Data Driven Exploration of the Gaming Landscape in 2021

Shreyash Sahare
University of Colorado Boulder, shsa7246@colorado.edu

Adwait Mahajan
University of Colorado Boulder, adma4717@colorado.edu

Junsoo Jung
University of Colorado Boulder, juju6944@colorado.edu

In the year 2021, the gaming industry had unprecedented growth and development. To manage this dynamic market, advanced analytics techniques were used to conduct a data-driven investigation of significant industry trends and participant actions. In this project, we aim to gain more insights about player behavior, game dynamics, and industry trends in order to plan and formulate strategies for making decisions that will spur innovation in the gaming business. There were several stages involved in this project which mainly comprised of data collection (using BeautifulSoup), data pre-processing techniques and implementation of machine learning models to predict the number of owners for each game based on its attributes. The models deployed for this purpose were K-Nearest Neighbors (KNN) regressor, Support Vector Regressor (SVR), Random Forest, linear regressor and artificial neural networks (ANN). The above models were trained and evaluated keeping the mean square logarithmic error (MSLE) as the metric. Among the models tested, ANN emerged as the top performer, exhibiting the lowest MSLE value of 0.48, followed by random forest, KNN regressor, SVR and linear regression having values of 0.67, 0.75, 1.09 and 2.54 respectively.

Additional Keywords and Phrases: Data cleaning, Data Pre-Processing, Visualization, Data Modeling, SVR, KNN, linear regression, ANN, random forest, MSLE

1 INTRODUCTION

The gaming industry has demonstrated resilience and ingenuity in the face of extraordinary difficulties posed by the global epidemic of 2021. It has done so by navigating a terrain marked by fast expansion and radical change. When communities around the globe struggled with the effects of mass lockdowns and strict social distancing policies, gaming became a vital source of comfort, entertainment, and a much-needed break from the stresses of social isolation.

The apparent increase in player spending and participation highlighted the gaming industry's lasting appeal and economic importance. Even with the current state of health and economic uncertainty, people's enthusiasm for video games increased, leading to a significant increase in video game sales. The much-awaited release of next-generation gaming consoles, such the PlayStation 5 and Xbox Series X/S, which sparked intense customer expectation and boosted industry economic activity, added to this momentum [13].

Furthermore, gaming has assumed a complex role as a facilitator of virtual social connectivity and community participation, surpassing its traditional position as a mere kind of entertainment. Virtual worlds and online gaming platforms become popular means of building a common digital culture, creating and maintaining relationships, and promoting camaraderie across geographic barriers. The widespread availability of gaming video content (GVC) on websites such as Twitch and YouTube highlights the growing cultural influence of gaming, drawing in a wide range of viewers and creating significant income streams [17].

Modern technologies, such as real-time analytics and artificial intelligence (AI), have brought about a new phase of creativity and innovation in the gaming industry. Developers were able to quickly adapt gaming experiences in reaction to changing trends thanks to real-time analytics, which provided them with crucial insights into player behaviors and preferences. Simultaneously, AI-driven procedural generation approaches transformed game design by creating dynamically created game landscapes that improve replicability and immersion. Notably, sophisticated behavioral analytics strengthened player performance assessment and strategy planning by enabling educated decision-making in esports.

Our study aims to provide an in-depth overview of the gaming business in 2021 by carefully analyzing its core characteristics and emerging trends. We hope to shed light on the factors that contribute to the popularity and financial success of gaming through a comprehensive analysis that covers a wide range of topics, from game expenses and release dates to user interaction data and developer insights. We hope to provide statistical insights as well as shed light on the complex interactions between technological, cultural, and economic factors that shape the gaming industry by utilizing the analytical capabilities of our large dataset.

2 RELATED WORK

Technology breakthroughs and changing societal norms have driven the gaming industry's extraordinary expansion and change in recent years. An in-depth analysis of the literature on data-driven exploration in the gaming industry is presented in this section, along with insights into the approaches, resources, and software used to comprehend and take advantage of the intricacies of this dynamic field.

According to *Muhammad Jawad Hamid Mughal* (2018) [1], web data mining has become a crucial instrument for gleaning important insights from the massive amounts of data produced by the gaming industry. An enormous amount of unstructured data has been generated by the growth of online platforms and digital interactions, which presents opportunities and problems for academics and business professionals alike. By categorizing web content, web structure, and web usage mining, researchers explore various methods for obtaining relevant data from user interaction logs, content repositories, and hyperlinks. With the use of these methods, trends, preferences, and patterns in player behavior, game dynamics, and industry trends can be found.

