Projet SY09 - Étude du jeu "Animal Crossing: New Horizons"

Reda Sarehane, Ismail Kadiri

6 juin 2022

Résumé

Dans le cadre d'un projet académique réalisé au cours de l'UV SY09, nous avons pu analyser les données du jeu *Animal Crossing : New Horizons*, dont l'abréviation *ANCH* sera utilisée pour y référer, afin d'y appliquer les méthodes étudiées au cours du semestre. Ce rapport a pour but de rendre compte du travail effectué et du cheminement intellectuel suivi tout au long de la réalisation de ce projet. L'entièreté du projet est disponible dans le GitHub.

1 Introduction

ANCH est un jeu vidéo de simulation de vie développé par Nintendo, sortit le 20 Mars 2020 sur la console Nintendo Switch. Dans ce jeu, le joueur est maître d'une île déserte qu'il peut explorer et en recueillir les ressources afin de fabriquer des outils. Le joueur peut s'adonner à diverses activités comme la pêche ou le jardinage, et personnaliser dans son entièreté son île et son avatar. Le joueur partage son île avec des personnages dits villageois de caractéristiques différentes.

Les données relatives au jeu récoltées à partir du GitHub de TidyTuesday[1] permettent d'avoir des données d'une part relative au jeu à travers des informations sur les villageois (leur sexe, espèce, personnalité) et sur les objets du jeu (leur prix d'achat et de vente et leur catégorie), et d'autre part relatives aux avis des utilisateurs et de la presse ainsi que la note attribuée. Ce jeu de données initial a été complété par différentes informations complémentaires provenant de sources détaillées dans la suite du rapport. Cet apport a été fait dans le but d'apporter une analyse plus précise et d'élargir le spectre des méthodes applicables sur le jeu de données. Nous avons notamment apporté de nouvelles données relatives aux villageois tel que le signe astrologique, le hobby, le nombre d'apparitions dans les différents jeux de la saga, ainsi que le rang de popularité. Des données relatives à la compatibilité entre villageois et aux différentes personnalités ont également été ajouté.

L'ensemble de ces données nous ont permis d'axer notre étude selon 2 problématiques :

1. est-il possible d'inférer la ou les différentes caractéristiques qui permettent de déterminer la popularité d'un villageois parmi tous les autres ?
2. quelles sont les raisons poussant un utilisateur à aimer ou à ne pas aimer le jeu ?

Nous avons donc choisi de ne pas utiliser le jeu de données relatif aux objets du jeu car il n'apporte aucune information significative à notre analyse.

Pour répondre à ces problématiques, nous avons structuré le rapport tel qu'il suit : tout d'abord, nous allons présenter la phase d'exploration des données et les analyses. Cette première phase nous a mené vers des problèmes que nous allons expliquer et dont nous détaillerons les choix arbitraires que nous avons fait pour les résoudre. Ensuite, nous présenterons les résultats des différentes méthodes utilisées avant de conclure et présenter d'éventuelles perspectives d'études que nous aurions pu entreprendre pour répondre à notre problématique à travers une méthode alternative.

2 Popularité des villageois

2.1 Complétion du jeu de données

Le jeu de données initial est assez réduit, limitant les perspectives d'analyse. De plus, la présence majoritaire de variables uniquement catégoriques limite le nombre de méthodes applicables ainsi que la précision des résultats. Pour aboutir à notre problématique, nous avons été contraint d'enrichir ce jeu de données.

Pour réaliser une analyse sur la problématique de la popularité d'un villageois nous avons dû tout d'abord compléter ce jeu par des informations relatives au rang des villageois. Ce jeu de données complémentaires, disponible grâce à un utilisateur sur Kaggle[2], renseigne pour chaque villageois son rang et son appartenance à une catégorie comprise entre 1 et 6, 1 étant la catégorie la plus populaire de villageois et 6 la moins populaire. Il est nécessaire de préciser que l'information de popularité est arbitraire et découle d'un classement réalisé par

des joueurs du jeu. Suite à cet ajout, le jeu de données décrit ainsi 413 villageois.

