University of Leeds

COMP5122M: Data Science Coursework

School of Computing

Group Coursework

Authors:
Matthew Gordon
SID - [redacted]
Isaac Unsworth
SID - [redacted]

Authors:
Xavier Caine
SID - [redacted]
George McGuinness
SID - [redacted]

Group Name: Abra
December 11, 2024

1 Introduction

This report explores the relationship between self-isolation and in-game behaviours in *Animal Crossing: New Horizons (ACNH)*, using data from 640 players worldwide that includes socio-demographics, COVID-19 concerns, environmental perceptions, gaming habits, and related feelings. The analysis focusses on examining the distribution of self-isolation lengths across regions and gaming frequencies, identifying key in-game behaviours linked to self-isolation duration, and developing a Random Forest model to predict self-isolation length based on in-game actions. By leveraging this approach, the study aims to provide insights into how gaming behaviours reflect real-world social isolation, offering a deeper understanding of the connection between virtual activities and mental well-being during social distancing.

2 Detailed Analysis

The data set was checked for missing or duplicated values, which were appropriately addressed. Column A4 had 86 missing values, identified as responses of "Neither" to the question about having a pet or garden, based on data collection methods. Other columns had minimal missing values, so the corresponding rows were removed as they represented a negligible portion of the dataset.

2.1 Investigating The Distribution of the Players' Length of Being Self-isolated/Social Distancing

In order to make this part of the analysis easier, all responses between 1 and 7 days were compiled into a new category 'Under a Week'.

Figure 1 Observations

The majority of players (around 400) reported self-isolating for more than a month, reflecting the prevalence of extended isolation, likely due to strict lockdowns or personal precautions. In contrast, significantly fewer players reported shorter isolation periods (under a week or none). Categories like 'more than 3 weeks,' 'more than 2 weeks,' and 'more than a week' had smaller, more balanced counts, indicating that while some experienced moderate isolation lengths, most had undergone prolonged self-isolation.

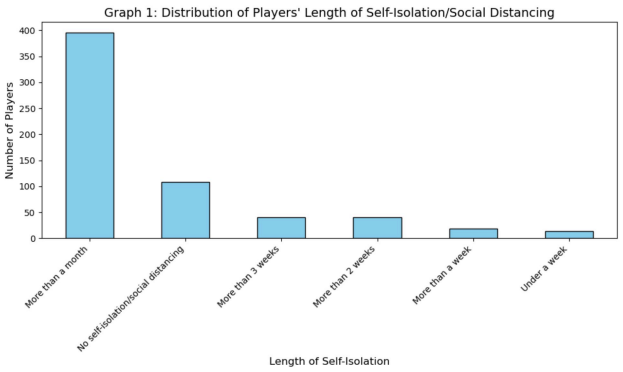


Figure 1: Histogram Showing the Number of Players in Each Social Distancing/Self-Isolation Length Bracket

Figure 1 Interpretation

The predominance of "more than a month" reflects the widespread prolonged isolation during the pandemic, with fewer players reporting shorter durations, possibly due to adherence to isolation measures or a sample bias toward those isolating longer. As the data set is tied to Animal Crossing, extended isolation could correlate with increased engagement with the game, potentially as a coping mechanism.

2.2 Investigating The Distribution of the Players' Length of Being Self-isolated/Social Distancing According To Regions

In this section, the distribution of self-isolation lengths of players in different regions will be investigated.

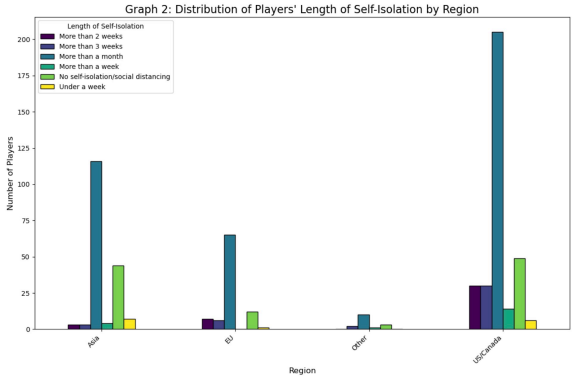


Figure 2: Grouped Histogram Showing the Number of Players in Each Social Distancing/Self-Isolation Length Bracket for Different Regions.