Furthermore, customer relationship management, or CRM, has become a reality in the gaming business thanks to the convergence of technology and marketing ideas. According to *Chris Rygielski et al.* (2002) [2], data mining is essential to relationship marketing because it can be used to identify important clients, forecast their behavior in the future, and encourage proactive decision-making. Organizations can acquire meaningful insights into player preferences through the use of data mining methods like neural networks, decision trees, and rough sets. This allows for the customization of game experiences and the optimization of marketing tactics. By means of customized messaging, targeted promos, and personalized suggestions, gaming companies may cultivate deeper relationships with their player base, thereby increasing player engagement and loyalty.

An in-depth analysis of data mining and analytics in games illuminates the revolutionary nature of data-driven methods in game production and study. According to *Gunter Wallner* (2019) [3], the development of data analytics tools has transformed the game production industry by enabling the ongoing assessment and improvement of gaming experiences. Data mining has become essential for improving game technology and maximizing player engagement in a variety of contexts, including player behavior analysis and esports performance evaluation. Developers can learn a great deal about player preferences, game mechanics, and market trends by examining large datasets that include player interactions, in-game activities, and user comments. This allows for incremental improvements and innovation.

Additionally, the use of data mining techniques in multiplayer online battle arena (MOBA) game item recommendation systems highlights how flexible data-driven approaches can be when handling challenging gaming problems. *Vladimir Araujo et. al.* (2019) [4] suggests a framework for in-game item recommendations based on contextual match dynamics by utilizing dataset analysis and machine learning algorithms. This enhances player decision-making and strategic advantage during gameplay. Gaming firms can personalize each player's gaming experience to increase engagement and happiness by using content-based filtering, collaborative filtering, and hybrid approaches.

Likewise, data mining techniques used to the examination of online game evaluations provide insightful information about the attitudes and tastes of players. *Ha-Na Kang et al.* (2017) [5] examined community data from gaming platforms using sentiment analysis and machine learning techniques, revealing elements that affect how useful game evaluations are and offering a sophisticated picture of how customers view the gaming industry. Gaming firms can improve player experiences, resolve player problems, and improve their game development plans by recognizing patterns, sentiment trends, and influential elements impacting player attitudes.

Moreover, the application of data mining techniques for the predictive study of gaming patterns highlights the predictive power of machine learning algorithms in predicting future events and market dynamics. *Mohini Chakerverti et. al.* (2019) [6] show how data mining may be used to anticipate game popularity and market trends by utilizing clustering and regression algorithms. This can provide independent game companies with useful market insights. Gaming companies can anticipate player preferences, adjust to shifting market conditions, and tailor their product offerings to match changing player wants by using predictive modeling, trend analysis, and market segmentation.

To summarise, the literature study clarifies the various uses of data-driven exploration in the gaming industry, ranging from item recommendation systems and predictive analysis to web data mining and CRM. In the dynamic gaming landscape of 2021 and beyond, gaming organizations may open up new opportunities for innovation, optimization, and player engagement by utilizing data analytics and machine learning.

3 METHODOLOGY

A systematic approach is employed in conducting the data-driven exploration for the gaming dataset, particularly focused on predicting the number of owners for each game based on several parameters that will be discussed later in this section. This section aims to uncover the key steps taken to gather, preprocess, analyze and interpret the dataset, as well as the methodologies utilized for model implementation and evaluation.

3.1 Data Gathering

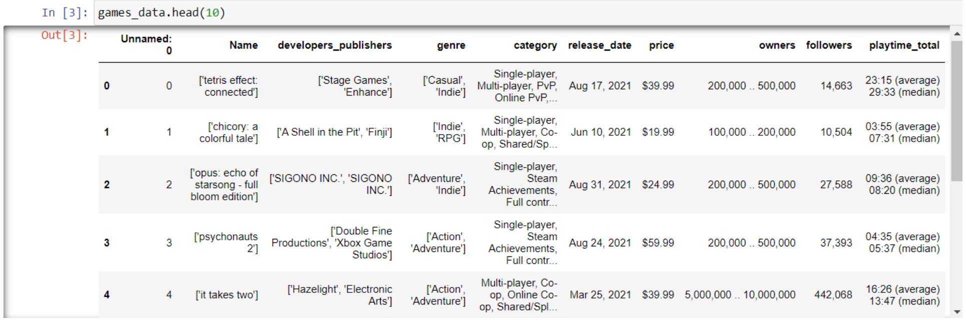
In the data collection stage, we chose to utilize SteamSpy, a well-known software known for its vast gaming data collection [5]. Our objective was to extract relevant information about games, such as publishers, creators, the number of owners, and other aspects. The problem we ran into, though, was how differently each game's details were presented on different web pages. As a result, the data collection process became more complicated since we had to carry out distinct web scraping operations for every game. Because different games have varying levels of attribute availability, this strategy resulted in null values and discrepancies in our dataset. We conducted thorough data cleaning operations in the following phase to address these problems [11]. Our method involved carefully extracting and organizing the data from the web pages using tools such as BeautifulSoup and a basic understanding of HTML. Our work produced a dataset with 9,901 rows and 9 columns in spite of these difficulties. It's important to note that the dataset's values were organized in a list-like manner, requiring extra processing to make interpretation and analysis easier.

