# Abstract

This project explores the effectiveness of machine learning and econometric models in forecasting video game sales across various regions. By comparing models such as Linear Regression (LR), Random Forest (RF), XGBoost (XGB), and Artificial Neural Networks (ANN), the study assesses their predictive accuracy and ability to capture the complex, non-linear relationships inherent in video game sales data. The findings indicate that while traditional econometric models like LR provide a foundational understanding of sales dynamics, they struggle to match the precision and adaptability of advanced machine learning models. ANN emerged as the most accurate model overall, particularly excelling in North American and European data, while RF showed strong performance in Japanese data and the Rest of the World. The research suggests a strategic approach to model selection based on the stage of game development, with LR being more useful during the planning and design phases and machine learning models becoming essential for refining sales strategies. These insights offer valuable contributions to the video game industry, highlighting the need for sophisticated forecasting techniques in an increasingly complex market.

Table of Contents

[Abstract i](#_Toc184090150)

[List of Figures iv](#_Toc184090151)

[List of Tables v](#_Toc184090152)

[List of Equations v](#_Toc184090153)

[List of Abbreviations v](#_Toc184090154)

[1. Introduction 1](#_Toc184090155)

[1.1 Background 1](#_Toc184090156)

[1.2 Importance and Motivation of the Research 1](#_Toc184090157)

[1.3 Research Questions and Scope 2](#_Toc184090158)

[1.4 Research Aims and Objectives 3](#_Toc184090159)

[1.5 Structure of the Research 3](#_Toc184090160)

[2. Literature Review 5](#_Toc184090161)

[2.1 Video Game Demand Forecast 5](#_Toc184090162)

[2.2 Demand forecasting comparison across all industries 5](#_Toc184090163)

[2.3 Demand forecasting comparison in gaming industry 7](#_Toc184090164)

[2.4 Regional Studies on Video Game Sales Forecasting 7](#_Toc184090165)

[2.5 Factors Impacting Video Game Sales 8](#_Toc184090166)

[2.6 Best performing models 10](#_Toc184090167)

[3. Methodology 12](#_Toc184090168)

[3.1 Philosophy 12](#_Toc184090169)

[3.2 Approach 12](#_Toc184090170)

[3.3 Strategy 12](#_Toc184090171)

[3.4 Choice 13](#_Toc184090172)

[3.5 Time Horizon 13](#_Toc184090173)

[3.6 Techniques & Procedure 13](#_Toc184090174)

[3.7 Data Understanding and Cleaning 14](#_Toc184090175)

[3.8 Data Exploration 16](#_Toc184090176)

[4. Results 20](#_Toc184090177)

[5. Discussion 28](#_Toc184090178)

[5.1 Residual Analysis 28](#_Toc184090179)

[5.2 Model Analysis 29](#_Toc184090180)

[5.3 Model Selection 31](#_Toc184090181)

[5.4 Answers to Research Questions 32](#_Toc184090182)

[5.5 Comparison with Existing Literature 33](#_Toc184090183)

[5.6 Limitations of the Study 34](#_Toc184090184)

[5.7 Future Scope of the Study 35](#_Toc184090185)

[6. Conclusion 36](#_Toc184090186)

[References 38](#_Toc184090187)

[Appendices 43](#_Toc184090188)

[Ethical Approval Form 43](#_Toc184090189)

[Research Proposal 50](#_Toc184090190)

# List of Figures

[Figure 1: Onion Model 12](#_Toc176298097)

[Figure 2: Impact of Platform launches on Sales Trends (1980-2016) 16](#_Toc176298098)

[Figure 3: Sales Distribution by Top Franchises (1980-2016) 17](#_Toc176298099)

[Figure 4: Sales Distribution by Region (1980-2016) 17](#_Toc176298100)

[Figure 5: Sales Distribution by Platform (1980-2016) 18](#_Toc176298101)

[Figure 6: Sales Distribution by Genre (1980-2016) 19](#_Toc176298102)

[Figure 7: Sales Distribution by Franchise (1980-2016) 19](#_Toc176298103)

[Figure 8: Sales Distribution by ESRB Rating (1980-2016) 20](#_Toc176298104)

[Figure 9: Training Residuals 21](#_Toc176298105)

[Figure 10: Test Residuals 22](#_Toc176298106)

[Figure 11: Feature Impact on Global Sales 22](#_Toc176298107)

[Figure 12: LR Feature Importance 23](#_Toc176298108)

[Figure 13: RF Feature Importance 23](#_Toc176298109)

[Figure 14: XGB Feature Importance 24](#_Toc176298110)

[Figure 15: ANN Feature Importance 24](#_Toc176298111)

[Figure 16: Global Model Measures 26](#_Toc176298112)

[Figure 17: Regional Model Measures 26](#_Toc176298113)

# List of Tables

[Table 1: Data Dictionary 15](#_Toc176300359)

[Table 2: Model Measures 25](#_Toc176300360)

# List of Equations

[Equation 1: Mean Absolute Error (MAE) 14](#_Toc176298122)

[Equation 2: Mean Squared Error (MSE) 14](#_Toc176298123)

[Equation 3: Root Mean Squared Error (RMSE) 14](#_Toc176298124)

[Equation 4: R-squared (R2) 14](#_Toc176298125)

# List of Abbreviations

|  |  |
| --- | --- |
| **Abbreviation** | **Description** |
| ANN | Artificial Neural Network |
| ARIMA | Autoregressive Integrated Moving Average |
| DL | Deep Learning |
| ESRB | Entertainment Software Rating Board |
| EU | Europe |
| JP | Japan |
| LR | Linear Regression |
| MAE | Mean Absolute Error |
| ML | Machine Learning |
| MSE | Mean Squared Error |
| NA | North America |
| NDS | Nintendo DS |
| PC | Personal Computer |
| PCA | Principal Component Analysis |
| PS3 | PlayStation 3 |
| PSP | PlayStation Portable |
| RF | Random Forest |
| RMSE | Root Mean Squared Error |
| RPG | Role-Playing Game |
| SVR | Support Vector Regression |
| XGB | XGBoost |

# 1. Introduction

## 1.1 Background

Since the 1970s the video game industry has grown significantly, starting from basic arcade games and home console systems to a global phenomenon. Initially the home-video game industry started with 8-bit consoles such as the Atari 2600, whereas now the gaming industry has exponentially grown with several genres and gaming platforms including advanced gaming PCs and virtual and augmented reality systems. This progress has not only enhanced the popularity of video games but also significantly improved their economic influence, thus establishing the industry as a major player in the global entertainment market.

Technology has developed rapidly over the years, from basic home consoles to the creation of modern online multiplayer games. This has changed the overall gaming experience of consumers and has played a significant role in shaping the video game industry. In the modern world, distribution channels such as Steam, Xbox Live, and PlayStation Network has redefined how games nowadays are promoted, sold, and played by consumers. These platforms have enabled users to access a wide array of games from the comfort of their homes, thus increasing market share and boosting sales.

Simultaneously, gaming has now become a social activity with the advent of online gaming communities. Global gaming communities have been established by platforms supporting online multiplayer games, resulting in users from all over the world to communicate, team up, and compete. This social feature of gaming has contributed to its popularity and commercial success. As a result, the video game industry currently generates more revenue than the music and film industries combined, showcasing the significance of the gaming industry in the entertainment sector.

## 1.2 Importance and Motivation of the Research

Despite the industry’s rapid expansion, precisely estimating video game sales remains a complicated and challenging task. Forecasting future trends becomes a challenge due to the continuous shift of consumer preferences along with the frequent innovations in technology. In order to make important business decisions that cover everything from production and inventory management to marketing and sales strategies, forecasting sales effectively becomes crucial for all stakeholders associated with the video game industry, including developers, publishers, investors, and marketers.

Traditional econometric models have often been used to forecast video game sales by looking at historical data and industry trends. Although these theory-based models provide insightful analyses, they mostly fail to capture the dynamic and nonlinear nature of the current gaming industry. Machine learning models, on the other hand, provide a viable substitute owing to their capacity to handle large datasets and recognise complex patterns. In comparison to traditional econometric models, recent research shows that machine learning methods might provide a better predictive accuracy.

The motivation for this study arises from the need to fill this gap in the existing literature by examining the effectiveness and accuracy of econometric and machine learning models in forecasting video game sales across multiple regions. By performing a thorough comparative analysis, this study attempts to determine the most effective forecasting methods that considers the distinctive features of the video game industry across various regions.

## 1.3 Research Questions and Scope

To address the complexities of video game sales forecasting, this study is guided by the following primary research question:

*How do econometric models compare to machine learning techniques in predicting video game sales in terms of fit and informing design?*

And more specifically, the following two research questions:

1. *Which specific algorithms demonstrate the highest accuracy in forecasting video game sales across different regions?*
2. *What insights can be gained from the comparative analysis of forecasting models to enhance sales forecasting strategies in the video game industry?*

The scope of this research encompasses a global perspective, analysing major regions such as North America, Europe, and Japan. These regions have distinct market dynamics and consumer behaviour, making them ideal for a comparative study. By examining a broad range of forecasting techniques, this research seeks to provide a comprehensive understanding of the most suitable methods for estimating video game sales in different parts of the world.

## 1.4 Research Aims and Objectives

The primary aim of this research is to enhance the understanding of forecasting video game sales by conducting a comparative analysis of econometric and machine learning models. The specific objectives of this study are as follows:

1. *Evaluate the performance of econometric models:* To assess the accuracy of traditional econometric models in forecasting video game sales.
2. *Analyse machine learning techniques:* To examine various machine learning algorithms and determine their accuracy in predicting video game sales.
3. *Identify the best-performing models:* To compare the performance of econometric and machine learning models in terms of their accuracy.
4. *Provide practical insights:* To offer actionable insights for industry stakeholders to optimize production, marketing, and sales strategies based on the most effective forecasting models.
5. *Address regional variations:* To analyse how different forecasting models perform across major regions and understand the unique market dynamics in North America, Europe, and Japan.

