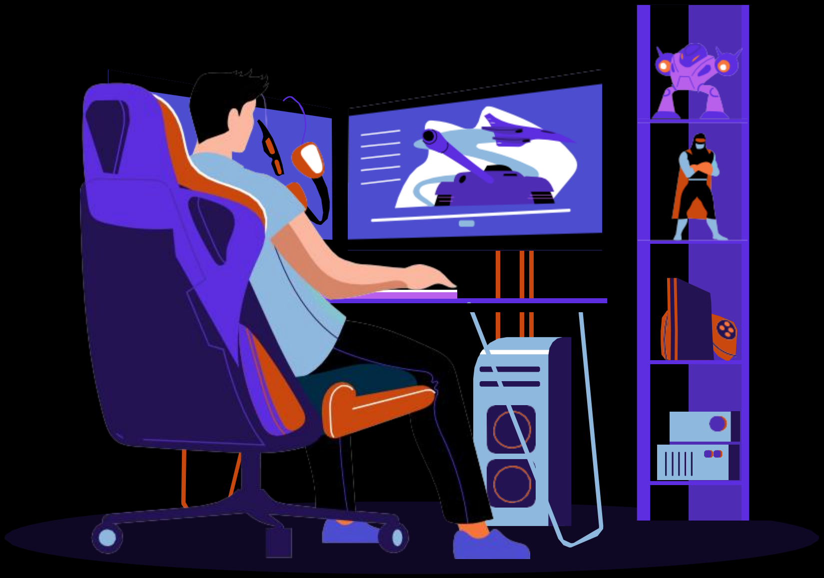
RETENTION PREDICTION FOR F2P GAME

Mahrukh



Introduction and Data Overview

User retention is a crucial metric for mobile game success and monetization. By accurately modeling retention, game companies can target interventions and offers to improve sticky-ness. They can also better align business plans to expected player lifespans.

Project Goal

Build a model and accurately predict whether a player will retain after 14 days

Data Overview

The dataset provided includes usage telemetry and attributes for the first 14 days after install for recently acquired players. Features capture profile information like country and device type, as well as key engagement metrics like playtime, sessions, spend, and churn risk.

Prepping Data

Handled missing values and ensuring a proper train-test split strategy with stratification based on retention columns.

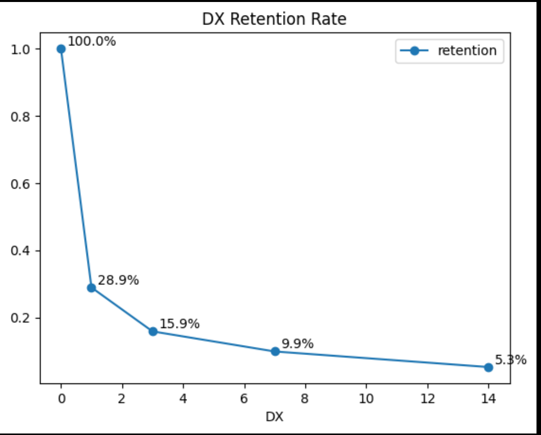
Feature Engineering

This includes target encoding, creating new game-related features, introducing binary retention and conversion features, computing percentage changes and rate changes, and evaluating progress in game chapters.

RETENTION RATE

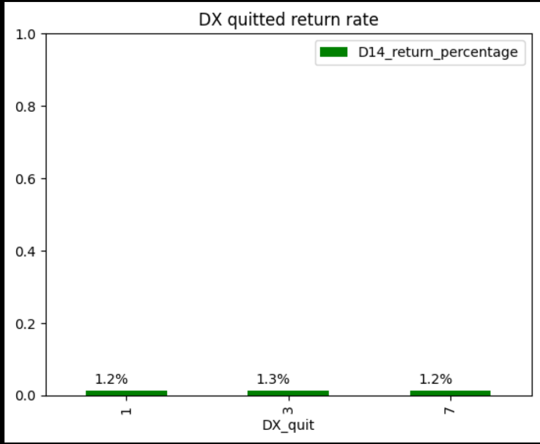
Although model accuracy improves with time, the trade off is missing the opportunity to retain more players who have already dropped off. There is more value in taking earlier action, even if based on less data

DX Retention Rate



Waiting longer means having more data points so can predict 14 day retention more accurately However, there are fewer players still active later on. Less players to impact by taking action closer to day 14.