3.2 Data Cleaning

Data refining plays an essential role and is considered to be the cornerstone of good data analysis, particularly in the fast-paced gaming industry. Here, our data's consistency provides a basis for interpreting important insights among rapidly shifting trends and player preferences.

Our gaming dataset, which is brimming with insightful data about player behavior, market dynamics, and the gaming industry in 2021, has several issues with anomalies, missing values, and redundancies. Ignoring these problems could compromise our analysis's accuracy and produce biased results. Therefore, data cleaning is more than just regular; it's a careful process meant to make sure that our dataset accurately captures the changing nature of the gaming ecosystem. Every step is carefully carried out to improve the quality of our dataset, from fixing hidden missing numbers to fixing inconsistencies and getting rid of duplicates. This meticulous procedure creates a solid basis for our next research and analysis, where the caliber of our dataset determines the breadth and accuracy of our findings.

3.2.1 Dataset before cleaning:



```
In [3]: games_data.head(10)
Out[3]:
```

	Unnamed: 0	Name	developers_publishers	genre	category	release_date	price	owners	followers	playtime_total
0	0	['tetris effect: connected']	['Stage Games', 'Enhance']	['Casual', 'Indie']	Single-player, Multi-player, PvP, Online PvP,...	Aug 17, 2021	$39.99	200,000 .. 500,000	14,663	23:15 (average) 29:33 (median)
1	1	['chicory: a colorful tale']	['A Shell in the Pit', 'Finji']	['Indie', 'RPG']	Single-player, Multi-player, Co-op, Shared/Sp...	Jun 10, 2021	$19.99	100,000 .. 200,000	10,504	03:55 (average) 07:31 (median)
2	2	['opus: echo of starsong - full bloom edition']	['SIGONO INC.', 'SIGONO INC.']	['Adventure', 'Indie']	Single-player, Steam Achievements, Full contr...	Aug 31, 2021	$24.99	200,000 .. 500,000	27,588	09:36 (average) 08:20 (median)
3	3	['psychonauts 2']	['Double Fine Productions', 'Xbox Game Studios']	['Action', 'Adventure']	Single-player, Steam Achievements, Full contr...	Aug 24, 2021	$59.99	200,000 .. 500,000	37,393	04:35 (average) 05:37 (median)
4	4	['it takes two']	['Hazelight', 'Electronic Arts']	['Action', 'Adventure']	Multi-player, Co-op, Online Co-op, Shared/Spl...	Mar 25, 2021	$39.99	5,000,000 .. 10,000,000	442,068	16:26 (average) 13:47 (median)

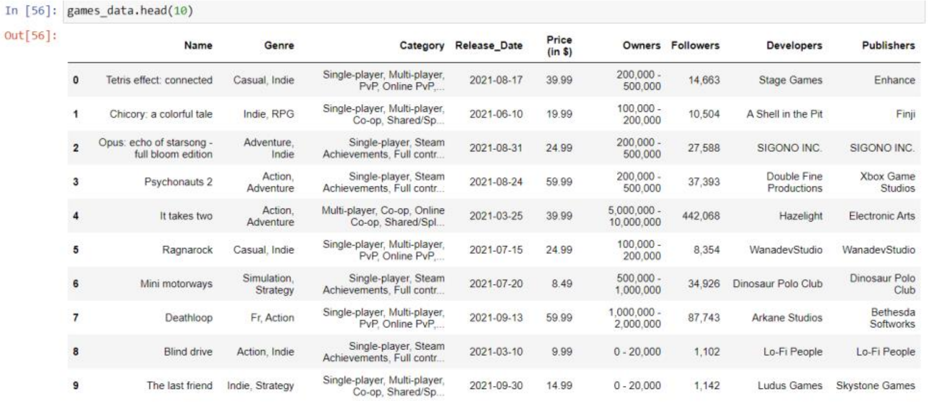
Figure 1: Dataset before cleaning

The accompanying image illustrates the various discrepancies and flaws present in the dataset that was retrieved by web scraping from the specified domain. One significant issue found in the dataset is the existence of an unnecessary column called "Unnamed: 0," which only contains indexes for each record and makes no meaningful contribution to the dataset's informative value. Moreover, data analysis procedures are complicated by the presence of prepared strings in columns like "Name," "developers_publishers," and "genre" that include unnecessary features like single quotes and square brackets.

