Prediction of Popularity of the Steam Store Games to Increase Sales

Vikas Kishanrao Thamke
National College of Ireland
Dublin, Ireland
x19180080@student.ncirl.ie

***Abstract*—This paper presents the design document of project to be worked on in the module of Domain Application of Predictive Analysis. The domain selected for the analysis is video game industry which is the highest revenue generating industry in entertainment sector. The dataset used is generated by Steam, a popular digital distribution platform of video games. Using intrinsic and extrinsic factors of games as features, the scope of this project is assumed to have the potential to accurately predict the range of users of a game. Taking ethical concerns of data into consideration, the business strategies are developed to boost the sales. The exploratory data analysis is performed on the dataset to comprehend emerging trends, patterns, relationship between variables, etc. Visualizations of these insights are unveiled using various plots such as bar plot, scatter plot, distribution plot, etc. Finally, the machine learning classification techniques that are applicable for the predictive analysis are listed.**

***Index Terms*—Steam Games, Predictive Analysis, Visualization, Trends, Machine Learning**

I. Introduction

A. Background of the domain

The provision of entertainment has never been considered as a subject of great interest by either economists or economic historians – at least in the working hours. Yet in the ever-evolving world of digital technology of current century, the presence of entertainment such as music, movies, sports, games, etc. has created profound psychological and economic impact. The giant gaming industry with 1.2 million video games worldwide [1], which are played for 3 billion hours each week [2] tops the entertainment sector. The revenue of this industry was $131 billion in 2018, which is expected to reach $300 billion by 2025 as claimed by Forbes [3]. Apart from economic impact, on analysing 116 scientific studies, it has been observed that video games cause changes to the region of brain, which is responsible for attention, competency, etc [4]. Also, being responsible for creating the feeling of empathy and compassion between players, World Economic Forum claims that video games are transforming the way of communication and has a potential to fix several other global issues [5].

Undeniably the developer community analyses intrinsic and extrinsic indicators that affect sales of video games [6], and keeps an eye on the recent engaging activities in games [7] to strengthen collaborative and social experience of the users [8], a much research is yet to be done on the popularity of games attributed to language, ratings, platform, publisher, genre, playtime, price, etc. combined. This project is an attempt to bring such distinct and independent factors to attention through the medium of predictive analysis using machine learning algorithms.

B. An Overview of the Data

The data picked for analysis is available on *Kaggle for exploration purposes, which is created from scratch using Steam Store and SteamSpy **APIs.

Column Name	Datatype	Description
appid	INTEGER	Id of the game
name	STRING	Name of the game
release_date	DATETIME	Date when the game is released
english	INTEGER	Is English used as a language
developer	STRING	Name of the developer
publisher	STRING	Name of the publisher
platforms	STRING	Compatible Operating System
required_age	INTEGER	Required age of the player
categories	STRING	Categories (multiplayer, online multiplayer, etc.)
genres	STRING	Genre of the game
steamspy_tags	STRING	Tag used by SteamSpy
achievements	INTEGER	Achievements of the game
positive_ratings	INTEGER	Number of positive ratings
negative_ratings	INTEGER	Number of negative ratings
average_playtime	INTEGER	Average playtime of the game
median_playtime	INTEGER	Median Playtime of the game
owners	STRING	Range of players who play the game
price	FLOAT	Price of the game

Fig. 1. Description of columns in the dataset

The Fig. 1 shows the descriptive information of columns of the dataset along with their datatypes. The column 'owners' is a categorical which holds the information of range of users of various games, considered as the target variable in this project.

C. Scope

The intrinsic factors such as developer, publisher, genre, platform can be useful for prediction of success before the

*https://www.kaggle.com/nikdavis/steam-store-games
**https://partner.steamgames.com/doc/webapi_overview

release of a game. However, the factors such as the cost spent on development, advertisement, and ratings for game's poster, trailer/teaser before release are not included in the dataset. This reduces the accuracy of prediction if the scope is considered before release of games.

