**Predictive Analytics and Machine Learning Approaches to Predict Global Video Game Trends**

RACHANON CHOMPHOO

Student Number: 22080474

**SUPERVISOR**

EJAY NSUGBE

UNIVERSITY OF THE WEST OF ENGLAND

DECEMBER 2023

**ACKNOWLEDGEMENT**

While pursuing an MSc in Data Science at the University of the West of England, this research project was carried out. All sources utilized or referenced are duly acknowledged and credited.

RACHANON CHOMPHOO

**ABSTRACT**

This study thoroughly examines the trends in video game sales for various genres. Using a comprehensive dataset of global sales figures, the research employs Linear, Ridge, and Lasso Regression models to explore the connection between the year of release and median sales. The study offers insights into the market dynamics driving sales trajectories within the video game industry. The results show a general decrease in median sales across multiple genres, suggesting changes in consumer preferences and market saturation. However, this trend is countered by the rise of mobile gaming, which has significantly impacted the gaming industry. The study also delves into the effect of mobile gaming on traditional gaming platforms, highlighting a stark decline in PC and console game sales compared to the booming mobile gaming sector. Statistical findings indicate that the models accurately capture the linear relationship between time and sales, with consistently low Mean Squared Error (MSE) metrics.

Keywords: **Video Game Sales, Regression Analysis, Market Trends, Mobile Gaming, Consumer Behavior, Gaming Industry.**

TABLE OF CONTENTS

1. INTRODUCTION 6
   1. PROBLEM STATEMENT 6
   2. AIMS AND OBJECTIVES 7
   3. RESEARCH QUESTION 7
   4. SUCCESS CRITERIA 7
   5. ETHICAL, LEGAL, AND   
      PROFESSIONAL CONSIDERATIONS 8
   6. CODE AVAILABILITY 8
2. LITERATURE REVIEW 9
   1. THE EVOLUTION OF GAMING 9
   2. PREDICTIVE ANALYTICS 11
   3. PREDICTIVE ANALYTIC IN GAMING INDUSTRY 12
   4. MACHINE LEARNING 13
   5. CURRENT LANDSCAPE OF DATA ANALYTICS IN GAMING 14
   6. RELATED RESEARCH 15
3. METHODOLOGY 17
   1. DATASET OVERVIEW 17
   2. DATA PREPARATION 17
   3. DATA EXPLORATION 18
   4. DATA TRANSFORMATION 18
   5. DATA SPLITTING 19
   6. MACHINE LEARNING ALGORITHMS 20
      1. LINEAR REGRESSION 21
      2. RIDGE REGRESSION 21
      3. LASSO REGRESSION 21
   7. MODEL PERFORMANCE AND EVALUATION 21
4. ANALYSIS AND EVALUATION 22
   1. EXPLORATORY ANALYSIS (EDA) 22
      1. UNIVARIATE ANALYSIS 22
      2. BIVARIATE ANALYSIS 26
      3. MULTIVARIATE ANALYSIS 27
   2. DATA TRANSFORMATION 29
   3. LINEAR REGRESSION 29
   4. RIDGE REGRESSION 34
   5. LASSO REGRESSION 37
   6. MODEL EVALUATION 39
   7. MODEL INTERPRETATION 41
5. CONCLUSION 47
   1. CONCLUSION 47
   2. LIMITATION AND FURTHER WORK 47
   3. REFLECTION 48
   4. RESEARCH CONTRIBUTION 48
6. REFERENCES 49

**CHAPTER 1**

**INTRODUCTION**

* 1. **INTRODUCTION**

In recent decades, the gaming industry has experienced tremendous growth and has become a dominant force in the entertainment world. By the close of 2020 and 2021, the worldwide gaming market had reached remarkable financial milestones. With a wide range of platforms, including computers, consoles, and mobile devices, it has evolved into a multi-billion-dollar market with a constantly expanding audience. Technological advancements, changes in consumer preferences, and the emergence of online gaming have created a dynamic landscape with opportunities and challenges for stakeholders. The global video game industry's revenues will exceed 200 billion U.S. dollars in 2022 (Clement J., 2023).

To thrive in the competitive gaming industry, it is crucial to have accurate data-driven insights to make strategic decisions. Companies are constantly striving to meet the changing needs of gamers and gain market share, which means anticipating market trends, and optimizing their product offerings (Eventyr, 2023). This dissertation analyses the published games data in the global gaming industry using predictive analytics and machine learning techniques. The goal is to provide valuable insights to industry players so they can make informed decisions and stay ahead of the competition.

* 1. **PROBLEM STATEMENT**

The global video game industry is characterized by its fast-paced and ever-evolving nature, presenting significant challenges in accurately predicting market trends. Traditional forecasting methods often need to catch up to capture the complexity and dynamism of this industry. This study seeks to address the problem of developing a more effective forecasting methodology to adapt to the rapid changes in the video game market. It aims to investigate various machine learning techniques, following an extensive Exploratory Data Analysis (EDA), to identify models that can efficiently and accurately forecast future market trends. The research will contribute to understanding the limitations and potential of machine learning in the context of a highly volatile and innovation-driven industry, offering insights for better-informed strategic planning and decision-making in the video game sector.

* 1. **AIMS AND OBJECTIVES**

**1.3.1 AIMS**

This study aims to analyze video game sales changes across genres using advanced regression techniques to understand market dynamics and sales evolution.

**1.3.2 OBJECTIVES**

1. Perform an extensive exploratory data analysis (EDA) on the available data for the video game market. Use this analysis to identify key trends, patterns, and anomalies, providing a foundational understanding of the market dynamics and critical factors influencing it.
2. Explore different machine learning techniques to determine their appropriateness and effectiveness in predicting video game market trends.
3. Construct and fine-tune machine learning models based on the insights gained from EDA. These involve training the models on historical data and validating their predictive accuracy using appropriate metrics.
4. Translate the findings from the predictive models into actionable insights for video game industry stakeholders, including game developers, publishers, and marketers, facilitating data-driven decision-making.
   1. **RESEARCH QUESTION**

How can machine learning techniques be effectively utilized to predict future trends in the global video game market, and what is their comparative performance in accuracy and reliability against traditional forecasting methods?

* 1. **SUCCESS CRITERIA**
* Identified and accurately interpreted sales trends across various video game genres and the industry.
* Effectively interpreted the coefficients to understand the impact of the year on sales.
* Ensured that the research methodology is transparent and reproducible.
* Offered practical recommendations and strategies for industry stakeholders.
  1. **ETHICAL, LEGAL, AND PROFESSIONAL CONSIDERATIONS**

This research uses publicly available sales data of video games that are aggregated and do not contain any personal or confidential information. As a result, no significant ethical or legal concerns are associated with its use. The data adheres to privacy norms and protection regulations due to its absence of personal data.

* 1. **CODE AVAILABILITY**

Gitlab: <https://gitlab.uwe.ac.uk/r2-chomphoo/dissertation>

**CHAPTER 2**

**LITERATURE REVIEW**

Chapter two of the report investigates the thriving gaming industry and its profound influence on the worldwide market. The chapter commences by providing a concise overview of the industry's evolution before delving into the factors that have propelled its expansion.