Figure 2 Observations

Across all regions, 'more than a month' was the most reported duration of self-isolation, reflecting the global prevalence of prolonged isolation due to lockdowns and public health measures. Notably, the second largest category was 'no self-isolation/social distancing,' indicating that a significant proportion of respondents did not isolate. Shorter durations, such as "under a week," "more than a week" and "more than 2 weeks", were consistently less common across all regions.

Graph 2 Interpretations

The similar distribution of the duration of self-isolation in all regions suggests consistent attitudes towards the pandemic and comparable demographics among survey respondents around the world.

2.3 Investigating The Relationship Between The Players' Length of Being Self-isolated/Social Distancing and the Game-Playing Frequency of the Players.

In this section, the relationship between the players length of self-isolation and their game-playing frequency will be investigated.

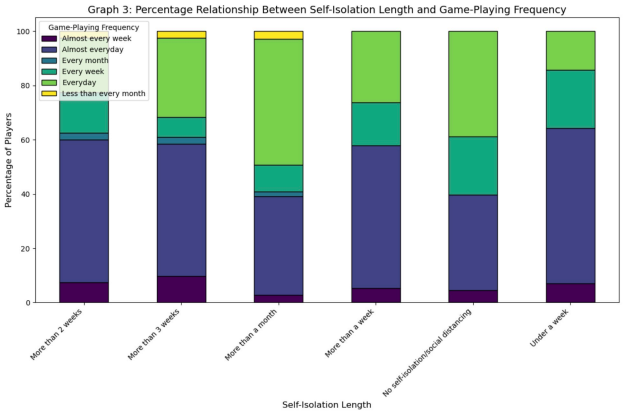


Figure 3: Stacked Histogram Showing the Percentage of Players in Each Game Playing Frequency Category for Each Social Distancing/Self-Isolation Length Category.

Graph 3 Interpretation

The "Everyday" category dominates across all self-isolation durations, especially for longer periods like "More than a month," indicating frequent daily gaming was common during prolonged isolation. "Almost every day" consistently ranks second, while infrequent gaming ("Less than every month" or "Every month") is under-represented, suggesting fewer casual gamers participated or turned to gaming during the pandemic. Even for shorter or no isolation periods, many respondents reported playing daily or almost daily, implying the survey reflects the gaming population more than the general population.

Longer isolation durations appear to correlate with more consistent and frequent gaming habits, whereas shorter durations or no isolation are associated with more variability in the frequency of gaming.

2.4 A 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".

This section examines the relationship between responses to the statement "I lost connection with the outside world" (rated on a 1-5 scale, with 1 being "Not at all" and 5 "Extremely") and the length of self-isolation. The mean response values were calculated for each isolation length to understand attitudes between categories.

"No self-isolation/social distancing" has the lowest mean response, around 2.5 whilst "Under a week" also showed a relatively lower mean response compared to other isolation lengths. This indicates that shorter isolation periods or no isolation were associated with a reduced feeling of disconnection. This could be because they were still partaking in their usual daily activities (e.g. going to work, meeting family).

"More than 2 weeks," "More than 3 weeks," and "More than a week" had the highest mean responses (around 3.5), suggesting greater disconnection during moderate isolation periods. However, the slightly lower mean for "More than a month" (around 3) indicates that prolonged isolation may have led to adaptation or reconnection efforts, resulting in only moderate feelings of disconnection.