The extrinsic factors that are included in the dataset such as achievements, steamspy tags, positive ratings, negative ratings are combined with intrinsic factors has the potential to increase the accuracy of prediction. Using these factors altogether, it is possible to predict success of game at different stages after release, such as after a week, month, year, etc.

II. Goal

The main objective of the project is to find the algorithm that will predict the range of number of owners of a game with maximum accuracy using intrinsic and extrinsic features considered for the analysis.

The secondary goal of the project is to find insights of the business using exploratory data analysis techniques and visualizations. These insights obtained would eventually assist to boost sales and revenue of the gaming business.

III. Ethical Concerns

The result of scientific methods is an outcome of rigorous hypothesis testing using empirical evidences. In the field of data science, the data obtained from various sources corroborates the finding. However, it is not always possible to develop datasets from scratch which enables reuse and sharing of existing datasets that are available publicly. While dealing with such open datasets, there are certain ethical concerns as mentioned in paper [9] that should be considered.

A. Identify the stakeholders

The primary stakeholders are the developers and publishers as they are connected directly to the data. The secondary stakeholder is the Steam platform which is going to benefit or suffer in the process of releasing game.

B. Informed consent

As the data is made publicly available through API by the Steam community, the consent of secondary stakeholders are satisfied. While making the data open, the Steam community has considered the consent of primary stakeholders as mentioned in the **documentation.

C. Identification of harms

The dataset does not include the reviews/ratings of users which eliminate the chances of harms to or from the users. Had the ratings provided by the Steam community kept confidential from developers/publishers, it would have been a harm as they are made available for the analysis. But that is not the case here.

D. Safeguards

As the data made available for the exploration after screening, the confidential information has been removed and safeguard policies are taken into consideration.

E. Justice

As the goal of this research does not aim toward unfair advantage or dis-advantage of any specific cultural, social or demographic group, the research does not cause injustice to them.

F. Public interest

The data considered for analysis is picked from public domain where necessary precautions are already taken before publishing data. This proves that the data is socially acceptable and does not form any ethical issues.

IV. Business Strategies

Using appropriate advance technological tools, the legacy and current data can be used to find insights of the business. The results obtained from predictive analysis can be obtained to construct various business strategies to enhance the popularity and sales of gaming applications discussed as follows.

A. Marketing using Cross Sales Penetration and Seasonal Information

To gather data, nowadays, most of the digital services mandate the customers for signing up and filling their basic information such as name, age, country, interest, etc. Combining this data with playing history and total time spend by a user, it is possible to find the probability of customer's potential to purchase other gaming products [10]. This marketing technique is termed as cross sales penetration. Also, the ideal season can be deduced from the metadata for campaigning and public attention, which would help to increase the popularity and sales, and maximize the profitability [11].

B. Game Monetization to Collect Revenue

The type of process for generating revenue from a video game using various models which can be used by the publisher is termed as video game monetization [12]. Predictive analysis can be used for the prediction of compatible model of monetization using the data collected from customers by investigation that utilizes a set of questionnaires [13]. Also, the prediction of percentage of all the players who would eventually subscribe the paid membership of game [14] can be implemented.

There are six fundamental models for monetization of a game [15]:

1) Free with In-App Advertising: This model allows users to download the game app for free and displays relevant business advertisements in break sessions to generate revenue from the businesses.

2) Free with Gated Features: The games where certain features could be kept as gated can use this model. The users are allowed to download the app for free however require money to unlock certain premium features.

3) Sponsorship: This model requires sponsorship to publishers to provide real time rewards to the users on various achievements. The publisher can generate revenue from sponsors through the advertisement.

4) Paid Games: The users require to purchase the game applications in this business aspect. The key to get succeed in this model is the ability of developer to develop a killer app and showcase its perceived values that differentiates the app from similar free apps.

5) Free with In-App purchasing: Selling physical or virtual goods to decorate the characters of game and generate revenue is the strategy of this model.

6) Paywall Subscriptions: This model is same as Free with Gated Features model except it focuses on gating the content, and not features. This model allows users to play a game for certain amount of time for free and prompts to signing up for paid membership.