Afin d'apporter plus de matière à notre analyse, nous avons mis-à-jour le jeu de données avec de nouvelles données relatives aux villageois (le signe astrologique, le hobby, le nombre d'apparitions dans les différents jeux de la licence à l'aide de l'API proposée par Nookipedia[3], un wiki communautaire autour de l'univers d'Animal Crossing. Nous avons également pu traduire une table de compatibilité[4] présente sur le Wiki en une matrice de similarité. Pour cette table, nous avons fait l'hypothèse que la compatibilité entre villageois est la même que pour le précédent opus de la série, *New Leaf*.

Enfin, nous avons essayé de traduire les différentes personnalités en introduisant des variables binaires représentant différentes caractéristiques plus ou moins communes entre les villageois à l'instar de l'amabilité, la maturité ou encore l'égocentrisme. Cette implémentation a été réalisée grâce aux différents descriptifs présents sur un autre wiki communautaire[5]. Il est primordial de garder à l'esprit qu'il est difficile de caractériser une personnalité dans le jeu à partir simplement de quelques variables binaires.

2.2 Exploration des données

2.2.1 Analyse préliminaire

Une première exploration des données brutes permet d'obtenir quelques aperçus sur les caractéristiques des villageois. On peut notamment remarquer qu'il y'a quasiment autant de villageois mâle que femelle et que la répartition des hobbies et des signes astrologiques est la même. Ces critères peuvent donc sembler à première vue non discriminants quant à l'appartenance d'un villageois à une certain catégorie de popularité ou tier. Quant à la distribution des différentes espèces de villageois (au nombre de 35), elle diffère puisque certaines espèces sont plus présentes que d'autres (l'espèce de villageois la plus présente est le chat, contrairement à la vache et à la pieuvre qui sont les moins dénombrées). Ce premier critère semble ainsi intéressant à investiguer pour la suite de l'analyse. La répartition des différentes personnalités (au nombre de 8) est aussi particulière car les personnalités ne sont pas partagées par les deux sexes du jeu. En effet, une villageoise peut être grande soeur, arrogante, vive ou normale. Un villageois peut être versatile, sportif, paresseux ou chic. Nous n'avons pas trouver de raison pour cette implémentation. Leur distribution est plus ou moins égale sauf pour les grandes soeurs et les chics qui sont moins nombreux.

2.2.2 Analyse de la popularité

L'analyse de la distribution du tier (cf. figure 1) permet de tirer quelques conclusions quant à la suite du travail : les tier n'ayant pas une distribution égale, la tâche d'identifier le niveau de popularité d'un villageois sera compliquée et la tâche de prédiction imprécise. Cela découle du nombre d'individus insuffisants et du nombre d'individus relativement faible des tiers supérieurs (1,2,3 et 4). Une analyse plus poussée nous permet d'obtenir les espèces et personnalités les plus populaires en moyenne. Nous avons considérer la moyenne pondérée par le nombre d'individus appartenant à la catégorie car certaines espèces sont représentées en faible effectif, à l'instar de la pieuvre qui ne compte que 4 individus. Ainsi, les chats sont les plus populaires, à l'opposé des vaches qui sont les moins populaires parmi la communauté de joueurs. De même, les villageoises normales sont les plus populaires, à l'opposée des villageoises grandes soeurs. Nous pouvons donc penser que ces deux critères jouent un rôle essentiel dans la détermination de la popularité d'un villageois dans le jeu.

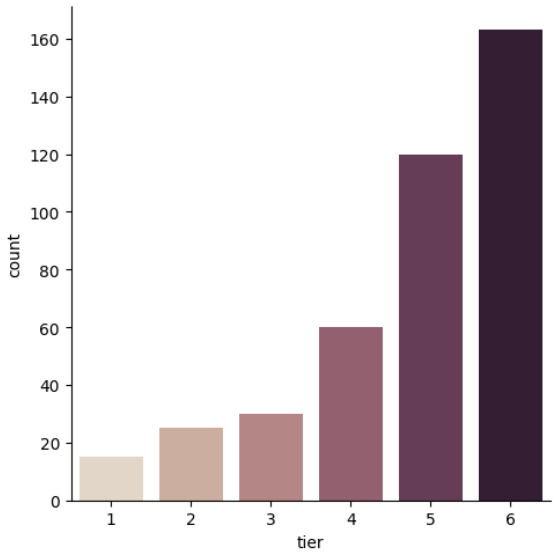

FIGURE 1 – Distribution des différents tier

2.2.3 Analyse de la personnalité

Grâce aux variables binaires caractéristiques de la personnalités, chacune arbitrairement choisie à partir du descriptif des personnalités présent dans les wiki communautaires, nous avons pu réaliser une ACP afin de visualiser les différences entre personnalités et transformer la variable catégorie de la personnalité en variable numérique. La représentation sous 2 axes (cf. fi-