Additionally, the "release_date" column only contains string format entries, which restricts the dataset's analytical and usable potential. These records should be converted to datetime format in order to improve the dataset's dependability and enable more thorough data analysis and interpretation. In a similar vein, the "price" column's items are now kept as strings; changing them to a float data type will improve computing performance.

More than 8000 records in the "playtime_total" column contain a large number of missing values, which is cause for concern. An important obstacle to the dataset's processing is the high frequency of missing data, which could lead to incorrect interpretations. For the dataset to remain accurate and reliable for use in later analytical projects, careful attention must be paid to fixing these missing values.

3.2.2 Dataset after cleaning



```
In [56]: games_data.head(10)
Out[56]:
```

	Name	Genre	Category	Release_Date	Price (in $)	Owners	Followers	Developers	Publishers
0	Tetris effect: connected	Casual, Indie	Single-player, Multi-player, PvP, Online PvP,...	2021-08-17	39.99	200,000 - 500,000	14,663	Stage Games	Enhance
1	Chicory: a colorful tale	Indie, RPG	Single-player, Multi-player, Co-op, Shared/Sp...	2021-06-10	19.99	100,000 - 200,000	10,504	A Shell in the Pit	Finji
2	Opus: echo of starsong - full bloom edition	Adventure, Indie	Single-player, Steam Achievements, Full contr...	2021-08-31	24.99	200,000 - 500,000	27,588	SIGONO INC.	SIGONO INC.
3	Psychonauts 2	Action, Adventure	Single-player, Steam Achievements, Full contr...	2021-08-24	59.99	200,000 - 500,000	37,393	Double Fine Productions	Xbox Game Studios
4	It takes two	Action, Adventure	Multi-player, Co-op, Online Co-op, Shared/Spl...	2021-03-25	39.99	5,000,000 - 10,000,000	442,068	Hazelight	Electronic Arts
5	Ragnarock	Casual, Indie	Single-player, Multi-player, PvP, Online PvP,...	2021-07-15	24.99	100,000 - 200,000	8,354	WanadevStudio	WanadevStudio
6	Mini motorways	Simulation, Strategy	Single-player, Steam Achievements, Full contr...	2021-07-20	8.49	500,000 - 1,000,000	34,926	Dinosaur Polo Club	Dinosaur Polo Club
7	Deathloop	Fr, Action	Single-player, Multi-player, PvP, Online PvP,...	2021-09-13	59.99	1,000,000 - 2,000,000	87,743	Arkane Studios	Bethesda Softworks
8	Blind drive	Action, Indie	Single-player, Steam Achievements, Full contr...	2021-03-10	9.99	0 - 20,000	1,102	Lo-Fi People	Lo-Fi People
9	The last friend	Indie, Strategy	Single-player, Multi-player, Co-op, Shared/Sp...	2021-09-30	14.99	0 - 20,000	1,142	Ludus Games	Skystone Games

Figure 2: Dataset after cleaning

The dataset was cleaned using a number of methodical procedures designed to remove errors and inconsistencies and improve the dataset's dependability and usability for further research. To reduce any potential distortions in the analysis, columns like "playtime_total" that had a significant prevalence of null values were first removed from the dataset. Rows with null values in any column were then removed, and the indexes were reset to guarantee that the remaining data entries would continue to be indexed. To further simplify the dataset, the unnecessary column "Unnamed: 0" was eliminated.

The string entries in particular columns, like "Name," "developers_publishers," and "genre," were further refined to eliminate unnecessary characters like square brackets, double quotes, and single quotes. Using string manipulation techniques, undesirable characters were methodically replaced or removed during this process. To enable simpler organization and analysis, entries in the "developers_publishers" column were additionally divided into separate lists for developers and publishers. These lists were then sorted and added as new columns to the dataset.

To maintain consistency and enable chronological analysis, transforming date values in the "Release_Date" column from string format to datetime format was another essential step in the cleaning process. Similarly, to enable numerical computations and analysis, items in the "Price (in $)" column were transformed from string format to float data type by deleting the dollar sign and converting the values to float.

Finally, the dataset's coherence and interpretability were improved by cleaning the "Owners" column to get rid of unnecessary characters and guarantee consistent formatting. After addressing a number of data quality concerns methodically, the cleaning procedure produced a polished dataset that was prepared for in-depth examination and interpretation. This technique should provide insightful information about the underlying trends and patterns in the dataset.

3.3 Modeling Approach

Regression approaches were utilized in our modeling approach to forecast continuous numerical values, with a specific focus on estimating the number of prospective owners for a given game based on different input variables. For this goal, five different regression models were chosen, each based on its own advantages and applicability to the given task.

