# 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 report of the academic project in the module "Domain Application of Predictive Analysis". The domain considered for the analysis, in this project, is related to the business of Video Game industry, the highest revenue generating industry from entertainment sector. The project is attributed to predicting game's popularity by finding key business features using Machine Learning techniques. The analysis is carried out with the historical data of games from 'Steam', a popular platform for digital distribution of video games. The past researches in the field are studied to understand the applicable techniques and key features. The exploration of key business features from the dataset is carried out with the help of visualizations such as bar plots, scatter plots, distribution plots, pie charts, etc. to comprehend patterns, emerging trends, and relationship between variables. Popularity of games is decided in terms of the feature 'owners', which is classified in four classes. The machine learning algorithm used for the classification predictions is Decision Tree, and found that the accuracy of the model is 86%.**

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

## I. Introduction

"The provision of entertainment has never been a subject of great interest either to economists or to economic historians – at least in their working hours." [1]. yet, in the current century of spectacular inventions in fastest-growing digital technology, the presence of digital entertainment essentially audio/video songs, movies, video games, etc. has established profound psychological and economic impact. Unsurprisingly, the entertainment sector is lead by the giant gaming industry with $131 billion revenue in 2018 [2], having 1.2 million video games worldwide [3] that are being played for 3 billion hours in each week [4]. The Forbes claims that the revenue of this industry is expected to reach to $300 billion by 2025 [2].

Despite the success of industry in terms of revenue, not all the games are successful commercially. The reasons that affect the popularity of games could be the high selling price, uninteresting plots, user age constraints, unavailable platforms, genres, playtime etc. If the exact impact of these reasons is found, the developers can focus precisely which would increase the popularity of games leading to increase in sales. This paper is an attempt to investigate such reasons using predictive analysis approach. The dataset considered for the analysis is picked from *Kaggle, which is created from scratch using Steam Store and SteamSpy **APIs. The Decision Tree machine learning algorithm is considered for the analysis.

The remaining organization of the research is as follows - the literature of analyzed techniques along with the key features, is presented in section II. The key features from the dataset are explored using visualizations in section III-A. The data processing, before feeding data to the model, is explained in section III-B. The training and designing of the model is illustrated in section III-C. The model is evaluated with various metrics in section IV.

## II. Research of Applicable Techniques and key features

The research [5] has the prediction of success of video games using machine learning regression and classification techniques. The performance is compared using Logistic Regression, Decision Tree and Neural Nets. The data is given to be collected from the sources that are commonly available and partitioned in two subsets: train and test, in 75% and 25% proportion, respectively. As usual, train subset is used for training the model and test subset is used to test and compare the prediction of machine learning model with the actual result. The Logistic Regression model is observed to be the best performing model. Also, the predictor variables "Metacrtitic", "users ratings", and "Types of games" are pointed out to be the significant variables.

The research from Business Analytics field [6], demonstrates an effective way of prediction of sales of video games in terms of software and hardware sales in European Market. The data used for the research is ensembled data from various sites such as vgchartz, Wikipedia, Metacritic, etc. It is observed that sales are boosted on holidays, and get declined in every subsequent week after week 1. Unfortunately, the formulas such as Logarithmic, Exponential, and Polynomial were found to be not much useful. Hence, third-degree polynomial formula with low sum squared error is used. The second-degree polynomial formula was also useful in some cases.

The research [7] predicts the churn of gamers of Destiny game, the most expensive digital game ever released. The focus of the paper is on multinomial Hidden Markov Model (HMM). The performance of HMM is compared with Theoretical Random Classifier, Bagging, Naïve Bayes, Nearest

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

Neighbour, Gradient Boosting, Decision Tree, Ada Boost, Logistic Regression, Random Forest, and observed that HMM outperformed with 92% precision. The precision of all other models is around 55%, except Theoretical Random Classifier which has 75% precision.

The paper [8] examined the prediction of sales of video games using Regression and Classification methods. The attribute 'player' which is continuous ranging from 0 to 135,300 is predicted using Linear Regression, Random Forest, Gaussian Process, and Support Vector Machines (SVM). The Relative Root Squared Error is found to be around 73% for all the models, where Linear Regression has 75.4% and Random Forest has 71.8% value. The attribute 'players' is then divided in three classes to perform the classification using General Linear Model (GLM), Random Forest, SVM, and Naïve Bayes. The maximum accuracy of 73.2% is achieved using Random Forest model, whereas Naïve Bayes gave least accuracy of 54.9%.