DX Quitted Players Return Rate



Very few players (about 2%) who quit before day 14 eventually come back on day 14. This indicates it may not be efficient to focus retention efforts on re-engaging quitters. Retaining existing players is likely a better use of resources than trying to win back players who already quit

Algorithm Selection

Testing different complex and simple models helps determine the best approach for this dataset and business problem. We have chosen Logistic Regression, Random Forest, and XCBoost as the algorithms

Logistic Regression

A good baseline model for binary classification problems like this. It is interpretable and fast to train. Useful benchmark.

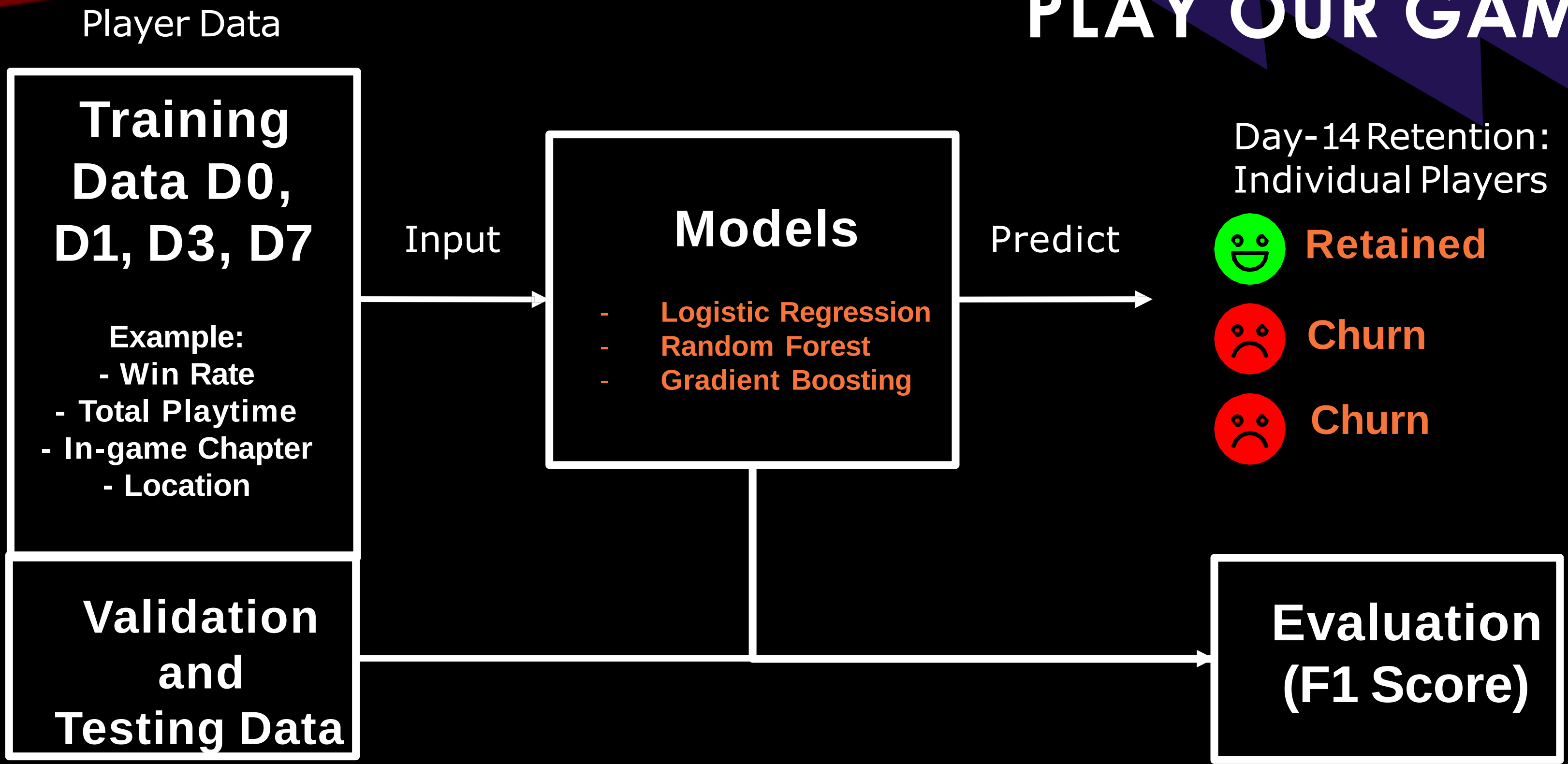
Random Forest

An ensemble method suited for tabular data that can capture nonlinear relationships and handle many input variables. Helps avoid overfitting.

Gradient Boosting

A powerful gradient boosted decision tree algorithm known for high predictive accuracy. Handles imbalanced data well.