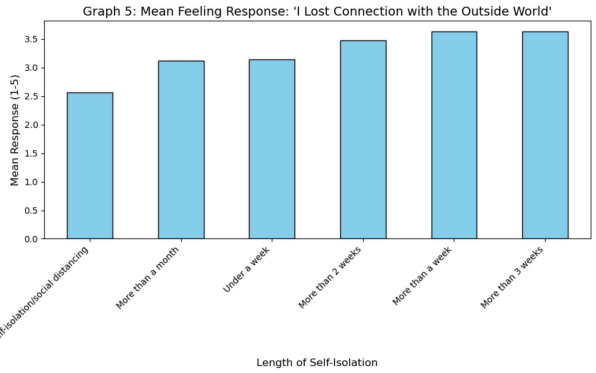
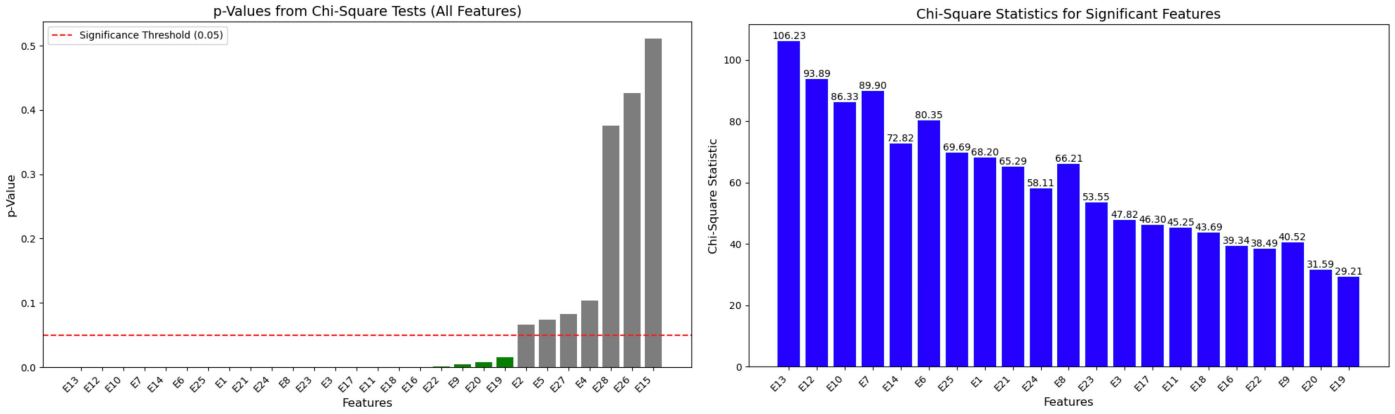


Figure 4: Histogram Showing The Mean Response (1-5) To The Statement 'I lost connection to the outside world' For Each of The Self-Isolation Length Categories.

3 Identifying the most important in-game behaviours that indicate the player's length of being self-isolated/ social distancing.

3.1 Chi-Squared Tests

The Chi-Square test was applied to assess the relationship between independent variables and the target variable, identifying statistically significant features for further analysis. By calculating the Chi-square statistic and the p-value, the features with $p < 0.05$ were prioritised, offering a robust basis for feature selection in modelling. Visualisations highlight significant features with strong associations, such as E13 and E12, which exhibit the highest Chi-Square values. These features, ranked by significance, are critical for predicting the duration of self-isolation and will inform the subsequent development of the model.



(a) p-Values

(b) Chi-Squared

Figure 5: a) p-values from Chi-Squared Tests for all features. b) Chi-Squared Values for significant features.

3.2 Mutual Information Tests

This section evaluates the relationship between independent variables and the target variable using mutual information, which captures both linear and non-linear dependencies to quantify how much one variable informs another. Mutual information is particularly suited for ordinal data, complementing the Chi-Square analysis by offering a broader perspective on feature importance.

The analysis highlighted **E13 (0.085) and E12 (0.071**) as the most informative features, aligning with Chi-Square results and reinforcing their significance. While many features contributed minimally, the combined insights from both methods provide a robust foundation for feature selection and predictive modelling.

