COMP5112M: Data Science

Module leader: Duygu Sarikaya

Coursework 1:

Understanding The Impact of Social Isolation and Loneliness in a Game Environment

Group ID:

Skitty

Group Members:

Yuhao Shen, 201900003, bthh1050

Jianlin Yao, 201902801, dzdl5424

Haolan Gao, 201899584, qbnk0918

Xilong Liu, 201851112, tmwg2894

Section 1: Group Members

- Yuhao Shen, 201900003, bthh1050
- Jianlin Yao, 201902801, dzdl5424
- Haolan Gao, 201899584, qbnk0918
- Xilong Liu, 201851112, tmwg2894

Section 2: Introduction

2.1 Overview of the aims and dataset

The primary aim of this project is to conduct some specific data analysis tasks on the given dataset, a multinational dataset of game players' behaviors in a virtual world and environmental perceptions, including exploratory data analysis, important variable selection, and classification modeling based on Machine Learning models. This dataset provides various resources about environmental worldviews and gameplayers' behaviors in the virtual game world, Animal Crossing: New Horizons (ACNH), with 640 records in six categories of attributes collected from players all over the world. In this project, only part of the attributes will be employed to conduct data analysis and eventually complete our work, since the scale of the attributes is distinctly enormous.

2.2 Detailed Objectives

Objective 1. Data preprocessing

Objective 2. Exploratory data analysis

- Distribution of the players' length of being self-isolated/social distancing.
- Distribution of the players' length of being self-isolated/social distancing according to regions.
- Relationship between the players' length of being self-isolated/social distancing and the game-playing frequency of the players.
- Comparison of the frequency of the different lengths of being self-isolated/social distancing and game-playing feeling response "I lost connection with the outside world".

Objective 3. Important variable selection

Objective 4. Prediction Modeling with ML models

Section 3: Detailed analysis

Objective 1: Data preprocessing

The dataset ought to be properly preprocessed so that it could be employed to carry on the tasks afterwards. So, the procedure including checking missing data, encoding categorical data (for modeling) is conducted after considerate observation of the data.

The statistics of missing values are as follows:

Column	Missing Values	Column	Missing Values
A4	86	D2	5
D7	13	D4	1
D1	6	D3	1

Since the specific variables participating in analysis of this project are: B2、A1、D3、F31、E1-E28, and only D3 has 1 missing record, the method to handle this problem is to dicard this row.

Meanwhile, the variable B2 is going to be included in modeling and the values of it are categorical ones, encoding is needed to quantify this variable for modeling.

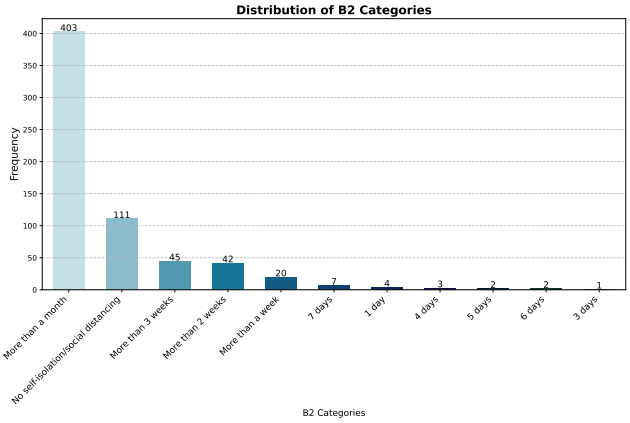
No self-isolation/social distancing	0	7 days	1
1 day	1	More than a week	2
3 days	1	More than 2 weeks	3
4 days	1	More than 3 weeks	4
5 days	1	More than a month	5
6 days	1		

Notice: the ‘n-days’ values are combined as a new value ‘Less than a week’ in that the count of these values are extreamly few to do any analysis tasks and modeling, changing the categorizing rules.

