Identification of At-Risk Gamers through Gamers' Engagement Levels

Project Presentation

INTRODUCTION

THE STRAITS TIMES

FOR SUBSCRIBERS

Gaming addiction on the rise among children in S'pore amid pandemic: Counsellors

Counsellors say they have seen a stark increase in reports from parents about their children being hooked on online gaming since the pandemic hit, with the number of cases rising by to 60 per cent. ST PHOTO ILLUSTRATION: GIN TAY

- **Overview of gaming addiction as a global concern**
  - Adverse impact on cognitive, behavioural and emotional well-being
- **Recognition of gaming disorder by World Health Organisation, added to International Classification List in 2022**
  - World Health Organization. (n.d.). *Gaming disorder*. Retrieved 26 August 2024, from https://www.who.int/standards/classifications/frequently-asked-questions/gaming-disorder
- **Gaming addiction trend in Singapore**
  - *MCI's response to PQ on Measures in Place to Manage Gaming Addiction among Youths*. (n.d.). Retrieved 26 August 2024, from https://www.mddi.gov.sg/media-centre/parliamentary-questions/gaming-addiction-among-youths/
  - Teng, H., Zhu, L., Zhang, X., & Qiu, B. (2024). When Games Influence Words: Gaming Addiction among College Students Increases Verbal Aggression through Risk-Biased Drifting in Decision-Making. *Behavioral Sciences*, *14*(8), 699. https://doi.org/10.3390/bs14080699
- **Importance of identifying at-risk gamers**
  - Enable early intervention
- **Objectives of the project**
  - Construct an end-to-end machine learning pipeline to identify at-risk gamers

RECAP OF EXPLORATORY DATA ANALYSIS

- **Dataset**
  - Obtained from Kaggle
- **Purpose of EDA**
  - Understanding the dataset
- **Steps involved**
  - Data extraction
  - Exploration and cleaning
  - Subset analysis and visualization
  - Key findings from the EDA

EXPLORATORY DATA ANALYSIS

df.head()

Target Variable

	PlayerID	Age	Gender	Location	GameGenre	PlayTimeHours	InGamePurchases	GameDifficulty	SessionsPerWeek	AvgSessionDurationMinutes	PlayerLevel	AchievementsUnlocked	EngagementLevel
0	9000	43	Male	Other	Strategy	16.271119	0	Medium	6	108	79	25	Medium
1	9001	29	Female	USA	Strategy	5.525961	0	Medium	5	144	11	10	Medium
2	9002	22	Female	USA	Sports	8.223755	0	Easy	16	142	35	41	High
3	9003	35	Male	USA	Action	5.265351	1	Easy	9	85	57	47	Medium
4	9004	33	Male	Europe	Action	15.531945	0	Medium	2	131	95	37	Medium

df.tail()

	PlayerID	Age	Gender	Location	GameGenre	PlayTimeHours	InGamePurchases	GameDifficulty	SessionsPerWeek	AvgSessionDurationMinutes	PlayerLevel	AchievementsUnlocked	EngagementLevel
40029	49029	32	Male	USA	Strategy	20.619662	0	Easy	4	75	85	14	Medium
40030	49030	44	Female	Other	Simulation	13.53928	0	Hard	19	114	71	27	High
40031	49031	15	Female	USA	RPG	0.240057	1	Easy	10	176	29	1	High
40032	49032	34	Male	USA	Sports	14.017818	1	Medium	3	128	70	10	Medium
40033	49033	19	Male	USA	Sports	10.083804	0	Easy	13	84	72	39	Medium

df.sample(n=5)

	PlayerID	Age	Gender	Location	GameGenre	PlayTimeHours	InGamePurchases	GameDifficulty	SessionsPerWeek	AvgSessionDurationMinutes	PlayerLevel	AchievementsUnlocked	EngagementLevel
38406	47406	31	Female	USA	Sports	0.087878	0	Hard	2	96	31	5	Low
1875	10875	32	Male	Other	Action	10.612119	0	Easy	1	75	63	21	Low
39848	48848	37	Male	Europe	Simulation	22.208582	0	Hard	8	152	4	7	Medium
32650	41650	20	Female	Europe	RPG	20.482825	0	Hard	14	123	81	3	High
19555	28555	24	Male	Other	Sports	10.613646	0	Medium	9	94	37	4	Medium