gure 6) nous permet d'établir des clusters de pairs de personnalités. L'axe des abscisses semble traduire des caractéristiques positives pour les valeurs négatifs, à savoir l'amabilité et l'attention, et des caractéristiques négatives pour les valeurs positives, à savoir l'impolitesse, l'entêtement et l'excès de colère. L'axe des ordonnées traduit quant à lui l'arrogance et l'égocentrisme.

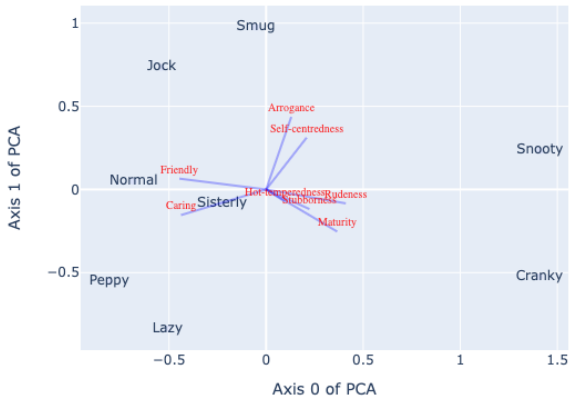

FIGURE 2 – Représentation de la personnalité sur 2 axes

2.2.4 Analyse de la compatibilité

La matrice de similarité explicite le niveau de compatibilité entre chaque paire de villageois. Ce niveau prend des valeurs discrètes, 0 pour signifier que les villageois ne sont compatibles, 1 pour une compatibilité intermédiaire et 2 pour une compatibilité élevée. D'après la source de la matrice, cette compatibilité est obtenue en fonction de l'espèce, la personnalité et le signe astrologique d'un villageois. Afin de réaliser une AFTD sur cette donnée et pouvoir la représenter, nous avons fait le choix arbitraire de transformer les données en appliquant la similarité cosinus (1), qui calcule la similarité entre deux échantillons X et Y :

$$K(X,Y) = \frac{< X,Y >}{|X| * |Y|} \quad (1)$$

La matrice résultante traduit la compatibilité pour chaque paire de villageois en fonction de la compatibilité de chacun des individus composant la paire avec les autres villageois. Nous avons tout d'abord fait une analyse des proximité afin de pouvoir représenter les données dans un espace de plus faible dimension (la matrice initiale ayant 413 colonnes et 413 lignes). La représentation optimale se fait sous 5 axes (on obtient l'optimum de stress que l'AFTD cherche à minimiser). Nous avons ensuite cherché à détecter des clusters. Nous avons utilisé cette matrice de similarité pour faire de la classification ascendante hiérarchique et détecté des similarités entre les individus. Après analyse, nous avons remarqué que cette classification permet de mettre en valeur les différences entre personnalités : chaque cluster, au nombre total de 4 (cf. figure **??** réunit des villageois de 2 personnalités distinctes. Ainsi des clusters se forment entre les personnalités normales et grincheux, les sportifs et les vives, les paresseux et les grandes soeurs et enfin entre les arrogants et les chics. D'après la table de construction de la matrice de compatibilité, ces paires de personnalités sont très compatibles (si on ne prend en compte que l'information de la personnalité). Une autre conclusion plutôt intéressante et que ces clusters réunissent une personnalité de villageois et une autre de villageoises. Cette analyse montre des résultats opposés à ceux obtenus dans l'analyse de la sous-partie précédente. Il semble nécessaire de préciser que cette table de compatibilité n'est pas une table officielle. Cependant, elle reste plus précise que la table que nous avons voulu implémenter en traitant les différentes personnalités.