**2.1 THE EVOLUTION OF GAMING**

The gaming industry has its roots in the 1950s when computer scientists and engineers began experimenting with rudimentary games on early computers. One of the first noteworthy creations was "Spacewar!" by Steve Russell at MIT in 1962 (Bellis M., 2019), which laid the foundation for multiplayer gaming. The late 1960s marked the rise of arcade games such as "Pong" and "Space Invaders," marking the dawn of the arcade gaming era. The introduction of the Atari 2600 in the 1980s ushered in the age of home gaming consoles, making gaming more accessible (Opus Web Design, 2016). During the 1980s and 1990s, personal computers (PCs) emerged as a significant gaming platform. PC gaming was characterized by text-based adventures, early graphical games such as "King's Quest," and the emergence of graphical user interfaces. With the advent of CD-ROMs, enhanced storytelling and multimedia experiences became possible. The mid-1990s witnessed a 3D revolution with the introduction of 3D graphics accelerators. Sony's PlayStation and Nintendo's N64 became dominant players, while Sega's Genesis continued to compete. This era marked the beginning of the console wars, intensifying competition and driving innovation in hardware and game development.

A chart of a game

Description automatically generated with medium confidence

**Figure 1**. The Rise of Gaming Revenue Visualized (Visual Capitalist, 2020)

The 2000s witnessed the growth of broadband internet and the proliferation of online gaming. Massively Multiplayer Online Games (MMOs) the launching of "World of Warcraft" in 2004, which made a standard for a subscription-based massively multiplayer online game and online multiplayer shooters such as "Counterstrike" gained immense popularity, fostering global gaming communities and esports. The advent of smartphones in the late 2000s brought about the mobile gaming boom. Games such as "Angry Birds" or "Candy Crush Saga" reached enormous global audiences, reshaping the industry's demographics and revenue streams, and become the most downloaded freemium game in 2009 as shown in Figure 1. The indie game development scene also thrived during this period, with titles such as "Minecraft" in 2010 and becoming cultural phenomena.

As we move into the early years of the 2020s, the gaming industry is undergoing a rapid transformation with the advent of Virtual Reality (VR) technologies. Game developers are increasingly exploring this emerging technology to create innovative gaming prototypes (Mattoo S., 2022). VR's immersive experience transports players to a universe they could only imagine. With over 230 gaming companies jumping into the virtual reality market since the release of Oculus Rift CV 1, the global VR gaming market has surged to USD 6.26 billion in 2020 (Caicedo J., 2023).

Cloud gaming services, such as Google Stadia, have the potential to make high-quality games accessible to a broader audience. However, the industry is facing connectivity-related challenges. Many users have reported frame rate drops and frequent crashes, which has hindered their ability to enjoy games. Additionally, several countries require more infrastructure to run Stadia successfully. These technical issues have adversely impacted the gaming experience, making it difficult for Stadia to compete with other gaming platforms. As such, the growing societal influence of Stadia remains in question (Osborne H., 2021).

**2.2 PREDICTIVE ANALYTICS**

Predictive analytics is a sophisticated method of data analysis that leverages current and past data to anticipate future behavior and trends. Using statistical, modelling, data mining, and machine learning techniques, it can assess the likelihood of forthcoming outcomes based on historical data (Project Pro, 2023). Its primary aim is not simply to comprehend previous events but also to deliver the most precise assessment of what is probable to occur. Fundamentally, predictive analytics relies on three key elements:

* **Data Collection:** Data acquisition is a pivotal yet time-intensive aspect of any project. Stakeholders may supply the data from an external database or require extraction from various sources. Nonetheless, the gathered data may sometimes need to be revised to develop a solution, thereby requiring the procurement of supplementary information from multiple sources. It is essential to factor in the level of access available to the requisite datasets.
* **Exploratory Data Analysis (EDA):** Collecting data may lead to the acquisition of unnecessary information, which, if used as input for the model, could result in inaccurate predictions. To mitigate this risk, Exploratory Data Analysis (EDA) is a valuable tool. We can detect and address outliers, null values, and other extraneous data in the dataset through EDA (Ane Y., 2022). By recognizing patterns in the data, we can more efficiently determine the model's parameters, resulting in enhanced accuracy even before constructing the model.
* **Model Building and Evaluation:** Developing a model requires careful evaluation and continuous improvement of its accuracy. The quality of a model depends on its ability to predict outcomes based on the data used to train. Predictive analytics, powered by artificial intelligence and machine learning, can analyze vast amounts of data, identify patterns, and improve their predictions over time. By leveraging these insights and continuously refining the model, it becomes more reliable and proficient at forecasting future trends and outcomes (Qualtrics, 2023). Therefore, developing and assessing models carefully, considering factors such as data quality, algorithm selection, and performance metrics, is crucial to ensure that the resulting predictions are as precise as possible.

**2.3 PREDICTIVE ANALYTIC IN GAMING INDUSTRY**

The gaming industry has experienced a remarkable surge in market and revenue in recent years. However, it has evolved into a data-driven sector that collects and analyzes customer behavior data. Developers meticulously monitor various aspects, including playtime and in-game purchases. With the emergence of predictive analytics, the gaming industry is significantly transforming. Predictive analytics harnesses statistical algorithms and machine learning techniques to scrutinize extensive data sets and anticipate future outcomes (Ankleshwariya Y., 2023). This translates to leveraging data to predict player behavior, enhance game design, and optimize business outcomes in the gaming industry. There are a few examples of how predictive analytics can make an impact on the gaming industry:

1. **In-game purchases:** The implementation of predictive analytics has the potential to aid game developers in devising a comprehensive marketing strategy. By scrutinizing player data on item purchases, the analytics can effectively determine the most prevalent and favored items among players. Developers can utilize this information to optimize their marketing campaigns and boost player engagement.
2. **Fraud Prevention:** Identifying cheating or bot-players in games used to be a time-consuming task. However, thanks to new technology, developers can swiftly detect and act against such players to maintain the game's balance. Predictive analytics can also be leveraged to identify any anomalous activity by players, and appropriate measures can be implemented to prevent such activity (Ankleshwariya Y., 2023).
3. **Enhanced User Experience:** Steam's gaming platform is an excellent example of an enhanced user experience. It utilizes predictive models to recommend games to users based on their playing habits. Additionally, Steam has a built-in friend’s system that can suggest games to other users who share similar interests in game categories. This helps users discover new games and connect with others who have similar gaming preferences.
4. **Community Engagement:** Using predictive tools in gaming analytics can offer valuable insights into the most active periods of players. This information can be harnessed to guide the optimal timing of community events and game updates, resulting in heightened engagement and enhanced user experience. By leveraging predictive analytics, game developers can strategically plan their marketing and promotional campaigns and other key initiatives to take advantage of the most opportune moments to maximize user engagement.

**2.4 MACHINE LEARNING**

Machine learning (ML) is significant in artificial intelligence. It involves creating algorithms that enable computers to improve their task performance through experience, primarily by processing and learning from data. This differs from traditional programmatic methods, where computers are explicitly programmed to perform tasks. Instead, they use a more flexible framework where computers use data to make decisions or predictions. ML allows computers to learn from experience and improve their decision-making capabilities (Brown S., 2021). In its essence, Machine Learning (ML) is all about identifying patterns in data. It utilizes various techniques, including neural networks, deep learning, decision trees, and support vector machines. The learning process in ML can be broadly categorized into three types:

* **Supervised Learning:** In this type of learning, algorithms are trained on labelled data, which provides the algorithm with input-output pairs. This type of learning aims to create a mapping from inputs to outputs and use this mapping to make predictions on new, unseen data.
* **Unsupervised Learning:** Algorithms in this type of learning explore unlabeled data to find underlying patterns or structures. This can include identifying clusters or groupings in the data.
* **Reinforcement Learning:** In this type of learning, algorithms learn by interacting with an environment and receiving feedback as rewards or penalties. Through this process, the algorithm learns to optimize its behavior to achieve the best possible outcome.