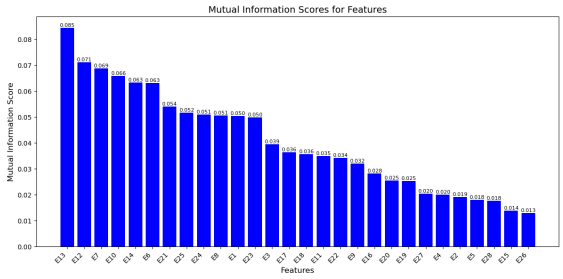


Figure 6: Mutual Information Test Results

3.3 Why Chi-Squared and Mutual Information?

The **Chi-Square test** and **mutual information** were chosen for feature importance evaluation due to their complementary strengths. The Chi-Square test identifies statistically significant associations between ordinal variables with computational efficiency, making it ideal for small datasets. **Mutual information** captures both linear and non-linear dependencies, offering a broader perspective on feature importance. Together, these methods provide a robust analysis, combining statistical significance with dependency measurement for effective feature selection.

3.4 Excluded Methods

Target Encoding and **Cramér's V** were excluded to prioritise more relevant techniques. Target Encoding, which replaces categories with the target variable's mean, was avoided due to potential bias from class imbalance. Cramér's V, derived from the Chi-Square test, was excluded as it offered no additional insights beyond the Chi-Square statistic.

3.5 Conclusion

Key features, particularly E13 and E12, were strongly associated with the target variable, as confirmed by Chi-Square and Mutual Information analyses. Robust data cleaning and careful consideration of the dataset's ordinal nature ensured the validity of the feature selection process, providing a solid foundation for predictive modelling and deeper insights.

4 Developing a Machine Learning Model To Predict Players' Length of Being Self-Isolated Based on In-Game Behaviours Only

4.1 Introduction

This section develops a machine learning model to predict players' length of self-isolation based on in-game behaviours. It is framed as a multi class classification problem with 6 target classes and 28 ordinal features, measured on 1–4 or 1–5 scales which represent various in-game behaviours identified during data exploration. The objective is to train a model that accurately classifies players into the six categories, addressing class imbalance and delivering meaningful predictions across all target classes.

4.2 Understanding the Dataset

Here we examine the distributions of the feature and target variables to identify class imbalance, skewness, and variability, informing preprocessing and modelling decisions.

4.2.1 Target Variable Distribution

The target variable, representing self-isolation duration, is dominated by the 'More Than a Month' category (63%), with 'No Self-Isolation' at 17% and the remaining categories under 10%. This imbalance risks biasing the model toward majority classes at the expense of smaller ones.

Implications For Modelling

Addressing class imbalance requires strategies like class weighting to penalise misclassification of smaller classes. Metrics such as F1-score, precision, and recall are critical for fair evaluation, as overall accuracy may favour majority classes. Random Forests, with hyper-parameters like `class_weight`, are well-suited to handle such imbalances.

4.2.2 Feature Variables Investigation

The independent variables, measured on 1-4 or 1-5 scales, exhibit varying distributions. Variables like **E10, E12, E21, and E25** show little variability, reflecting universal behaviours, while balanced variables (**E18, E22, E23, and E24**) show weaker correlations but may add value through interactions. Notably, **E13** ('Never') is a key predictor, likely linked to extended isolation periods.

Similarly, variables like **E6, E7, E8, and E9** are dominated by "Very Likely," indicating strong correlations with the target variable but limited variability. Mid-scale responses such as "Maybe" and "Likely" are rare, reducing overall balance.

Implications for Modelling

Skewed variables, while informative, may require adjustments to limit their dominance. Balanced features, though weaker individually, could prove valuable through interactions with other predictors during model development.

4.3 Chosen Model & Methodology

4.3.1 Introduction

Here we outline the process of selecting the most suitable machine learning model for predicting players' length of self-isolation. It begins by evaluating various potential models, highlighting why certain approaches were deemed unsuitable for this task. The chosen model is then presented, along with the motivations for its selection, focusing on its alignment with the dataset's characteristics and the project's objectives.

4.3.2 Chosen Model

After evaluating various approaches, **Random Forests** was selected as the most suitable model for this task. Random Forests are capable of handling ordinal features without extensive preprocessing, utilise the `class_weight` parameter to address class imbalance, and are robust against overfitting on small datasets. Their ensemble nature also enables them to capture complex, non-linear relationships, making them a reliable choice for this dataset.

