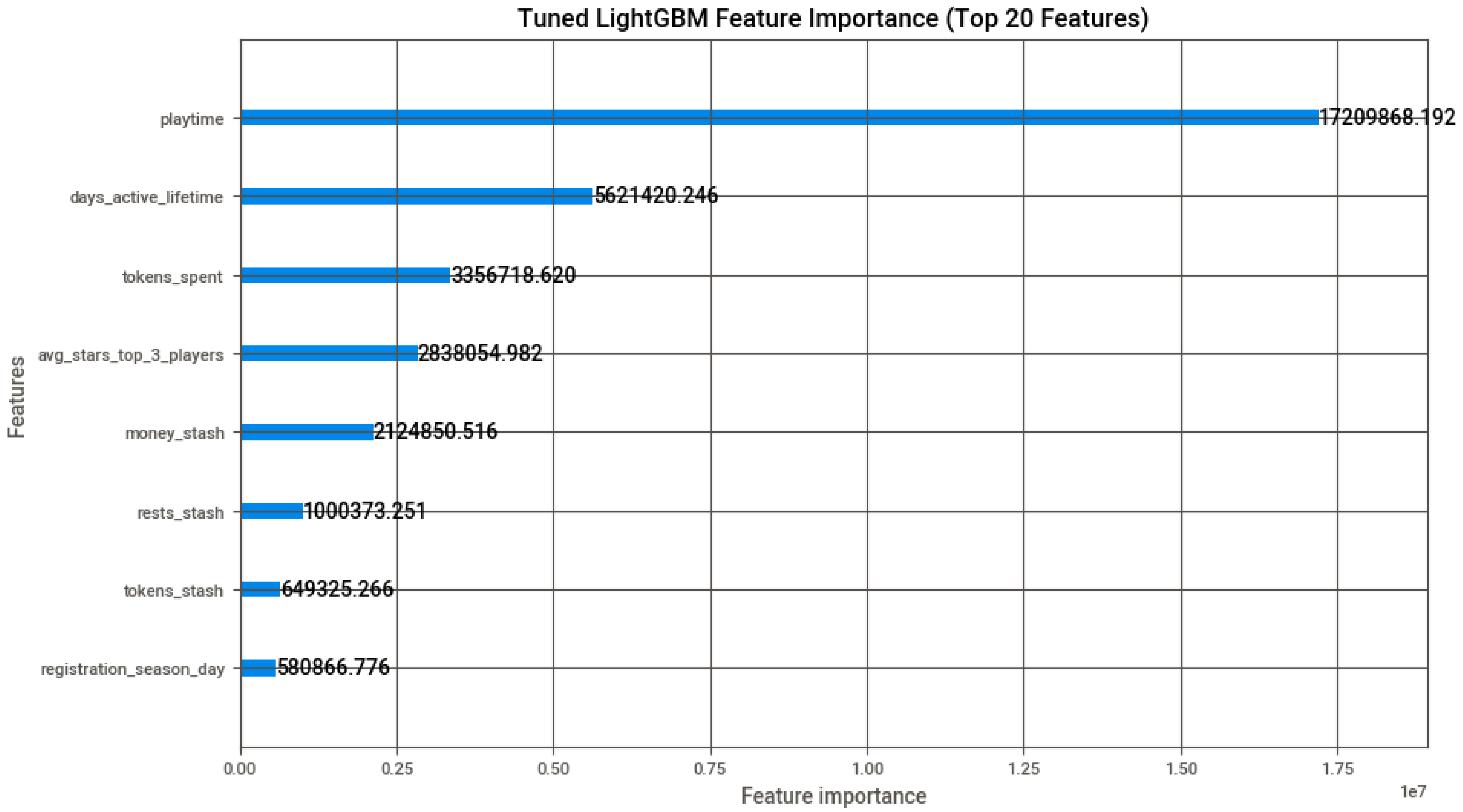
Overall Results
XGBoost and Stacking Regressor had higher MAE, with FFNN emerging as the most accurate predictor.