EXPLORATORY DATA ANALYSIS

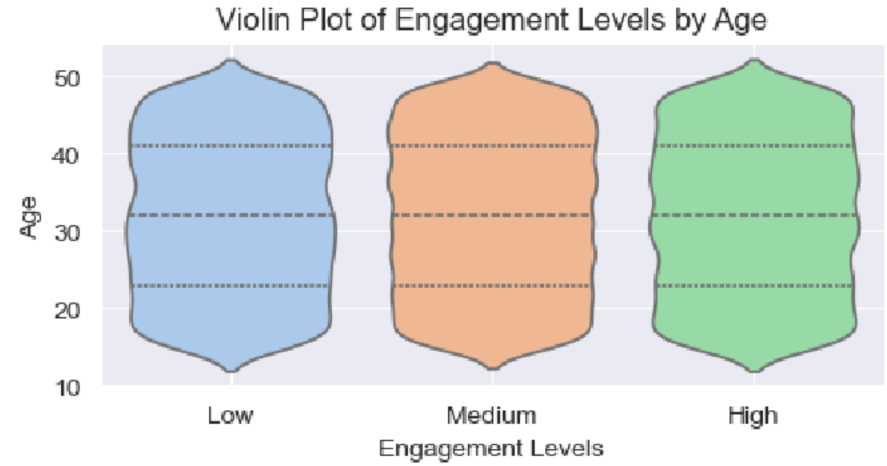
Bivariate Analysis

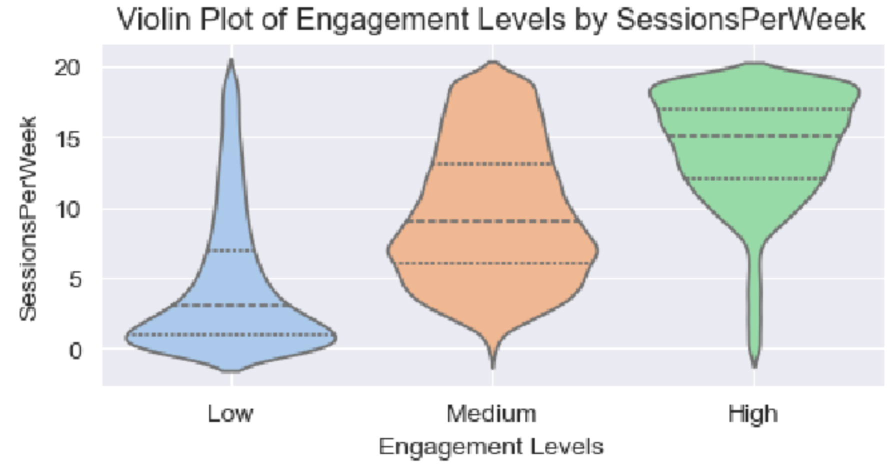
Numerical Fields vs Engagement Levels

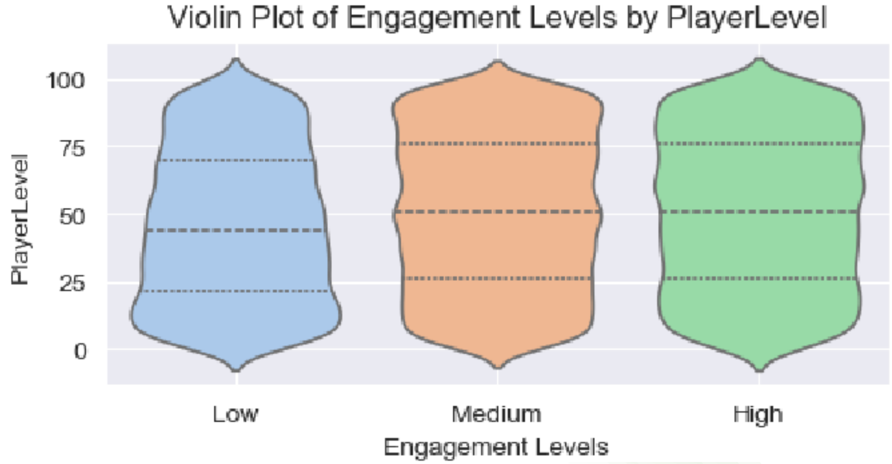
Engagement Levels vs Age:
- Even data distribution across all variables.

Engagement Levels vs Sessions per Week:
- Players with High engagement levels generally spent more time gaming every week.

Engagement Levels vs Player Level:
- The median of Low engagement level is slightly lower than Medium and High levels.
- More players with Low engagement levels also have lower levels of playing skills.
- Converse is true for players with High engagement levels.







EXPLORATORY DATA ANALYSIS

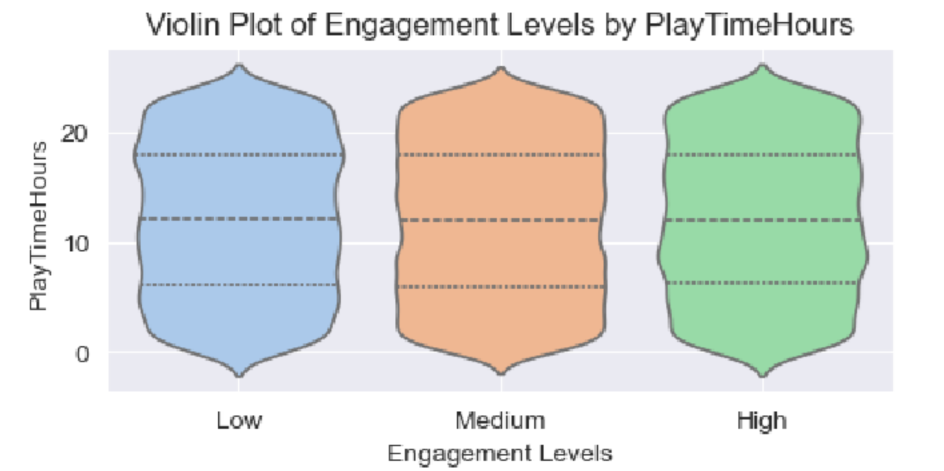
Bivariate Analysis

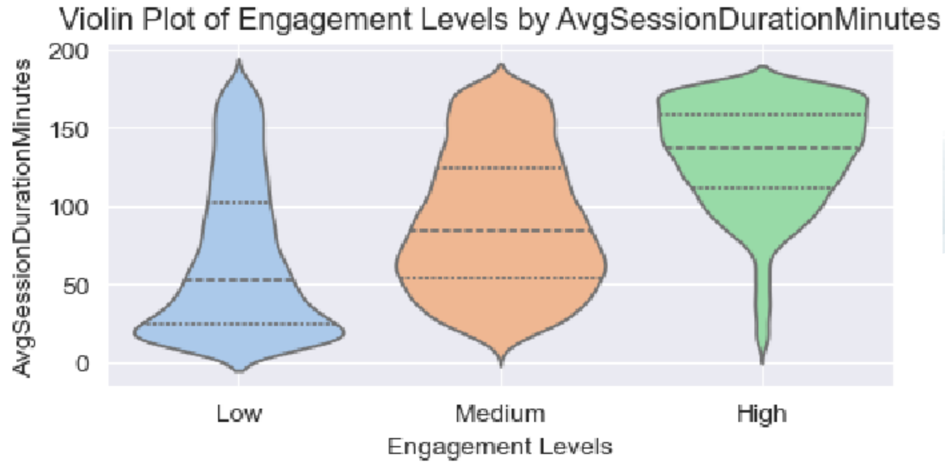
Numerical Fields vs Engagement Levels

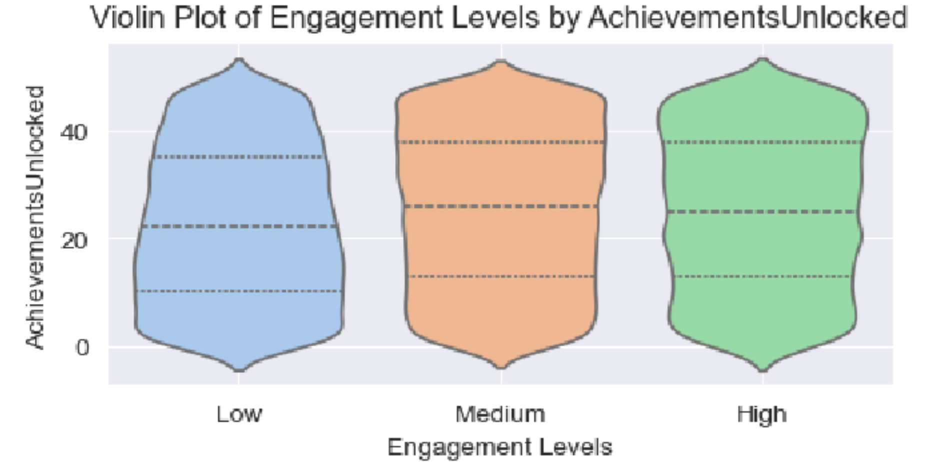
Engagement Levels vs Play Time Hours:
- Even data distribution across all variables.