The impact of intrinsic and extrinsic cues on the sales of the games is analyzed in the paper [9]. The intrinsic and extrinsic cues such as newness, company reputation, retro features and games reviews, user engagement, game price, product popularity respective are considered for the study. It is found that newness, positive reviews, and price are the most affecting factors, whereas reputation of the company has negligible significance. High price signifies the superior product quality for popular game, however, sales of less popular games decrease with increasing price.

The creation and use of game mod are studied in the paper [10], using data from Nexus website. The popularity of game mod is decided by the number of downloads. The research found various tags of collected data, and the features are used to categorize the game mod as popular or non-popular, using classification algorithms. It is observed that few features were popular among users and others were in creators. Most of the popular features were incorporated by the mod creators which confirms that mod market is highly driven by the user expectations.

The paper [11] uses Knowledge Discovery in Databases methodology. Heat maps and graphical representation are developed to seek the collected user data. The findings were based on Geographical Distribution of Steam Users, Geographical Distribution of Genre, Genre Popularity per Country, and Contrast Genre Groups. This research reveals several interesting patterns, trends, and the correlation between industry trends. The patterns in the data also reveal some clues about the current market practice strategies, which are changing.

The paper [12] focuses on the various implementations of prediction of games modeled after fantasy sports, an example is demonstrated using weather forecasting properties. Ten participants were selected for the research and data is collected for seven consecutive days. The researchers noted that the players got more engaged with games than usual with the news of weather data.

The research [13] is focused on the analysis of Steam community network for 2011. The paper claims to be the first to analyse the Steam network and characteristics of gaming platforms. The data was collected by crawling the networks of the Steam users. The graphs have depicted the Node Degree Distribution. The platform has 1824 active number of games in regards to 1.98 million numbers of groups. The activity of the groups and the games did not directly put relevance on the user strength in the community. Hence, the conclusions are drawn on the basis of activities of users in the games and their connectivity.

Modern game developers encounter difficulties because more and more people are distracting due to unique personal preferences. The paper [14] gave information of preferences on Consequence of Interactivity, Player-Centric Solution, Psychological Characteristics and Player-Centric Gaming. The survey was done to collect the pre and post gaming data regarding trial preferences and player characteristics. The correlations were found. A rise of new relationships based on the in-game perception of enjoyment is observed. Also, the player characteristics could be associated with both play styles and preference for goals and rewards.

The research [15] has designs of sub-samples of research related to game engagement of users subjected to subjective experience and psychological concomitants. It is found that the user interest of playing games has been surged in the last 10 years attributed to increase in digital technology and availability of games. Fun enjoyment and challenges are the most important reasons with the contribution of narratives and suspense in games.

## III. Employed Technique

The machine learning technique used in this project is Decision Tree, where predictions are done by mapping observations about an item, to derive conclusion class of the target value. The following subsections are the steps followed.

### A. Exploration of Business Features using Visualizations

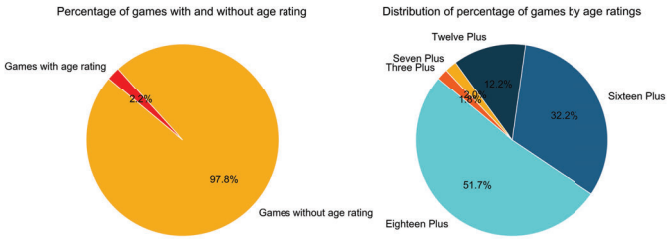


Fig. 1. Distribution of Age Ratings using Pie Chart

*1) Age Ratings:* The Fig. 1 displays the distribution of age ratings using two different pie-charts. The first pie-chart shows that only 2.2% games has age rating condition, whereas all other games are open for everyone. The second pie-chart shows the distribution of age rating of those 2.2% games. 57.8% games falls in Eighteen Plus category, 32.2% games are in Sixteen Plus category, 12.2% games has Twelve Plus age criteria, and only 1.8% and 2% games have Three Plus and Seven Plus age rating, respectively.