By achieving these objectives, this research aims to make a meaningful contribution to both academic literature and practical applications in the video game industry. The findings have the potential to enhance the competitive advantage of industry stakeholders in an ever-changing market by assisting them to make more informed decisions.

## 1.5 Structure of the Research

This project is designed to provide a comprehensive analysis of video game sales forecasting techniques, systematically addressing the research questions and objectives listed above. The structure is as follows:

* Introduction: The present section provides the background, importance, and motivation for the research, outlines the research questions and scope, and states the research aims and objectives.
* Literature review: The following section reviews existing studies on video game demand forecasting, comparing econometric and machine learning techniques not only within the gaming industry but also across other major industries. It also highlights the key findings and gaps in the existing literature.
* Methodology: This section describes the research design, and analytical methods used to compare the forecasting models. It details the data collection process, the selection of models, and the criteria for evaluating their performance.
* Results: The results section presents the findings of the comparative analysis, showcasing the performance of different forecasting models in predicting video game sales across various regions.
* Discussion: This section interprets the results in the context of the broader literature, discussing the implications on the gaming industry. It also addresses the limitations of the study and suggests areas for future research.
* Conclusion: The final section summarises the key insights from the research, highlights the contributions to the field, and provides recommendations for industry stakeholders.

Through this structured approach, the project aims to provide a comprehensive analysis of forecasting techniques in the video game industry, offering valuable contributions to both academic knowledge and practical applications. The findings presented in this study can be used to guide future research and help industry stakeholders in managing the complexities and intricacies of video game sales forecasting.

# 2. Literature Review

## 2.1 Video Game Demand Forecast

The video game industry has grown at an exponential rate over the past few decades, with the market expected to generate $282.30 billion in revenues in 2024, increasing by 8.76% per year to reach $363.20 billion in revenues by 2027. This rapid growth requires reliable demand forecasting models to assist industry stakeholders in making educated decisions. In a report on the global video game market, LinkedIn Pulse focussed on technology improvements and consumer expenditure on gaming to analyse growth trends between 2024 and 2032. Furthermore, Kevuru Games highlighted some technological developments in gaming, such as augmented reality and virtual reality, which are expected to have a significant impact on demand.

Several studies have looked into different methods for forecasting sales of video games. For instance, Liang (2022) forecasted future sales by looking at user engagement indicators, historic sales data, and social media trends. Toivonen and Sotamaa (2010) conducted a survey among Finnish gamers and examined the effects of online distribution channels on video game sales, underscoring the significance of platforms such as Steam and PlayStation in the dynamic market.

A detailed report on analysing video game sales using data science techniques provided insights into forecasting top-selling video games in North America from 1983 to 2016. This study analysed past sales information and developed models to predict future sales, showcasing the significance of historical trends and consumer behaviour in predicting market trends.

## 2.2 Demand forecasting comparison across all industries

With the use of both econometric and machine learning models, demand forecasting has become a crucial area of research across various industries. Econometric models, which are based on economic theory, often use historical data to detect patterns and relationships between variables. For example, Kohli et al. (2020) used Linear Regression to accurately predict the sales of the German drugstore chain Rossmann.

Machine Learning models, such as Random Forests (RF), XGBoost (XGB), and Artificial Neural Networks (ANN), automatically detect and model complex patterns within large datasets, enabling accurate predictions. These models excel at capturing non-linear relationships between variables, adapting to the underlying data structure. For example, RFs use multiple decision trees to enhance prediction accuracy, while XGB optimizes gradient boosting for efficient forecasting. ANNs mimic the human brain's neural structure, learning intricate patterns over time, making them highly effective for dynamic environments.

Machine learning models have grown popular due to their capacity to handle large datasets and complex non-linear relationships. Hyndman and Athanasopoulos (2018) evaluated traditional econometric models against machine learning methods, such as Random Forests and Neural Networks, across different industries. According to the findings of their study, machine learning models mostly surpassed traditional approaches in terms of accuracy and flexibility.

Ahmed et al. (2010) conducted a comprehensive analysis of econometric and machine learning models in predicting retail sales. The research findings revealed that whereas econometric models generated results that could be interpreted easily, machine learning methods exhibited a higher predictive accuracy, particularly when managing large and unstructured datasets.

The study by Suman Basuroy et al. (2003) examined the effects of critical reviews, star power, and budgets on box office results, highlighting how econometric models can be utilised to evaluate such impacts in the film industry. Similarly, Chintagunta et al. (2009) used an advanced econometric model to analyse the significance of marketing-mix elements—namely prices and software availability—in the video game console industry. By utilising a flexible estimation framework and addressing the potential endogeneity of prices through instrumental variables, the study effectively captured the intrinsic growth patterns of 32/64-bit video-game consoles. The research revealed that while prices heavily influenced sales in the early stages, software availability had a greater impact in later periods.

In a recent research paper, Pérez-Pons et al. (2021) analysed the latest applications of machine learning and econometric methodologies in multiple industries. The study discovered that while combining both methodologies often produced the best results, machine learning methods typically performed better than econometric models. For instance, in the investment industry, Support Vector Regression (SVR), a machine learning technique, was combined with ARIMA, a traditional econometric model, to predict time series data. Similarly, in tourism, the combination of Deep Learning (DL) with regression model was used to enhance the prediction of future performance. The research found that these hybrid approaches often yielded superior results, suggesting that integrating the strengths of both methodologies could lead to more accurate and robust predictions across different sectors.

## 2.3 Demand forecasting comparison in gaming industry

Forecasting demand in the video game industry has become really difficult due to the ever-changing consumer preferences and innovative technological developments. Traditionally, econometric models have been used to forecast sales based on past data and market trends. Zhang (2022), for instance employed a multiple LR model to estimate the number of players, using variables such as the number of reviews, discount strength, release timing, and language support to explain the variability in game ownership. This econometric approach highlighted the significant factors influencing game sales on the steam platform.

Machine learning models have increasingly been used in recent research to forecast sales and demand more accurately in the video game industry. Manel González-Piñero (2017) performed a comparative analysis with machine learning methods, such as Support Vector Machines and Gradient Boosting, and classical econometric models. The findings of the study indicated that both in terms of predictive accuracy and the ability to capture complex interactions between variables, machine learning methods performed better than traditional econometric models.

In another study, Liedtke (2023) explored the use of deep learning techniques, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), in forecasting video game sales. In comparison to traditional econometric models, these models produced more robust and reliable forecasts as they had the ability to capture a large volume of data from social media, user reviews, and gameplay statistics.

An in-depth analysis of video game sales employing data mining techniques explored factors that impact sales of video games such as the genre, platform, and release season of video games. The research findings established that machine learning models, especially those based on ensemble methods, performed well with a better accuracy that that of traditional econometric models.

## 2.4 Regional Studies on Video Game Sales Forecasting

Video game demand demonstrates significant variations across different regions, and hence region-specific forecasting models become a crucial part of demand forecasting. Geethanjali et al. (2020) conducted research in North America which focused on forecasting the sales of popular video games. The study used historical sales data and developed a LR model that can accurately predict future sales in the region, highlighting the significance of market trends and regional preferences in this process.

A research conducted by Marchand and Hennig-Thurau (2013) in Europe shed light on the process of value generation in the video game industry, with an emphasis on customer benefits and industry economics. The study focussed on how econometric and machine learning models can be used to understand and analyse the European market dynamics. The research concludes that the video game industry's inherent creativity and innovation will continue to drive the development of diverse game types, formats, and business models, presenting ongoing challenges and opportunities for future research.

In order to investigate the demand of video games in Japan, Harada (2007) conducted a household data analysis of consumers in that region. The study looked at factors like income of household, genre of the game, and popularity of the platform on which the game was released, and then analysed how each of these factors impacted video game sales using logarithmic regression models. The research findings showcased the unique market dynamics in Japan thus highlighting the need for region-specific forecasting.

## 2.5 Factors Impacting Video Game Sales

There are multiple features that affect the sales of video games, such as genre, platform, release season, online reviews and ratings, consumer behaviour, economic conditions, technological advancements, and marketing strategies. A study by Chintagunta et al. (2009) analysed the impacts of marketing-mix in the video game console industry using econometric models, which showcased how price, advertising, and promotional strategies influence the sales across multiple regions. The researchers measured price in terms of its effect on sales elasticity, meaning how sensitive the sales volume was to changes in price. Advertising was analysed by looking at expenditure levels and their effectiveness in driving sales, while promotional strategies were assessed by the extent to which they boosted sales through short-term incentives like discounts or bundling. The best strategy identified varied by region, but a common finding was that price cuts had a significant positive impact on sales, particularly when combined with strong advertising support. Thus, the interplay between these factors—especially the synergy between price reductions and advertising—proved to be the most effective in driving sales across different markets.

Similarly, the study by Zhu and Xiaoquan (Michael) Zhang (2010) focussed on product and consumer attributes as it explored the effect of online user reviews on sales. The study concluded that online user reviews are extremely vital in determining sales, particularly for products like video games.

A recent study by Liang (2022) analysed past sales data, user engagement metrics, and social media trends for forecasting future sales. The study reported that social media interactions and online reviews are two main aspects of consumer behaviour that significantly impact video game sales. The research highlighted the importance of using big data analytics to accurately capture user sentiment and estimate market trends.

Video game sales are significantly influenced by economic factors such as overall economic environment and consumer spending power. Marchand and Hennig-Thurau (2013) conducted an analysis on the economic aspects of the video game industry and examined how consumer purchasing trends and economic downturns impact sales. According to the research, when the market is in recession consumers tend to focus more on purchasing essential items which results in a negative impact on video game sales. On the other hand, at times when the economy is booming, consumers will have a higher purchasing power which can lead to greater spending on personal entertainment including video games.