Engagement Levels vs Avg Session Duration Minutes:
- Players with High engagement levels generally spent more time gaming every session.

Engagement Levels vs Achievements Unlocked:
- The median of Low engagement level is slightly lower than Medium and High levels.
- More players with Low engagement levels also unlocked fewer game achievements.
- Converse is true for players with High engagement levels.



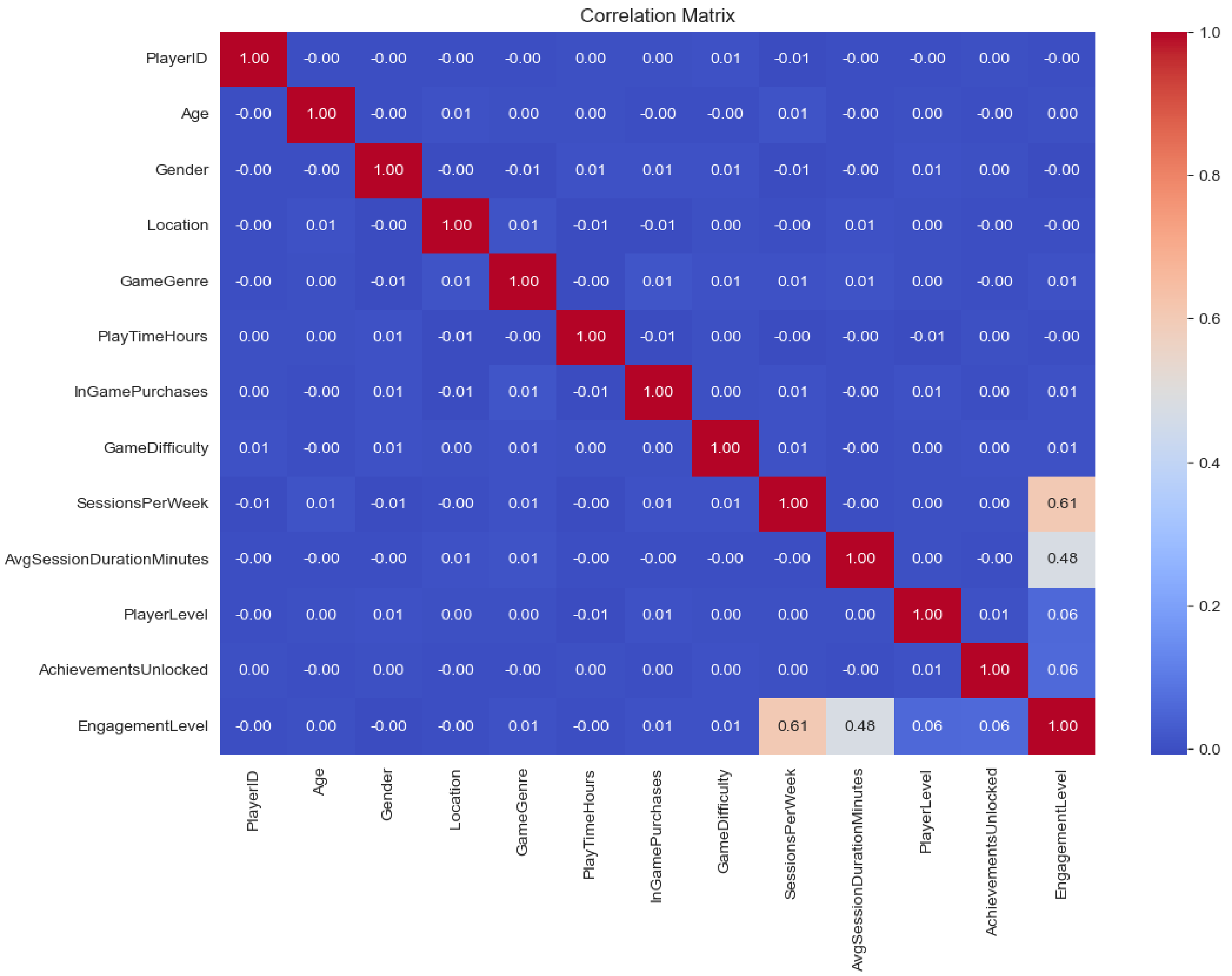




EXPLORATORY DATA ANALYSIS

Correlation Matrix

- +ve correlation between Engagement Level and Sessions Per Week (0.61)
- +ve correlation between Engagement Level and Avg Session Duration Minutes (0.48)