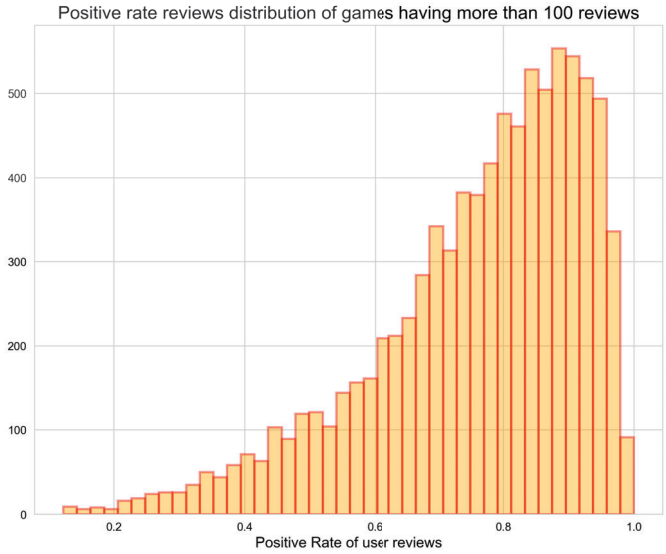


Fig. 2. Distribution of Positive Rate Reviews of Games

*2) Positive Rate Reviews:* The Fig. 2 illustrates the distribution of positive rate reviews of the games that has more than 100 reviews. There are 7492 records with more than 100 reviews which counts to approximately 28% of all the records. It can be observed that the distribution is skewed towards 1.0, which signifies the increasing count of high positive rated games.

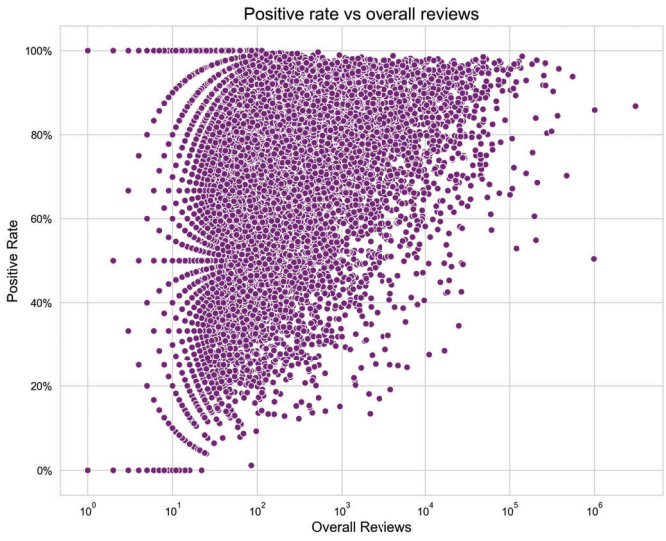


Fig. 3. Scatter Plot of Positive Rate vs Overall Reviews

*3) Positive Rate vs Overall Reviews:* The Fig. 3 shows the the distribution of Positive Rate and Overall Reviews using scatter plot. Log transformation of Overall Reviews is applied on x-axis for re-scaling. Positive Rating is seen to be saturated from 20% to 100% for around 1000 reviews, which later increases with the increase in overall reviews.

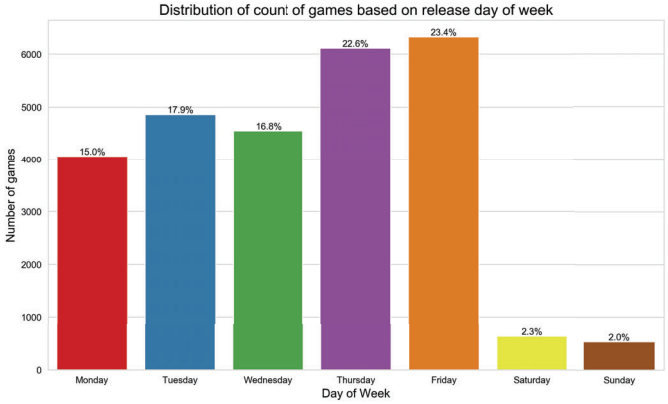


Fig. 4. Distribution of Games Released on Various Days of Week

*4) Release Day of Week:* The Fig. 4 shows the count of games that has been released on various days of week. Unsurprisingly, highest percentage of games, counting to approximately 56% are released on Thursdays and Fridays combined, because of upcoming weekend holiday. However, the least percentage of 2.3% and 2% are released on Saturdays and Sundays, respectively. The games released on Mondays, Tuesdays, and Wednesdays counts to 15%, 17.9%, and 16.8%, respectively.