Technological improvements are crucial to the video game industry, and it has a significant impact on both video game production and sales. Using the meta-frontier approach – a method used to compare the efficiency of decision-making units (DMUs) that operate under different technologies or in different environments – Xi et al. (2022) looked at features that impact technological innovation efficiency – the ability to maximize outputs (such as advanced video game production and sales) relative to the inputs used (like human capital, financial resources, and infrastructure) in the innovation process – in the Chinese video game market. The research findings depict that greater technological innovation efficiency is shown by regions with higher levels of human capital, financial investment, economic development, and infrastructure support. The study also highlighted the importance of governments to encourage and support technological innovation as that might lead to increase in interest and sales for advanced video games.

Apart from economic conditions, there are various other factors that impact the sales of video games. Miller (2016) found that video game sales are highly impacted by the number of platforms where the game is accessible and also on the number of weeks since the game was released. It is typically seen that there are higher sales during the game launch time and then gradually the sales declines over time. Additionally, games which are launched on multiple platforms lead to a higher sales volume as the game becomes accessible to more consumers across the world. One interesting finding of the research was that the Entertainment Software Rating Board (ESRB) ratings, where the games are categorized by their content, has no significant impact on the sales indicating that violent or mature game content does not necessarily translate to higher or lower sales.

One of the other factors that significantly impact video game sales is the platform on which the game is launched. In a research, Babb and Terry (2013) discussed how the popularity of the platform and the number of devices where the game is being made available influence the sales of the game. The study concluded that games launched on popular platforms like Wii, NDS, Xbox 360, PlayStation result in higher sales as those platforms are accessible to a large user base. Moreover, games that provide cross-platform compatibility and have multiplayer gaming options can benefit from the overall social effect as the popularity of the game increases as more people from the gaming community start playing it.

## 2.6 Best performing models

When it comes to forecasting video game sales, it becomes evident from the existing literature that some machine learning models perform better than others. In an extensive research, Huang (2023) used factors such as release time, ranking, publisher, and game genre and used that to compare three different machine learning models—Neural Networks, XGB, and Light Gradient-Boosting Machine (LightGBM). The research results show that XGB performed the best, followed by LightGBM, whereas Neural Networks fared poorly compared to the other two models.

In another research, applied ANNs (ANNs) to forecast video game sales. The research concluded that prediction accuracy can be significantly improved if ANN model is combined with feature engineering technique such as PCA-based preprocessing. The research further highlighted that ANN, due to their non-linear nature, is the best suitable model taking into consideration the complex relationships present in video game sales data.

Using YouTube trends data, Blomgren (2022) analysed multiple machine learning models including Decision Trees, RFs, Bagging, and Boosting ensemble models to predict initial sales of video games. RF was initially used as the baseline model due to their speed and robustness, but it was outperformed by other ensemble models. The best performing model, with the highest accuracy, was the Boosting ensemble model, which was modelled using Google trends data, demonstrating that Google trends data leads to a better prediction than YouTube trends data. However, the combined data did not always improve the predictive accuracy across all models.

In the study by Pei Pei Chen et al. (2018), deep learning models including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) along with parametric models like the Pareto/Negative-Binomial Distribution model and its extensions were investigated to forecast customer lifetime value of video games. According to the study, CNNs performed better than other models as they are adept at processing huge amounts of sequential data. The study demonstrated that CNNs lead to more accurate predictions as they are better at capturing temporal relationships in sales data.

Eulerich et al. (2023) applied multiple regression models including multiple regression and Quantile regression to analyse the factors that contribute to successful video game sales. The Quantile regression model provided the most detailed insights, particularly in capturing the disproportionate impacts on the variables such as region, console type, game genre, and other factors on the sales of video games.

Despite extensive research on forecasting video game sales, there still remain some significant gaps, particularly in the comparative analysis of econometric and machine learning models within this domain. Existing research has often focussed on single-model approaches or specific techniques without any detailed comparison on multiple models across several regions. Additionally, most studies have not fully explored variations in video game sales forecasting based on locations, which are crucial for developing tailored marketing and promotional strategies. This project aims to address these gaps by applying and comparing multiple models, including LR, RF, Boosting algorithms, and ANNs, to determine the most accurate forecasting techniques. By performing a detailed comparative analysis, this research will provide actionable insights that can enhance forecasting strategies in the video game industry, thereby addressing the shortcomings of previous studies.

# 3. Methodology

The research methodology for this project follows Saunders' Research Onion model, which provides a structured approach to address the various layers of research methodology systematically as shown in Figure 1 below.

A diagram of a diagram

Description automatically generated

Figure 1: Onion Model

## 3.1 Philosophy

This research is grounded in **pragmatism**, which integrates both positivist and interpretivist perspectives to address real-world problems. Pragmatism supports flexibility in research design, allowing for the use of multiple methods to achieve a comprehensive understanding of effective forecasting techniques in the video game industry.

## 3.2 Approach

The study employs a **deductive** **approach**, starting with a review of existing literature on econometric and machine learning models. Hypotheses are formulated based on theoretical frameworks and tested using empirical data, ensuring a logical progression from theory to practice.

## 3.3 Strategy

An **archival research strategy** is used, leveraging existing datasets from Kaggle to compare econometric and machine learning models. This strategy is well-suited for analysing historical sales data and identifying patterns, supporting the study’s goal of evaluating different forecasting methods.

## 3.4 Choice

A **quantitative research design** is adopted, focusing on numerical data and statistical methods to analyse relationships and test hypotheses. This approach aligns with the study’s objective to evaluate multiple forecasting models.

## 3.5 Time Horizon

The research uses a **cross-sectional time horizon**, analysing data collected at a single point in time across various video games. This approach provides a snapshot of the market, facilitating a comparative analysis of how different factors influence sales.

## 3.6 Techniques & Procedure

The techniques and procedures used in this research involve several key steps, from data collection and preprocessing to model development and evaluation. Each step is designed to ensure that the research is conducted systematically and rigorously.

1. **Data Collection and Preprocessing:** The open-source licensed data is sourced from Kaggle, which includes various numerical and categorical predictors related to video game sales. The preprocessing steps include:
   1. For categorical variables, missing values were replaced by a new category “Unknown”. For numerical variables, missing values were not imputed and instead kept as blank as imputations could lead to inaccurate predictions.
   2. Encoding Categorical Variables: Categorical variables were encoded using techniques such as one-hot encoding and Numerical variables were normalized to ensure that they have a consistent scale.
2. **Model Development:** In this study, we developed a range of models to analyse the data and predict outcomes effectively, including both econometric and advanced machine learning approaches such as Linear Regression (LR), Random Forest (RF), XGBoost (XGB), and Artificial Neural Networks (ANN). LR serves as a baseline, while machine learning models capture complex relationships in the data.
3. **Model Training and Validation:** The dataset was split into training and validation sets to evaluate the models. Cross-validation techniques such as k-fold cross-validation are used to ensure that the models generalize well to new data.
4. **Model Evaluation:** The performance of each model is evaluated using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2) as shown in Equations 1, 2, 3, and 4 below.

Equation : Mean Absolute Error (MAE)

Equation : Mean Squared Error (MSE)

Equation : Root Mean Squared Error (RMSE)

Equation : R-squared (R2)

1. **Comparative Analysis:** Models are compared based on performance metrics to identify the most effective forecasting techniques.
2. **Reporting and Interpretation:** Results are documented, interpreted, and recommendations are made for industry stakeholders, providing actionable insights.

## 3.7 Data Understanding and Cleaning

The dataset for this study was sourced from Kaggle and underwent an extensive phase of data understanding and cleaning before any analytical procedures were applied. The dataset comprises 16,719 entries, with each entry representing a unique release of a video game on a particular platform. The dataset includes various variables that capture critical aspects of each game as shown in Table 1:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Type** | **Description** |
| Name | Categorical | The name of the video game. |
| Platform | Categorical | The platform on which the game was released. |
| Year\_of\_Release | Categorical | The year in which the game was released. |
| Genre | Categorical | The genre of the video game. |
| NA\_Sales | Numerical | The sales of the game in North America (Million USD). |
| EU\_Sales | Numerical | The sales of the game in Europe (Million USD). |
| JP\_Sales | Numerical | The sales of the game in Japan (Million USD). |
| Other\_Sales | Numerical | The sales of the game in other regions (Million USD). |
| Global\_Sales | Numerical | The total sales of the game across the world (Million USD). |
| Critic\_Score | Numerical | The average score given to the game by professional critics. |
| Critic\_Count | Numerical | The number of critics who reviewed the game. |
| User\_Score | Numerical | The average score given to the game by users. |
| User\_Count | Numerical | The number of users who reviewed the game. |
| Rating | Categorical | The ESRB rating assigned to the game. |

Table 1: Data Dictionary

The initial examination of the dataset involved assessing its structure and format, ensuring that all features were appropriately categorised, and identifying any inconsistencies or missing values. Missing data were handled through imputation methods or, where appropriate, removal of incomplete entries. To enhance the dataset, a new categorical column, 'Franchise', was derived from the combination of the game name and genre, providing a more granular level of analysis. The year of release was also converted into bins of 5 years to facilitate trend analysis over time. Additionally, both the dependent and independent variables were log-transformed to address the skewness in the original data, making the distribution more normal and improving the linearity between the dependent and independent variables. Correlation analysis was conducted to identify relationships between features, which informed the subsequent model-building process.

## 3.8 Data Exploration

The exploration of the dataset provided critical insights into how video game sales has been impacted by several phenomena across the past few decades.

The analysis of sales trends over time as seen in Figure 2, reveals distinct periods of growth corresponding to the launch of major gaming consoles. Notably, there were significant spikes in global video game sales following the release of popular platforms such as the PlayStation and Xbox series. These trends indicate that the launch of new gaming platforms plays a pivotal role in driving sales across all regions.

A graph of a graph of a graph

Description automatically generated with medium confidence

Figure 2: Impact of Platform launches on Sales Trends (1980-2016)

The contribution of individual franchises to global video game sales was dominated by a select few as shown in Figure 3. Franchises such as "Mario", "Call of Duty", and "Pokémon" emerged as the top contributors, each generating sales exceeding several hundred million dollars globally. This underscores the importance of established franchises in maintaining consistent sales, highlighting a form of brand loyalty among consumers.