Objective 2: Exploratory data analysis

(1) Distribution of the players' length of being self-isolated/social distancing (B2)

To analysis the distribution of the players' length of being self-isolated/social distancing, the sorted bar chart is well organized to display the properties of the distribution as below.



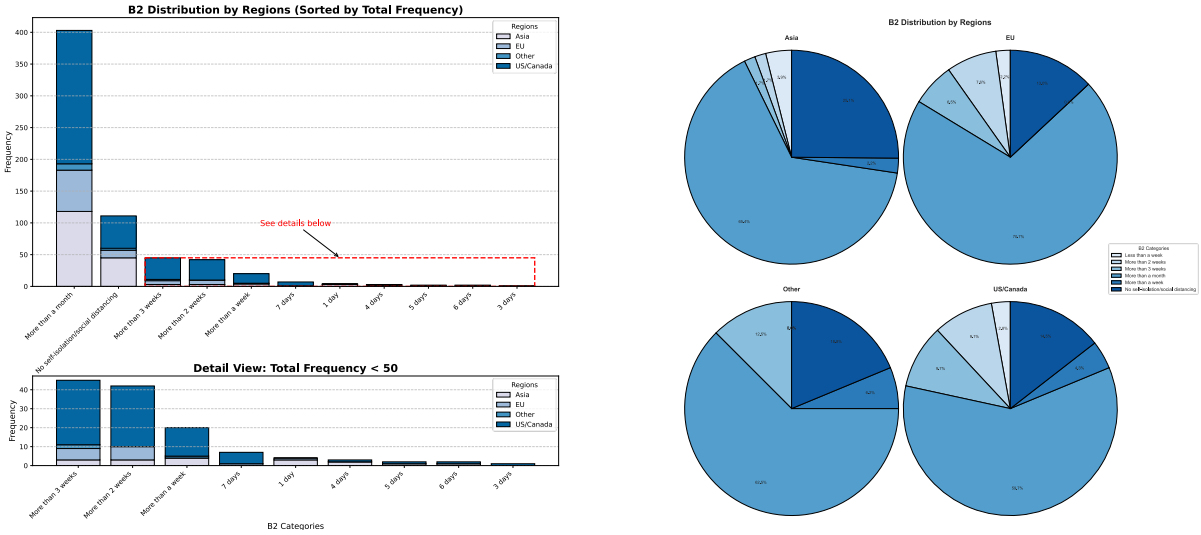
According to the bar chart, the findings can be summarized as follows:

- The dataset exhibits a clear class imbalance, with "More than a month" making up 63% of the records, while other B2 values are infrequent, forming a long-tailed distribution.
- The dominance of "More than a month" highlights the global prevalence of prolonged isolation during COVID-19, reflecting the widespread impact of restrictive policies and challenging times.

Consequently, in regard to the class imbalance problem, methods like resampling, adding focal loss are considered in the prediction modeling in the last objective. That is also the reason for encoding.

(2) Distribution of the players' length of being self-isolated/social distancing (B2) according to regions (A1_2)

To explore the distribution of the players' length of being self-isolated/social distancing (B2) classified by regions (A1_2), the sorted stacked bar chart and pie chart are well organized as below.



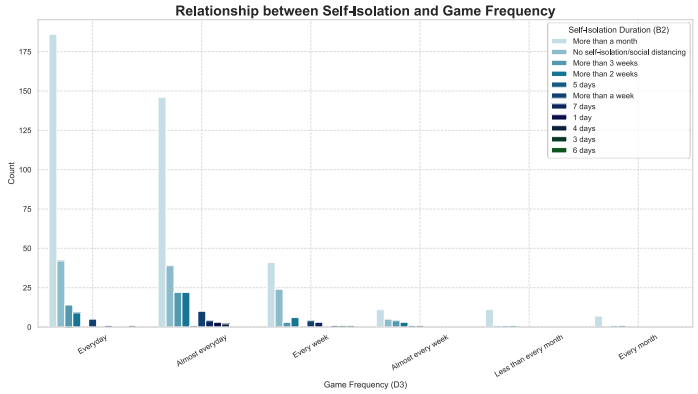
According to the stacked bar chart and the pir chart, the distribution pattern of the gameplayer from different regions classified by their length of being self-isolated/social distancing can be observed (A1_2 v.s. B2). To be specific, it can be concluded as follows:

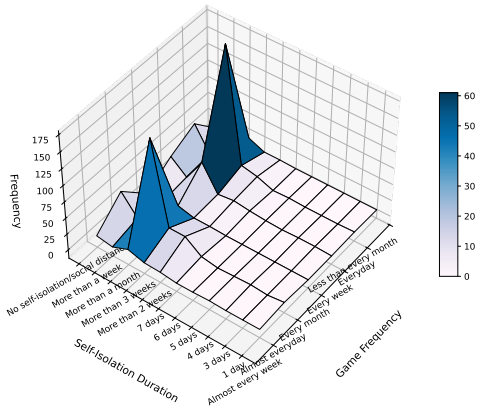
- Players from US/Canada formed the largest proportion of respondents, while other regions had smaller sample sizes, likely due to the survey's North American focus and limited global reach.
- Over 60% of respondents experienced "More than a month" of isolation, reflecting a shared global challenge of prolonged social distancing during the pandemic.

- A minority of players avoided isolation, possibly due to living in rural, less-affected areas or fortunate circumstances.

3) Relationship between the players' length of being self-isolated/social distancing and the game-playing frequency of the players

To explore the relationship between the players' length of being self-isolated/social distancing (B2) and the game-playing frequency of the players (D3), the combined histogram and 3-D heatmap are well designed as below.



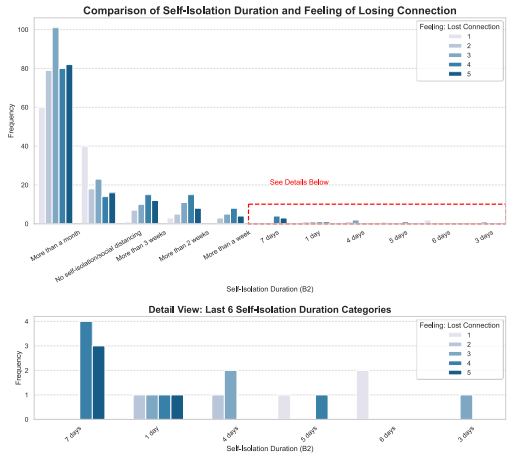


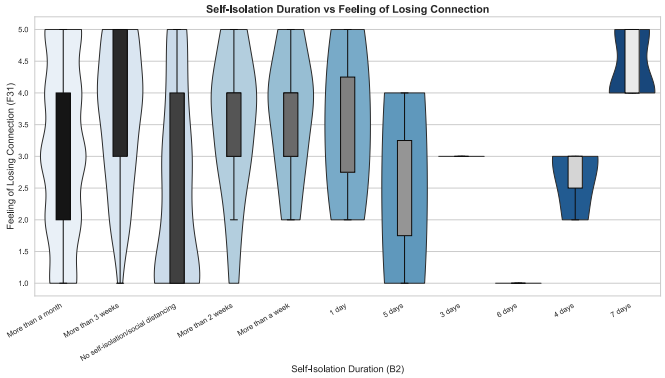
According to the histogram and 3-D heatmap, the patent relationship between the players' length of being self-isolated/social distancing (B2) and the game-playing frequency of the player (D3) can be observed and concluded as follows:

- Most players reported playing the game "Everyday" or "Almost every day," making up the largest portion of the player base.
- The majority of self-isolation durations were "More than a month," with similar distribution patterns observed across gaming frequency groups.

4) Comparison of the frequency of the different lengths of being self-isolated/social distancing and game-playing feeling response “I lost connection with the outside world”

To make comparison between frequency of duration of being self-isolated/social distancing (B2) and game-playing feeling of disconnection (F31), the violin box plot and histogram are designed as below.