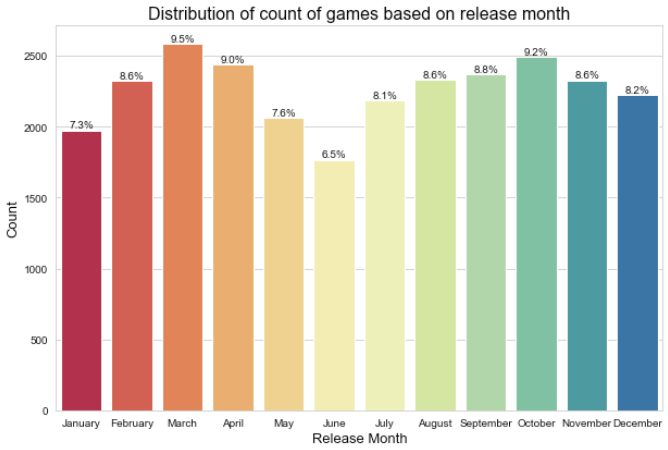


Fig. 5. Distribution of Games Released in Various Months

*5) Release Month:* The Fig. 5 illustrates the games released in various months. The variation in the percentage of games released ranges from 6.5% to 9.5%, where least and most games are released in June and March, respectively. The pattern from the figure shows the games releasing starts increasing from June till October, then starts decreasing till January, then again starts increasing till march, and then decreases till June.

*6) Games per Year:* The Fig. 6 demonstrates the count of games in the form of bar chart. The log transformation is applied on y-axis for re-scaling as negligible percentage of games (less than 9) are released in initial years till 2005. The games released are increased in each year, except a few occasions, which infers the production of games is increased

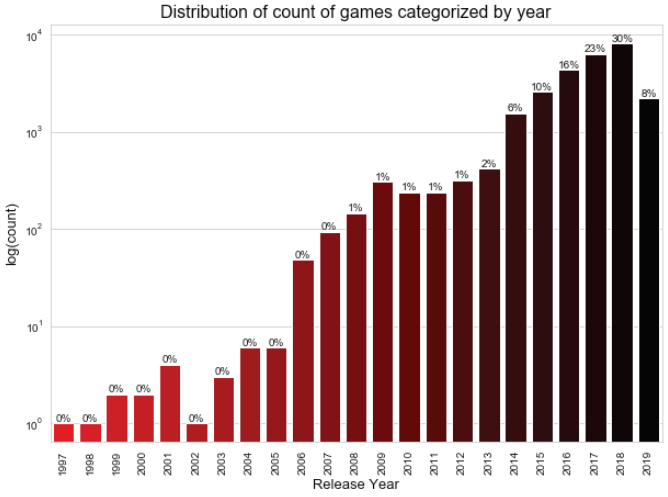


Fig. 6. Distribution of Games Released from 1997 to 2019

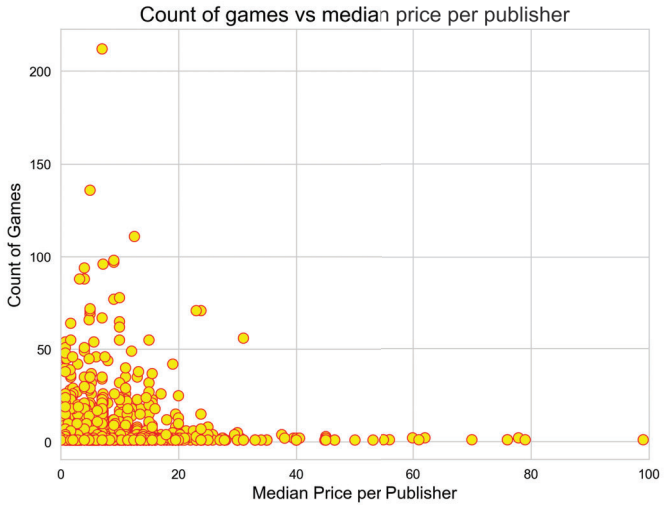


Fig. 8. Scatter Plot of Count of Games vs Median Price per Publisher

in each subsequent year. The highest games are released in the year 2018. It would have been 2019, if the data for whole year of 2019 had been considered, and not till May 2019.