Firstly, the openness and interpretability of the linear regression model made it a foundational technique. Based on the supposition of a linear correlation between the predictor variables and the target variable, Linear Regression is a useful technique for preliminary analysis and understanding of the information.

Secondly, the Support Vector Regression (SVR) model was incorporated, providing adaptability in managing non-linear associations between the target variable and features. Through the use of a kernel function to transform the data into a higher-dimensional space, SVR is able to identify complex patterns that linear models can miss.

The K-Nearest Neighbors (KNN) Regressor model was selected due to its capacity to locate specific, discernible patterns or clusters within the dataset. KNN Regressor performs best in situations where spatial relationships are important for prediction, as it computes the target variable by averaging its k nearest neighbors.

The Random Forest Regressor model was included for its robustness and effectiveness in managing large, high-dimensional datasets. As an ensemble learning method, Random Forest reduces the risk of overfitting by combining predictions from multiple decision trees, resulting in accurate and reliable forecasts.

Lastly, Neural Networks were employed for their capacity to capture complex and non-linear relationships inherent in the data. With multiple hidden layers, Neural Networks possess the ability to learn intricate patterns and interactions among features, making them well-suited for modeling the intricate nuances involved in predicting game ownership.

By incorporating a diverse array of regression models, our approach aims to leverage the strengths of each technique to provide comprehensive insights into the factors influencing game ownership, thereby facilitating informed decision-making within the gaming industry.

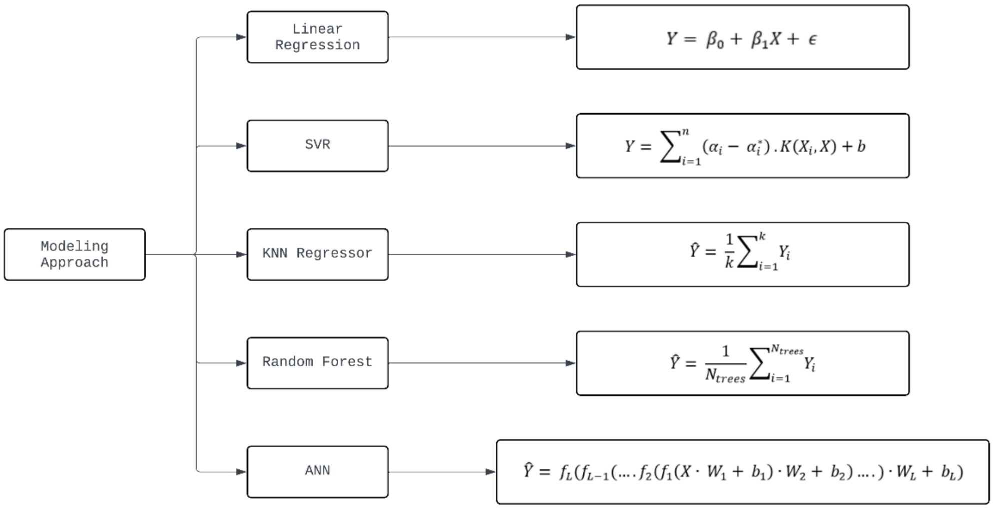

Modeling Approach
Linear Regression
SVR
KNN Regressor
Random Forest
ANN
$Y = \beta_0 + \beta_1 X + \epsilon$
$Y = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) . K(X_i, X) + b$
$\hat{Y} = \frac{1}{k} \sum_{i=1}^{k} Y_i$
$\hat{Y} = \frac{1}{N_{trees}} \sum_{i=1}^{N_{trees}} Y_i$
$\hat{Y} = f_L(f_{L-1}(\dots . f_2(f_1(X \cdot W_1 + b_1) \cdot W_2 + b_2) \dots .) \cdot W_L + b_L)$

Figure 3: Block diagram of different statistical modeling approaches

3.4 Model Implementation

3.4.1 Feature Engineering:

The following stage involved feature engineering after the dataset was collected and cleaned via web scraping. An important part of data analysis and machine learning is feature engineering, which is the process of creating new features or variables from preexisting ones in order to increase a model's performance or forecasting ability.

Feature engineering took numerous phases in the framework of our investigation using game data from SteamSpy. The game title, publisher, developer, number of owners, and other pertinent data were among the first features we assessed in the dataset. We next evaluated the ways in which these characteristics may be modified or combined to produce new features that could more effectively identify trends or insights in the data.

Since the "name," "publishers," "category," and "developers" columns had a large number of unique values and had no relationship to the objective column for the models, "Number of Owners," we removed them. Since the machine learning algorithms demand integer or float-based values, they were transformed to integers in the "Followers" column, which contained only categorical or string-based values. We took the month out of the release date in order to utilize it, and we found that it correlated well with the target column, making it a useful characteristic for machine learning.