Other models were considered but deemed unsuitable due to various limitations. **Logistic Regression** assumes linear relationships and struggles with multi-class and imbalanced data. **SVC** is computationally expensive and biased toward majority classes. **Naïve Bayes** fails due to unrealistic feature independence assumptions, while **k-NN** requires unsuitable distance metrics for ordinal features. **Decision Trees** overfit small datasets, and **Neural Networks** are overly complex for this task. **LDA** assumes multivariate normality, making it unsuitable for categorical or imbalanced data. While **XGBoost** and **LightGBM** manage non-linear relationships, XGBoost lacks direct class imbalance support for multi class target variables, and LightGBM is sensitive to hyper-parameter tuning and less interpretable.

These limitations highlight Random Forests as the most robust option for addressing class imbalance, ordinal data, and small sample sizes.

4.3.3 Methodology

The Random Forest model will be developed through a structured approach to ensure robust predictions. Data preprocessing will address class imbalance, split the data into training, validation, and test sets, and represent ordinal features appropriately. The model will be trained using stratified splits to maintain class proportions, with hyper-parameter tuning to optimise performance and balance under-fitting and over-fitting. Evaluation metrics, including F1 score, precision, and recall, will focus on performance across all classes, especially under-represented ones. This approach aims to create an accurate model for predicting players' self-isolation durations based on in-game behaviours.

4.3.4 Data Preprocessing

Stratified sampling was applied during data splitting to maintain the target variable's class proportions across training, validation, and test sets, ensuring a balanced and representative distribution. Key features identified through Chi-Square and Mutual Information analyses were combined with results from a model-driven feature importance test to determine the top 10 most predictive features. These features were prioritised during training and utilised in the feature reduction section to minimise noise and enhance model performance.

4.4 Hyper-parameter Tuning

Hyper-parameter tuning is essential for optimising the Random Forest model, balancing under-fitting and over-fitting while addressing challenges like class imbalance and a small dataset. Through a randomised search over a well-defined parameter space, the optimal configuration was identified, achieving a balance between performance and computational efficiency for the predictive task.

4.5 Evaluation on the Validation Set

After completing hyper-parameter tuning, the validation set is used as an additional step to confirm the model's generalisation ability, providing an unbiased assessment of its performance before testing on the final test set.

The model achieved a weighted F1-score of 0.634, indicating difficulties in balancing precision and recall, while the accuracy of 0.661 is misleading due to class imbalance favouring the dominant class, "More than a month" (Class 5). Class 5 is well-predicted with an F1-score of 0.82, but minority classes perform poorly, with Classes 2 and 4 entirely missed (F1-scores of 0.00). Misclassification's frequently assign minority classes to Class 5, reflecting the model's bias.

4.5.1 Challenges and Solutions

Severe class imbalance, limited data (628 rows), and noisy features hinder generalisation. To address this, feature selection will prioritise high-importance features to reduce noise and over-fitting. Additionally, merging under-represented classes (Classes 1–4) into "Short Isolation" (less than one month) creates a more balanced target variable, simplifying the classification task and improving performance metrics like F1-score.

4.6 Performance Evaluation

4.6.1 Introduction

This evaluation assesses the performance of the optimised Random Forest model on the test set. The model's performance is evaluated incrementally, testing baseline performance, the impact of feature reduction, class reduction, and the combined adjustments.

4.6.2 Evaluation Metrics

To evaluate the performance of the Random Forest model on this imbalanced multi-class dataset, metrics were chosen to ensure fairness and relevance. The **weighted F1-score**, **precision** and **recall**, while metrics like accuracy and ROC-AUC were excluded due to their limitations.

The **weighted F1-score** balances precision and recall by assigning greater weight to classes with more instances, making it ideal for imbalanced datasets. It evaluates the model's ability to handle class imbalance while maintaining predictive strength.

Precision measures the proportion of correct predictions for each class, evaluating the avoidance of false positives, while **recall** measures the proportion of correctly identified instances, particularly for minority classes. Together, these metrics provide a granular understanding of model strengths and weaknesses.

Metrics like **accuracy** and **ROC-AUC** were excluded, as accuracy is skewed by majority classes in imbalanced datasets, and ROC-AUC is less interpretable for multi-class tasks.