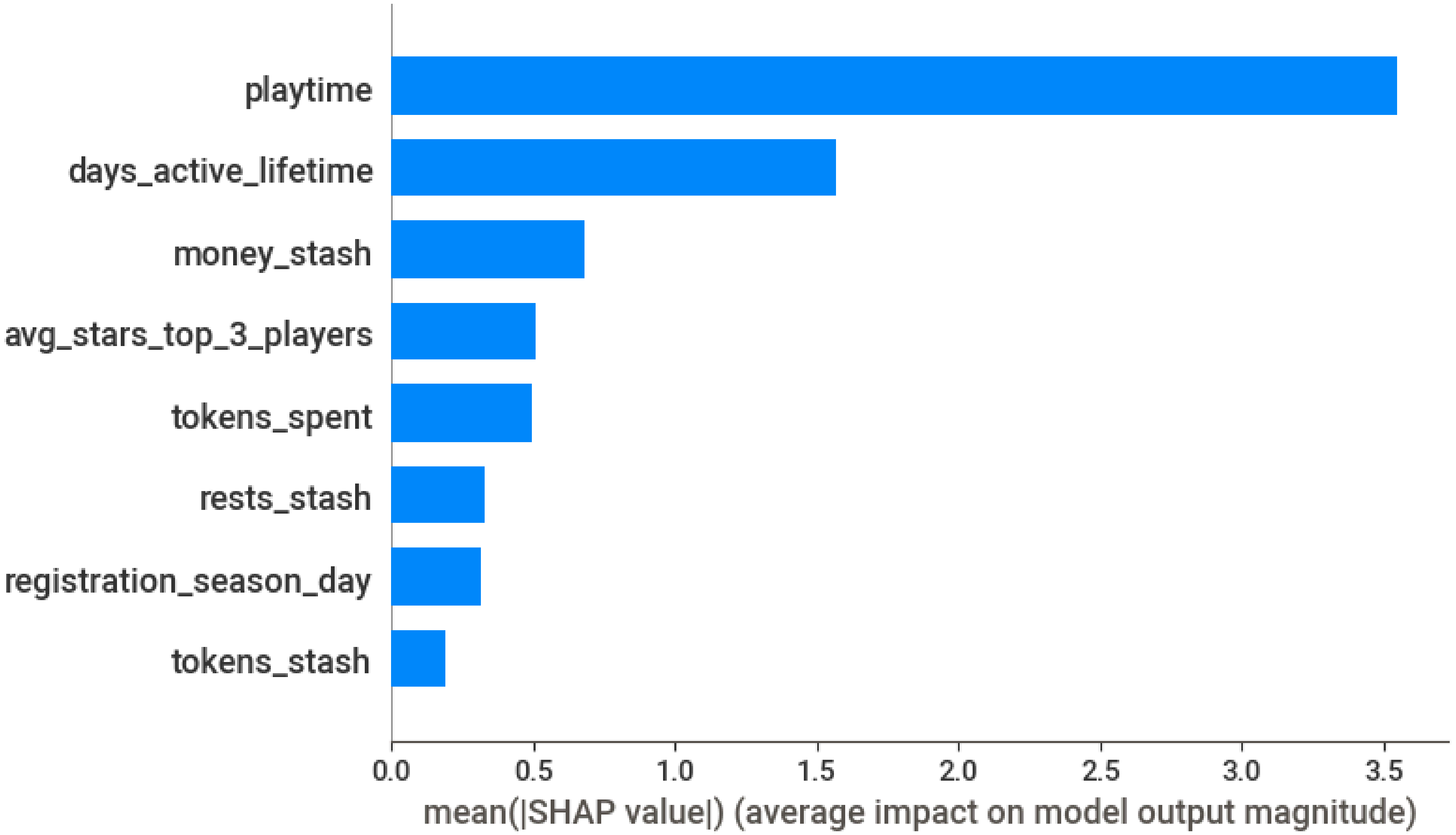
Model Interpretability

- **SHAP Analysis:** Identified key features influencing predictions, such as playtime and session count.

- **Feature Importance Plot:** LightGBM feature importance indicates 'playtime' as the top predictor.

- **SHAP Summary Plot:** SHAP values highlight 'playtime' and 'days_active_lifetime' as the most impactful features.





Recommendations

Personalized Content
Use predictions to offer customized experiences, targeting users with high predicted engagement.

Targeted Marketing
Focus re-engagement efforts on users predicted to have lower activity, with special offers and incentives.

Feature Optimization
Leverage feature insights to enhance game mechanics and increase user satisfaction.