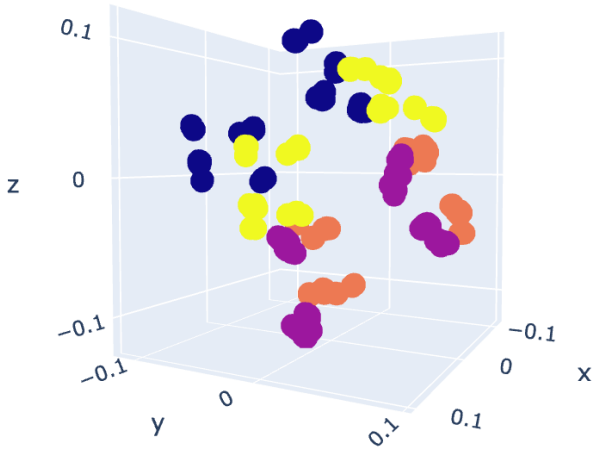

FIGURE 3 – Représentation de la matrice de compatibilité sur 3 axes et des 4 clusters

2.3 Classification des données

La tâche de classification semble compliquée de prime abord. Le jeu de données ne contient que 413 observations, et nous cherchons à prédire une information subjective et arbitraire qu'est la popularité d'un villageois (plus précisément son appartenance à un tier en fonction de ses caractéristiques)

2.3.1 Variable catégorique

Un des problèmes majeurs du jeu de données est la présence majoritaire de variables catégoriques (personnalité, espèce, signe astrologique, sexe). La majorité des méthodes de classification nécessite un jeu en entrée qui soit numérique (en principe, la méthode des arbres peut traiter des données catégoriques, mais son implémentation ne le permet pas. Une des solutions est d'encoder ces variables. Deux méthodes sont possibles, l'encodage dit ordinal mais qui suppose une relation d'ordre entre les variables, et l'encodage one-hot qui créera une variable binaire pour chaque différente catégorie. Étant donné que nous n'avons pas de relation d'ordre parmi les différentes variables catégoriques, nous utilisons ici l'encodage one-hot. Cet encodage entraîne en revanche le problème du fléau de la dimension puisqu'on obtient

2.3.2 Prédiction

Pour la prédiction, nous avons choisi quelques modèles que nous avons appliqué sur deux jeux de données :

1. les données catégoriques encodées (donc le jeu de données incluant le sexe, l'espèce, la personnalité, le hobby, le signe astrologique ainsi que le nombre d'apparitions dans les différents jeux de la licence)
2. les données issues de l'AFTD de la matrice de compatibilité

Nous avons utilisé comme modèle prédictifs les K plus proches voisins, les arbres, les forêts, AdaBoost et la régression logistique. Pour chacun des modèles nous avons tout d'abord découpé l'ensemble des données en ensemble d'apprentissage et de validation. Ce découpage a été réalisée avec de la stratification afin de préserver le pourcentage d'individus de chaque classe. Nous avons également optimisés les hyper-paramètres pour chaque méthode en utilisant une grille de recherche (GridSearch). Enfin, nous avons réaliser une validation croisée afin d'évaluer les performances des différents modèles

2.3.3 Résultats

Comme ce que nous avions prévus, les modèles ne sont pas très performants (cf.figure 4). Nous pouvons cependant en tirer plusieurs conclusions :

1. les modèles performent mieux sur le jeu de données encodés, contrairement au jeu de données liés à la compatibilité. Cela semble logique car la seule information de la compatibilité d'un villageois avec les autres n'est probablement pas suffisante pour qu'un utilisateur le préfère plus qu'un autre.
2. la précision limitée des modèles est en parti dû au manque crucial de données. Nous avons relativement peu de villageois, et encore moins de villageois appartenant aux catégories (tier) les plus populaires. Ce jeu de données ne se prête donc pas à une tâche prédictive.
3. les villageois d'ACNH sont un peu plus que quelques catégories et nombres : c'est des avatars uniques, chacun ayant son comportement et ses dialogues. Cela a déjà été précisé mais la popularité des villageois parmi les utilisateurs est un critère arbitraires.

2.4 Conclusion

Lorsqu'on cherche à comprendre pour quelles raisons un joueur préfère un villageois, on découvre très vite que c'est plus une question de feeling. En se balandant sur le Reddit du jeu, on comprend assez vite que chaque histoire est unique, ce qui fait la beauté du jeu. Certains joueurs créent un réel lien entre eux et un ou plusieurs villageois car le réconfort apporté par le jeu leur permet de créer une bulle leur permettant d'échapper à la réalité. L'exercice auquel nous nous sommes prêter ne sera peut-être jamais réalisable avec une précision parfaite. Cependant certains éléments peuvent venir enrichir cette analyse, à l'instar d'une analyse de l'avatar des villageois ou une analyse de sentiments appliqués sur les différents dialogues qu'ils peuvent entamer. Néanmoins, la beauté du jeu est intiment liée à son imprévisibilité.