PREDICTING WHICH PLAYERS WILL PLAY OUR GAME



How classifying works in detail

Algorithm was used to predict retention_14 for each player

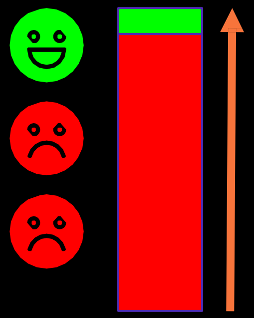
↓

Top 6% of training population was used to determine the retained players

42 % likely to retain

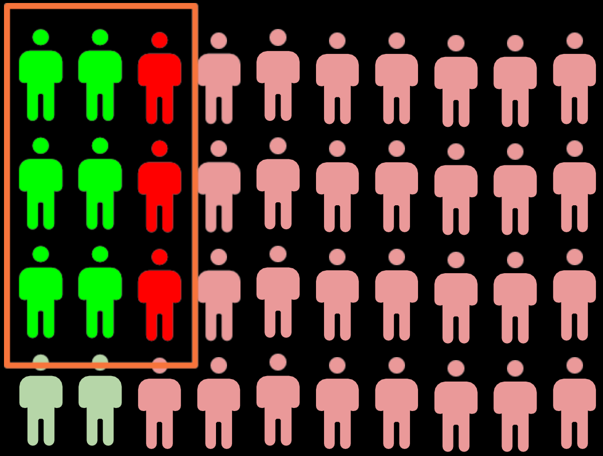
14 % likely to retain

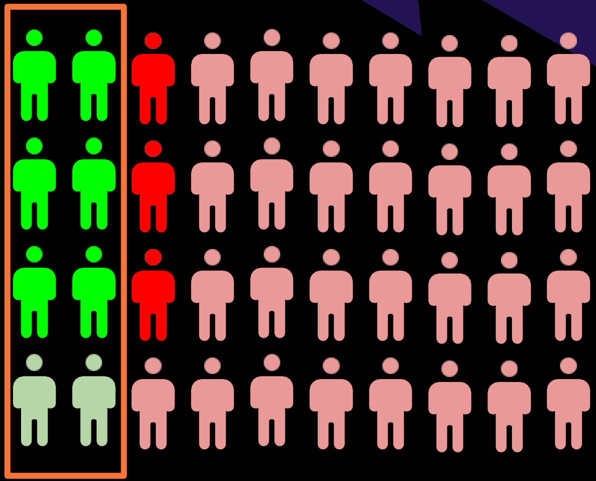
3 % likely to retain



Predicted Retention Cutoff >= 22.6%

The model is measured by Precision and Recall





Precision:
% of correct retention predicted

+

Recall:
% of all retention covered

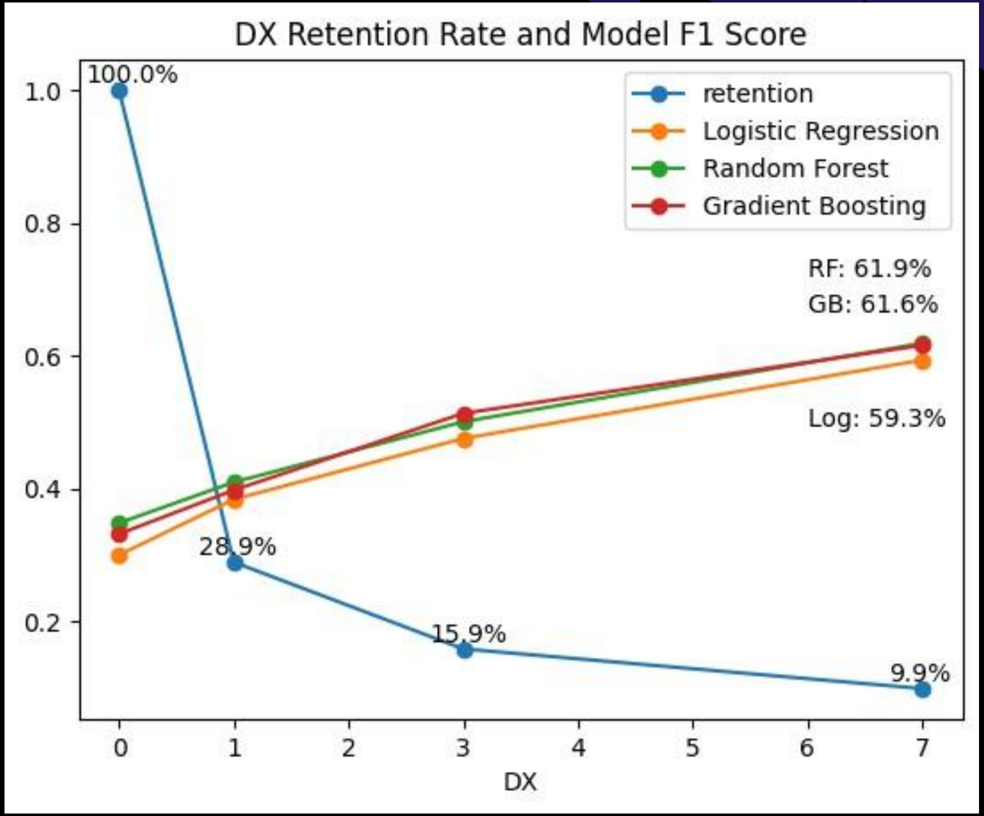
Using day-7 player data might be too late

Model Performance (F1) - D7 data

Logistic Regression	59.3%
Random Forest	61.9%
Gradient Boosting	61.6%

The model will perform better as we collect more player's data.

However, most players will leave after the first day of installation.



Feature Importance

DX_END_ISIN_ CHAPTER_7

0.33

DX_TOTAL_ PL AYTIME_7

0.31

DX_TOTAL_ SESSION_CT_7

0.24

CITY

0.08

Prediction with New Data

Day 14 Retention Likelihood 5.9 %

Testing Data F1_Score 0.62

Key Takeaways and Findings

- For the industry as a whole (or a game of this genre), European is not a market to be neglected.