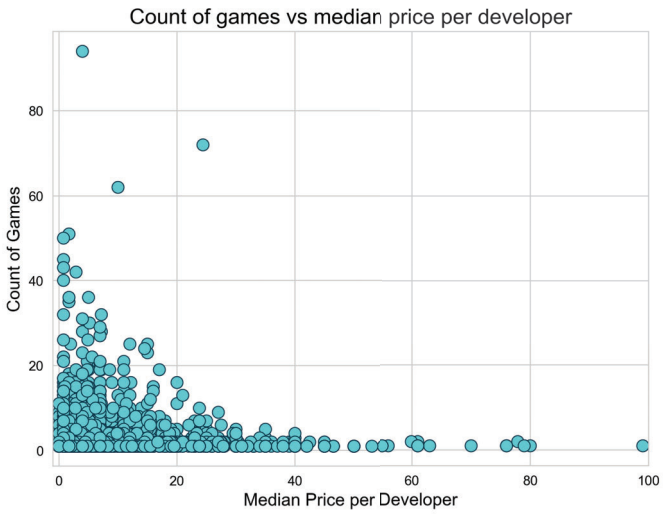


Fig. 7. Scatter Plot of Count of Games vs Median Price per Developer

*7) Developer and Price:* Relationship between the count of games and median price per developer is visualized in Fig. 7 using dot plot. The size of the dots is increased in the plot for better visualizations. The plot is fatty at zero and gets thinner along both the axes, which shows that the count of games with high median price are very less. Most of the games has low median price, and the games with moderate median price are in the middle-of-the-road.

*8) Publisher and Price:* The Fig. 8 shows the relationship between the count of games and median price per publisher using scatter plot. The plot is fatty at zero and gets thinner along both the axes. It infers that most of the games has low median price, the games with moderate median price are in the middle-of-the-road, and the count of games with high median price is very less.

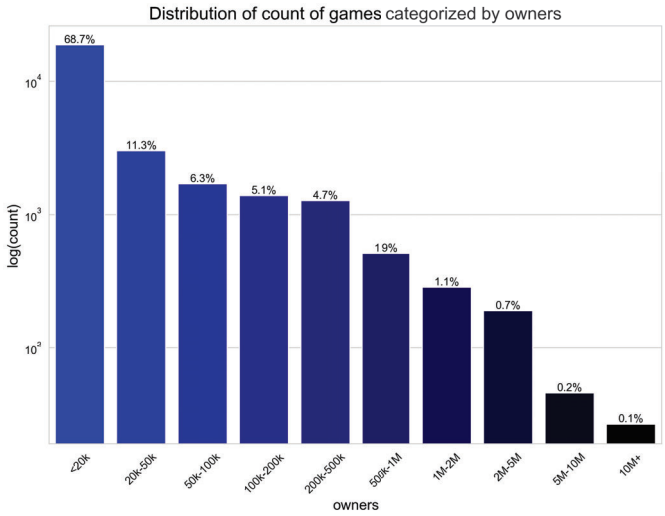


Fig. 9. Distribution of Games in Various Categories of Owners

*9) Categories of Owners:* The owners are categorized in ten classes as shown in Fig. 9. The categories initiate from the class where owners are less than 20000 and end to the class having more than 10 million owners. The log scaling is used on the y-axis to reduce the variation. As the range of the users increases, the count of games seems to be decreasing. The highest percentage of games of around 69% have only around 20000 customers' fan base, on the other hand more than 10 million customers play very popular games counting to only 0.1%

*10) Supported Platforms of Games:* The games are observed to be played on three platforms viz; Windows, Mac, and Linux as shown in Fig. 10. The important features included in the graph involve Release Year, Supported Platforms, and