Although Machine Learning and Statistical Methods share a common goal of identifying patterns within data, they differ in their philosophies and approaches. Statistical methods rely on hypothesis testing to predict outcomes and find relationships between variables, assuming specific data distributions based on predetermined models. On the other hand, Machine Learning uses a data-driven approach to optimize prediction accuracy without a predetermined model form. Rather than relying on explicitly programmed instructions, Machine Learning aims to build systems that can learn from data (Stewart M., 2019). Both statistical methods and machine learning share a common obstacle, namely overfitting. This phenomenon arises when a model becomes overly intricate and leans too heavily on the training data. Practitioners employ cross-validation, regularization, dropout, and early stopping techniques to tackle this issue. These methods simplify the model, encourage the network to learn more resilient features and prevent the model from overfitting. The resulting model can better generalize to new data by utilizing these approaches, leading to more precise insights and predictions.

**2.5 CURRENT LANDSCAPE OF DATA ANALYTICS IN GAMING**

The gaming industry has come a long way from being just a source of recreation. It has become a multifaceted ecosystem with various stakeholders, such as developers, publishers, gamers, and advertisers. The latest advancements in analytics and big data have brought about a revolution in the industry's future. The growing popularity of Esports has led to a surge in demand for data scientists, highlighting the industry's potential and the need for specialized talent (Ozyazgan, 2019).

Data analytics plays a crucial role in the gaming industry and provides significant benefits, especially in assessing the player experience. Developers can recognize players by using data analysis's valuable insight when they lose interest in the game and take prompt action to re-engage them through personalized incentives or communication. Moreover, the data insight is crucial for improving the gaming experience and ensuring user satisfaction (Eng D., 2020). The data insights can also help many companies create targeted marketing campaigns by profiling player demographics and preferences, resulting in increased ROI (Game Marketing Genie, 2021) and Analyzing player engagement during in-game and eSports tournaments to identify peak activity times. Use this information to create a balanced monetization model that maximizes profitability while maintaining player satisfaction.

The use of data analytics in the gaming industry has witnessed a remarkable surge in the last decade. Currently, data analytics has become more serverless in the gaming industry. Numerous game analytics platforms have introduced a new feature that allows teams to work on near real-time data and minimize the burden of server maintenance. According to a report by Ben Weber (2018), Riot Games has utilized Apache Spark for its large volume of data from their online game with custom ETL processes and machine learning initiatives. Notably, the company requires the ability to initiate infrastructure on demand promptly. Apache Spark has a real-time processing capability for Riot's developers to analyze player behavior and game performance, which lets them quickly identify issues affecting game performance and player experience. Sharma, Mohammad and Parashar (2023, p.33) discuss the architecture of Riot's game data processing and analysis from the start of distributed server architecture, where multiple servers work together to run game sessions to prevent fault tolerance. The game session will be extracted to a data warehouse where analysts can use the data to analyze the performance of the games, as shown in Figure 2.

A diagram of a data warehouse

Description automatically generated

**Figure 2**. Riot Game Data Processing and Analysis Architecture (2023)

With precise data analytics, game developers and marketers can glean invaluable insights to make informed decisions rather than relying solely on guesswork and intuition (More J., 2023). This, in turn, leads to the creation of better games, more precise targeting of their intended audience, and improved player engagement and retention.

**2.7 RELATED RESEARCH**

Predictive analytics has proved to be a powerful tool for the gaming industry, like sports predictions, and has now been applied to eSports. In a noteworthy study conducted by Shim, Sharan, and Srivastava (2010), it was found that baseball prediction methods, PECOTA, and MARCEL, can be used to forecast the performance of eSports players. The study presented a prediction framework using bins for discretization, where histograms had a better prediction accuracy. The research revealed that a player's past performance can significantly influence their future performance, and the framework can provide valuable insights into players across various dimensions, such as archetypes.

Moreover, there are many approaches which use machine learning techniques to predict the eSport match outcome. One study conducted by Ong, Deolalikar, and Peng (2015) developed a prediction framework that accurately predicted the outcomes of League of Legends matches. The framework utilized unsupervised learning techniques to categorize player behaviors in-game through clustering. Later, it selected features for game outcome predictors based on classification models. The study's findings suggested that including time-dependent player statistics features significantly improved the accuracy of predictive models.

The evolution of the gaming industry has been remarkable, experiencing both triumphs and setbacks like any other field. According to Future Market Insight (2023), the video game market is projected to increase from US$ 227 billion to US$ 805.3 billion by 2033, primarily due to the advent of smartphones, which has been a game changer for the industry. Report Linker (2023) predicts that the mobile gaming market will expand from USD 141.71 billion in 2023 to USD 300.47 billion by 2028. The success of many current games is a testament to the industry's rapid growth. Developers can leverage various factors, such as game mechanics and storyline, to determine which features or strategies will make their game successful. Additionally, analyzing game trends can help developers select game categories to generate profits.

The study by Moreira, Filho, and Ramalho (2014) utilized supervised learning in their research, using a linear regression model to extract insights into the relationships between predictive and class attributes in a database. Their objective was to discover the factors contributing to the success of mobile games in the gaming industry, such as in-game events that increase popularity and the potential for players to convert from free-to-play to pay-to-play, which benefits game monetization. Additionally, providing players with a PVP mode can have a positive impact on the game's appeal since players are more likely to convert to pay-to-play if they are attracted to the progress of other players, as Katkoff (2012) notes.

**CHAPTER 3**

**METHODOLOGY**

This chapter describes the methodology approach used in the study, which involves data collection and preparation, model selection, development, and implementation, hyperparameter tuning, and model validation. Additionally, it discusses potential areas for future research, such as investigating genre-specific dynamics or integrating consumer sentiment analysis to enhance the predictive models.

**3.1 DATASET OVERVIEW**

The dataset's information was obtained from Kaggle, initially scraped from Metacritic. The dataset Comprises 16,719 entries (rows) and 16 attributes (columns). It encompasses a diverse array of variables, including:

* **Descriptive attributes**: Game Name, Platform, Genre, Publisher, and Developer.
* **Temporal attribute**: Year of Release.
* **Sales data**: Regional sales (North America, Europe, Japan, Other) and Global Sales.
* **Evaluative metrics**: Critic Score, Critic Count, User Score, User Count.
* **Content rating**: ESRB Rating.

The dataset includes video game sales data from 1980 to 2020, focusing on modern gaming eras, averaging around 2006. It covers a wide range of sales figures, with Global Sales reaching a peak of 82.53 million units, showcasing the substantial market reach of the industry.

**3.2 DATA PREPARATION**

In the dataset, a comprehensive assessment for missing values in crucial columns, namely 'Year\_of\_Release' and 'Genre,' data filtration was executed for 1991-2016, ensuring temporal relevance. The three processes of data ingestion known as ETL (Extract, Transform, Load) have been applied: first gathering data from Kaggle, importing required libraries, and then importing datasets through Pandas Library using Python and Google Collab environment.

**3.3 DATA EXPLORATION**

Exploratory Data Analysis (EDA) is essential in discovering hidden patterns, detecting trends, and developing hypotheses. EDA is a crucial component of data science, as it empowers researchers to comprehend the raw data by summarizing, visualizing, and interpreting it.