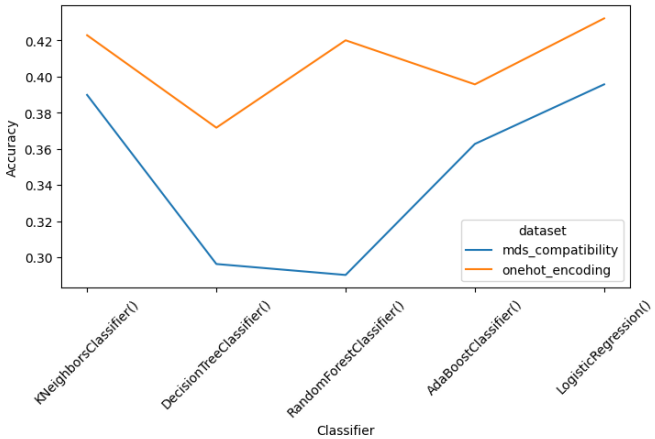

FIGURE 4 – Scores des différents modèles prédictifs sur les 2 jeux de données

3 Approche alternative

Le jeu de données étant particulièrement hétérogène, nous avons choisi d'explorer une approche alternative quant à l'analyse des jeux de données de commentaires. Ces jeux de données contiennent des informations textuels sous la forme d'un commentaire et des informations quantitatives sous la forme d'une note sur dix. Nous avons utilisé des méthodes NLP, ou natural language processing, afin d'en tirer des données des différents commentaires que nous avons ensuite traité.

3.1 Analyse initiale

Lors de notre analyse globale du jeu de données nous avons pu tirer plusieurs premiers éléments de réponse à notre problématique initiale. En effectuant un premier histogramme sur nos données bruts nous remarquons une distribution particulière.

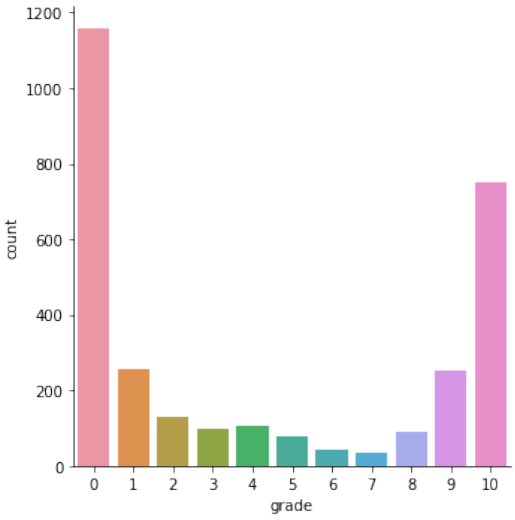

FIGURE 5 – Distribution des notes du jeu de données

Nous pouvons remarquer le grand nombre de 0 attribués au jeu par les joueurs. En triant les commentaires en fonction de leurs notes nous pouvons remarquer que la majorité des joueurs ayant attribué une note de 0 au jeu à cause de l'unicité du terrain de jeu généré initialement. En effet, il n'y a qu'une seule île générée par console et cela pose problème pour un grand nombre de joueurs.

3.2 Approche NLP

Afin de confirmer nos hypothèses tirées de notre analyse initiale nous avons décidé d'utiliser le principe de tokenisation pour vérifier qu'il existe bien une corrélation entre les notes de 0 et l'unicité du terrain de jeu. La tokenisation nous permet de séparer un corpus de textes en une liste de mots analysables. Ces mots sont souvent nettoyés pour ne garder que des mots pertinents au sens et au contexte du texte. Les mots de liaison, comme les déterminants, sont ignorés car inutiles à la compréhension du sens d'une phrase. Nous remarquons alors que les mots les plus utilisés dans les commentaires notés 0 sont "one", "island", "game", "switch" et "play". Ces mots étant bien dans le contexte de notre hypothèse nous avons décidé de creuser cette piste plus loin. Afin d'avoir un point de vue plus global, nous avons généré un nuage de mot correspond au corpus de texte.