According to the histogram and violin boxp plot, it can be observed and demonstrated that:

- Longer durations of isolation increased the feeling of disconnection from the outside world.
- Isolation exceeding one month saw many players alleviating their sense of disconnection, suggesting gaming's positive role in reducing stress and isolation.
- After a month of isolation, the levels of disconnection feelings tended to distribute more evenly.

Objective 3: Important Variable Selection

Methodology: Variable Importance Calculated by XGBoost Model and Compared with SHAP

(1) Methods Introduction and Motivation Statement

- **XGBoost Feature Importance**:

Methodology: Importance scores were derived from the XGBoost model by evaluating each feature's contribution to the split points in decision trees (gain-based importance).

Motivation: This provides a global view of how often and effectively each feature contributes to the model's predictive ability, offering a model-centric perspective. **XGBoost instead of Random Forest was chosen for** its superior handling of feature interactions, gradient boosting optimization, and seamless integration with SHAP for interpretability."

- **SHAP Analysis (use normalized mean absolute SHAP as feature importance):**

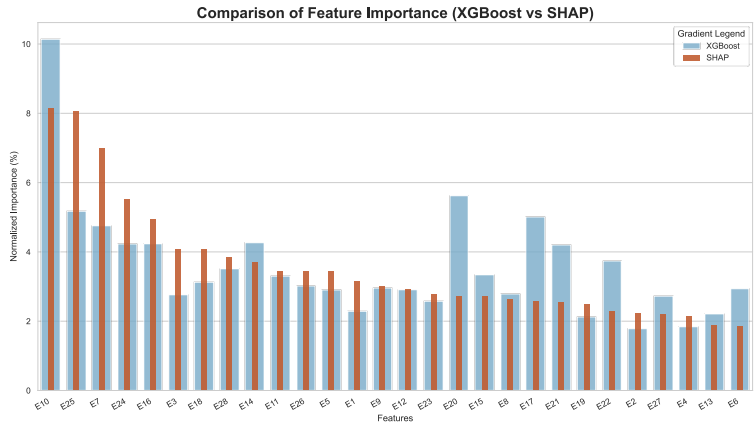
Methodology: SHAP values quantified the marginal contribution of each feature to the prediction for every sample. Mean absolute SHAP values were used to rank features globally.

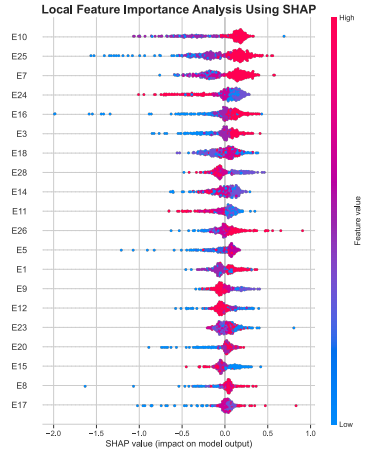
Motivation: SHAP is better at explaining individual predictions and provides insights into both global and local importance absolute quatification.

- **Comparison and Valuation**:

By comparing XGBoost and SHAP results, whether global importance aligns with sample-specific contributions, Variability in feature importance ranking for methodological differences can be explained.

(2) Results and Explanations





According to the charts above, the comparison of the results from these two methods is displayed in the table below (with first 7 significant variable importance compared).

Feature	XGBoost Importance	SHAP Importance (%)	Rank Agreement
E25	7.25	9.18	yes
E10	6.85	6.05	yes
E16	5.74	5.79	yes
E24	3.17	4.86	yes
E18	3.09	4.65	yes
E21	4.85	2.30	no
E3	2.15	4.11	no

- **Explaination for Results**

Top Features Across Both Methods: E25, E10, E16, E24, E18: Consistently ranked among the top features in both methods, indicating strong relevance to predicting self-isolation/social distancing duration. These features likely capture behaviours that are highly indicative of prolonged isolation.