**3.4 DATA TRANSFORMATION**

The data has been cleaned and transformed to remove irrelevant features and missing values that could affect accuracy following these steps below:

* The dataset had a few columns, such as 'Publisher', 'Critic Score', 'Critic Count', 'User Score', and 'User Count' that contained incomplete information. To address this, we replaced the minimal missing data in 'Publisher' with the term 'Unknown.' This ensured that the dataset remained accurate while handling the gaps in data. More sophisticated techniques, such as imputation using median values or model-based imputation, could be explored for the other variables with more significant missing data.
* In this dataset it consisted of categorical variables like 'Genre', 'Platform', and 'Publisher'. Machine learning models generally require numerical input, so these categorical variables must be converted into a numerical format. To accomplish this, one-hot encoding was used. One-hot encoding is a common technique that converts categorical variables into a binary matrix. This transformation creates a new binary (0 or 1) column for each category in the original variable, allowing models to use categorical data effectively.
* The process involved creating new and informative features from the existing data, which improved the dataset's potential for revealing insights or enhancing model performance. For instance, the sales data was aggregated by year and genre to provide a clearer view of the temporal sales trends. Additionally, computing median sales offered a robust measure of central tendency that remained unaffected by outliers.

Each data transformation step is crucial to the overall data analysis process. These steps ensure that the dataset is clean, comprehensive, and structured to maximize the efficiency and effectiveness of the subsequent analyses and modeling efforts.

**3.5 DATA SPLITTING**

The available dataset was divided into two sets to adhere to standard machine learning practices: a training set that accounted for 80% of the data and a testing set for the remaining 20%. This approach ensures that the model is trained on a particular subset of the data and then assessed on a different, independent subset to evaluate its ability to generalize to new and unseen data. In other words, the model is tested to check whether it can make accurate predictions on data it has not seen before.

**3.6 BUILD A MACHINE LEARNING MODEL**

**Figure 3**. Modeling pipeline

1. **Data Cleaning and Preprocessing**:
   * Handling missing values, such as filling or removing them.
   * Filtering the data to include relevant years (e.g., from 1995 onwards).
   * Converting data types if necessary (e.g., converting 'Year\_of\_Release' to an integer).
2. **Exploratory Data Analysis (EDA)**:
   * Visualizing different aspects of the data to understand trends and patterns includes creating plots like line plots, box plots, or heat maps.
   * Calculating summary statistics to get an overview of the data.
3. **Feature Selection and Engineering**

* **Pivot Table Creation**: Created a pivot table for median sales by genre and year, which was used for subsequent analyses.
* **Feature Engineering**: The pivot table effectively transformed the data into a format suitable for regression analysis.

1. **Model Development**

* **Model Selection**: Choose Linear Regression, Ridge Regression, and Lasso Regression models for the analysis.
* **Cross-Validation and Hyperparameter Tuning**: The cross-validation technique has been chosen for the best alpha value for the Ridge and Lasso Regression.

1. **Model Training and Evaluation**

* **Training**: Each model was trained on the dataset.
* **Performance Metrics**: Using coefficients and Mean Squared Error (MSE) to evaluate the models.
* **Visualization**: Plotted the actual versus predicted sales to represent model performance visually.

1. **Model Interpretation and Comparison**

* **Interpretation of Results**: Analyzed the coefficients and MSEs to understand the relationship between the year and median sales for each genre.
* **Model Comparison**: Compare the three models' performance to determine which best fits the data for predicting global sales.

**3.7 MACHINE LEARNING ALGORITHM**

The dataset that has been selected comprises sales figures from the past, and this data has been labeled for straightforward interpretation. To understand the correlations between the various input features and the target variable, supervised learning algorithms have been chosen to make the most of the labeled data. This approach will allow the algorithms to make precise predictions based on historical sales data.

**3.7.1 LINEAR REGRESSION**

Linear Regression is a basic algorithm used in supervised learning, mainly for regression tasks. It creates a model that demonstrates the connection between a dependent variable and one or more independent variables by fitting a linear equation to the given data. The linear equation maps the input features (independent variables) to the target (dependent variable).

**3.7.2 RIDGE REGRESSION**

Ridge Regression is a type of linear Regression that includes a regularization component. The critical feature of Ridge Regression is that it adds a penalty term to the ordinary least squares (OLS) regression formula. This penalty term is the square of the magnitude of the coefficients multiplied by a parameter known as alpha.

**3.7.3 LASSO REGRESSION**

Lasso Regression (Least Absolute Shrinkage and Selection Operator) is another variant of linear Regression that includes regularization. Lasso adds a penalty to the OLS regression like Ridge Regression, but the penalty term differs. In Lasso Regression, the penalty is the absolute value of the coefficients, again multiplied by an alpha parameter.

**3.8 MODEL PERFORMANCE AND EVALUATION**

One of the most used methods is Mean Squared Error (MSE) when evaluating and assessing a model's performance. This method involves calculating the average of the squared differences between the predicted values and the actual values. It provides a quantitative measure of how well the model fits the data and can help identify areas where it may need improvement. Overall, MSE is a powerful tool in the analysis of model performance.

**3.8.1 MEAN SQUARED ERROR (MSE)**

Mean Squared Error is a widely used metric for quantifying the accuracy of a regression model. It measures the average squared difference between the observed actual outcomes and the predictions made by the model. MSE is the mean of the squared differences between predicted and actual values. It is calculated using the formula:

where y is the actual value, ŷ is the predicted value, and n is the total number of observations.

**CHAPTER 4**

**ANALYSIS AND EVALUATION**

This section begins by elucidating the analytical methodology employed in the study. It encompasses a detailed description of the linear regression model implemented, outlining the rationale behind using this specific model given the dataset's characteristics and the research objectives. This overview sets the stage for a deeper examination of the results and their implications.

**4.1 EXPLORATORY DATA ANALYSIS (EDA)**

**UNIVARIATE ANALYSIS**

In Figure 4, the plot shows the number of video games released yearly from 1991 to 2016. There's a clear growth trend in the number of games released annually from the early 1990s until around the mid-2000s. This reflects the expansion and growing popularity of the video game industry during this period.

A graph of a number of video games released

Description automatically generated

**Figure 4**. Graph of total game released each year.

The peak is around the late 2000s, specifically between 2000 and 2009. This could be due to several factors, such as technological advancements in the shift from 2D to 3D. The PlayStation 2 from Sony marked a game-changing moment in the industry by offering an immersive 3D experience that surpassed the previous 2D graphics. Grand Theft Auto III debuted in 2001 and quickly amassed a massive following among the platform's most popular titles. Its success paved the way for the release of GTA San Andreas in 2004, which many consider the series' pinnacle. Additionally, online roleplaying games were a phenomenon in the 2000s. Blizzard's World of Warcraft took the industry by storm upon its announcement in 2004 and earned a place in the record books as one of the highest-grossing video games ever. Following this peak there was a distinct drop in the number of games released annually. The leading cause of this decline can be attributed to the surge of mobile gaming, which was brought about by the launch of the first iPhone in 2007. This innovation completely transformed the mobile industry thanks to its expansive, high-quality screen and dominant processor, making it an optimal gaming platform. As a result, developers swiftly began creating games for the device, leading to the rise of mobile gaming. Most titles were simple and casual in the initial stages of smartphone gaming. Games like Angry Birds and Fruit Ninja were among the first smash hits in this new era of gaming (Academy Contributor, 2020).

A graph of a number of video games

Description automatically generated

**Figure 5**. Distribution of Video Games by Genre

The above bar chart (Figure 5) depicts the genre distribution of video games between 1991 and 2016. Specific genres have a more substantial representation in the number of game releases. Action, Sports, and Shooter genres stand out with significantly higher game releases than others.