C. Customer Churn Prevention to Avoid Losses

Most of the gaming companies started subscribing to systematic customer relationship management (CRM) tools to analyse the behaviour of customers [16]. From the data generated by CRM tools, it is observed that preventing a customer from churning is cost-effective than earning a new customer [17]. A churn prediction system can be built using machine learning algorithms to find out the user's customer lifetime [18]. In this way, the risk associated with a set of customers in terms of profit can be figured. Also, on taking appropriate initiatives, the customers attrition can be prevented [19].

V. Preliminary Visualisations

Data visualization is preliminary stage to find the insights of data. It helps to get the clear understanding of data quickly using few lines of code. The emerging trends, relationship between attributes, patterns of sales can also be found using simple visualizations such as bar graph, scatter plot, distribution plot, line graph, etc. as shown in the plots below.

A. Count of games categorized by users

As the variation between count of games in each category is very high, the log transformation of count is taken to reduce the scale of values, as shown in Fig. 2.

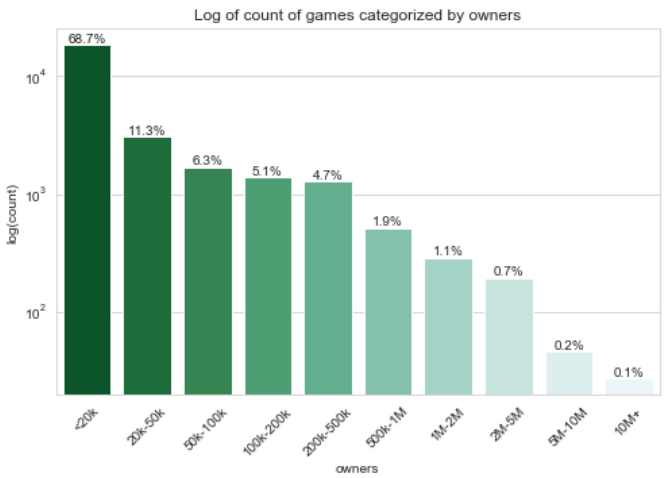

Fig. 2. Count of games categorized by owners

It can be observed that the count of games decreases with increasing range of users. The biggest percentage of games (around 69%) has fan base of around 20000 customers, whereas very few games (around 0.1%) are played by more than 10 million customers.

B. Count of games based on the year of release

The bar plot of count of games released in each year is visualized as shown in Fig. 3, where log transformation of count is taken for re-scaling. The games released till 2005 can be counted in a single digit. After that, the count is seen to be increased in each subsequent year except on few occasions, which informs the increase in production of games in successive years. The maximum games are released in the year 2018 which accounts to around 30% of total games.