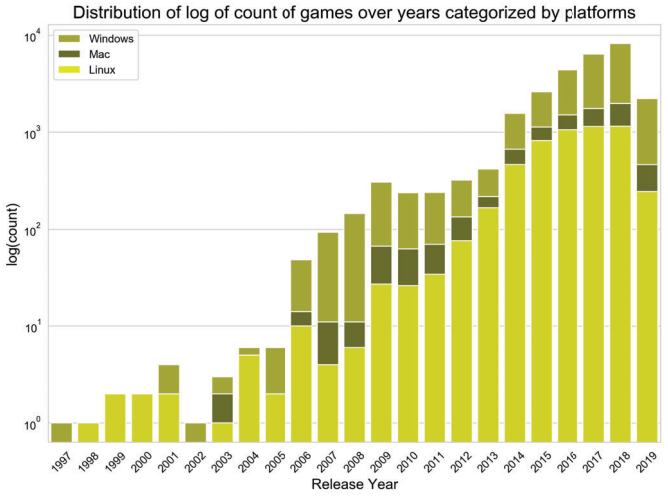


Fig. 10. Distribution of Supported Platforms of Games Over the Years

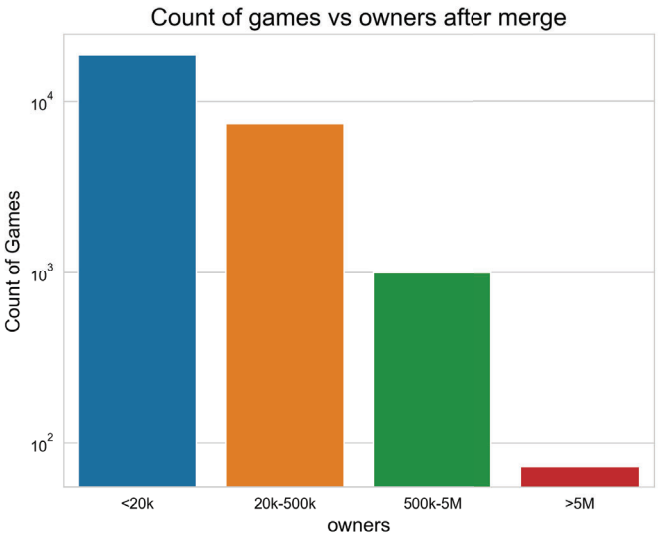


Fig. 11. Reduced Categories of Owners

Count of Games. The platform Linux has the highest share of games followed by Windows and Mac. Most of the games, released after the year 2007, are observed to be supported by Linux, increasing the share of Linux in each subsequent year. The share of Windows and Mac, after 2007, is almost constant except the few occasions where it is increased or decreased slightly.

### *B. Data Pre-processing*

The attribute `'release_date'` can not be used directly as it has categorical datatype. Thus, day of week, month, and year are extracted from the release_date to form three new columns `'weekday'`, `'month'`, and `'year'`. The figures Fig. 4, Fig. 5, and Fig. 6 visualizes the distribution of these three features. The feature `'platform'` has values separated by semicolon. For example, the game 'X-Blades' is supported by all the three platforms, and has value of 'windows;mac;linux' in the column platform. To extract this information, three new columns named after platforms are created, values in terms of binary digits are stored in them. The process is same as creating dummy variables with a slight difference. The `'category'` also has so many values in it separated by semicolons, with 'Single-player' and 'Multi-player' as leading categories. Thus, the values of `'category'` column are mapped in four main values such as 0 for 'Single-player', 1 for 'Multi-player', 2 for 'Both', and 3 for 'Other'.

The target variable is `'owners'`, which is classified in ten different classes. As visualized in the Fig. 9, the distribution is biased towards the first two classes viz; '<20k' and '20k-50k', where 80% of the records are present. Thus, to balance the count of games in each class, the classes are merged in four main classes that are '<20k', '20k-500k', '500k-5M', and '>5M' as shown in the Fig. 11. Even though the class '>5M' has less number of records, it is required to keep it separate as it has the most popular games.

The null values in the dataset are checked, and found that, not a single null value is present in the dataset. As the data is clean and all the important features are extracted, the features which does not play any role in the model are dropped from the dataset. The features that are dropped involve 'appid', 'name', 'release_date', 'developer', 'publisher', 'platforms', 'categories', 'genres', and 'steamspy_tags'. The final features that are selected for the prediction involve `'english'`, `'required_age'`, `'achievements'`, `'positive_ratings'`, `'negative_ratings'`, `'average_playtime'`, `'median_playtime'`, `'price'`, `'weekday'`, `'month'`, `'year'`, `'windows'`, `'mac'`, `'linux'`, and `'category'`.