A graph with green line

Description automatically generated

**Figure 6**. Total Yearly Global Video Game Sales

The data presented in Figure 6 reveals the video game industry's history and how it has adapted to changing market trends and technological advancements. It highlights the years with the highest global sales, which could be attributed to the release of successful games like Wii Sports, Mario Kart, and Grand Theft Auto IV. Another factor is the introduction of new gaming consoles, such as Xbox 360 in 2003 and Wii and PlayStation 3 in 2006, marked the beginning of the console wars (MCV, 2009)

A graph of sales

Description automatically generated

**Figure 7**. Global Video Game Sales by Genre (1991-2016)

In Figure 7, Gaming enthusiasts worldwide have strongly preferred the Action, Sports, and Shooter genres, which have consistently proven to be the most commercially successful. These genres' broad appeal and popularity could be attributed to their accessibility and the excitement they offer. The Role-Playing genre also boasts significant global sales, which can be attributed to the deep storytelling and immersive experiences they provide. This genre is prevalent in regions like Japan, with a rich role-playing game history. While genres like Strategy and Puzzle have dedicated followings, they tend to have lower global sales than their more mainstream counterparts. This may be due to their specific appeal or the nature of these games, which cater to different gamers.

**BIVARIATE ANALYSIS**

A graph with different colored bars

Description automatically generated

**Figure 8**. Distribution of Global Sales by Genre in boxplot

Figure 8 illustrates the variability in sales among different genres. Notably, specific genres, such as Action or Shooter, display a broader range of total sales, indicating more significant shifts in popularity over time. In contrast, other genres demonstrate more consistent sales with a narrower spread. The median line in each box plot reflects the typical annual sales performance for each genre, with higher medians associated with more successful sales, as demonstrated by the Action genre. While some genres exhibit outliers, representing years with exceptionally high sales, these instances are relatively infrequent.

**MULTIVARIATE ANALYSIS**

A graph of a bar chart

Description automatically generated with medium confidence

**Figure 9**. Regional Video Game Sales by Genre (1991-2016)

As shown in. Figure 9, Most genres have a strong presence in North America, with sales often surpassing those of other regions. This underscores the region's crucial role in the global video game industry. Specific genres such as Action and Sports enjoy widespread popularity across all markets. However, it's worth noting that Japan's preferences for genres like Role-Playing differ from those of Western markets (NA and EU). As we said above, RPG tend to be more popular among the people of Japanese players, with the former showing a greater affinity for this category.

A green and white chart

Description automatically generated with medium confidence

**Figure 10**. Global Sales (in millions) of Games Released Each Year by Genre

Figure 10 shows the heatmap of global video game sales by year and genre from 1991 to 2016. Specific genres like Action or Sports show consistent popularity, while others like Adventure or Strategy have more fluctuating sales. The heatmap highlights peak periods in particular genres. A concentration of higher sales in 2008 and 2009 indicates successful game releases or a surge in the popularity of the action and sports genres.

**4.2 DATA TRANSFORMATION**

After selecting the years for analysis, the next step is determining the most appropriate method for representing the yearly global sales for each genre. Based on the KENDALLGILLIES method in 2017, utilizing the total global sales for each genre would not be a consistent measure over time, as the number of games released each year varies. Furthermore, a plot of the global sales show significant right skewness in the data, rendering the yearly average for each genre's global sales an unreliable measure. Hence, we will employ the median number of games sold per genre yearly as a suitable measure for highly skewed data.

**4.3 LINEAR REGRESSION**

A screenshot of a graph

Description automatically generated

**Figure 11**. Linear Regression plots using the median sales game each genre.

Figure 11 depicts the evolution of the video game genre's popularity and market performance over the years. Notably, the Shooter genre has demonstrated a steady incline, suggesting heightened consumer interest or the triumph of titles such as Call of Duty and Battlefield. Conversely, several other genres have displayed either a stagnant trend or a decline in popularity, which could be attributed to shifts in consumer preferences or the emergence of new competitors, such as mobile game developers.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Genre | Coefficient | MSE |
| 8 | Shooter | 0.002897 | 0.090635 |
| 10 | Sports | -0.003609 | 0.009422 |
| 11 | Strategy | -0.007181 | 0.005257 |
| 7 | Role-Playing | -0.014518 | 0.018141 |
| 2 | Fighting | -0.017509 | 0.068727 |
| 1 | Adventure | -0.017771 | 0.031397 |
| 3 | Misc | -0.019046 | 0.182028 |
| 4 | Platform | -0.023665 | 0.094100 |
| 5 | Puzzle | -0.026877 | 0.041042 |
| 9 | Simulation | -0.030133 | 0.130467 |
| 0 | Action | -0.033838 | 0.426196 |
| 6 | Racing | -0.046752 | 0.595651 |

**Table 1**. Coefficients and MSE for each genre

As shown in Table 1 which presents each genre's Coefficient and Mean Squared Error (MSE) values. The positive coefficient of the Shooter genre implies a rise in popularity or successful releases over the years. Lower MSE values indicate a better fit for the linear regression model. The genres like racing or action have MSE values of 0.596 and 0.426, respectively, suggesting a notable variability in sales data that the linear model could not accurately predict. However, specific genres experienced significant outliers before 1995, affecting the trend slope. These outliers are less relevant to recent sales, resulting in significantly higher MSE values for Action and Racing. It is recommended that the same analysis be conducted starting from 1995 to address this issue.

A screenshot of a graph

Description automatically generated

**Figure 12**. median sales data pivot table for years starting from 1995.

Based on the data presented in Figure 12, a filtering process was implemented in 1995. Despite this, the overall trend of the plot remained the same even after removing data from before 1995. However, there are notable variations in the mean square error for each genre, as demonstrated in Table 2. Each genre significantly reduces MSE with a more distinct coefficient trend.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Genre | Coefficient | MSE |
| 8 | Shooter | 0.018275 | 0.021797 |
| 4 | Platform | -0.003721 | 0.024200 |
| 10 | Sports | -0.004128 | 0.010125 |
| 2 | Fighting | -0.004989 | 0.006724 |
| 11 | Strategy | -0.005553 | 0.002682 |
| 1 | Adventure | -0.006220 | 0.001072 |
| 3 | Misc | -0.006533 | 0.005597 |
| 5 | Puzzle | -0.007899 | 0.002222 |
| 0 | Action | -0.008961 | 0.003407 |
| 6 | Racing | -0.009430 | 0.014980 |
| 7 | Role-Playing | -0.011652 | 0.003489 |
| 9 | Simulation | -0.012230 | 0.020250 |

**Table 2**. Coefficients and MSE for each genre start from 1995.

According to Linear Regression, The Shooter genre stands out with a growth trend, which is relatively rare among the other genres, most experiencing declines in median sales. This analysis reveals significant insights into the evolving trends of the video game industry, particularly the shifts in popularity among various genres. In 2007, several AAA games were launched, which has since been regarded as the start of the golden age of shooting games. Call Of Duty 4: Modern Warfare played a significant role in revolutionizing the gaming industry, and it is now recognized as one of the most popular first-person shooter games at that time (Jensen K., 2007). Valve Corporation, an American game developer, released Team Fortress 2. It features a cartoony art style and remains one of the most popular FPS titles. German developers Cry Team unveiled their CryEngine and continued to evolve the engine with Crysis in 2007. Its high graphic quality surpassed the technological capabilities of its time, making Crytek a trailblazer in game engine creation. Crysis's high hardware specifications sparked a cultural phenomenon, making it a benchmark and pop culture icon in gaming. Phrases like "Can it run Crysis?" are used to gauge the power of a computer system jokingly.

After the specific genre analysis, let’s move to the global sales to see the trend of the gaming industry.

A graph with blue lines and red dots

Description automatically generated

**Figure 13**. Overall Median of Global Sales Over the Years

As shown in Figure 13, the analysis reveals a slight but consistent decline in the median global video game sales over the years. This trend indicates various market factors, such as changes in consumer preferences or the impact of mobile gaming platforms.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Metric | Coefficient | MSE |
| 0 | Global Sales | -0.006283 | 0.001838 |

**Table 3**. display the coefficients and MSE for Global Sales

The negative coefficient of -0.006283 indicates a declining trend in the median global video game sales per year. This suggests that, on average, the median sales across all genres have decreased annually since 1995. The relatively low MSE of 0.001838 also indicates that the linear regression model fits the data well.