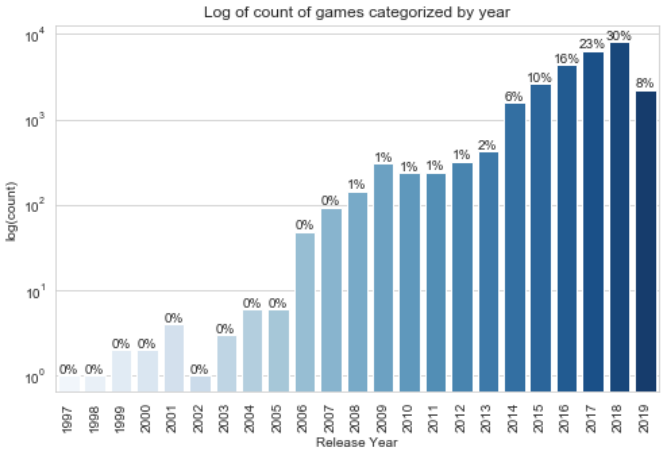

Fig. 3. Count of games categorized by year of release

C. Count of games based on the month of release

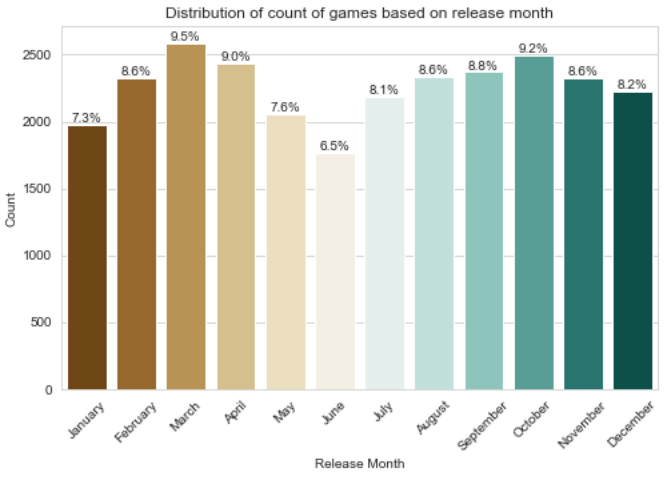

Fig. 4. Count of games categorized by month of release

The Fig. 4 shows the count of games released in each month. The variation in the count of games released in each month is

not very high. However, March and October are the preferred months as more than 9% games are released in these months, whereas January and June are the least preferred months in which around 7% games are released.

D. Percentage of games falling in various genres

The percentage of games that fall under various genres are visualized in the Fig. 5. There are total 29 genres present in the dataset. More than 70% games make up Indie genre keeping it at the top and more than 45% games constitute Action genre placing it at the second position. Negligible count of games represent genres such as Education, Video Production, Software Training, Audio Production, Web Publishing, Game Development, Photo Editing, Accounting, Tutorial, and Documentary.

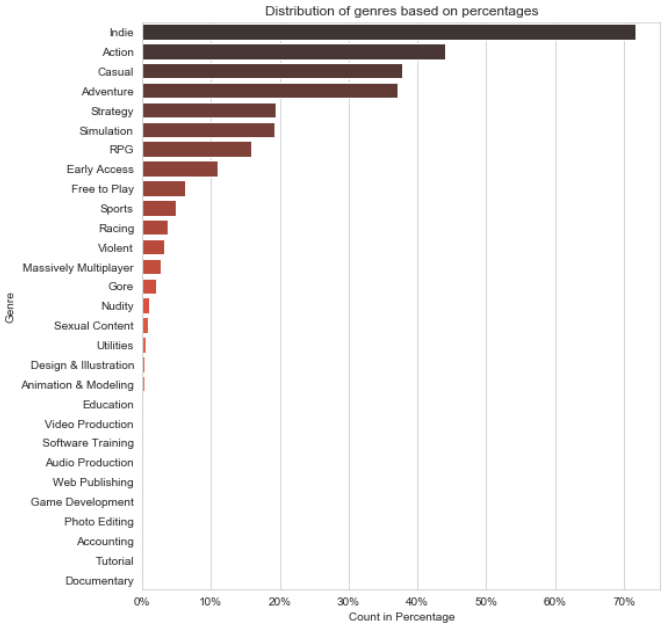

Fig. 5. Distribution of genres based on percentages

E. Percentage of games representing various categories

The categories of games with the count of percentage of games is presented in the Fig. 6. Total 29 categories of games are there in the dataset where the Single-player is the most famous category under which more than 90% games fall. There are around six categories at bottom side of plot which acquires very few games.

F. Most famous developers

The count of various developers present in the dataset is 17113, out of 16477 developers have developed less than five games. The top 25 developers are visualized in decreasing order using horizontal bars as shown in Fig. 7. The Choice of Games has the highest number of games counting to 94. There are total 9 publishers that have developed more than 40 games.