### *C. Model Design for Predictive Analysis*

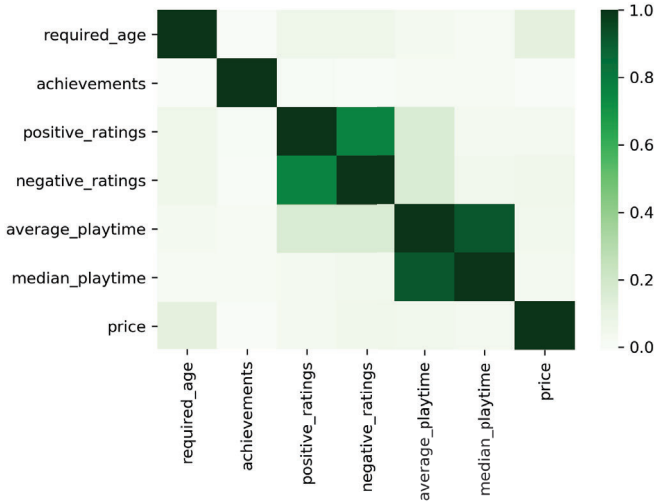


Fig. 12. Heat Map of Correlation Between Variables

The features that are selected for the model design may or may not be helpful. A way to check which features could be helpful is to find the correlation between them, as one or multiple features depend on another features or a cause for another features. Also, the features could be associated with one another with positive correlation or negative correlation. The Fig. 12 shows the correlation between continuous features from the dataset using heatmap. The correlation is weak between most of the variables and has a value of around 0.2. The features 'average_playtime' and 'median_playtime' has highest correlation of 0.9 between them followed by 'positive_ratings' and 'negative_ratings' with 0.7 value.

All the machine learning models require data to get trained. Also, a dataset identical to the training dataset, in terms of features, is required for testing the model. The given dataset is thus divided in two subsets, train and test in the proportion of 7:3 with the selection of random records from the data. Thus, the newly formed train subset has 70% records and test subset has 30% records. The Decision Tree model works on the principle of creating flowchart having multiple pathways, which ultimately leads to the the target variable using the decisions of each pathway.

### *D. Feature Importance by the Model*

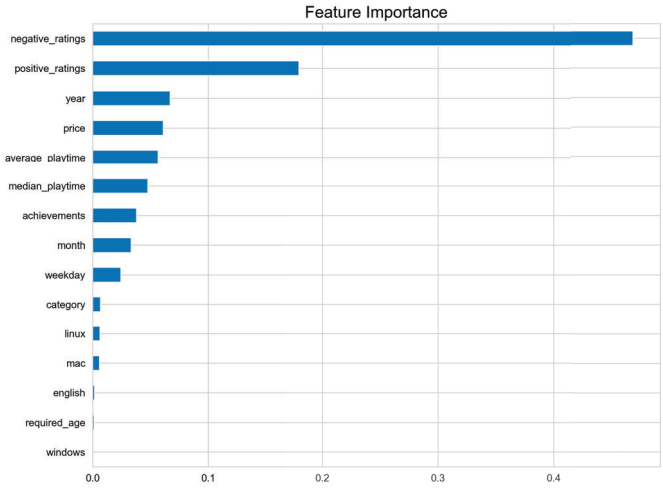


Fig. 13. Feature Importance derived by the Decision Tree model

The Fig. 13 shows the model considered important features while training the model. The features 'negative_ratings', 'positive_ratings' are considered as the most important features, followed by 'year', 'price', 'average_playtime', 'median_playtime', 'achievements', 'month', 'weekday'. The features 'category', 'linux', 'mac', 'english', 'required_age', 'windows' are the least important features.

The Fig. 14 shows the tree generated by the model, at the depth of 3 levels. It can be observed that the model initially looks for the feature 'negative _ratings', then checks