**4.4 RIDGE REGRESSION**

A screenshot of a graph

Description automatically generated

**Figure 14**. Ridge Regression plots using the median sales game each Genres.

In Figure 14, The analysis using Ridge regression has revealed a positive coefficient, indicating an increase in the median sales of Shooter games. Although the magnitude of the coefficient is relatively small, it still represents a significant trend, particularly when compared to other genres.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Genre | Coefficient | MSE | Alpha |
| 8 | Shooter | 0.018071 | 0.048383 | 10.0 |
| 4 | Platform | -0.003680 | 0.025302 | 10.0 |
| 10 | Sports | -0.004127 | 0.011496 | 0.1 |
| 2 | Fighting | -0.004933 | 0.008705 | 10.0 |
| 11 | Strategy | -0.005491 | 0.005137 | 10.0 |
| 1 | Adventure | -0.006213 | 0.004183 | 1.0 |
| 3 | Misc | -0.006460 | 0.008995 | 10.0 |
| 5 | Puzzle | -0.007811 | 0.007190 | 10.0 |
| 0 | Action | -0.008861 | 0.009799 | 10.0 |
| 6 | Racing | -0.009324 | 0.022059 | 10.0 |
| 7 | Role-Playing | -0.011521 | 0.014296 | 10.0 |
| 9 | Simulation | -0.012217 | 0.032278 | 1.0 |

**Table 4**. Coefficients, MSE, and Alpha value of Ridge Regression for each genre.

Table 4 highlights the Shooter genre as the only one with a positive coefficient, indicating an increased trend in median sales over time. However, compared to the Linear Regression model (MSE = 0.021797), Ridge Regression (MSE = 0.048383) tends to have a higher MSE, implying considerable fluctuation in its sales.

A graph with green lines and red dots

Description automatically generated

**Figure 15**. Ridge Regression plots using the median global sales.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Metric | Coefficient | MSE | Alpha |
| 0 | Global Sales | -0.006229 | 0.004989 | 7.675 |

**Table 5**. display the coefficients, MSE, and Alpha value for global sales.

The mean squared error (MSE) in global sales equals 0.004989, as presented in Table 5. MSE measures the average squared difference between the actual and predicted values, with a relatively low value indicating that the Ridge regression model fits the data reasonably well. The trend is further illustrated in Figure 15, where the green line represents the actual overall median sales per year, and the red dashed line shows the Ridge regression prediction. The plot depicts a general decline in sales over time.

**4.5 LASSO REGRESSION**

A screenshot of a graph

Description automatically generated

**Figure 16**. Lasso Regression plots using the median sales game each Genres.

The best alpha values have been applied in the lasso regression as shown in Figure 16 using 'LassoCV' with five-fold cross-validation, the regularization parameters in Lasso regression varying significantly across genres. Some genres, like Adventure (0.0469) and Racing (0.1889), have higher alpha values, suggesting a more substantial regularization effect.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Genre | Coefficient | MSE | Alpha |
| **8** | Shooter | 0.018257 | 0.048655 | 0.000736 |
| **2** | Fighting | -0.000000 | 0.007726 | 0.200795 |
| **4** | Platform | -0.000000 | 0.024757 | 0.149773 |
| **10** | Sports | -0.000000 | 0.010810 | 0.166136 |
| **6** | Racing | -0.004737 | 0.019462 | 0.188901 |
| **1** | Adventure | -0.005054 | 0.003658 | 0.046909 |
| **3** | Misc | -0.005391 | 0.008485 | 0.045952 |
| **11** | Strategy | -0.005548 | 0.005162 | 0.000224 |
| **5** | Puzzle | -0.007415 | 0.006947 | 0.019509 |
| **0** | Action | -0.008952 | 0.009865 | 0.000361 |
| **7** | Role-Playing | -0.011640 | 0.014406 | 0.000469 |
| **9** | Simulation | -0.012218 | 0.032280 | 0.000492 |

**Table 6**. Coefficients, MSE, and Alpha value of Lasso Regression for each genre.

In Table 6, The shooter genre shows a positive coefficient (0.0183), suggesting an increasing trend in median sales over time. Some genres like Fighting, Platform, and Sports have coefficients close to zero or exactly zero, suggesting a relatively stable trend with little to no change over time.

A graph with orange and blue lines

Description automatically generated

**Figure 17**. Lasso Regression plots using the median global sales.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Metric | Coefficient | MSE | Alpha |
| **0** | Global Sales | -0.004585 | 0.004273 | 0.068355 |

**Table 7**. display the coefficients, MSE, and Alpha value for global sales.

As presented in both Figure 17 and Table 7, the Lasso coefficient is -0.004585. This implies a minor negative trend in the median video game sales overall, indicating a gradual decline in median sales over the years. An MSE value of 0.004273 suggests that the model fits the data well, with relatively small prediction errors.

**4.6 MODEL EVALUATION**

|  |  |  |
| --- | --- | --- |
| Model | Coefficient | MSE |
| Linear Regression | 0.018275 | 0.021797 |
| Ridge Regression | 0.018071 | 0.048383 |
| Lasso Regression | 0.018257 | 0.048655 |

**Table 8**. Summary of Model Performance Results of Shooter Genre.

Based on the provided Linear Regression, Ridge regression, and Lasso regression results for the Shooter genre as shown in Table 8. All three models have very similar coefficients, indicating a consistent estimate of the impact of the predictor variable on the target (median sales for the Shooter genre). The slight differences can be attributed to the regularization in Ridge and Lasso. The Linear regression model has a significantly lower MSE than the Ridge and Lasso models. This suggests that for the Shooter genre, a simple linear model might fit the data better than the regularized models. The higher MSE in Ridge and Lasso indicates that the penalty on coefficients may not benefit this dataset.

|  |  |  |
| --- | --- | --- |
| Model | Coefficient | MSE |
| Linear Regression | -0.006283 | 0.001838 |
| Ridge Regression | -0.006229 | 0.004989 |
| Lasso Regression | -0.004585 | 0.004273 |

**Table 9**. Summary of Model Performance Results of Global Sales.

The same method was applied to Table 9, as shown in Table 8. The Linear regression model is the best fit for predicting global sales, given its lowest MSE. Ridge and Lasso's regressions do not improve the model performance for global sales. The regularization slightly degrades the model's accuracy. The consistent negative coefficients across all models reinforce the conclusion of a general decline in global sales.

**4.7 MODEL INTERPRETATION**

A graph with blue lines and numbers

Description automatically generated

**Figure 18**. Global Sales of Shooter Games.

Let's talk about the rise in popularity of shooting games. Our analysis indicates that several factors have contributed to this shift in consumer interest towards this genre as shown in Figure 18. Since the early 2000s, many new games have been developed, including Looter Shooters like Borderlands (2009), known for their procedurally generated weapons and loot, such as armor or weapon parts. These games offer fast-paced gunplay and encourage players to focus on grinding to progress their characters. They combine long-term character growth, the satisfaction of upgrading your gear, and exciting high-energy shooter gameplay (Karthikeyan K., 2023). It's worth noting that 2007 began a golden age for shooting games. Many excellent shooting games were released during this time and gained global popularity. From Left 4 Dead 2 in 2009 to the Counter Strike Global Offensive launch in 2012, these games have significantly impacted the gaming industry.