The scikit-learn `train_test_split()` function is used to divide the dataset into training and testing sets. Eighty percent of the data are in the testing set (X_test, y_test), and the remaining twenty percent are in the training set (X_train, y_train). This guarantees that performance of the model may be assessed with non-visual data. The models were put into practice with great care and specialized methods to guarantee the precision and effectiveness of every model.

Each model was put into practice using a methodical process that was customized to meet the unique needs and specifications of the corresponding algorithms. Here, we provide a comprehensive overview of the implementation process for each of the five models:

3.4.2 Linear Regression Model:

In order to standardize the feature values between 0 and 1, the linear regression model's development started with an emphasis on feature scaling. The `MinMaxScaler()` function from scikit-learn was utilized to accomplish this, since it hindered features with greater scales from having an excessive impact on the model. The linear regression model was then initialized and trained on the training set of data using the {LinearRegression()` class. An ideal fit was established by adjusting the model parameters by minimizing the residual sum of squares between the observed and anticipated values.

3.4.3 Support Vector Regressor (SVR) Model:

To ensure a uniform contribution to the model's learning process, the `StandardScaler()` function was used to standardize the features before beginning the implementation of the SVR model. After this preprocessing phase, the `SVR()` class was used to initialize an SVR model using default hyperparameters. The `fit()} method was then used to train the model on the training set of data, optimizing the parameters to reduce the error between the observed and predicted values.

3.4.4 K-Nearest Neighbors (KNN) Regressor Model:

The KNN regressor model's implementation required careful consideration of feature selection and preprocessing methods. In line with earlier models, the dataset was split into features (X) and the target variable (y). To guarantee equal contributions to the distance calculation, features were then standardized using the `StandardScaler()` function. The `fit()` technique was then used to initialize and train a KNeighborsRegressor model with a predetermined number of neighbors on the training set. This method improved the model's predictive capability by making it easier to find localized patterns or clusters within the dataset.

3.4.5 Random Forest Regressor Model:

The random forest regressor model was implemented using a thorough procedure that included feature selection, preprocessing, hyperparameter tweaking, and model selection. Standardized features were employed to guarantee uniformity in their input to the model, and a pipeline was established to integrate feature scaling and model training. In order to find the ideal set of hyperparameters and improve the model's predicted performance, hyperparameter tuning was done using GridSearchCV cross-validation. Ultimately, grid search was used to determine the ideal hyperparameters, which allowed for the selection of the best model and reliable game ownership forecasts.

3.4.6 Artificial Neural Networks (ANN) Model:

The implementation of the Artificial Neural Networks (ANN) model involved a series of steps to construct, compile, train, and validate the model. Here, we provide a detailed overview of the implementation process specific to the ANN model:

- Model Architecture:

The Sequential API from Keras, a high-level Python neural network API, is used to build the ANN model. Several layers make to the architecture of the model, including dropout layers for regularization and dense (completely connected) layers. In our architecture, three dense layers are incorporated:

- The Rectified Linear Unit (ReLU) activation function, used by the 64 neurons in the first layer, adds non-linearity to the model and enables it to recognize intricate patterns in the data.
- The model's ability to capture complex correlations in the data is further enhanced by the second layer, which consists of 32 neurons with ReLU activation.
- The output layer is made up of a single linearly activated neuron that generates continuous numerical predictions and is ideally suited for regression applications.

- **Model Compilation:**

The `compile()` method is used to compile the model after the model architecture was constructed. The loss function from TensorFlow, MeanSquaredLogarithmicError (MSLE), is given during compilation. Because MSLE equally penalizes overestimation and underestimation, it is especially well-suited for regression tasks where it is needed to minimize the difference between observed and projected values.

The optimizer is also configured to 'adam,' an effective optimization technique that is frequently used in neural network training. Faster convergence and better performance are achieved during training using the 'adam' optimizer, which modifies the learning rate.

- **Early Stopping:**

The `EarlyStopping()` callback from Keras is used to implement early stopping, which reduces overfitting and enhances generalization efficiency. This callback keeps track of the validation loss during training and halts it if, after a predetermined number of epochs (patience), the validation loss ceases to decrease.

Early stopping prevents the model from memorizing noise in the training data and encourages the learning of more broadly applicable patterns by terminating training when the model's performance starts to deteriorate on the validation set.

- **Training and Validation:**

Using the `fit()` technique, the model is trained on the training set (X_train, y_train). To make sure the model doesn't overfit the training set, early halting is used to check the training process.