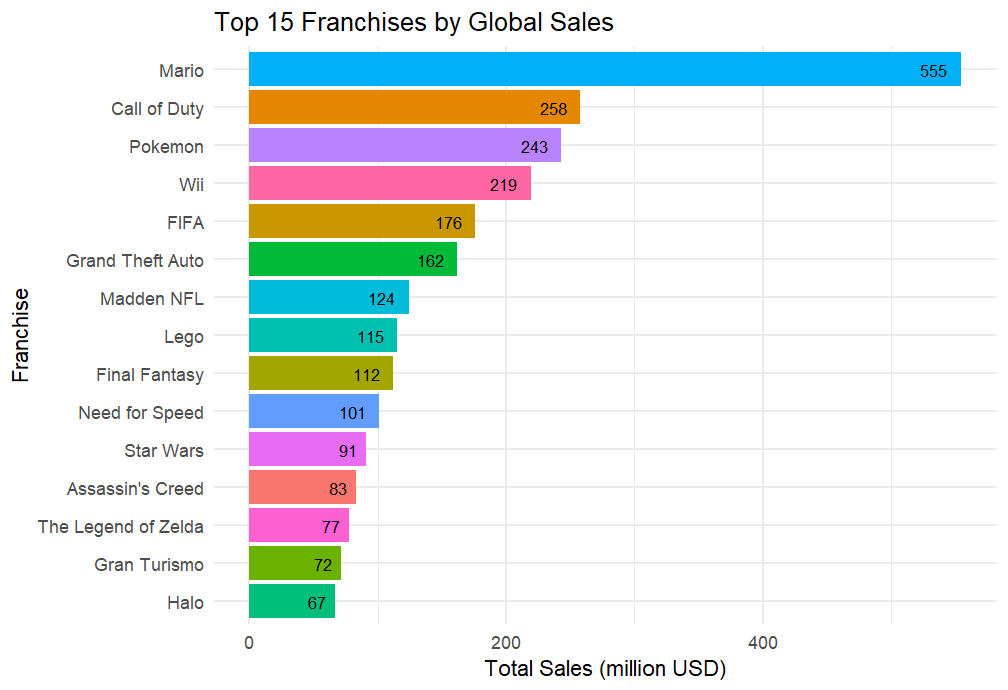


Figure 3: Sales Distribution by Top Franchises (1980-2016)

Sales were broken down by region, including North America, Europe, Japan, and other regions. As we can see in Figure 4, North America and Europe led in sales contributing to 86.5% of global sales, with North America being the standout region contributing 49.3% of global sales, indicating a strong market presence in these regions. Japan, while smaller in total sales, showed strong performance relative to its population, indicating a high per capita consumption of video games.

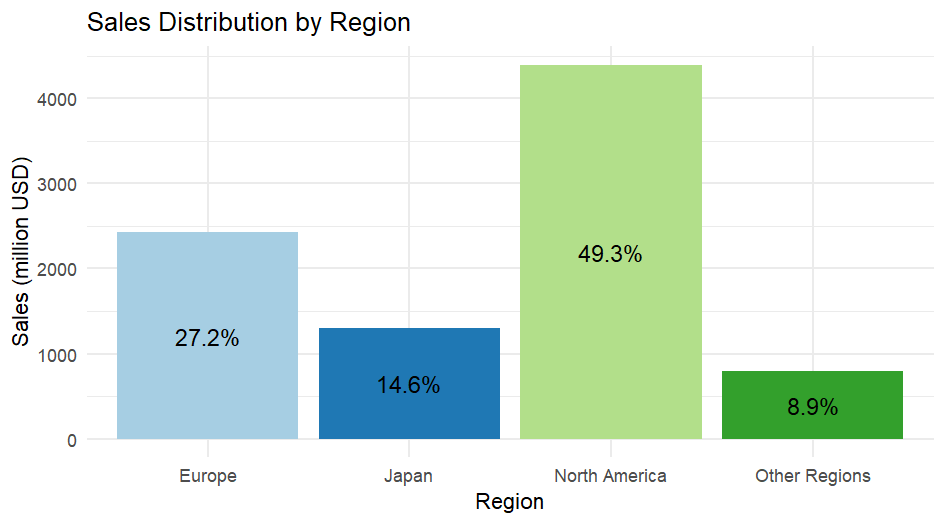


Figure 4: Sales Distribution by Region (1980-2016)

The distribution of overall sales across different platforms in Figure 5 revealed that certain consoles, such as Nintendo, PlayStation, and Xbox dominated the market during their respective eras. The data also showed that in Japan and to some extent North America, Nintendo stands out to be the most popular gaming console whereas PlayStation consoles dominate the European and Rest of the world markets. Across all regions, platforms like PC, Atari, Sega, and others have negligible sales, highlighting the dominance of the major three – Nintendo, PlayStation, and Xbox.

A graph of sales distribution by platform

Description automatically generated

Figure 5: Sales Distribution by Platform (1980-2016)

Analysis of overall sales by genre in Figure 6 showed that action, sports, and shooter games were the most popular globally, with role-playing games (RPGs) also commanding a significant share. This trend was consistent across most regions, though some, like Japan, showed a high preference for RPGs over other genres, reflecting regional differences in gaming preferences.

A graph of sales distribution

Description automatically generated

Figure 6: Sales Distribution by Genre (1980-2016)

The analysis of the total video game sales distribution across different regions reveals distinct preferences and dominance of certain franchises. In North America, "Mario" and "Call of Duty" are the most popular individual franchises contributing to the total sales in that region as shown in Figure 7. Europe shows a similar pattern, but "FIFA" makes a noticeable impact at 14.6%. In Japan, "Pokémon" and "Final Fantasy" stands out at 24.8% and 16% respectively. The rest of the world presents a more balanced distribution across most of the popular franchises. These variations highlight the regional differences in gaming preferences and the varying success of gaming franchises across the globe.

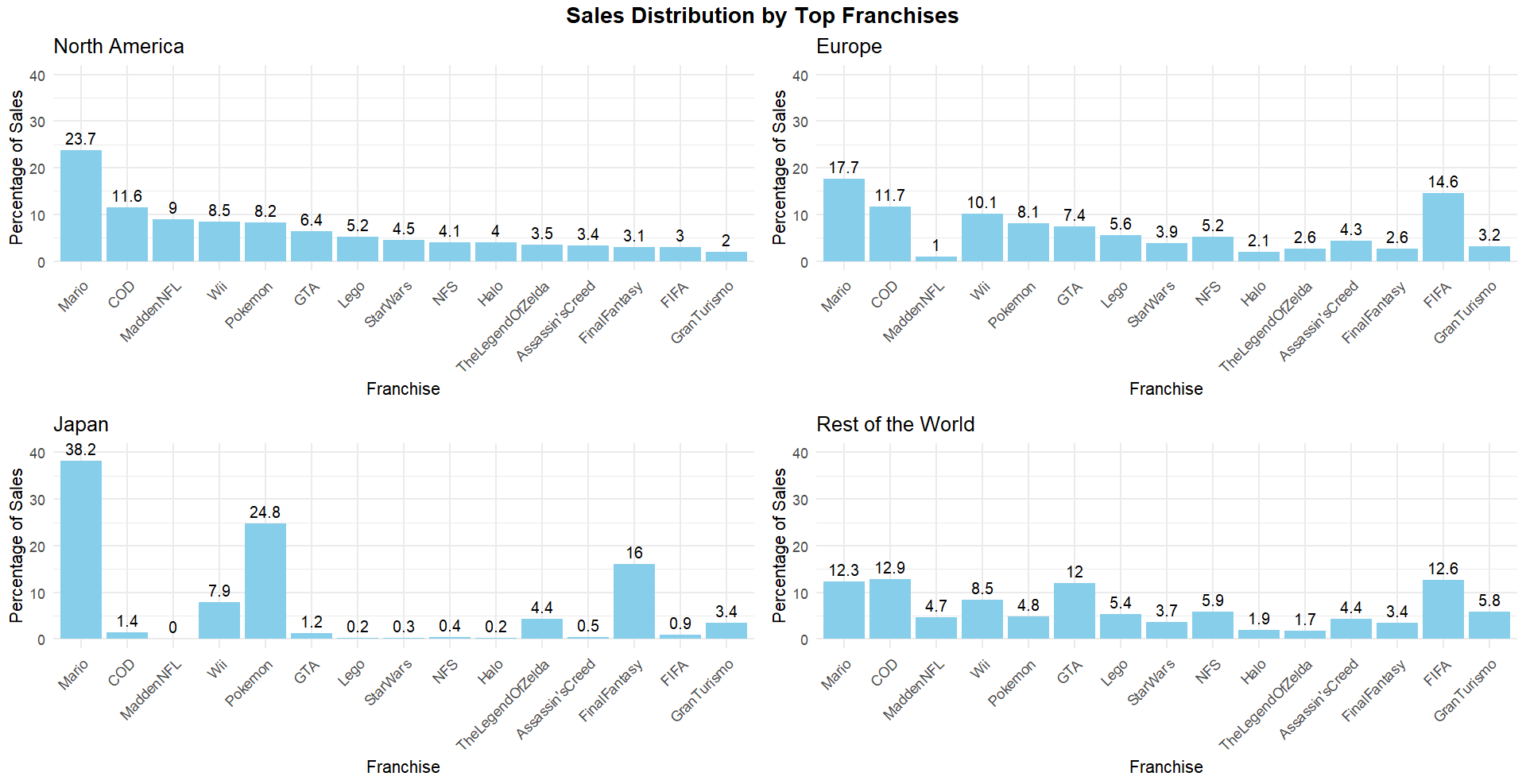


Figure 7: Sales Distribution by Franchise (1980-2016)

The Entertainment Software Rating Board (ESRB) ratings provided insights into the target demographics for video games. We can see from Figure 8 that games rated "E for Everyone" had the highest total sales followed by “T for Teens”, as those contribute to a bigger market along with a preference for family-friendly content. However, titles rated "M for Mature" also showed substantial sales, particularly in western markets, suggesting a good market for mature games.