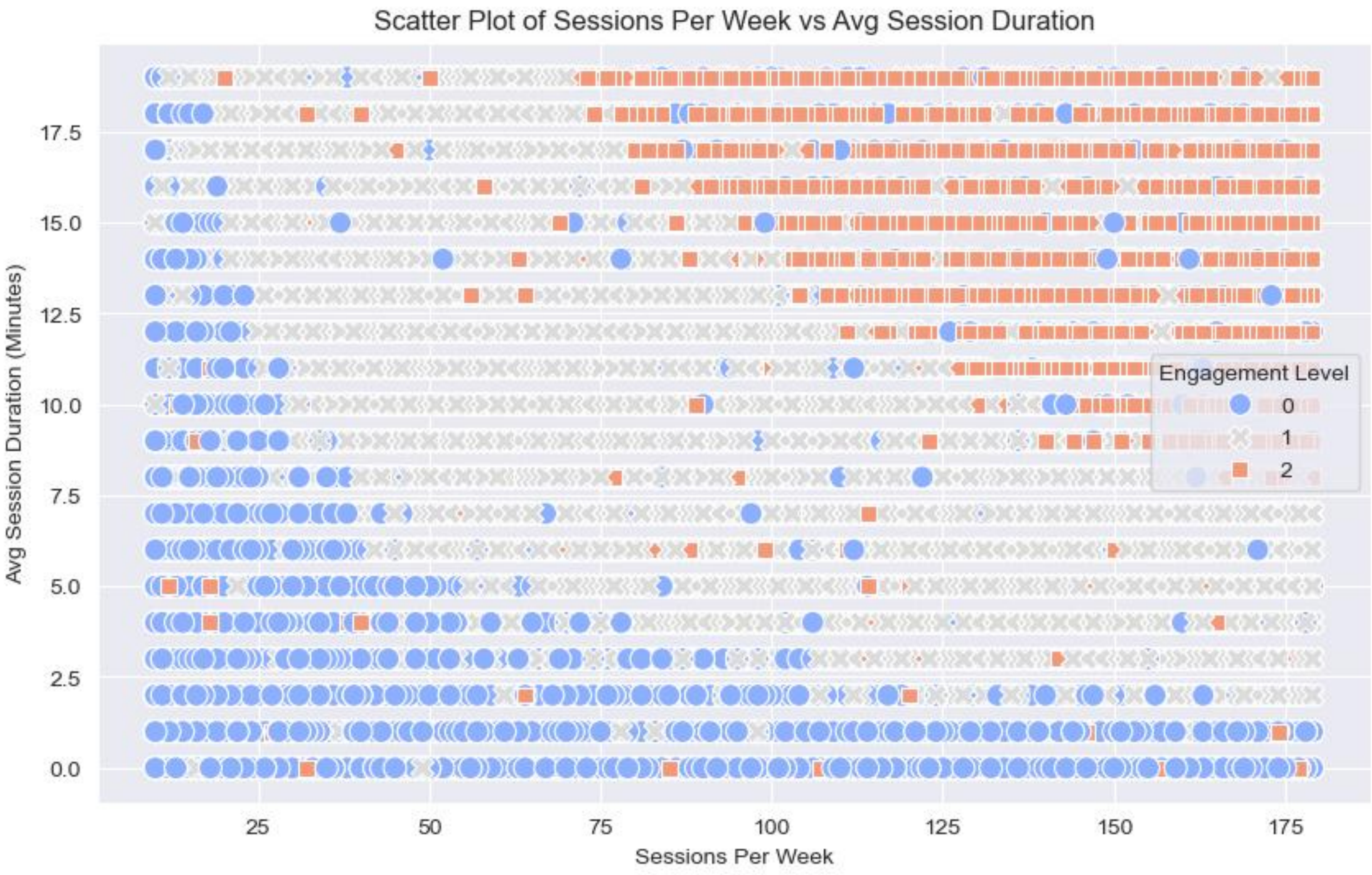


EXPLORATORY DATA ANALYSIS

Multivariate Analysis

EngagementLevel
AvgSessionDuration
AvgSessionPerWeek

- Distinct groups identified
- Players who spent more time are likely to be highly engaged



The Machine Learning Pipeline

Model Development → Model Evaluation → Feature Importance → Hyperparameter Tuning → Model Deployment

- Decision Tree
- Random Forest
- Logistic Regression
- kNN
- Support Vector Machines
- Gradient Boosting

MACHINE LEARNING MODELS

Objective: To predict Engagement Levels (Low, Medium, High)

Classification Models

- **Decision Tree**
  - Easy to interpret and visualise
  - Handles categorial features well
- **Random Forest**
  - Ensemble method combining multiple decision trees
  - Robust to overfitting and feature correlations
- **Support Vector Machines (SVM)**
  - Effective for high-dimensional data
  - Robust to noise and outliers
- **K-Nearest Neighbours (kNN)**
  - Simple, intuitive and efficient
  - Sensitive to feature scaling
- **Logistic Regression**
  - Assume linear relationship, easy to interpret and efficient
  - Not effective when applied to complex datasets
- **Gradient Boosting**
  - Combines weak learners to create strong predictive model
  - Prone to overfitting

EVALUATION OF ML MODELS

Model Development

- Decision Tree
- Random Forest
- Logistic Regression
- kNN
- Support Vector Machines
- Gradient Boosting

DecisionTreeClassifier Model:
Accuracy: 0.84 +/- 0.01
Precision: 0.84 +/- 0.01
Recall: 0.84 +/- 0.01
F1 Score: 0.84 +/- 0.01
Runtime: 4.12 seconds

kNN Model Performance:
Accuracy: 0.81 +/- 0.00
Precision: 0.82 +/- 0.01
Recall: 0.81 +/- 0.00
F1 Score: 0.80 +/- 0.00
Runtime: 1.43 seconds

RandomForest Model Performance:
Accuracy: 0.90 +/- 0.00
Precision: 0.90 +/- 0.00
Recall: 0.90 +/- 0.00
F1 Score: 0.89 +/- 0.00
Runtime: 7.89 seconds

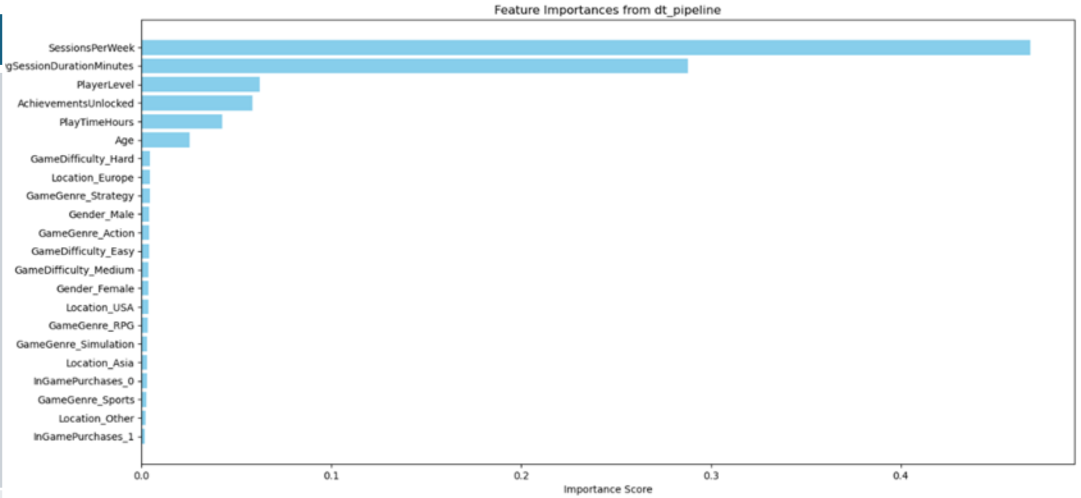
SVC Model Performance:
Accuracy: 0.90 +/- 0.00
Precision: 0.90 +/- 0.00
Recall: 0.90 +/- 0.00
F1 Score: 0.90 +/- 0.00
Runtime: 261.51 seconds