Furthermore, validation data (X_test, y_test) are supplied in order to assess the model's performance on data that was not encountered during training. This gives us insights into the model's overall performance and enables us to evaluate how well it generalizes to new data. In order to guarantee that the model converges to an ideal solution while avoiding pointless computation and overfitting, the training process is carried out for a maximum of 600 epochs or until the early stopping requirements are satisfied.

To ensure each model's resilience and effectiveness in accurately predicting game ownership, great care was taken during the implementation process with regard to preprocessing, model initialization, hyperparameter tuning, and training/validation protocols.

4 EVALUATION AND RESULTS

As the models learn the underlying patterns in the data during training, they adjust their parameters to minimize the chosen loss function, in this case the MSLE. The test set (X_test) was then predicted using the training models and the predict() function. To facilitate evaluation, the expected values (y_pred) are obtained and stored. Using the scikit-learn module, the mean squared logarithmic error (MSLE) between the true target values (y_test) and the projected target values (preds) was computed for assessment purposes. The MSLE is calculated to evaluate the performance of the models. The squared difference between the actual and expected values' logarithms is used to calculate this measurement. MSLE can be formulated as follows:

$$MSLE = \frac{1}{n}\sum_{i=1}^{n}(\log(1 + y_{true,i}) - \log(1 + y_{pred,i}))^2$$

where,
n = number of samples
$y_{true,i}$ = true value of the target variable for sample i
$y_{pred,i}$ = predicted value of the target variable for sample i

The five models were evaluated in order to determine how well they performed and how well they could predict game ownership using the attributes of the dataset. The outcomes of the evaluation of each model show its advantages and disadvantages and offer insightful information about how well it can predict outcomes.

1. Linear Regression Model (MSLE: 2.54):
   - The linear regression model exhibited the highest Mean Squared Logarithmic Error (MSLE) among all the evaluated models.
   - The linear regression model's capacity to explain variation in data may be hindered by its presumption of a linear relationship between the target variable and the characteristics.
2. Support Vector Regression Model (MSLE: 1.09):
   - The Support Vector Regression (SVR) model performed better than the linear regression model, with a lower MSLE.
   - SVR, while powerful for regression tasks, may struggle with capturing complex relationships and large datasets.
3. K-Nearest Neighbors (KNN) Regressor Model (MSLE: 0.75):
   - The KNN regressor model demonstrated further improvement over the SVR model, achieving a lower MSLE.
   - Complex patterns in the data can be captured by KNN because of its non-parametric character and dependence on similarity metrics.
4. Random Forest Regressor Model (MSLE: 0.67):
   - The random forest regressor model exhibited significant improvement over the previous models, with a notably lower MSLE.
   - Through its ensemble learning methodology, Random Forest successfully reduces overfitting and captures complicated interactions.
5. Artificial Neural Network (ANN) Model (MSLE: 0.48):
   - The ANN model outperformed all other models, achieving the lowest MSLE among the evaluated models.
   - The substantially lower MSLE highlights the ANN model's superior predictive performance, with predictions closest to the true target values.
   - The remarkable performance of ANNs in this work can be attributed to their capacity to recognize intricate patterns in high-dimensional data, which makes them ideal for regression tasks.

Table 1: Evaluation of models based on MSLE values

Model Name	MSLE Value
Linear regression	2.54
SVR	1.09
KNN	0.75

Random Forest	0.67
ANN	**0.48**

5 CONCLUSIONS

In conclusion, this study used sophisticated analytics methods to perform a data-driven analysis of important patterns and player behavior in the gaming sector. By means of extensive data gathering, preprocessing, and application of machine learning models, our objective was to acquire discernments into player conduct, game dynamics, and industry patterns. Our analysis was robust and effective because of the systematic approach taken by the approaches we used.

Our results highlight how crucial it is to use a variety of regression models in order to precisely estimate game ownership. Artificial Neural Networks (ANN) had the best prediction performance with the lowest Mean Squared Logarithmic Error (MSLE) value out of all the studied models. The ANN model's performance demonstrates how well deep learning methods can extract the intricate patterns present in game data.

Furthermore, our study illustrated the importance of data refining for improving the caliber and precision of analysis in the ever-changing gaming industry. We made sure our results were reliable and helped the gaming sector make well-informed decisions by taking care of abnormalities, missing numbers, and redundancies.

Overall, this study offers insightful information on how to forecast game ownership and has useful ramifications for those involved in the gaming industry. Going forward, more investigation and study in this area could stimulate growth and innovation in the always-changing gaming sector.