A graph with green lines and numbers

Description automatically generated

**Figure 19**. Counter Strike Global Offensive Players Online from Steam DB.

As illustrated in Figure 19, Since its debut in August 2012, this esports game has gained a significant following. It boasts an expansive player base and professional scene and has played a pivotal role in developing esports (Bresaola R., 2023). The game has gained widespread recognition and acclaim and has remained a staple in the industry, retaining a devoted fan base. Its influence can be seen in the evolution of modern esports, setting a high standard for other games to strive for.

A graph of growth and statistics

Description automatically generated with medium confidence

**Figure 20**. ESport Audience Growth (NEWZOO, 2019).

According to Figure 20, the esports industry has experienced significant growth in audience engagement, surpassing expectations and propelling the gaming industry forward. The rising popularity of shooter games and esports has transformed the gaming industry landscape, with an influx of investors expected to come forward and invest in this burgeoning sector.

A screenshot of a video game

Description automatically generated

**Figure 21**. Most watched shooter games on Twitch

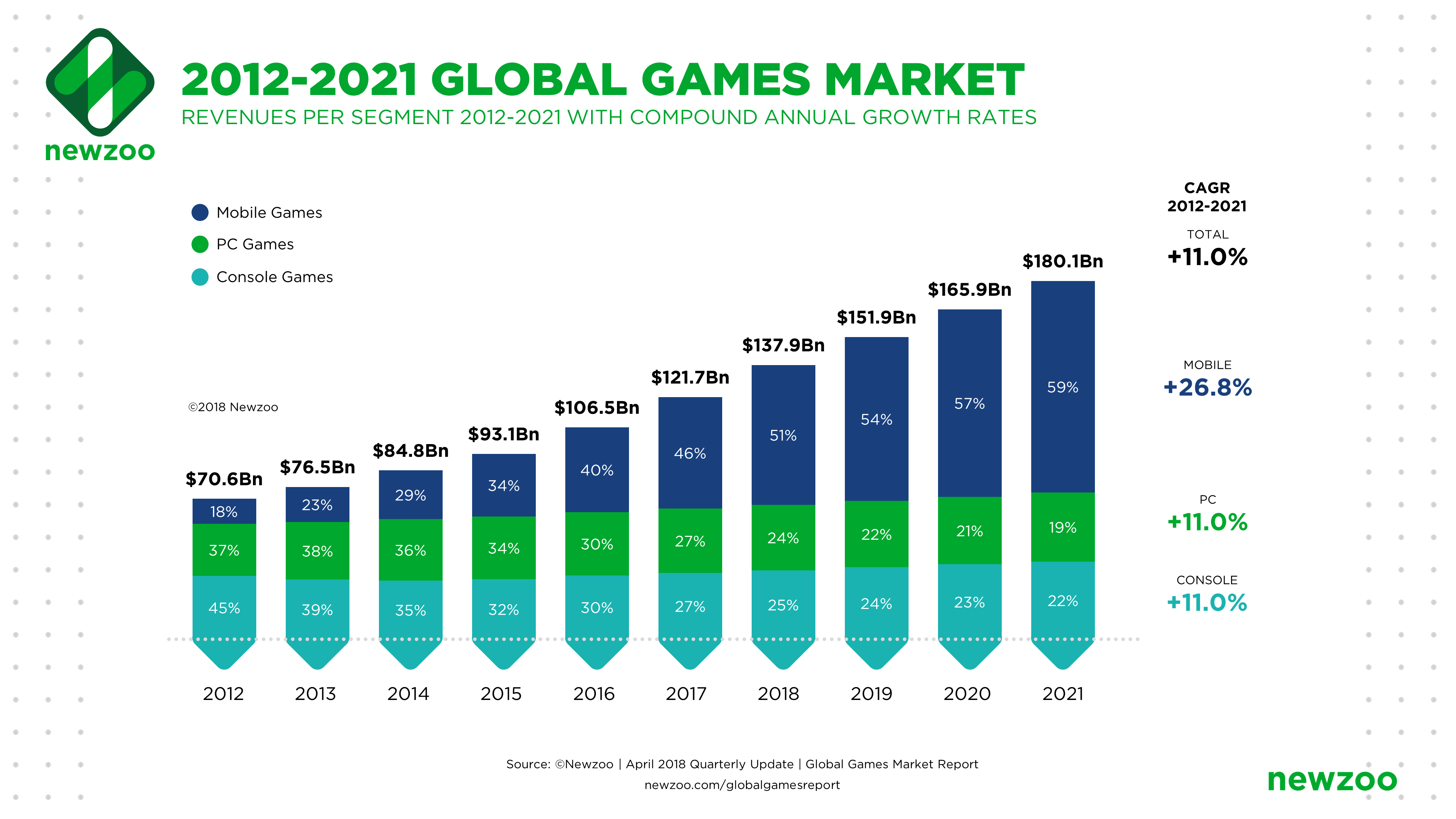
Carrico F. discusses the most watched shooter games on a streaming platform called Twitch in 2022; the number of people watching the streaming of shooter games has continued to increase yearly since the release of many iconic shooter games, as listed in Figure 21.

A graph with blue lines and red dots

Description automatically generated

**Figure 22.** Linear Regression Global Game Sales

The negative coefficients across various genres suggest a decreasing median video game sales trend. Many aspects cause the negative trend. The Mean Squared Error (MSE) metrics, which are consistently low in global sales, indicate a high degree of model accuracy in predicting median sales based on the year of release. This underscores the effectiveness of the Linear Regression model in capturing the linear relationship between time and sales as shown in Figure 22. The rise of mobile gaming has completely transformed the gaming industry, while PC and console games have experienced a decline.



**Figure 23**. NEWZOO Quarterly Update 2018

Figure 23 shows the Global Games Market report from the NEWZOO. The CAGR (Compound annual growth rate) has increased by around 11% over the years. But to deeply see the past data. The PC and Consoles have shown a declining train each year. The emergence of mobile gaming shows significant growth as many investors shift their target to the companies that develop mobile games.

A graph of a graph showing the cost of gaming

Description automatically generated

**Figure 24.** Worldwide Consumer Spending on Games (data.ai and IDC, 2022)

The mobile gaming industry has seen a significant rise in free-to-play games which can be downloaded and played without charge. These games often include in-game purchases for virtual items or upgrades. According to a recent study by data.ai and IDC, mobile gaming surpasses the gaming industry and is projected to capture approximately 61 percent of the market share by 2022, as depicted in Figure 24. In the first quarter of 2022, it was noted that mobile game consumers spent over US$1.6 billion per week on both the iOS App Store and Google Play Store (Wong D., 2022).

**CHAPTER 5**

**CONCLUSION**

This research provides a detailed analysis of the sales patterns of video games across different genres over time. The study involved extensive data preprocessing and using Linear, Ridge, and Lasso Regression models to understand video game industry trends comprehensively. The findings indicate that there has been a general decline in sales over the years and that sales trajectories vary significantly across genres. The comparative analysis of the regression models suggests that Linear and Ridge Regression are more effective than Lasso Regression in fitting the dataset, as Lasso Regression tends to over-regularize.

**LIMITATIONS AND FURTHER WORK**

This research provides valuable insights, but it has some limitations. Firstly, the analysis only considered median sales and was restricted to a specific time frame. Factors such as marketing strategies, consumer demographics, and game platforms that may affect sales trends were not considered. Secondly, the study primarily relied on regression models, which may not effectively capture non-linear relationships or interactive effects between various predictors. Finally, the choice of features, particularly in the Lasso Regression model, was limited to the year of release, which could have ignored other vital predictors that could impact sales. To further develop the findings of this research, future studies can consider the following:

* Expanding the scope of data by incorporating additional variables such as consumer age groups, digital versus physical sales channels, and platform-specific sales data. This can provide a more comprehensive analysis.
* Exploring advanced modeling techniques such as time series analysis, machine learning algorithms, or even deep learning to capture complex, non-linear relationships in the data.
* Conducting a longitudinal study to examine how introducing new gaming technologies and platforms affects sales trends over time.
* Comparing the video game industry's sales trends with other entertainment sectors to identify unique market dynamics and consumer behavior patterns.