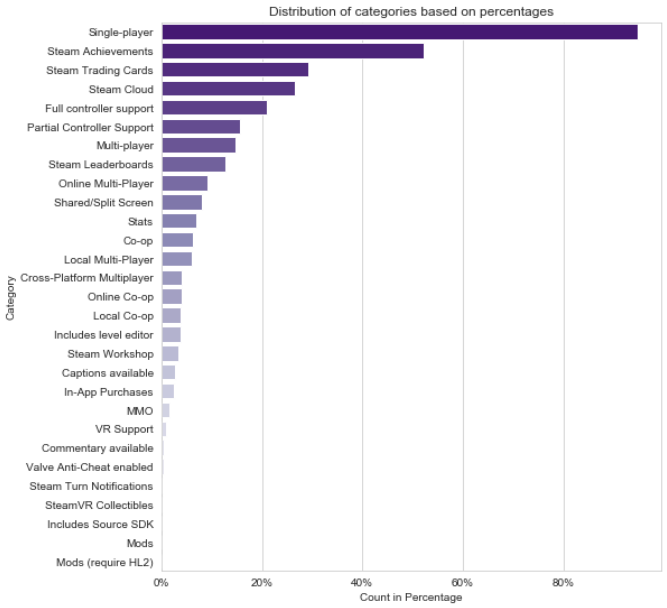

Fig. 6. Distribution of categories based on percentages

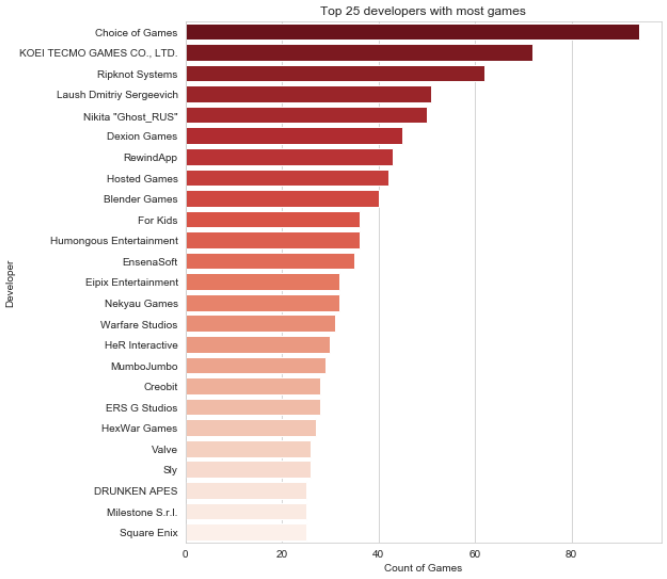

Fig. 7. Developers with most of the games developed

G. Most famous publishers

The top 25 publishers are visualized in a similar way to the top 25 developers, as shown in Fig. 8. There are total 14354 publishers present in the dataset. The count of publishers who published less than 5 games is 13683. The publisher Big Flash Games has published highest number of games counting to 212. Strategy First and Ubisoft has published 136 and 111 games positioning at second and third rank respectively. There are total 11 publishers that have published more than 75 games.

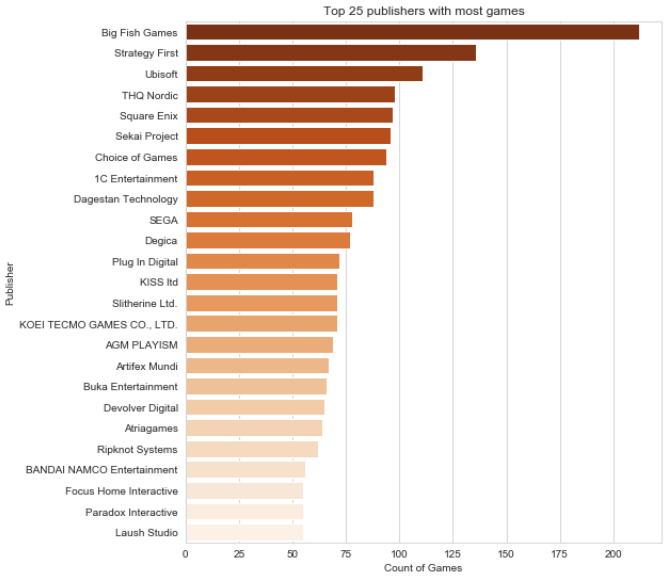

Fig. 8. Publishers with most of the games published

H. Relationship between count of games, release year and supported platform

The Fig. 9 visualizes the count of games released in years from 1997 to 2019 categorized by supported platforms i.e. Windows, Mac, Linux. It can be observed that the share of games supporting Linux is comparatively high than the Windows or Mac. The games supporting Mac are lowest in three platforms. Also, the Linux games are released in more numbers every subsequent year except few occasions which is not the case with Windows or Mac games.