- Making players retain until day 7 is more important than trying to retain them for day 1, as day 1 total playtime is even less important than most of the other features

- Side Mission Game Percentage seems to be more important than PVP and Campaign. However, PVP win rate is the most important among all the game modes

Recommendations

01 Marketing

Marketing in the European Region for games of this kind

02 Day 7 retention

Develop continuous operation strategy and make the early game experience abundant. Avoid click baiting

03 Side Mission

Focusing on the development of side missions to attract players to retain until day 14

04 Game Balance

Focus on the balance of the PVP mode so players don’t get discouraged because of their upset win rates

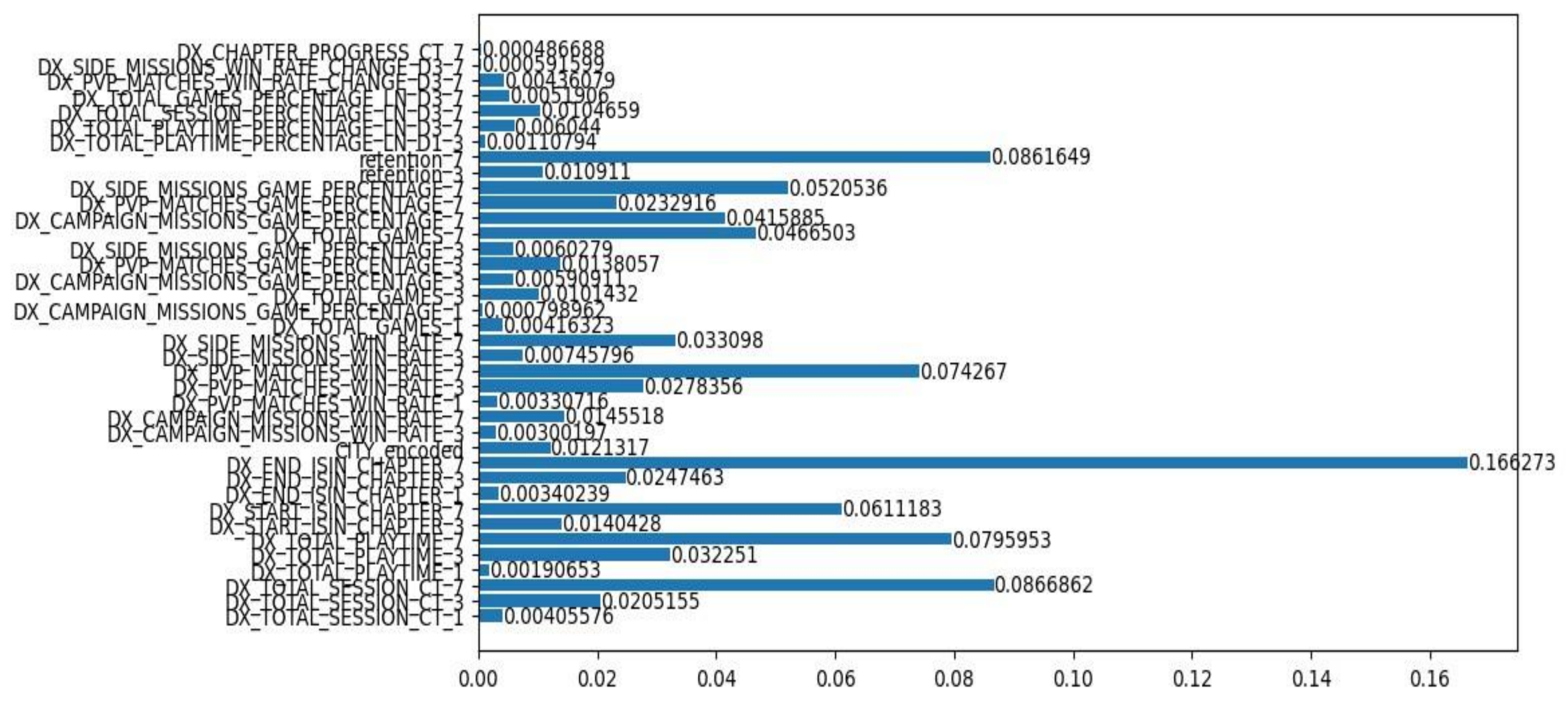
APPENDIX - ALL ALGORITHM RESULTS

Larger F1 score indicates better precision and recall of a model, meaning the prediction of a model is better when F1 score is higher. Below are the F1 score