6 FUTURE WORK

- *Enhanced Feature Engineering*: There may be more predictors that have a major impact on game ownership that can be found with more investigation and improvement of feature engineering methodologies. To further understand player preferences, this can entail gathering more detailed information from game platforms or adding user engagement measures.
- *Deep Learning Architectures*: Investigating more complex deep learning architectures, like recurrent neural networks (RNNs) or convolutional neural networks (CNNs), may provide light on how to best capture spatial or temporal correlations in game data. These architectural designs exhibit efficacy in representing spatial or sequential data and have the potential to yield additional enhancements in prediction precision.
- *Interpretability and Explainability*: The incorporation of interpretability and explainability techniques into predictive models has the potential to augment their transparency and reliability in the gaming sector. Techniques like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) enable stakeholders to adopt actionable insights by offering insights into the factors influencing model predictions.
- *User Segmentation and Personalization*: It can be possible to create customized marketing plans and recommendations for various player segments by looking into user segmentation strategies and personalized modeling methodologies. Personalized models can improve user experience and engagement by identifying unique player cohorts based on behavior and interests.
- *Ethical Considerations*: It is imperative that ethical concerns such as algorithmic fairness, bias mitigation, and data privacy be taken into account while developing and implementing predictive models in the gaming sector. In order to guarantee the ethical and fair application of predictive analytics in gaming, future research should give special attention to ethical frameworks and principles.

REFERENCES

[1] Mughal, Muhammd Jawad Hamid. "Data mining: Web data mining techniques, tools and algorithms: An overview." International Journal of Advanced Computer Science and Applications 9, no. 6 (2018)

[2] Rygielski, Chris, Jyun-Cheng Wang, and David C. Yen. "Data mining techniques for customer relationship management." Technology in society 24, no. 4 (2002): 483-502.

[3] Wallner, Günter. "A brief overview of data mining and analytics in games." Data analytics applications in gaming and entertainment (2019): 1-14.

[4] Araujo, Vladimir, Felipe Rios, and Denis Parra. "Data mining for item recommendation in MOBA games." In Proceedings of the 13th ACM Conference on Recommender Systems, pp. 393-397. 2019.

[5] Kang, Ha-Na, Hye-Ryeon Yong, and Hyun-Seok Hwang. "A study of analyzing on online game reviews using a data mining approach: Steam community data." International Journal of Innovation, Management and Technology 8, no. 2 (2017): 90.

[6] Chakarverti, Mohini, Nikhil Sharma, and Rajiva Ranjan Divivedi. "Prediction analysis techniques of data mining: a review." In Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE). 2019.

[7] Gombolay, Matthew C., Reed E. Jensen, and Sung-Hyun Son. "Machine learning techniques for analyzing training behavior in serious gaming." IEEE Transactions on Games 11, no. 2 (2017): 109-120.

[8] Prathama, Narendra Yogha, Rengga Asmara, and Ali Ridho Barakbah. "Game Data Analytics using Descriptive and Predictive Mining." In 2020 International Electronics Symposium (IES), pp. 398-405. IEEE, 2020.

[9] Ahmad, Muhammad Aurangzeb, and Jaideep Srivastava. "Behavioral data mining and network analysis in massive online games." In Proceedings of the 7th ACM international conference on web search and data mining, pp. 673-674. 2014.

[10] Delen, Dursun, and Ercan Sirakaya. "Determining the efficacy of data-mining methods in predicting gaming ballot outcomes." Journal of Hospitality & Tourism Research 30, no. 3 (2006): 313-332.

[11] Kumar, Anurag, and Ravi Kumar Singh. "Web mining overview, techniques, tools and applications: A survey." International Research Journal of Engineering and Technology (IRJET) 3, no. 12 (2016): 1543-1547.

[12] Lee, Ho Geun, and Hyun Kwak. "Investigation of factors affecting the effects of online consumer reviews." Informatization Policy 20, no. 3 (2013): 3-17.

[13] Arthur, Z. (2020, May 25). 50 VIDEO GAME STATISTICS: 2020/2021 INDUSTRY OVERVIEW, DEMOGRAPHICS & DATA ANALYSIS. https://comparecamp.com/video-game-statistics/

[14] Best Games on PC. (2024, April 30). https://www.metacritic.com/browse/game/pc/all/2021/metascore/?platform=pc&page=1

[15] Donny K. (2021, June 15). 2021 Gaming Spotlight: The Trends You Need to Know Across Mobile, Console, Handheld and PC/Mac Gaming. https://www.data.ai/en/insights/mobile-gaming/2021-gaming-spotlight/

[16] Video Game Insights 2021 Market Report. (2022, January 23). https://vginsights.com/insights/article/video-game-insights-2021-market-report

[17] Riad C. (2015, October 31). The History Of Gaming: An Evolving Community. https://techcrunch.com/2015/10/31/the-history-of-gaming-an-evolving-community/