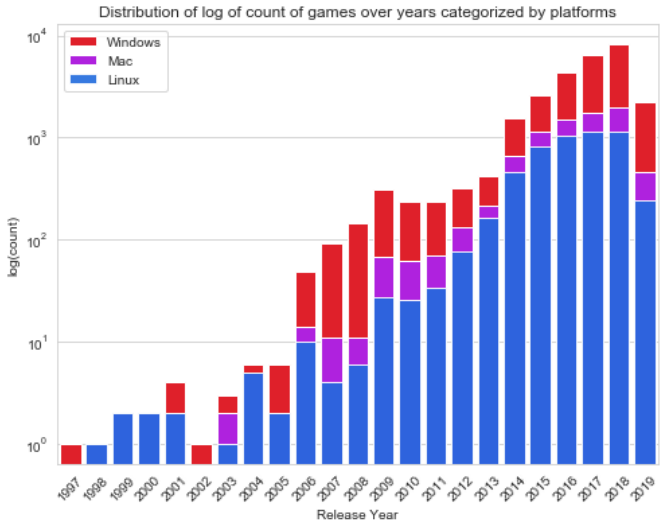

Fig. 9. Count of games over years categorized by platforms

I. Distribution of median price of developers with count of games

Fig. 10 shows the distplot which visualizes distribution of median price of dvelopers with count of games. Log transformation of count of games is used for rescaling as the variation was very high. Most of the developers have low median price. The price is measured using the currency United States Dollar (USD).

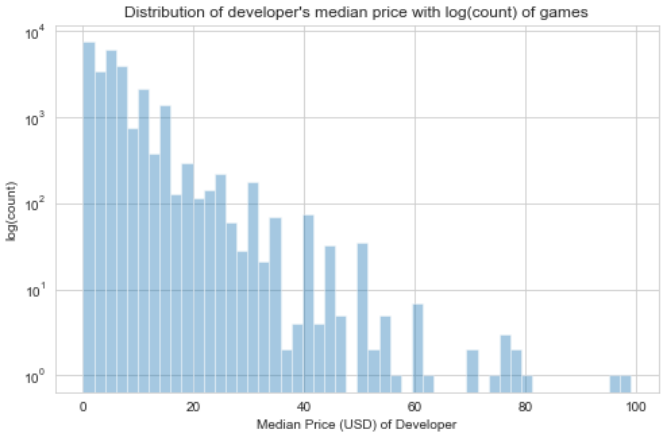

Fig. 10. Distribution of developer's median price with count of games

J. Relationship between price and positive ratings

The scatterplot as shown in Fig. 11 shows the relationship using price and positive ratings. As majority of games are not that costly, the points are observed to be saturated in the range 0-20 USD. The games with the price more than 50 USD have more than 60% positive ratings, which implies the costlier games are worth the money. The positive ratings can be noticed to be directly proportional to the price of games, except at certain occasions.

Fig. 11. Relationship between price and positive ratings

K. Relationship between price and release date

Fig. 12 shows the scatterplot of price and release date. As the number of games released before year 2005 are comparatively less, the plot has less points at the bottom and is saturated at the top. It can be noticed that there are no expensive games before 2005, which started to release in the subsequent years after 2005. Also, even after the year 2005, highest share of the games has price ranging between 0 and 20 USD.