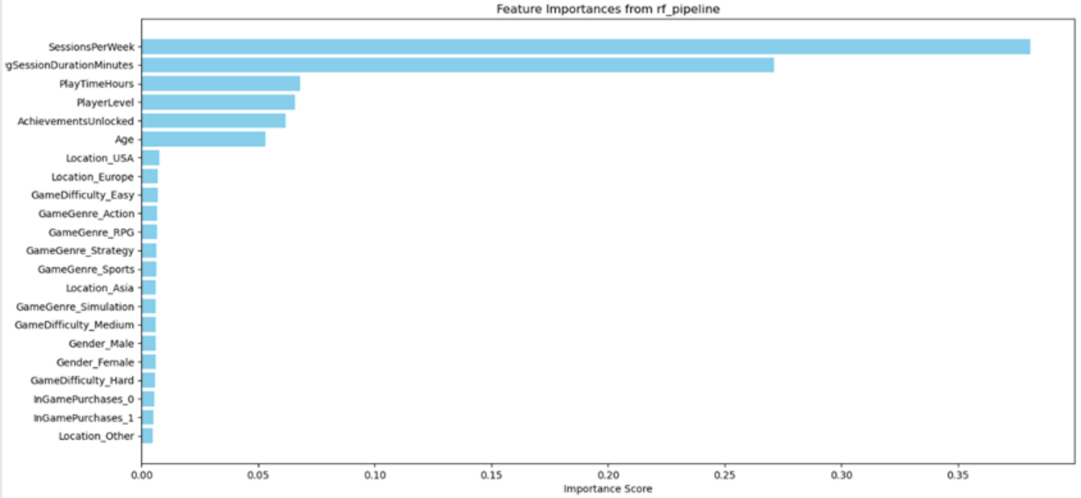
LogisticRegression Model Performance:
Accuracy: 0.82 +/- 0.01
Precision: 0.83 +/- 0.01
Recall: 0.82 +/- 0.01
F1 Score: 0.82 +/- 0.01
Runtime: 0.42 seconds

GradientBoosting Model Performance:
Accuracy: 0.91 +/- 0.00
Precision: 0.91 +/- 0.00
Recall: 0.91 +/- 0.00
F1 Score: 0.91 +/- 0.00
Runtime: 26.18 seconds

FEATURE IMPORTANCE ANALYSIS

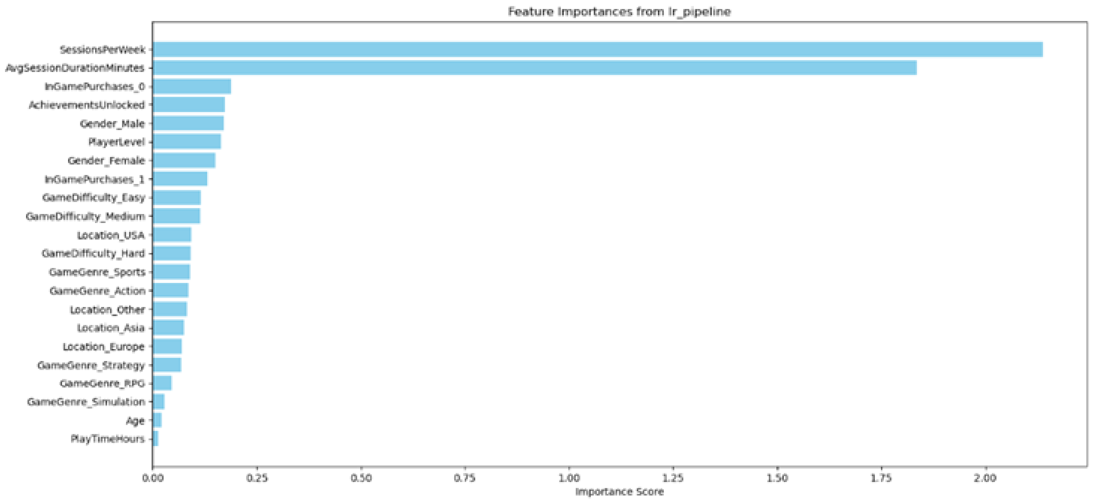
Model	Observations
Decision Tree	The ‘**SessionsPerWeek**’ and ‘**AvgSessionDurationMinutes**’ show high importance, indicating they are crucial for making splits in the tree structure.
Random Forest	The results show a broader distribution of importance scores, with features such as ‘**PlayTimeHours**’ and ‘**AchievementsUnlocked**’ and ‘**Age**’ also gaining prominence.