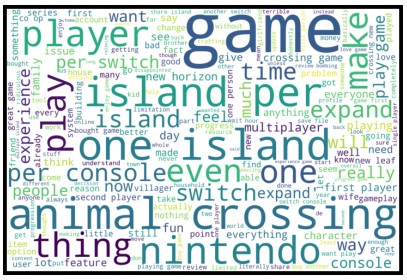

FIGURE 6 – Nuage de mot du corpus de textes

Nous pouvons voir que de manière générale, les mots qui ressortent le plus restent dans le contexte de notre hypothèse initiale. Nous ne pouvons donc pas réellement tirer des conclusions quant à la corrélation directe entre l'unicité du terrain de jeu et les mauvaises notes attribuées au jeu.

3.3 Méthode TF-IDF

Comme nous l'avons vu dans la section précédente, nous avons besoin de plus d'information sur le contexte des mots afin de tirer des réponses pertinentes à notre hypothèse. La pondération TF-IDF, ou term frequency-inverse document frequency en anglais, permet de nous donner une idée sur l'importance d'un mot dans le contexte d'un corpus de texte.

Nous avons d'abord attribué un contexte à chaque mot, pour chaque mot donné nous allons chercher à connaître le mot qui le précède et le mot qui le succède. C'est ainsi que nous avons constitué un premier dictionnaire python qui pour chaque mot nous donne

une liste de mot qui se trouve dans son contexte.

Ensuite, nous avons calculé la pondération TF-IDF de chaque mot dans le contexte de tous les mots du document. Voici les formules de ces deux pondérations :

$$idf(\boldsymbol{x}) = \log \frac{D}{C}, \tag{2}$$

Avec D le nombre total de documents dans le corpus et C le nombre de document dans lesquels le terme apparaît

$$tdf(\boldsymbol{x}) = \frac{N}{M}, \tag{3}$$

Avec N le nombre d'apparitions du terme dans le document et M le nombre total de termes dans le document.

Pour un mot A donné, nous avions un vecteurs de n dimensions, n étant le nombre total de mots uniques, dont la valeur de chaque colonne correspondant à la pondération TF-IDF du mot A dans le contexte du mot de la colonne.Nous avons ensuite crée une liste de distance entre les différents vecteurs afin d'essayer d'en tirer une conclusion ou trouver une corrélation entre les différents mots du texte. Cependant, étant donné la grande quantité de commentaires et de mots, le calcul de la matrice de distance nous a pris beaucoup de temps, et cette dernière est assez conséquente en terme de taille. La matrice de dissimilarité nous a pris plus de 45 minutes à calculer et cette dernière faisait plus de 650 Mbs.

3.4 Conclusion

Cette méthode nous a permis d'avoir un début de réponse à la deuxième partie de notre problématique, à savoir, ce qui pousse un joueur à aimer ou ne pas aimer le jeu. En effet, nous avons pu faire un lien entre les notes et le contenu des différents commentaire que nous avons commencé à démontrer par des méthodes statistiques. Cependant, il y a encore quelques axes d'amélioration. Nous aurions pu améliorer cette approche en procédant à un nettoyage des données au préalable en supprimant les stop-words inutiles et en calculant les distances et les contextes pour chaque note séparément. Cela nous permettrai de réduire considérablement notre temps de calcul puisque nous réduirons le nombre de dimensions. De plus, nous aurions pu prendre en compte les commentaires anglais seulement. En effet, il n'y a aucun contexte linguistique entre les mots de deux langues différentes et cela rajoute des calculs inutiles et des dimensions inutiles à notre problème.

Références

[1] TidyTuesday, Animal Crossing - New Horizons. *Disponible en ligne à cette URL.*

[2] Animal Crossing Portal Villager Popularity A dataset of animal crossing new horizons tier list, and ranking. *Disponible en ligne à cette URL.*

[3] Nookipedia, community-driven Animal Crossing wiki encyclopedia *Disponible en ligne à cette URL.*

[4] Compatibility table from Nookipedia *Disponible en ligne à cette URL.*

[5] Animal Crossing Fandom, community-driven Animal Crossing wiki encyclopedia *Disponible en ligne à cette URL.*