A graph of sales distribution

Description automatically generated

Figure 8: Sales Distribution by ESRB Rating (1980-2016)

# 4. Results

After thoroughly cleaning and understanding the dataset, the next step involved applying various econometric and machine learning models to forecast video game sales. The models included LR as the econometric model, and RF, XGB, and ANN as the machine learning models. For each of these models, the dataset was split into the same two parts: 80% for training and 20% for testing. This equal split across all models ensures that the models are trained on a substantial portion of the data while still allowing for robust evaluation on an unseen test set.

The LR model was used to understand the relationship between various predictors and the target variable. Key predictors, such as Critic\_Score, User\_Score, and Franchise\_Others, were found to have statistically significant coefficients, indicating their strong influence on global sales. For instance, Critic\_Score and User\_Score had negative coefficients, suggesting an inverse relationship with global sales. The coefficient estimates were statistically tested using t-tests, with values such as -12.655 for Critic\_Score and 39.848 for User\_Count, both being highly significant (p < 0.0001).

The RF model, utilizing 200 trees and an optimal mtry value of 6, was employed to capture complex interactions among predictors. The importance of each predictor was assessed, with the mean decrease in Gini impurity being used as a metric to rank the variables.

XGB was trained using 5-fold cross-validation with a random search approach for hyperparameter tuning, evaluating 20 random combinations. The model's performance was optimized for RMSE.

The ANN model was built with a 52-20-1 architecture, featuring a significant number of weights and biases that adjust to minimize error. This model, utilizing linear output units and decay regularization, demonstrated a complex interplay between the input features and hidden layers.

Figure 9 plots the training residuals for all the models on global video game sales, whereas Figure 10 showcases the residual plot on the test dataset. The residual plots along with a thorough analysis of each of the estimated models will be taken up in the next section.

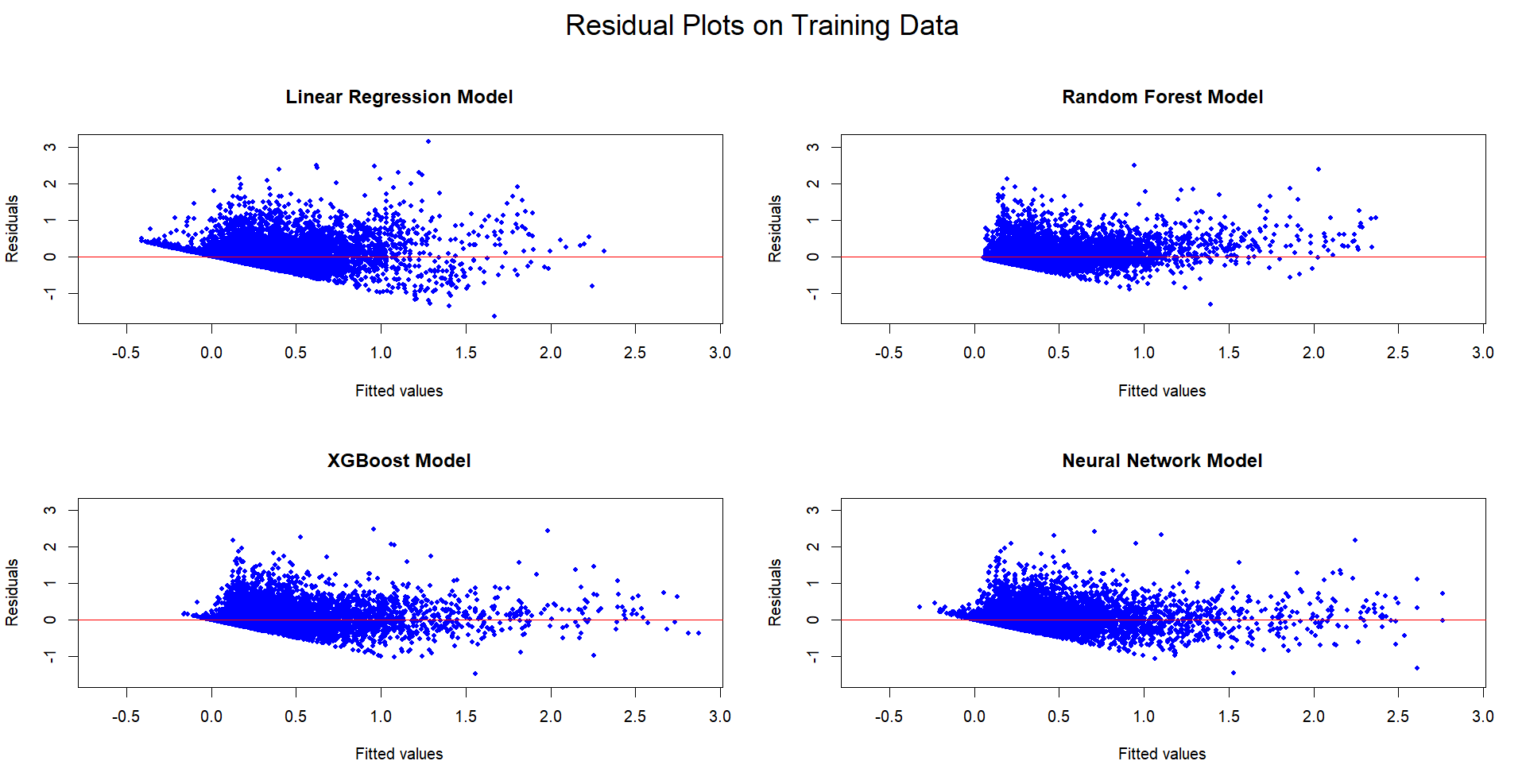


Figure : Training Residuals

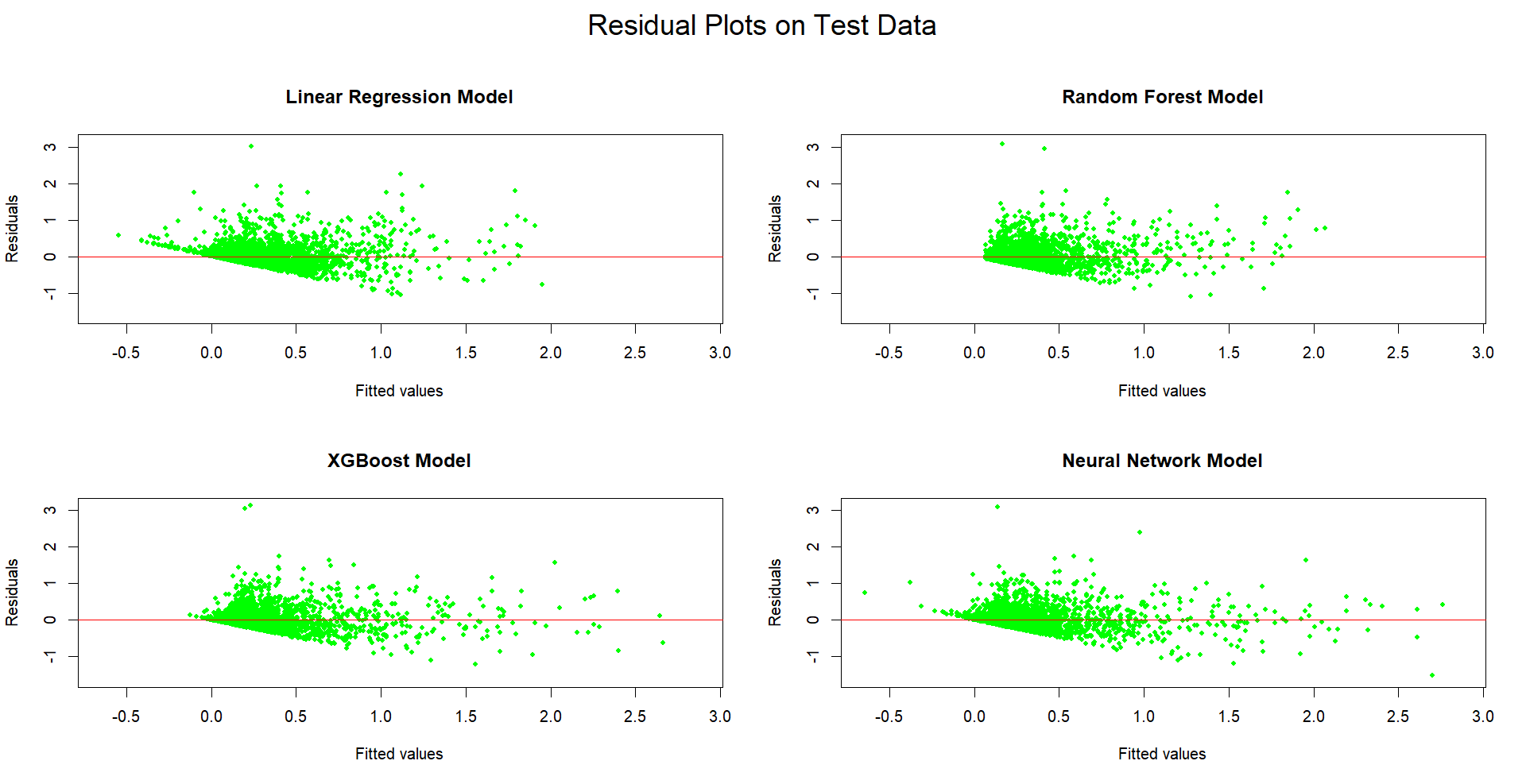


Figure : Test Residuals

Figure 11 shows the impact of each feature on Global Sales whereas Figures 12, 13, 14, and 15 depicts the feature importance of each model across different regions. Feature importance was measured differently for each model. In LR, importance was determined by the absolute value of the coefficients, whereas for RF and ANN, importance was evaluated based on how much each feature contributed to improving the predictive accuracy. Importance of XGB was assessed by how frequently and effectively each feature was used in the decision-making process across all trees.