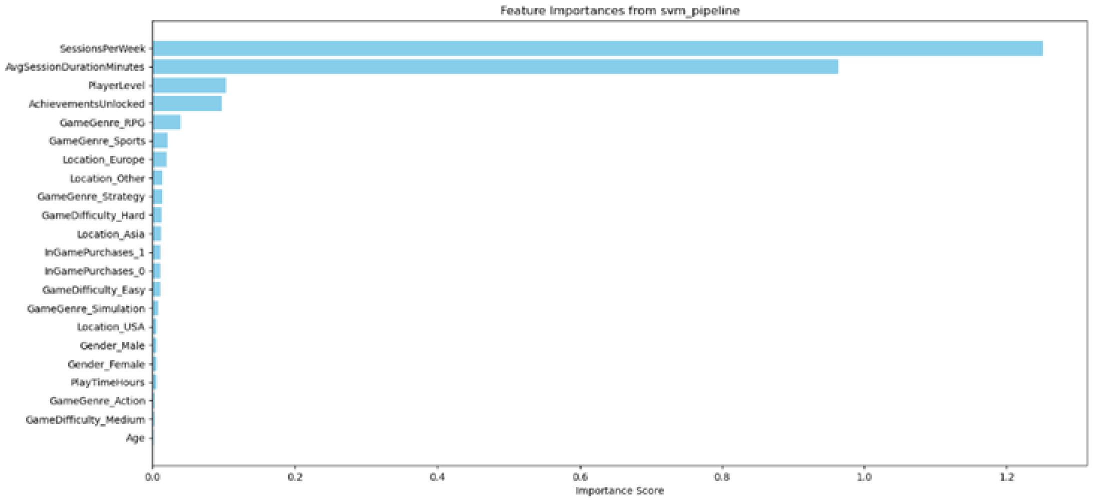




FEATURE IMPORTANCE ANALYSIS

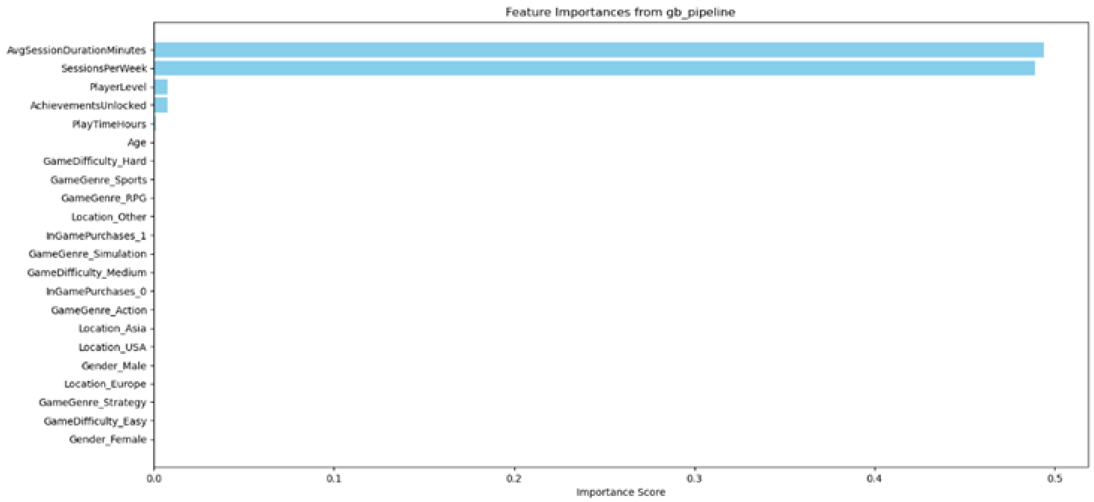
Model	Observations
Logistic Regression	‘**SessionPerWeek**’ and ‘**AvgSessionDurationMinutes**’ emerge as significant predictors. While a range of other features also gain prominence, they have lower absolute values compared to tree-based models.
SVM	‘**SessionPerWeek**’ and ‘**AvgSessionDurationMinutes**’ emerge as significant predictors. Similar to Logistic Regressions, a range of other features also gain prominence. The higher absolute values reflect the model’s ability to capture more interactions between features compared to Logistic Regression.





FEATURE IMPORTANCE ANALYSIS

Model	Observations
Gradient Boosting	This model narrows the variables to two key features, ‘**AvgSessionDurationMinutes**’ and ‘**SessionsPerWeek**’. In addition, the analysis also shows weaker interactions in ‘**AchievementsUnlocked**’ and **PlayerLevel**.



FEATURE IMPORTANCE ANALYSIS

Model	Observations
K-Nearest Neighbours	This model ranked ‘**Gender**’, ‘**Location**’, ‘**Age**’ and ‘**PlayerLevel**’ as prominent features. **However, as kNN performed badly in predicting engagement levels, its feature importance analysis will be ignored.**

Permutation Importance of Features for kNN model:

	Feature	Importance
0	Gender_Male	0.334036
1	Location_Asia	0.269925
2	Location_Other	0.016341
3	Location_Europe	0.013992
4	Age	0.005698
5	PlayerLevel	0.005082
6	AvgSessionDurationMinutes	0.001908
7	SessionsPerWeek	0.001647
8	PlayTimeHours	0.001168
9	Gender_Female	0.000454
10	AchievementsUnlocked	0.000142
11	GameGenre_Strategy	0.000000
12	GameDifficulty_Hard	0.000000
13	GameDifficulty_Easy	0.000000
14	InGamePurchases_1	0.000000
15	InGamePurchases_0	0.000000
16	Location_USA	0.000000
17	GameGenre_Sports	0.000000
18	GameGenre_Simulation	0.000000
19	GameGenre_RPG	0.000000
20	GameGenre_Action	0.000000
21	GameDifficulty_Medium	0.000000

FEATURE IMPORTANCE ANALYSIS

SUMMARY

Categories	Variables
Engagement Metrics	**SessionsPerWeek** **AvgSessionDurationMinutes** **PlayTimeHours**
Player Characteristics	**Age** **AchievementsUnlocked** **PlayerLevel**

HYPERPARAMETER TUNING

Objective

- To find a set of hyperparameters that minimises a predefined loss function on given data

Method

- Uses cross-validation to estimate generalisation performance and determine the best hyperparameter values

Grid Search	Random Search
Comprehensive search of every possible combination of hyperparameters	Random sampling of combinations of hyperparameters
Inefficient when searching in large spaces	Quicker exploration of hyperparameter space
Computationally expensive and time-consuming	Suitable for larger dataset with high dimensional hyperparameter spaces

HYPERPARAMETER TUNING

DECISION TREE CLASSIFIER

Best Parameters and Accuracy Score for Random Forest, 10-Fold	Best Parameters and Accuracy Score for Random Forest, 3-Fold
Fitting 10 folds for each of 50 candidates, totalling 500 fits **Best parameters for Random Forest:** {'classifier__n_estimators': 1000, 'classifier__min_samples_split': 2, 'classifier__min_samples_leaf': 1, 'classifier__max_features': 'sqrt', 'classifier__max_depth': None} **Best score for Random Forest:** 0.899710332136013	Fitting 3 folds for each of 50 candidates, totalling 150 fits **Best parameters for Random Forest:** {'classifier__n_estimators': 700, 'classifier__min_samples_split': 2, classifier__min_samples_leaf': 1, 'classifier__max_features': 'sqrt', 'classifier__max_depth': None} **Best score for Random Forest:** 0.8943078436344779

HYPERPARAMETER TUNING

GRADIENT BOOSTING

Best Parameters and Accuracy Score for Gradient Boosting, 10-Fold	Best Parameters and Accuracy Score for Gradient Boosting, 3-Fold
Fitting 10 folds for each of 50 candidates, totalling 500 fits **Best parameters for Gradient Boosting:** {'classifier__n_estimators': 900 'classifier__min_samples_split’: 2 'classifier__min_samples_leaf’: 2 'classifier__max_depth': 7, 'classifier__learning_rate': 0.01} **Best score for Gradient Boosting:** 0.9186690027434468	Fitting 3 folds for each of 50 candidates, totalling 150 fits **Best parameters for Gradient Boosting:** {'classifier__n_estimators': 900, 'classifier__min_samples_split': 2, 'classifier__min_samples_leaf': 2, 'classifier__max_depth': 7, 'classifier__learning_rate': 0.01} **Best score for Gradient Boosting:** 0.9159771821436102