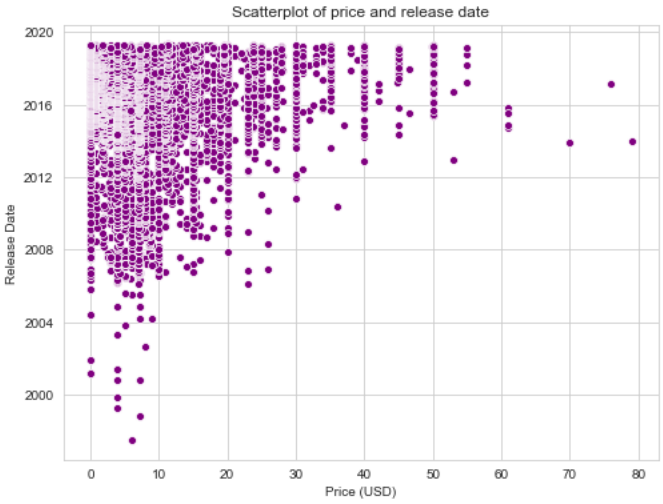

Fig. 12. Relationship between price and release date

VI. Applicable Techniques

The target variable which is to be predicted is 'owners' that has categorical values. Also, the dataset has features which can be helpful to determine the value of target variable. Thus, the classification algorithms of supervised learning can be applied for the prediction in this case. There are several machine learning classification algorithms such as Logistic Regression, Random Forest, Decision Tree, K Nearest Neighbor, Naive Bayes, K-Means Clustering, Support Vector Machine, etc. that are developed in various programming languages such as Python, R, etc. and are application as a solution in this project. The results of these techniques can be compared using metrics such as Precision, Recall, Accuracy to find the best suitable algorithm. Some of these algorithms are as discussed below.

A. Logistic Regression and Naive Bayes

Naïve Bayes algorithm uses the classification technique based on Bayes' theorem. There is an assumption that a particular feature in a class is independent to the other features, thus for implementation, the algorithm finds the probability of output variable by doing the product of all the individual probabilities of input features [20]. The Logistic Regression algorithm, on the other hand, uses sigmoid function for the prediction and probably gives best output for binary classification [21], hence not much useful for this dataset.

B. Decision Tree and Random Forest

Output of Decision Tree algorithm is the result of a tree created using decisions made on input variables. For example, if a player wants to play the football, he first checks if the weather is favourable, then checks if other players are also coming. If both the conditions are satisfied, the player will leave for the ground otherwise he will prefer to stay at home. Random Forest has a large number of individual decision trees which work as an ensemble. The combined effect of multiple trees results in the increased performance of Random Forest [22].

C. K Nearest Neighbour and K-Means Clustering

KNN algorithms works on the principle of finding nearest K number of nodes. The class of the resultant node is decided to be the class of maximum nodes in the nearest K-nodes. It is crucial to find proper boundaries of groups in the KNN technique. Also, it is required to bring the scales of all variables to the common ground, otherwise the features with high magnitudes dominates the output. K-Means Clustering forms clusters of homogenous non-overlapping groups such that data points are different in various groups. It has been observed that K-Means Clustering often yield better result of classification than KNN [23], [24].

D. Support Vector Machine

SVM works on the principle of finding the most accurate hyperplane from the hyperplanes of N-dimensional space that classifies the data identically. Hyperplanes are nothing but the decision boundaries which classifies the input data points. The dimensions of the hyperplane are same as the dimensions of independent variables of the dataset.

References

[1] "How many video games exist?" https://gamingshift.com/how-many-video-games-exist/, accessed: 2020-06-27.

[2] "How video games affect the brain," https://www.medicalnewstoday.com/articles/318345, accessed: 2020-06-27.

[3] "Video gaming industry & its revenue shift," https://www.forbes.com/sites/ilkerkoksal/2019/11/08/video-gaming-industry--its-revenue-shift/#70f684d4663e, accessed: 2020-06-27.

[4] M. Palaus, E. M. Marron, R. Viejo-Sobera, and D. Redolar-Ripoll, "Neural basis of video gaming: A systematic review," *Frontiers in human neuroscience*, vol. 11, p. 248, 2017.

[5] "Video games are transforming how we communicate with each other - and they could fix a range of other global issues too," https://www.weforum.org/agenda/2019/12/video-games-culture-impact-on-society/, accessed: 2020-06-27.