These selected metrics ensure a fair and clear evaluation, addressing the challenges of class imbalance and highlighting areas for improvement.

4.6.3 Evaluation of Model Performance

The performance of the optimised Random Forest model was evaluated under four scenarios: the baseline model, feature reduction, class reduction, and a combination of both. Tables 1 and 2 summarise the overall metrics and class-level F1-scores, respectively.

4.6.4 Overall Metrics

Table 1 summarises the weighted metrics across all scenarios. The baseline model performed well for the majority class but struggled with minority classes. Feature reduction improved precision by reducing noise, though recall declined slightly. Class reduction showed the largest improvement in recall, while the combined approach balanced performance across all metrics, achieving the best overall results.

Table 1: Overall Performance Metrics for Different Scenarios

Scenario	**Weighted F1-Score**	**Precision (Weighted)**	**Recall (Weighted)**	**Accuracy**
Optimised Model Only	0.5685	0.5489	0.5968	0.60
Optimised with Feature Reduction	0.5573	0.5817	0.5484	0.55
Optimised with Class Reduction	0.6023	0.5648	0.6452	0.65
Optimised with Both	0.6096	0.5972	0.6290	0.63

4.6.5 Class-Level F1-Scores

Table 2 highlights F1-scores for individual classes. The baseline model performed well for the majority class but struggled with minority classes. Feature reduction marginally improved Class 0, while Class reduction enhanced performance for minority classes by consolidating categories. The combined strategy offered the most balanced results, though Class 0 remained underrepresented, reflecting persistent challenges with extreme class imbalance.

Table 2: Class-Level F1-Scores for Different Scenarios

Scenario	**Class 0 (No Isolation)**	**Class 1 (Short)**	**Class 2 (Moderate)**	**Class 3 (Moderate-Long)**	**Class 4 (Long)**	**Class 5 (Very Long)**
Optimised Model Only	0.22	0.00	0.00	0.20	0.00	0.80
Optimised with Feature Reduction	0.25	0.00	0.40	0.25	0.00	0.75
Optimised with Class Reduction	0.00	0.43	0.81	-	-	-
Optimised with Both	0.12	0.48	0.78	-	-	-

4.6.6 Insights and Conclusion

The results demonstrate that the combined approach, which integrates feature and class reduction, delivers the most balanced performance across classes. It addresses some of the challenges posed by minority class predictions and noise in the data.

5 Conclusion

This project analysed survey data to explore the relationship between in-game behaviours in Animal Crossing and the duration of self-isolation or social distancing. The survey creators suggested that activities like "Enjoying the scenery" and "Participating in Mystery Island Tours" might appeal to players seeking escapism, while "Sending gifts to friends" could foster social connection during isolation.

Our analysis provided a nuanced picture. Feature E27 ("Enjoy the scenery") showed a high p-value, low mutual information score, and low feature importance, suggesting it has minimal impact on self-isolation duration. Feature E5 ("Participate in Mystery Island Tours") also showed weak direct correlation but ranked highly in feature importance, indicating potential indirect effects or interactions.

Feature E13 ("Send it to a friend as a gift") emerged as the most impactful variable, with strong statistical significance and the highest feature importance score. To further evaluate the relationship between E13 and the target variable, we conducted Spearman's rank correlation analysis, which is specifically suited for ordinal data. The results indicated a weak negative correlation (Spearman Correlation: -0.231) with a significant p-value (6.05e-08). These findings suggest that higher levels of E13 are weakly associated with shorter durations of self-isolation, contrary to the suggestions of the survey creators.

However, the study is limited by the separation between gaming behaviours and real-world coping mechanisms, as games primarily serve as entertainment. These findings highlight the complexity of linking virtual behaviours to real-world outcomes.

Overall, the report found limited evidence to support the survey creators' claim of a link between extended isolation durations and higher activity in **E13** and **E27**. While **E13** showed statistical significance, its weak negative correlation suggests no strong association with isolation duration. Similarly, **E27** exhibited minimal impact, indicating these variables are not reliable indicators of extended isolation.