Method-Specific Variations:

1) XGBoost-Specific Findings: Features like E23 (Importance: 3.97%) and E21 (Importance: 4.85%) are emphasized more by XGBoost but have relatively lower SHAP importance. This suggests their role may lie in early model splits rather than influencing individual predictions significantly.
2) SHAP-Specific Findings: Features like E3 (SHAP: 4.11%) and E5 (SHAP: 3.73%) exhibit higher SHAP importance but are less emphasized by XGBoost. These features may capture non-linear interactions with other features or have a more distributed impact across samples.

Alignment vs Divergence:

1) Alignment: The top features (E25, E10, E16) align across both methods, strengthening confidence in their importance.
2) Divergence: Lower-ranked features (E9, E6) show greater divergence, reflecting methodological differences in capturing global vs local importance.

Objective 4: Prediction Modeling with ML Methods

Methodology Initial: XGBoost Multi-class Classification with K-Fold Cross Validation

(1) Task Definition:

The task involves building a multi-class classification model to predict the self-isolation duration (B2) by incrementally increasing the number of features according to the results in **Objective 3**. This approach ensures a balance between model complexity and performance.

(2) Dataset Splitting:

Stratified K-fold cross-validation is employed to split the dataset into training and validation sets. This ensures that the class distribution remains consistent across folds, which is essential for imbalanced datasets. For the test set, 20% of the data were reserved to evaluate the final model's performance.

(3) Model Training:

XGBoost was selected due to its robustness, efficiency, and interpretability in handling tabular data. For this stage, the default "softmax" objective function was used for multi-class classification without additional weighting for class imbalance. The evaluation metric was multi-class log-loss (mlogloss).

(4) Evaluation Metrics:

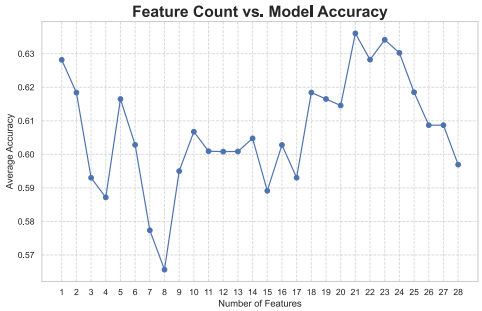
Accuracy: To measure the overall performance of the model in correctly predicting labels.

Confusion Matrix: To provide insight into specific misclassification patterns.

ROC and AUC: To evaluate the model's ability to distinguish among multiple classes.

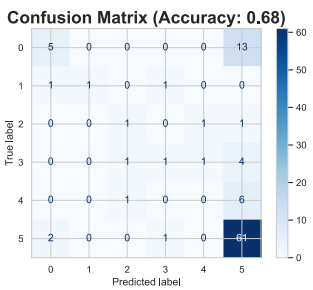
F1 Score: To balance precision and recall, especially for imbalanced classes.

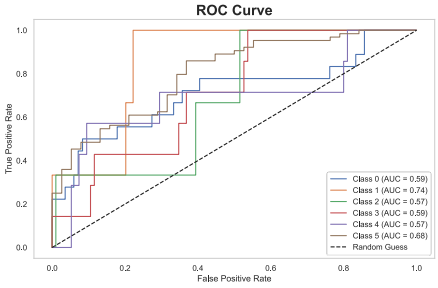
(5) Visualization:



Classification Report:

	precision	recall	f1-score	support
0	0.62	0.28	0.38	18
1	1.00	0.33	0.50	3
2	0.33	0.33	0.33	3
3	0.33	0.14	0.20	7
4	0.00	0.00	0.00	7
5	0.72	0.95	0.82	64
accuracy			0.68	102
macro avg	0.50	0.34	0.37	102
weighted avg	0.62	0.68	0.62	102





The initial results revealed a significant class imbalance problem, as evident from the confusion matrix and relatively low recall for minority classes. This necessitated further refinement of the methodology to address class bias and enhance predictive performance.