Future Work

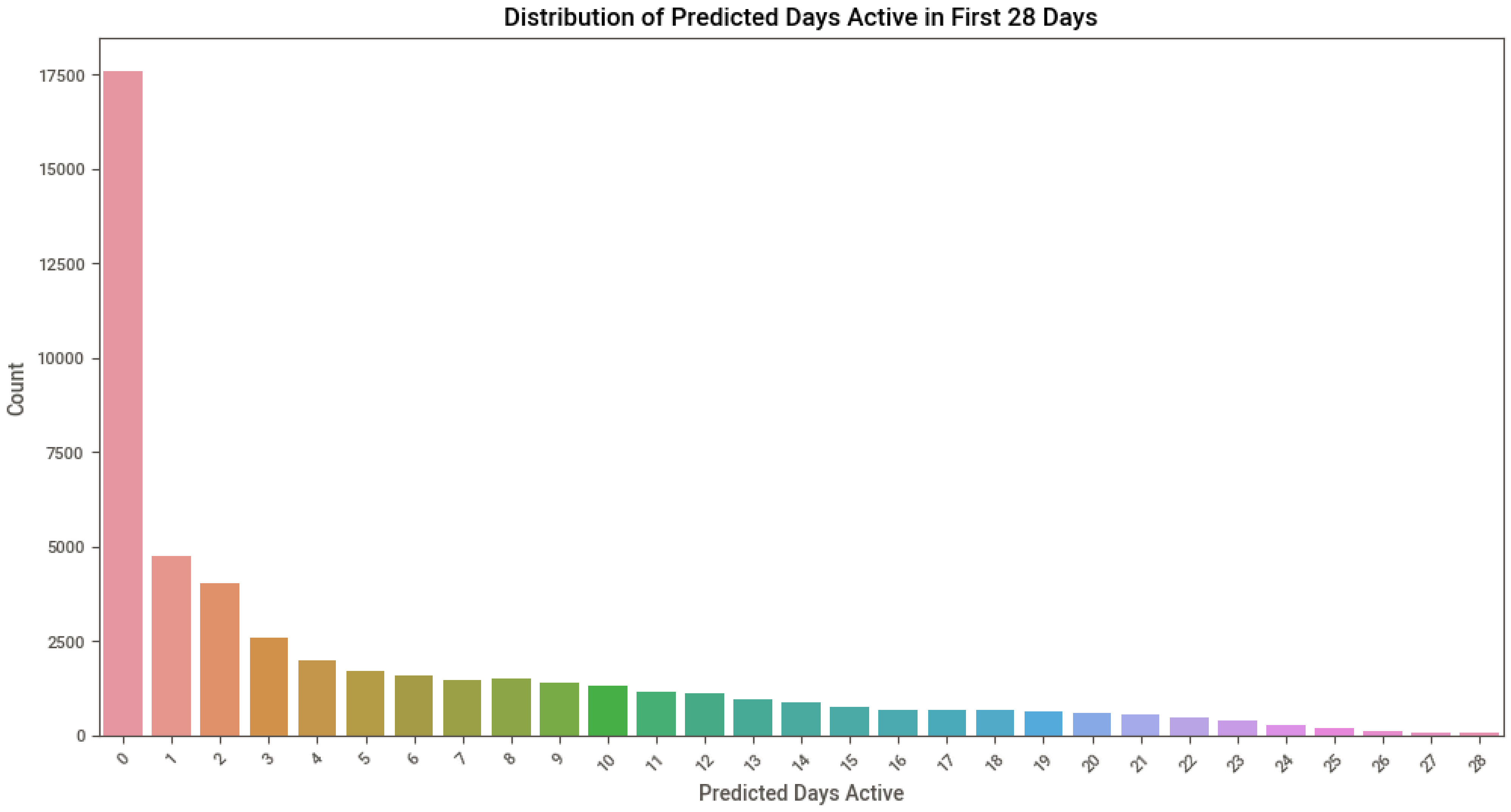
Incorporate More Data
Include additional features from user activity logs and external sources.

Automate Retraining
Implement an MLOps pipeline for automated model retraining with new data.

Real-Time Deployment
Deploy the model as a real-time API service for integration into production systems.

Key Takeaways

- **Best Model:** The tuned Feedforward Neural Network (FFNN) achieved the lowest MAE of 5.45, indicating strong predictive performance.
- **Business Impact:** Accurate predictions enable targeted re-engagement strategies, improving user retention and satisfaction.
- **Scalable Solution:** The project lays the foundation for future enhancements, including real-time predictions and automated retraining.



Q&A

Any questions? I’m happy to discuss further.

- **Feedback Welcome:** We value your insights and look forward to any suggestions for improvement.
- **Thank You:** Thank you for your time and attention throughout this presentation.
- **Contact Information:** LinkedIn: https://www.linkedin.com/in/konstantinos-soufleros Email: soufleros.kostas@gmail.com GitHub: https://github.com/kostas696

Photo by Abraham Serey on Unsplash