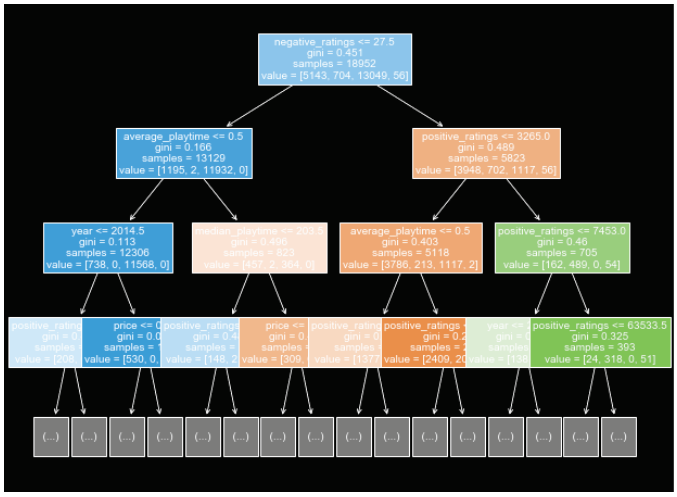


Fig. 14. Tree Generated by the Model

for 'average_playtime and 'positive _ratings' in the first level. The features 'year', 'median _playtime', 'average _playtime', 'positive _ratings', and 'price' are checked in the second and third levels followed by the remaining features in the subsequent layers.

## IV. Evaluation of Model

The metrics that can be used for the evaluation of classification model involve confusion matrix, recall, precision, F1-score, accuracy, etc. To check the values of these parameters, one can use classification report, which calculates the values of the parameters such recall, precision, F1-score, accuracy using the parameters derived by confusion matrix. Snapshot of classification report of the model is as shown in Fig. 15.

```
        * * * Classification Report * * *

              precision    recall  f1-score   support

     20k-500k       0.75      0.76      0.76      2269
     500k-5M        0.65      0.67      0.66       290
         <20k       0.92      0.91      0.92      5547
         >5M        0.53      0.47      0.50        17

     accuracy                           0.86      8123
    macro avg       0.71      0.70      0.71      8123
 weighted avg       0.86      0.86      0.86      8123
```

Fig. 15. Classification Report

The metrics values are highest for the class <20k, which are 0.91 for precision, 0.94 for recall, and 0.92 for F1-score. Whereas, the values are lowest for the class >5M. The overall precision of the model is 0.71, recall is 0.70, and f1-score is 0.71. The weighted averaged precision is 0.86, recall is 0.86, and f1-score is also 0.86. The accuracy of the model is found to be 0.86, which is 86%, if measured using percentage scale.

## V. Conclusion

The Decision Tree algorithm is found to have performed well giving the accuracy of 86%. The past researches are analysed intensively and Decision Tree algorithm is selected for the project. The features are extracted and the important business features are visualized using various categories of plots. Data pre-processing is carried out to check if null values are present in the data, and the features that does not play significant role are excluded from considering for the analysis. The correlation between the variables is visualized and the model is trained using the training data. The tree and the model considered important features are visualized to get the understanding of the most effective features that should be focused on in the business strategy. The model is evaluated using the classification metrics and found that it has enough accuracy which can help to increase sales of games.

## References


[1] A. Briggs, "Mass entertainment: The origins of a modern industry," *Welcome to the electronic edition of Australia's Economy in its International Context, volume 2. The book opens with the bookmark panel and you will see the contents page/s. Click on this anytime to return to the contents. You can also add your own bookmarks.*, p. 49, 1960.

[2] "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.

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

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

[5] P. Ghosh, "Prediction of the success rate of video games," *Oklahoma State University, Stillwater, 74078*.

[6] W. S. Beaujon, "Predicting video game sales in the european market," *Research Paper Business Analytics*, 2012.

[7] M. Tamassia, W. Raffe, R. Sifa, A. Drachen, F. Zambetta, and M. Hitchens, "Predicting player churn in destiny: A hidden markov models approach to predicting player departure in a major online game," in *2016 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, 2016, pp. 1–8.

[8] M. Trněný, "Machine learning for predicting success of video games," 2017.

[9] 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.

[10] T. Dey, J. L. Massengill, and A. Mockus, "Analysis of popularity of game mods: A case study," in *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*, 2016, pp. 133–139.

[11] E. J. Toy, J. V. Kummaragunta, and J. S. Yoo, "Large-scale cross-country analysis of steam popularity," in *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, 2018, pp. 1054–1058.

[12] G. Dzodom and F. Shipman, "Data-driven prediction games," in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2016, pp. 1857–1864.

[13] R. Becker, Y. Chernihov, Y. Shavitt, and N. Zilberman, "An analysis of the steam community network evolution," in *2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel*. IEEE, 2012, pp. 1–5.

[14] R. Staewen, P. Trevino, and C. Yun, "Player characteristics and their relationship to goals and rewards in video games," in *2014 IEEE Games Media Entertainment*. IEEE, 2014, pp. 1–8.

[15] 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.