A graph with red and blue lines

Description automatically generated

Figure : Feature Impact on Global Sales

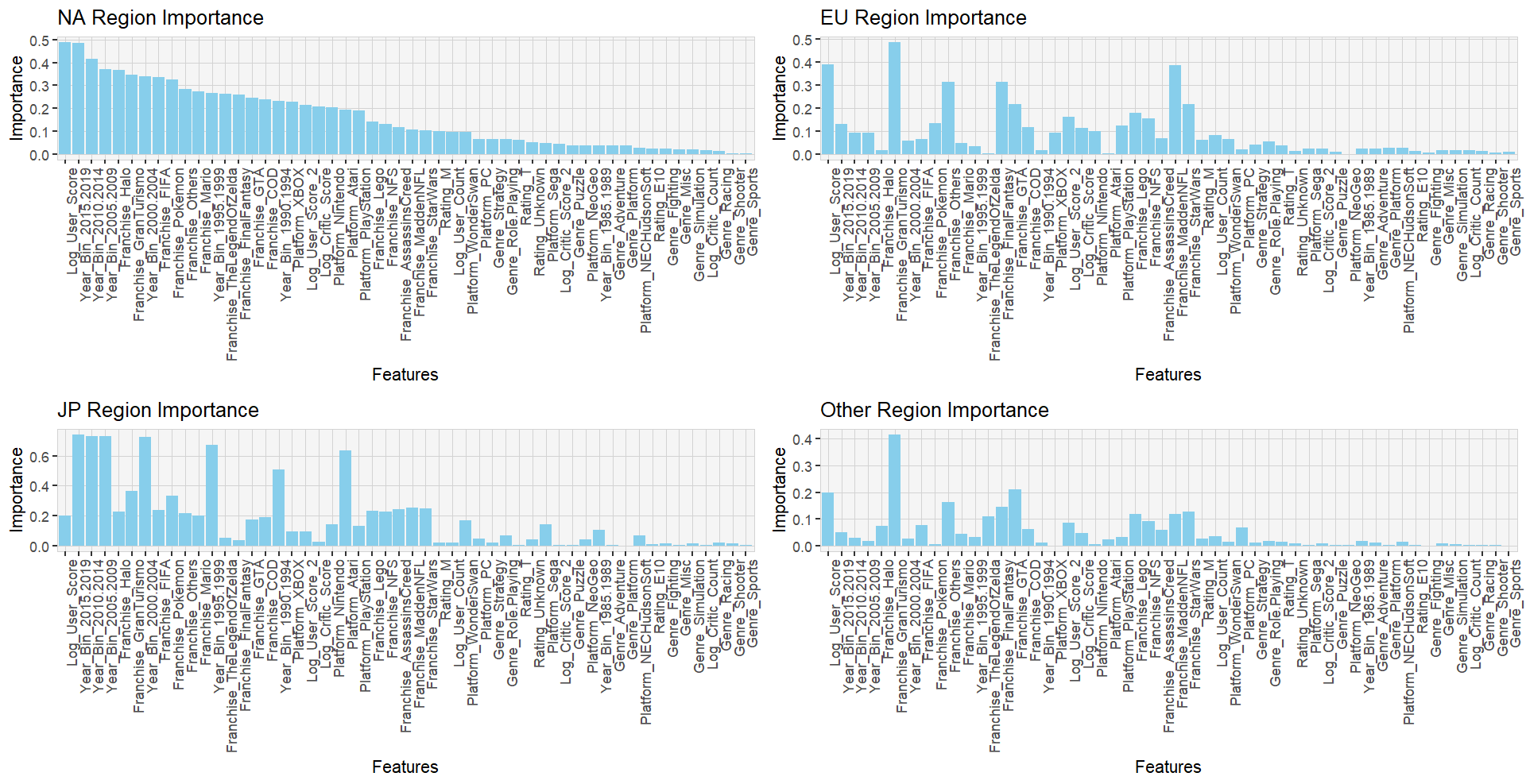


Figure : LR Feature Importance

A graph of different features

Description automatically generated with medium confidence

Figure : RF Feature Importance

A graph of different types of data

Description automatically generated with medium confidence

Figure : XGB Feature Importance

A graph of different features

Description automatically generated with medium confidence

Figure : ANN Feature Importance

A comparative analysis of the performance metrics across all models are shown below in Table 2 along with a graphical representation of the same in Figure 16, and Figure 17.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **TEST** | **R2-Test** | 0.379655 | 0.341199 | 0.385684 | 0.329145 | 0.429909 | 0.499604 | 0.479746 | 0.439944 | 0.493339 | 0.523028 | 0.447826 | 0.472087 | 0.399448 | 0.480654 | 0.534372 | 0.51062 | 0.495079 | 0.434988 | 0.490178 | 0.540286 |
| **RMSE-Test** | 0.2123713 | 0.1671684 | 0.1247372 | 0.0788347 | 0.2889139 | 0.2004403 | 0.151729 | 0.1207877 | 0.0649819 | 0.2727524 | 0.2079496 | 0.1514632 | 0.1239219 | 0.0655454 | 0.2667626 | 0.1957688 | 0.1481281 | 0.1201992 | 0.0649416 | 0.2650632 |
| **MSE-Test** | 0.04510155 | 0.02794529 | 0.01555936 | 0.00621491 | 0.08347123 | 0.04017633 | 0.02302168 | 0.01458967 | 0.00422265 | 0.07439385 | 0.04324304 | 0.0229411 | 0.01535664 | 0.0042962 | 0.07116226 | 0.03832544 | 0.02194193 | 0.01444784 | 0.00421742 | 0.07025849 |
| **MAE-Test** | 0.1353385 | 0.0953979 | 0.0631502 | 0.0397684 | 0.1962559 | 0.1192823 | 0.0837694 | 0.0564858 | 0.029678 | 0.1802432 | 0.1229223 | 0.0838943 | 0.0608537 | 0.0307109 | 0.1736817 | 0.1141663 | 0.0819185 | 0.0564541 | 0.0305058 | 0.1737925 |
| **TRAIN** | **R2-Train** | 0.393159 | 0.361918 | 0.345809 | 0.344994 | 0.435823 | 0.687726 | 0.706803 | 0.593271 | 0.747354 | 0.693193 | 0.494422 | 0.577963 | 0.394182 | 0.564106 | 0.642891 | 0.572735 | 0.552043 | 0.533844 | 0.536972 | 0.586043 |
| **RMSE-Train** | 0.221148 | 0.1710793 | 0.129508 | 0.0820599 | 0.2993833 | 0.1650919 | 0.1232042 | 0.1068369 | 0.0561328 | 0.2290445 | 0.2001099 | 0.1387396 | 0.1244802 | 0.0677385 | 0.236985 | 0.1839598 | 0.1429365 | 0.109193 | 0.069815 | 0.2551513 |
| **MSE-Train** | 0.04890645 | 0.02926812 | 0.01677232 | 0.00673383 | 0.08963037 | 0.02725533 | 0.01517927 | 0.01141411 | 0.0031509 | 0.05246139 | 0.04004398 | 0.01924868 | 0.01549532 | 0.0045885 | 0.05616191 | 0.03384121 | 0.02043084 | 0.01192312 | 0.00487414 | 0.06510217 |
| **MAE-Train** | 0.1384487 | 0.0969371 | 0.0643661 | 0.0398904 | 0.200105 | 0.1009741 | 0.0690189 | 0.0483726 | 0.0254575 | 0.1506035 | 0.1228751 | 0.0795471 | 0.0591801 | 0.0311496 | 0.1565427 | 0.1120342 | 0.0801929 | 0.0517941 | 0.0316918 | 0.1684971 |
| **TARGET** | | NA | EU | JP | Other | Global | NA | EU | JP | Other | Global | NA | EU | JP | Other | Global | NA | EU | JP | Other | Global |
| **MODEL** | | LR | LR | LR | LR | LR | RF | RF | RF | RF | RF | XGB | XGB | XGB | XGB | XGB | ANN | ANN | ANN | ANN | ANN |