Methodology Improved: Modeling with Resampling + Focal-Loss for Class Imbalance Problem

To address the limitations observed in the initial approach, the methodology was enhanced by incorporating resampling and a focal loss function to handle class imbalance.

(1) Task Definition: (The same as previous.)

(2) Dataset Splitting: (The same as previous.)

(3) Dataset Resampling after Splitting:

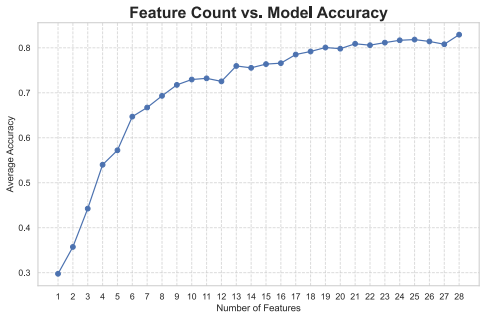
Synthetic Minority Oversampling Technique (SMOTE) was applied to balance the class distribution in the training data. This oversampling approach generates synthetic samples for minority classes to mitigate the class imbalance and improve model generalization.

(4) Model Training with Focal Loss:

The improved model leveraged a custom focal loss function during training. Focal loss down-weights the contribution of well-classified examples, focusing more on hard-to-classify samples. This adjustment aimed to reduce class bias and enhance recall for minority classes. The multi:softprob objective function was used to predict class probabilities, ensuring a smoother prediction distribution.

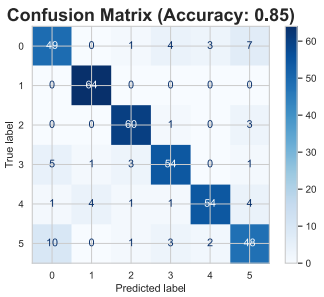
(5) Evaluation Metrics: (The same as previous.)

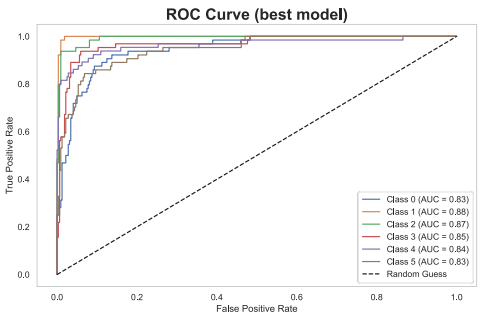
(6) Visualization:



Classification Report:

	precision	recall	f1-score	support
0	0.75	0.77	0.76	64
1	0.93	1.00	0.96	64
2	0.91	0.94	0.92	64
3	0.86	0.84	0.85	64
4	0.92	0.83	0.87	65
5	0.76	0.75	0.76	64
accuracy			0.85	385
macro avg	0.85	0.85	0.85	385
weighted avg	0.85	0.85	0.85	385





The model demonstrated notable improvements in recall and F1 scores for underrepresented classes, while preserving high accuracy for the majority class. This highlights the effectiveness of incorporating resampling and focal loss in addressing class imbalance.

Section 4: Conclusion

This project effectively combined advanced machine learning techniques with interpretability tools like SHAP to uncover meaningful insights about social isolation in gaming. It demonstrated the importance of handling class imbalance and leveraging robust models for predictive analysis in real-world datasets.

Appendix

References

[1] Vuong, Q.H., Ho, M.T., La, V.P., Le, T.T., Nguyen, T.H.T. and Nguyen, M.H., 2021. A Multinational Data Set of Game Players' Behaviors in a Virtual World and Environmental Perceptions. Data Intelligence, 3(4), pp.606-630.

[2] Schafer, J.L., 1997. Analysis of incomplete multivariate data. Chapman &Hall/CRC.

[3] Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[4] Lundberg, S., 2017. A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874.