[6] H. S. Choi, M. S. Ko, D. Medlin, and C. Chen, "The effect of intrinsic and extrinsic quality cues of digital video games on sales: An empirical investigation," *Decision Support Systems*, vol. 106, pp. 86–96, 2018.

[7] E. A. Boyle, T. M. Connolly, T. Hainey, and J. M. Boyle, "Engagement in digital entertainment games: A systematic review," *Computers in human behavior*, vol. 28, no. 3, pp. 771–780, 2012.

[8] G. Freeman, "Making games as collaborative social experiences: Exploring an online gaming community," in *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, 2016, pp. 265–268.

[9] D. R. Thomas, S. Pastrana, A. Hutchings, R. Clayton, and A. R. Beresford, "Ethical issues in research using datasets of illicit origin," in *Proceedings of the 2017 Internet Measurement Conference*, 2017, pp. 445–462.

[10] H. Nair, "Intertemporal price discrimination with forward-looking consumers: Application to the us market for console video-games," *Quantitative Marketing and Economics*, vol. 5, no. 3, pp. 239–292, 2007.
[11] "Video game sales are extremely seasonal," https://www.cnbctv18.com/market/data/video-game-sales-are-extremely-seasonal-1511161.html, accessed: 2020-06-27.
[12] "Video game monetization," https://en.wikipedia.org/wiki/Video_game_monetization, accessed: 2020-06-27.
[13] K. Samarnggoon and A. Kunkhet, "An investigation of monetisation models in digital games," in *2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON)*. IEEE, 2019, pp. 64–68.
[14] R. Numminen, M. Viljanen, and T. Pahikkala, "Predicting the monetization percentage with survival analysis in free-to-play games," in *2019 IEEE Conference on Games (CoG)*. IEEE, 2019, pp. 1–8.
[15] "Top 10 factors that impact game revenue," https://bidalgo.com/insight/top-10-factors-that-impact-game-revenue/, accessed: 2020-06-27.
[16] A. Drachen, M. S. El-Nasr, and A. Canossa, *Game Analytics: Maximizing the Value of Player Data*. Springer, 2013.
[17] T. Verbraken, W. Verbeke, and B. Baesens, "Profit optimizing customer churn prediction with bayesian network classifiers," *Intelligent Data Analysis*, vol. 18, no. 1, pp. 3–24, 2014.
[18] E. Lee, B. Kim, S. Kang, B. Kang, Y. Jang, and H. K. Kim, "Profit optimizing churn prediction for long-term loyal customer in online games," *IEEE Transactions on Games*, 2018.
[19] Á. Periáñez, A. Saas, A. Guitart, and C. Magne, "Churn prediction in mobile social games: Towards a complete assessment using survival ensembles," in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2016, pp. 564–573.
[20] F. Harahap, A. Y. N. Harahap, E. Ekadiansyah, R. N. Sari, R. Adawiyah, and C. B. Harahap, "Implementation of naïve bayes classification method for predicting purchase," in *2018 6th International Conference on Cyber and IT Service Management (CITSM)*. IEEE, 2018, pp. 1–5.
[21] J. Sánchez Monedero, A. Saez Manzano, P. A. Gutiérrez Peña, C. Hervas Martínez *et al.*, "Machine learning methods for binary and multiclass classification of melanoma thickness from dermoscopic images." *Ieee Transactions On Knowledge & Data Engineering*, no. ONLINE, 2016.
[22] K. Lavanya, S. Bajaj, P. Tank, and S. Jain, "Handwritten digit recognition using hoeffding tree, decision tree and random forests—a comparative approach," in *2017 International Conference on Computational Intelligence in Data Science (ICCIDS)*. IEEE, 2017, pp. 1–6.
[23] Y. Quek, W. L. Woo, and T. Logenthiran, "Dc equipment identification using k-means clustering and knn classification techniques," in *2016 IEEE Region 10 Conference (TENCON)*. IEEE, 2016, pp. 777–780.
[24] M. Manjusha and R. Harikumar, "Performance analysis of knn classifier and k-means clustering for robust classification of epilepsy from eeg signals," in *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*. IEEE, 2016, pp. 2412–2416.