Table 2: Model Measures

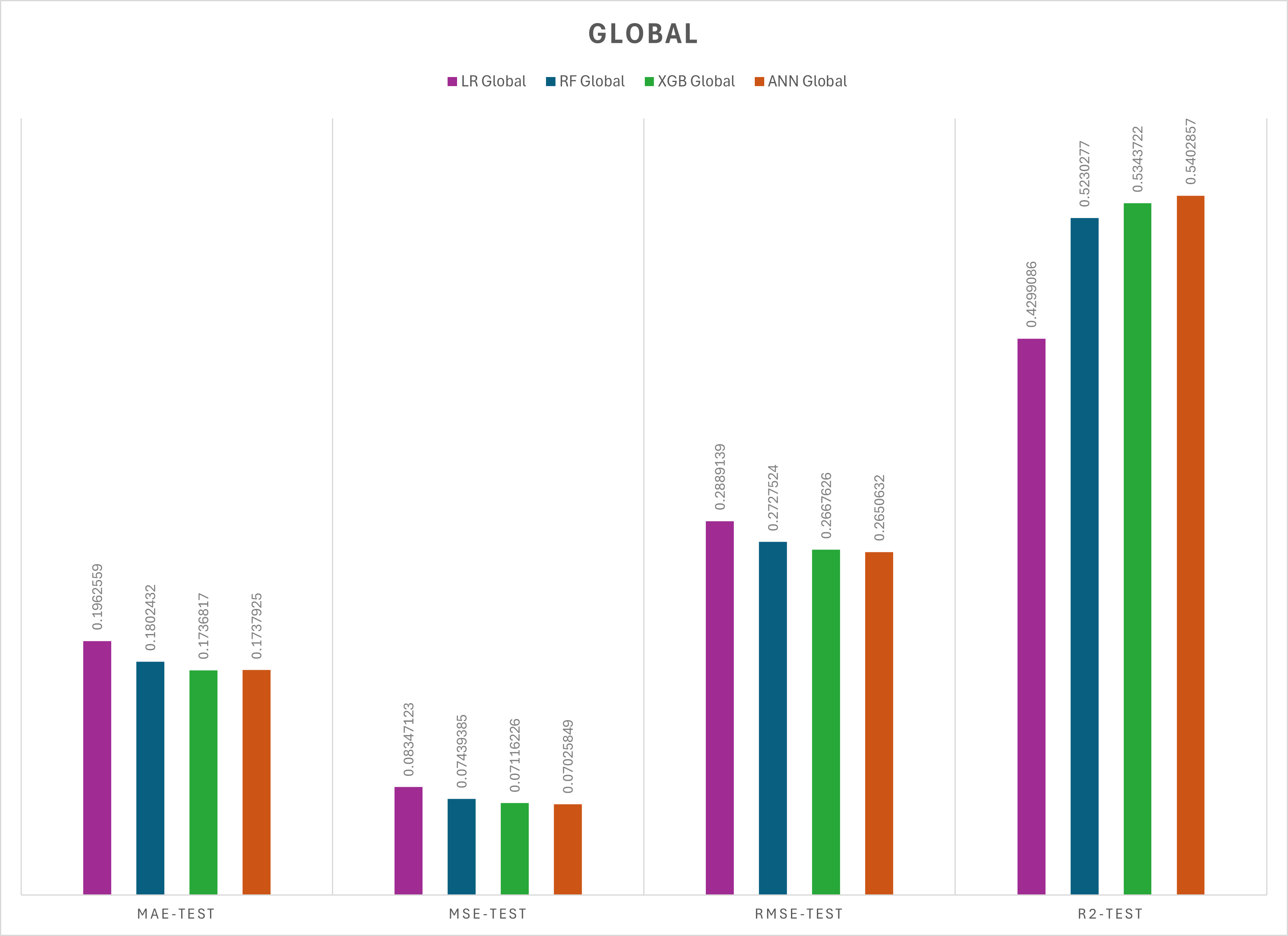


Figure : Global Model Measures

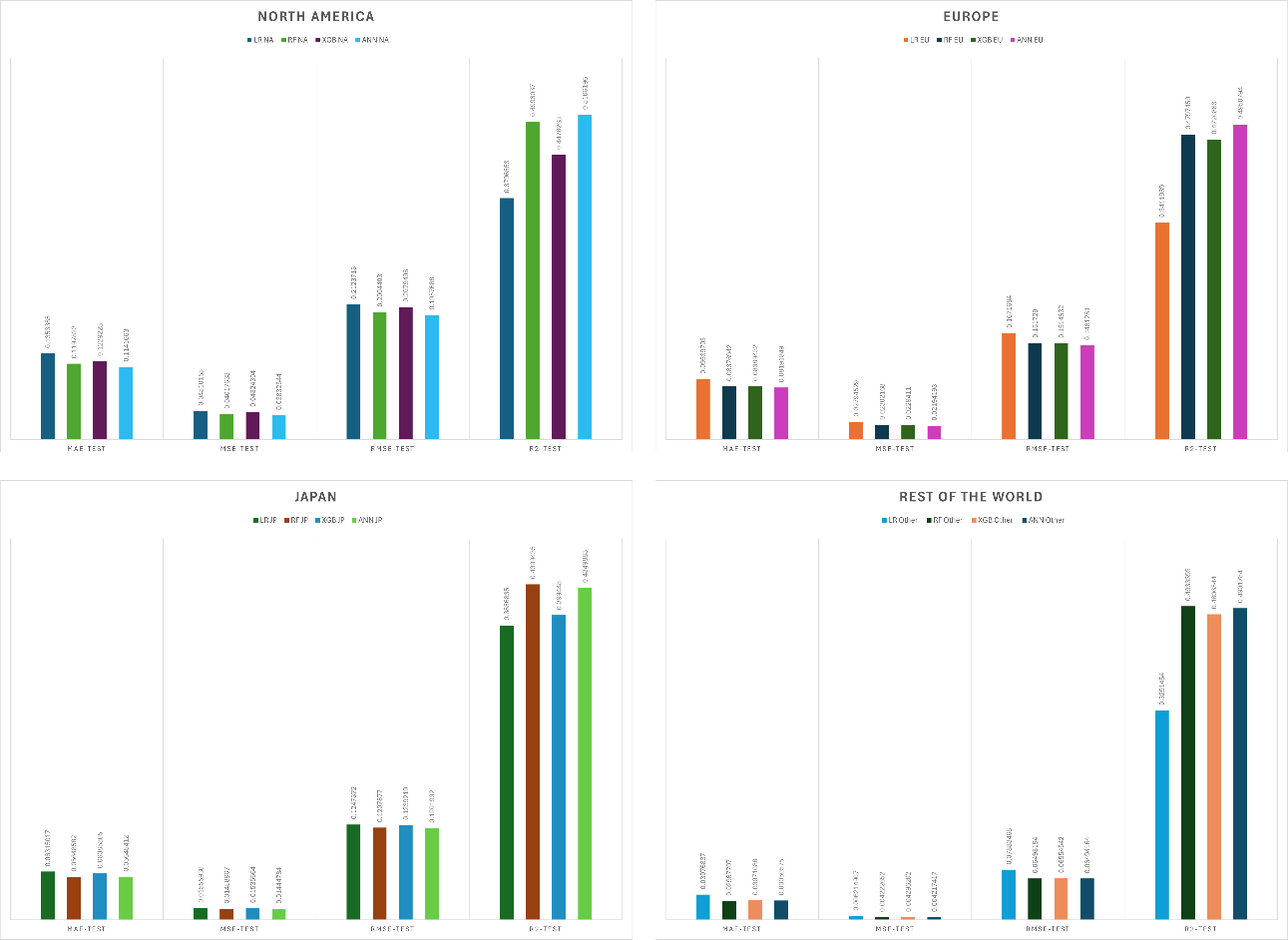


Figure : Regional Model Measures

The next section will delve into a detailed discussion of these results, exploring their implications and comparing them with findings from existing literature.

# 5. Discussion

## 5.1 Residual Analysis

The residual plots for the training dataset offer valuable insights into the performance and generalization of the models. LR exhibits a clear pattern of heteroscedasticity, with residuals spreading out as fitted values increase, indicating the model's struggle to capture variance, especially at higher sales values. This suggests potential underfitting, where LR might perform poorly on unseen data, particularly for extreme values.

The XGB model also shows signs of heteroscedasticity, though less severe than LR. The residuals are more evenly spread but still increase in variance with larger fitted values, implying challenges in predicting higher sales and possible overfitting on the training data.

In contrast, the RF model displays a more consistent distribution of residuals, with less pronounced heteroscedasticity, suggesting a better fit across the entire range of predicted values. This balance between complexity and variance control indicates that RF might generalize better on test data.

The most balanced residual pattern is observed in the ANN model, where residuals are evenly distributed around the zero line with minimal signs of heteroscedasticity. This suggests that ANN has likely captured underlying patterns effectively, indicating a well-fitted model that may generalize well to new data. However, the strong fit observed in ANN and RF raises concerns about potential overfitting, where models perform exceptionally on training data but might not generalize effectively to unseen data.

The residual plots for the test dataset further highlight how well each model generalizes. LR continues to show heteroscedasticity, with persistent funnel-shaped residuals suggesting underfitting and poor predictive performance on both high and low sales values. RF maintains a more uniform residual distribution with less pronounced heteroscedasticity, indicating relatively good generalization. However, some deviations suggest occasional prediction inaccuracies.

XGB exhibits a residual pattern with moderate heteroscedasticity, better than LR but still indicating challenges in higher sales predictions, suggesting a balance between complexity and generalizability. ANN shows a slight indication of heteroscedasticity at lower fitted values, but the residuals are generally evenly distributed, avoiding significant underfitting or overfitting.

Comparing the residual plots across training and test datasets reveals that RF and ANN maintain relatively consistent residual patterns, indicating better generalization. LR's persistent funnel shape highlights its underfitting, while XGB shows some signs of overfitting, particularly at higher fitted values. This comparison underscores the importance of balancing model complexity to ensure robust performance across both datasets.

## 5.2 Model Analysis

LR served as the baseline econometric model for this analysis. The model exhibited moderate predictive power across all regions, with R-squared values ranging from approximately 0.33 to 0.43 on the test datasets and similar values on the training datasets. The close alignment between train and test R² values suggests that LR is relatively stable and does not overfit the data. However, the MAE and RMSE values indicate that while LR captures the general trends, it fails to account for more complex, non-linear relationships in the data. For instance, in the North America region, the test MAE was 0.1353 and the test RMSE was 0.2124, which are relatively higher compared to the other models, reflecting the model's limitations in predictive accuracy.

Feature importance analysis revealed that variables such as User Score, Year of Release, and certain Franchises were heavily weighted, which suggests that LR can identify key drivers of sales. While looking at the regression we noticed that the periods 2015–2019 and 2010–2014 display strong negative coefficients, pointing to a significant downturn in sales. This trend can be linked to several factors, including the enduring impact of the 2008 global economic crisis, which dampened consumer spending and reshaped market behaviours. Additionally, the explosive growth of mobile gaming throughout the 2010s siphoned consumer attention and spending away from traditional gaming platforms. These patterns highlight the critical role of economic conditions and industry transformations, such as the rise of mobile gaming, in shaping the global video game sales landscape.