HYPERPARAMETER TUNING

DECISION TREE CLASSIFIER

GRADIENT BOOSTING

Best Parameters and Accuracy Score for Random Forest, 10-Fold	Best Parameters and Accuracy Score for Gradient Boosting, 3-Fold
Fitting 10 folds for each of 50 candidates, totalling 500 fits **Best parameters for Random Forest:** {'classifier__n_estimators': 1000, 'classifier__min_samples_split': 2, 'classifier__min_samples_leaf': 1, 'classifier__max_features': 'sqrt', 'classifier__max_depth': None} **Best score for Random Forest:** 0.899710332136013	Fitting 3 folds for each of 50 candidates, totalling 150 fits **Best parameters for Gradient Boosting:** {'classifier__n_estimators': 900, 'classifier__min_samples_split': 2, 'classifier__min_samples_leaf': 2, 'classifier__max_depth': 7, 'classifier__learning_rate': 0.01} **Best score for Gradient Boosting:** 0.9159771821436102

MODEL TRAINING & TESTING

1. Applying best hyperparameters onto Random Forest and Gradient Boosting models
2. Train models using X_train and y_train datasets
3. Evaluate using X_test and y_test datasets

Random Forest

```
Random Forest Model Evaluation:
Accuracy: 0.8996003996003996

Classification Report:
              precision    recall  f1-score   support

        High       0.93      0.84      0.88      1047
         Low       0.92      0.87      0.90      1045
      Medium       0.88      0.95      0.91      1912

    accuracy                           0.90      4004
   macro avg       0.91      0.89      0.90      4004
weighted avg       0.90      0.90      0.90      4004

Confusion Matrix:
 [[ 875   31  141]
 [  19  914  112]
 [  48   51 1813]]
```

Gradient Boosting

```
Gradient Boosting Model Evaluation:
Accuracy: 0.9215784215784216

Classification Report:
              precision    recall  f1-score   support

        High       0.93      0.89      0.91      1047
         Low       0.92      0.90      0.91      1045
      Medium       0.92      0.95      0.93      1912

    accuracy                           0.92      4004
   macro avg       0.92      0.91      0.92      4004
weighted avg       0.92      0.92      0.92      4004

Confusion Matrix:
 [[ 936   32   79]
 [  21  942   82]
 [  48   52 1812]]
```

MODEL SELECTION

Gradient Boosting before Tuning
Validation Dataset

GradientBoosting Model Performance:
Accuracy: 0.91 +/- 0.00
Precision: 0.91 +/- 0.00
Recall: 0.91 +/- 0.00
F1 Score: 0.91 +/- 0.00
Runtime: 26.18 seconds

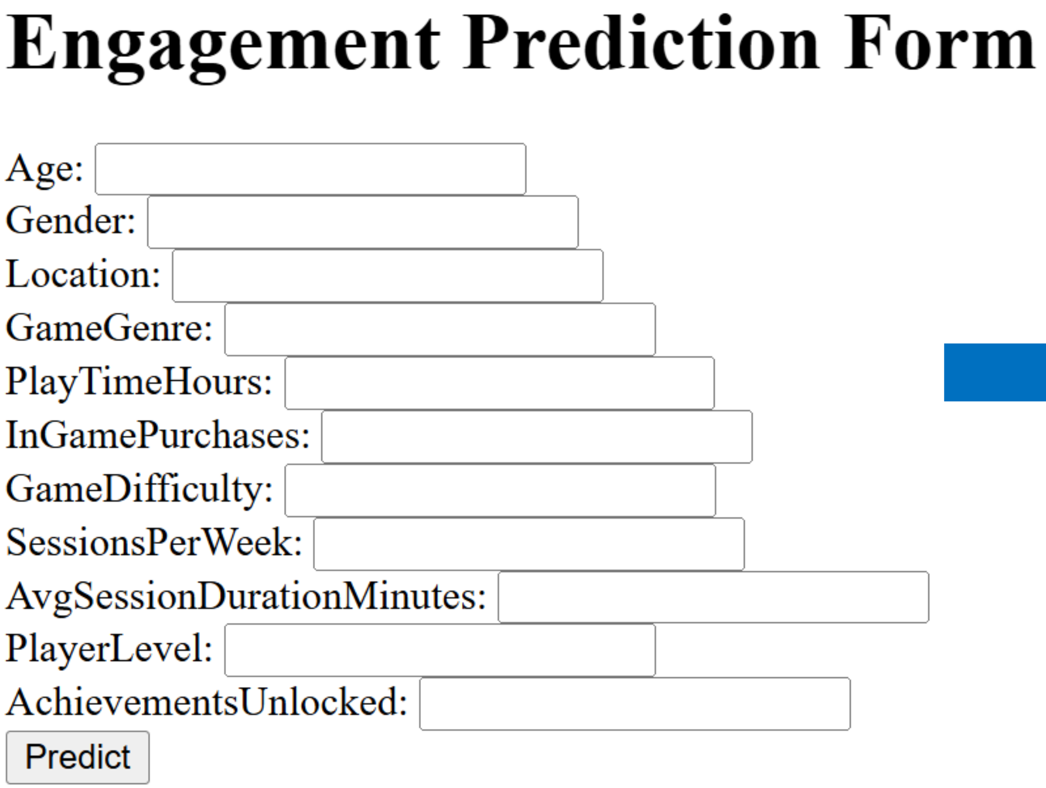
Gradient Boosting Tuned with Hyperparameters
Testing Dataset

```
Gradient Boosting Model Evaluation:
Accuracy: 0.9215784215784216

Classification Report:
              precision    recall  f1-score   support

        High       0.93      0.89      0.91      1047
         Low       0.92      0.90      0.91      1045
      Medium       0.92      0.95      0.93      1912

    accuracy                           0.92      4004
   macro avg       0.92      0.91      0.92      4004
weighted avg       0.92      0.92      0.92      4004

Confusion Matrix:
 [[ 936   32   79]
 [  21  942   82]
 [  48   52 1812]]
```