**REFLECTION**

This project is an enlightening journey, as it has allowed me to apply data science techniques to real-world data. While analyzing the video game sales dataset, I have gained a more robust understanding of the intricacies involved in predictive modeling. One key takeaway has been the importance of thorough data preprocessing and its impact on the outcomes of analytical models. I have also learned about the critical role of model selection in data science, as different regression models can provide unique perspectives on the same dataset. It was challenging to ensure the models' robustness and applicability to the dataset at hand, especially when balancing model complexity with the risk of overfitting in Ridge and Lasso Regression. Lastly, a significant challenge was interpreting the results meaningfully to provide practical insights for the video game industry.

**RESEARCH CONTRIBUTION**

This research provides valuable insights into the video game industry and its evolving sales landscape. The results can help game developers, marketers, and strategists understand historical trends, tailor marketing campaigns, and develop strategies that align with consumer preferences and market demands. Furthermore, the report highlights a general decline in sales over the years because of this shifting consumer behavior to mobile gaming, indicating the need for innovation and adaptation within the industry.

**REFERENCES**

1. Eventyr, (2023) Game Analytics and Data-driven Development: Leveraging Insights for Success. [online]. [Accessed 16 September 2023].
2. Clement, J. (2023) Video Gaming Worldwide - Statistics & Facts. [online]. [Accessed 16 September 2023].
3. Eng, D. (2020) What Is Player Behavior? [online]. [Accessed 17 September 2023].
4. Bellis, Mary (2020) The History of Spacewar: The First Computer Game [online]. [Accessed 14 October 2023].
5. Bellis, M. (2019) The History of Spacewar: The First Computer Game in 1962, Steve Russell Invented Spacewar. [online]. [Accessed 18 October 2023].
6. Opus Web Design, (2016) A Brief History of the Computer Games Industry. [online]. [Accessed 18 October 2023].
7. Mattoo, S. (2022) How Virtual Reality Is Revolutionizing the Video Gaming Industry. [online]. [Accessed 18 October 2023].
8. Caicedo, J. (2023) Potential Growth of VR and the Gaming Industry in the Next Five Years. [online]. [Accessed 18 October 2023].
9. Osborne, H. (2021) What’s the Problem with Cloud Gaming?​. [online]. [Accessed 18 October 2023].
10. Project Pro, (2023) Predictive Modeling Techniques- a Comprehensive Guide [2023]. [online]. [Accessed 18 October 2023].
11. Ane, Y. (2022) The 6 Steps of Predictive Analytics. [online]. [Accessed 18 October 2023].
12. Qualtrics, (2023) What Is Predictive Analytics? Your Ultimate Guide. [online]. [Accessed 18 October 2023].
13. Ankleshwariya, Y. (2023) Predictive Analytics: The Future of Gaming Industry Growth and Success. [online]. [Accessed 18 October 2023].
14. Brown, S. (2021) Machine Learning, Explained. [online]. [Accessed 19 October 2023].
15. Stewart, M. (2019) The Actual Difference Between Statistics and Machine Learning. [online]. [Accessed 24 October 2023].
16. Ozyazgan, (2019) The Data Science Boom in Esports. [online]. [Accessed 24 October 2023].
17. Game Marketing Genie, (2021) How to Improve Your Game Marketing Roi with Video Game Data Analytics. [online]. [Accessed 24 October 2023].
18. More, J. (2023) Tapping into the Data Analytics Landscape for Transforming the Gaming Industry. [online]. [Accessed 24 October 2023].
19. Jimmy, J. (2022) The Importance of Data Analytics and Machine Learning in Gaming. [online]. [Accessed 24 October 2023].
20. Weber, B. (2018) A History of Game Analytics Platforms. [online]. [Accessed 26 October 2023].
21. Sharma, K., Mohammad, F. and Parashar, D. (2023) Apache Spark in Riot Games: A Case Study on Data Processing and Analytics. [online]. 14, pp. 32-40. [Accessed 31 October 2023].
22. Shim, K., Sharan, R. and Srivastava, J. (2010) Player Performance Prediction in Massively Multiplayer Online Role-playing Games (MMORPGs). [online]. [Accessed 02 November 2023].
23. Ong, H., Deolalikar, S. and Peng, M. (2023) Player Behavior and Optimal Team Composition in Online Multiplayer Games. [online]. [Accessed 02 November 2023].
24. Future Market Insight, (2023) Video Game Market Overview (2023 to 2033). [online]. [Accessed 02 November 2023].
25. Report Linker, (2023) Mobile Gaming Market Size & Share Analysis - Growth Trends & Forecasts (2023 - 2028). [online]. [Accessed 02 November 2023].
26. Moreira, A., Filho, V. and Ramalho, G. (2014) Understanding Mobile Game Success: A Study of Features Related to Acquisition, Retention and Monetization. [online]. [Accessed 03 November 2023].
27. Katkoff, M. (2012) From Players to Payers. [online]. [Accessed 03 November 2023].
28. Academy Contributor, (2020) A Decade of Game Development and Innovation: Major Milestones From 2000 to 2010. [online]. [Accessed 19 December 2023].
29. MCV, (2009) UK Best Selling Game of 2008. [online]. [Accessed 19 December 2023].
30. KENDALLGILLIES, (2017) Video Game Sales by Genre. [online]. [Accessed 19 December 2023].
31. Thor Jensen, K. (2017) Knee Deep in the Dead: The History of First-person Shooters. [online]. [Accessed 20 December 2023].
32. Karthikeyan, K. (2023) The Evolution of First Person Shooter (Fps) Games. [online]. [Accessed 28 December 2023].
33. Bresaola, R. (2023) What Is the Significance of Counterstrike 2 and How Will It Impact Esports. [online]. [Accessed 28 December 2023].
34. Carrisco, F. (2022) Most Popular Twitch Shooters and Streamers of 2022. [online]. [Accessed 28 December 2023].
35. Newzoo, (2019) Global Esports Economy Will Top $1 Billion For the First Time in 2019. [online]. [Accessed 28 December 2023].
36. Wong, D. (2022) 60% of Entire Gaming Market Is Now Dominated by Mobile Gaming, Study Finds. [online]. [Accessed 28 December 2023].

**APPENDICES**

**Appendix A: Supervisory Meeting**

The following table depicts the important minutes of meeting (MOM) with supervisor (Mr. Ejay Nsugbe) and subsequent actions.

|  |  |  |
| --- | --- | --- |
| **No.** | **Date and Subject** | **Note** |
| MOM-1 | 25 Oct 2023 | Propose Project Proposal | * Write email to supervisor to propose the project scope and sending the project proposal. |
| MOM-2 | 8 Nov 2023 | Weekly Supervision | * Explain the scope of the project, the progress of finding dataset, and which machine learning algorithms will be applied in this analysis. |
| MOM-3 | 28 Nov 2023 | Weekly Supervision | * Update the progress of scraping the data and results of the ARIMA model. * Face the problem of dataset that does not satisfy the ARIMA model I’m choosing. |
| MOM-4 | 6 Jan 2024 | Last Supervision | * Update the change of dataset and ML algorithm into Regression model. * Show Overview of the final report. |

**Appendix B: Wore Count**

8160 words (excluding abstract, references and appendices)