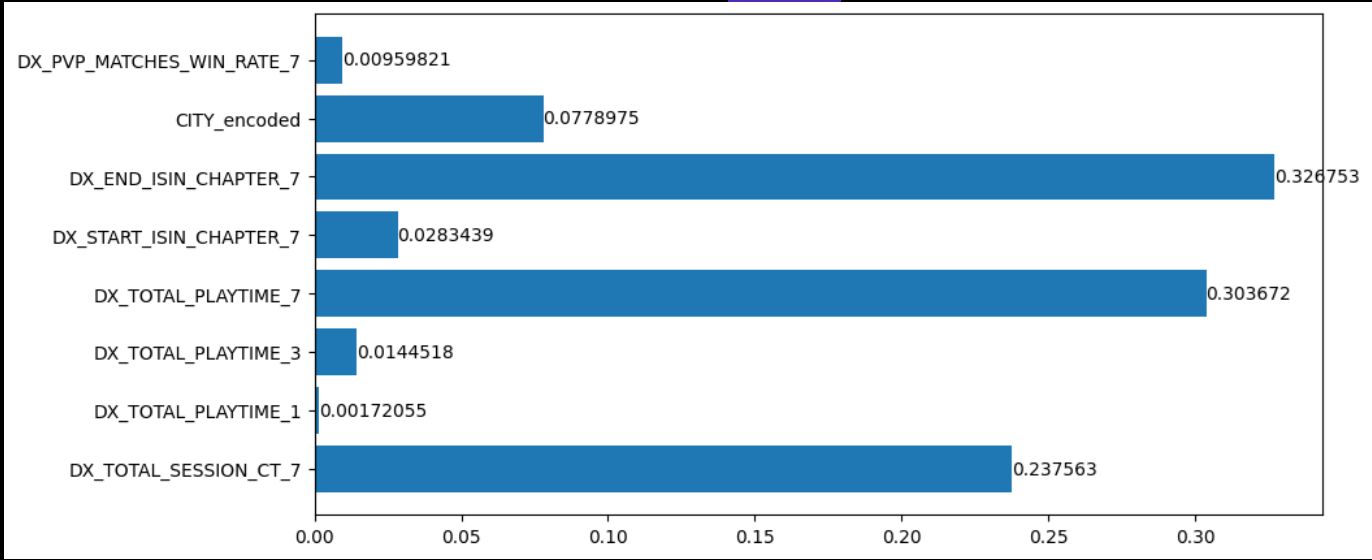
	Day 0	Day 1	Day 3	Day 7
Logistic Regression	30.0%	38.4%	47.5%	59.3%
Random Forest	34.8%	41.0%	50.1%	61.9%
Gradient Boosting	33.2%	39.8%	51.3%	61.6%

* The number shows how day 7 is always better because people have already been retained for 7 days and less noise or uncertainty (a.k.a more data).

APPENDIX - RANDOM FOREST FEATURE IMPORTANCE



APPENDIX - GRADIENT BOOSTING FEATURE IMPORTANCE



APPENDIX - CITY & COUNTRY FEATURE IMPORTANCE

COUNTRY	size	mean
FI	1778	0.091114
MY	185	0.086486
DK	1985	0.084131
NO	1263	0.079968
SE	3452	0.072422
NL	6076	0.058920

k

CITY	size	mean
Tampere	137	0.131387
Frederiksberg	180	0.094444
Helsinki	1244	0.088424
Oslo	362	0.082873
Gothenburg	223	0.076233
Jurong West	133	0.075188
Malmo	146	0.068493
Rotterdam	510	0.066667
Utrecht	182	0.065934
The Hague	213	0.065728
Copenhagen	294	0.064626
Brisbane	866	0.063510
Perth	470	0.055319
Sydney	1081	0.054579