MODEL DEPLOYMENT

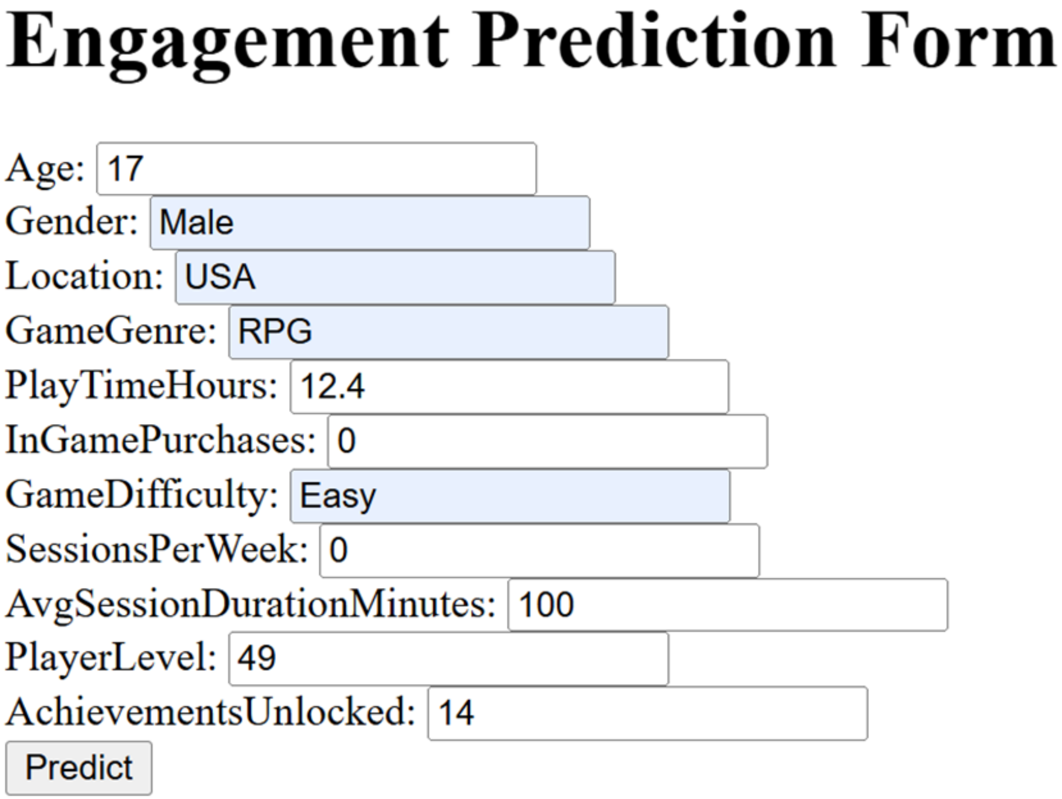
Local web application using Flask

User access http://127.0.0.1:5000



Engagement Prediction Form

Age:
Gender:
Location:
GameGenre:
PlayTimeHours:
InGamePurchases:
GameDifficulty:
SessionsPerWeek:
AvgSessionDurationMinutes:
PlayerLevel:
AchievementsUnlocked:
Predict

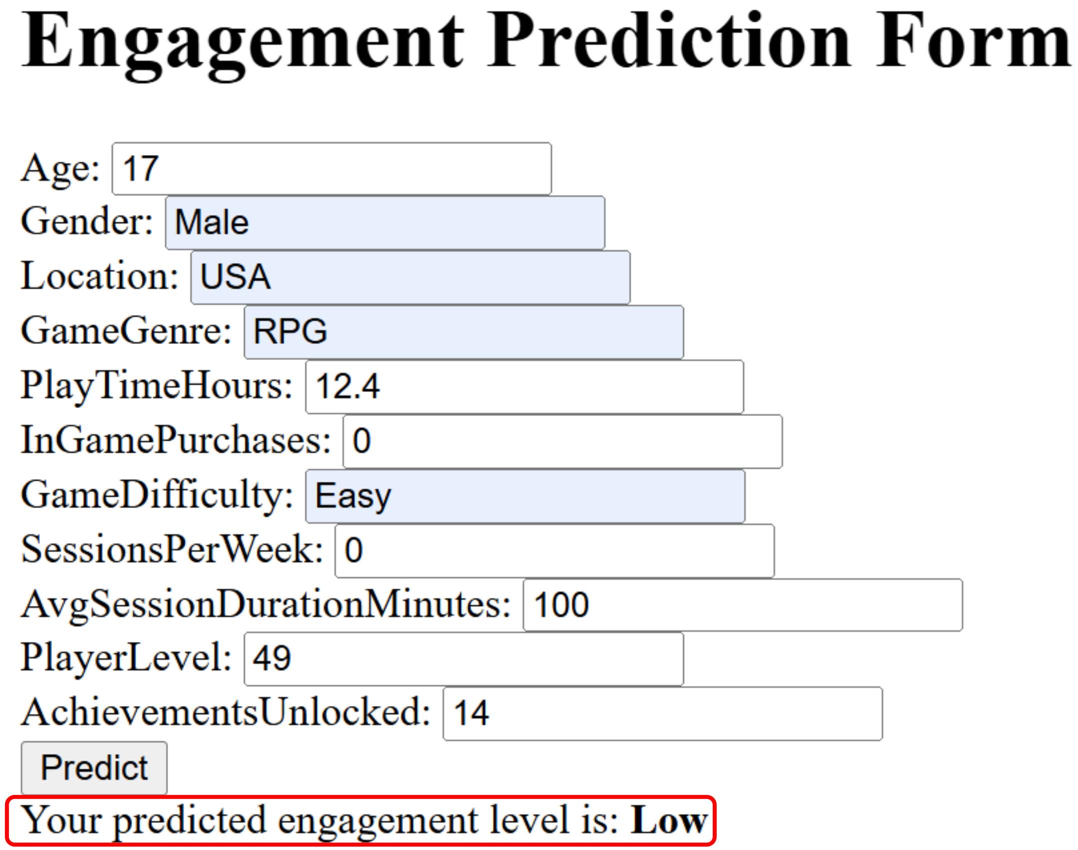


Engagement Prediction Form

Age: 17
Gender: Male
Location: USA
GameGenre: RPG
PlayTimeHours: 12.4
InGamePurchases: 0
GameDifficulty: Easy
SessionsPerWeek: 0
AvgSessionDurationMinutes: 100
PlayerLevel: 49
AchievementsUnlocked: 14
Predict

MODEL DEPLOYMENT

Local web application using Flask



Engagement Prediction Form

Age: 17
Gender: Male
Location: USA
GameGenre: RPG
PlayTimeHours: 12.4
InGamePurchases: 0
GameDifficulty: Easy
SessionsPerWeek: 0
AvgSessionDurationMinutes: 100
PlayerLevel: 49
AchievementsUnlocked: 14
Predict
Your predicted engagement level is: **Low**

Classification	Description
Low	Players interact minimally with gaming. Play infrequently and may abandon game after a short period.
Medium	Deeper involvement, may not have fully explored all aspects of the game.
High	Highly active and invested in gaming. Spend significant time in gaming.

LIMITATIONS OF THE PROJECT

Based on assumption that individuals suffering from <u>gaming disorder equates to highly engaged players</u>.

Project <u>primarily focuses on quantitative measures</u>. The qualitative factors, such as emotional and psychological aspects, are not measured.

- for example, escapism, social interaction.

It is critical to <u>collaborate with relevant medical professionals</u> to develop a standardised assessment tool for evaluating gaming disorder.

--- END ---