The RF model, a robust ensemble learning technique, significantly outperformed LR across all regions. The R-squared values on the test data ranged from 0.44 to 0.52, with the highest values seen in the Rest of the World (R² = 0.4933) and Japan (R² = 0.4399). Importantly, the RF model also maintained high R² values on the training data, ranging from 0.59 to 0.75, indicating that it captured a substantial portion of the variance in the training data. However, the slightly lower R² values on the test sets compared to the training sets suggest some degree of overfitting, particularly in regions like Japan where the train R² was 0.5933 but dropped to 0.4399 on the test set. Despite this, RF's ability to handle non-linear relationships and feature interactions contributed to its strong overall performance. For example, in North America, RF had a test MAE of 0.1193 and test RMSE of 0.2004, which were lower than those of LR, indicating better predictive accuracy. Feature importance analysis showed that RF relied heavily on User & Critic Counts and Scores across all regions. Additionally, region-specific features such as the Mario Franchise and particularly Nintendo platforms were highly influential in Japan, showcasing RF's strength in capturing regionally relevant predictors.

XGB, known for its efficiency and high performance, delivered competitive results similar to those of RF. The R-squared values for XGB ranged from 0.40 to 0.53 on the test datasets, with the highest performance observed in the Global region (R² = 0.5344). On the training data, XGB also showed strong R² values, although slightly lower than RF, ranging from 0.39 to 0.64. The lower test R² values compared to the train R² indicate a good balance between model complexity and generalization, with less overfitting than RF. In the North America region, XGB had a test MAE of 0.1229 and test RMSE of 0.2079, closely matching RF's performance but with a slightly higher MAE, indicating comparable predictive power. XGB's feature importance analysis highlighted similar key predictors as RF, such as User & Critic Scores, but with distinct regional trends, particularly in Japan, where platform-specific features were more pronounced. XGB's strong performance in the Global region suggests its robustness in handling diverse datasets with varying patterns.

The ANN model, designed to capture complex, non-linear patterns, produced results that were competitive with both RF and XGB. The R-squared values on the test datasets ranged from approximately 0.43 to 0.54, with the highest values seen in North America (R² = 0.5106) and Europe (R² = 0.4951). On the training data, ANN showed strong R² values (0.53 to 0.59), indicating its effectiveness in fitting the training data while maintaining good generalization to the test sets. In the Rest of the World region, ANN had a test MAE of 0.0305 and test RMSE of 0.0649, which were the lowest among all models, showcasing its superior predictive accuracy in that region. However, the performance was not as consistent across all regions; for instance, in Japan, the ANN model had a slightly lower test R² of 0.4350 compared to RF. Feature importance analysis revealed that ANN captured significant regional variations, with non-top Franchises contributing more to sales in regions like Japan and the Rest of the World, while platforms like Nintendo and Sony PlayStation had a stronger impact in Japan. The years 2005-2009 were particularly influential across regions except North America, indicating the global relevance of platforms like PSP, PS3, Xbox 360, and Wii during that period.

## 5.3 Model Selection

While comparing the feature importance scores across the models, we notice that LR consistently highlights temporal features (e.g., *Year\_Bins*) as the most important whereas, RF and XGB show more varied distributions with *User\_Count* and *Franchise\_Others* having higher importance. ANN also present a broader spread, valuing features like *Franchise\_Others, Platform\_PlayStation*, and *Year\_Bins* but with more emphasis on non-linear relationships. This diversity across models suggests that each algorithm captures and prioritizes different aspects of the data's underlying structure.

With regard to the metrics, ANN emerges as the best overall model due to its balance between training and test performance. ANN consistently demonstrates strong R² values and low error metrics across multiple regions, particularly in North America and Europe. Although RF shows higher R² values on the training data, it tends to overfit more than ANN, as evidenced by the drop in R² on the test data. XGB also performs well, particularly in the Global region, but its slightly lower R² values compared to ANN make it a close contender rather than the top performer.

**Regional Implications**

* North America and Europe: ANN's strong performance in these regions suggests that complex, non-linear relationships and interactions are crucial for predicting video game sales. Companies should consider using ANN-based models for forecasting in these markets to capture these complexities effectively.
* Japan: Although RF performs best in Japan, capturing region-specific trends such as the importance of the Mario Franchise and Nintendo platforms, ANN also performs well, making it a viable alternative depending on the specific focus of the prediction.
* Rest of the World: RF slightly outperforms ANN in this region, indicating that an ensemble approach may be more effective for capturing the diverse factors influencing sales. However, ANN's strong performance also suggests it can be effectively used in tandem with RF for a more comprehensive analysis.

**Practical Implications**

In practice, LR is useful during the early stages of game development for identifying promising genres, platforms, and other factors based on historical data. This helps guide strategic decisions. Once a game is released, ML algorithms like ANN and RF become essential for refining marketing strategies and improving sales forecasts. These models handle complex interactions and provide more accurate predictions, enabling better targeting and optimization of ongoing sales performance.

## 5.4 Answers to Research Questions

Primary Research Question: How do econometric models compare to machine learning techniques in predicting video game sales?

* Econometric models, such as LR, provide a basic and interpretable approach to understanding relationships between features and video game sales. However, my research indicates that LR models struggle with capturing the complexities and non-linearities inherent in the data, leading to moderate predictive accuracy and potential underfitting. While LR is useful for understanding general trends, it fails to account for the intricate interactions between variables that are crucial in accurately forecasting video game sales. In contrast, machine learning techniques like ANNs (ANN), RF, and XGB significantly outperform LR by effectively modelling non-linear relationships and complex interactions. These machine learning models exhibit a higher capacity for generalization across different regions, leading to more accurate and reliable sales predictions. The improved performance of these models, as demonstrated in the model analysis, underscores their superiority over traditional econometric approaches in the context of the video game industry.

Research Question 1: Which specific algorithms demonstrate the highest accuracy in forecasting video game sales across different regions?

* ANN and RF algorithms emerged as the top performers in forecasting video game sales across different regions, with ANN showing a slight edge in overall performance. ANN demonstrated the highest R-squared values in key regions such as North America and Europe, indicating its strong predictive accuracy and ability to generalize well from the training data to the test data. RF, however, performed exceptionally well in Japan and the Rest of the World, capturing region-specific trends and non-linear relationships effectively. XGB also showed competitive performance, particularly in the Global region, but slightly lagged behind ANN and RF in other regions. These findings suggest that while ANN is a highly effective model overall, the choice of the best algorithm may vary depending on the specific region, with RF and XGB also being reliable options depending on the market dynamics and data structure.

Research Question 2: What insights can be gained from the comparative analysis of forecasting models to enhance sales forecasting strategies in the video game industry?

* The study suggests a strategic approach to model selection based on the stage of game development. LR is particularly valuable during the initial phases, helping to identify promising genres, platforms, ESRB ratings, and franchises by leveraging historical data. This guidance is crucial for developers as they align new games with factors associated with higher sales. However, once a game main characteristics are set, machine learning algorithms like ANN, RF, and XGB become essential to forecast sales more precisely. These models offer superior precision and adaptability, capturing complex, non-linear relationships and regional variations. This enhances the accuracy of sales forecasts and supports detailed analysis for refining marketing and pricing strategies, ensuring that companies can effectively respond to market dynamics and maximize their sales potential.

## 5.5 Comparison with Existing Literature

In comparing the results of my analysis with the findings from existing literature, several key insights emerge that are valuable for both business and academic contexts. Huang (2023) highlighted the superior performance of eXtreme Gradient Boosting (XGB) and Light Gradient-Boosting Machine (LightGBM) over Neural Networks in forecasting video game sales. My results align with this finding to some extent, as XGB and RF indeed performed competitively, though the ANN model emerged as the overall top performer across most regions. This discrepancy suggests that while traditional boosting methods are robust, ANN's ability to model complex, non-linear relationships may offer additional advantages in certain contexts.

Marcoux and Sid-Ahmed Selouani (2009) emphasised the benefits of combining ANNs with feature engineering techniques like PCA for improved accuracy. My analysis corroborates this to an extent, as ANN demonstrated high predictive accuracy, though without the explicit use of PCA. This indicates that while ANN can perform well independently, integrating advanced feature engineering could potentially enhance its performance further.

Blomgren (2022) found that boosting ensemble models outperformed RF in predicting initial sales using YouTube trends data. However, in my study, RF outperformed other models in several regions, suggesting that while ensemble methods are effective, their performance may vary depending on the specific dataset and context used. Moreover, my findings indicate that combining multiple data sources or integrating models could yield more comprehensive insights, which aligns with the observation that model performance can be sensitive to data characteristics.

Pei Pei Chen et al. (2018) demonstrated that Convolutional Neural Networks (CNNs) were effective in capturing temporal relationships in video game sales data. While my study did not focus on CNNs, the success of ANN in capturing complex patterns suggests that similar deep learning techniques could be beneficial, particularly in scenarios involving sequential or temporal data.

Eulerich et al. (2023) highlighted the effectiveness of Quantile Regression in understanding the impact of various factors on sales. My analysis, which found RF and ANN to be the most effective models, underscores the importance of capturing diverse factors and interactions in sales predictions. The residual analysis from my study suggests that these models handle variations in the data more effectively than traditional regression approaches, reflecting a broader capability to manage complex datasets.

In summary, while my findings generally support the literature's observations on the efficacy of machine learning models like XGB and RF, the superior performance of ANN in my study suggests that deep learning models may offer additional benefits. This insight highlights the evolving nature of forecasting techniques and underscores the importance of selecting models based on specific data characteristics and forecasting needs. For future research and practical applications, it is crucial to consider the integration of various models and data sources to enhance prediction accuracy and strategic decision-making in the video game industry.

## 5.6 Limitations of the Study

While this research provides valuable insight into forecasting video game sales using econometric and machine learning methods, it does have some limitations. One of the major drawbacks is the reliance on past sales data, which might not account for new trends and changes in consumer behaviour. With consumer preferences changing and technology advancing rapidly, the video game industry has become increasingly dynamic and may affect sales in ways that are distinct from historical trends. As a result, models based on past data might fail to accurately predict future sales in a ever-changing market. For instance, techniques that depend on industry experts’ idea rather than historical data, such as Expert Judgement and the Delphi Method, might be more effective in predicting future market movements and trends.

Another constraint is the extent to which the data was used. This research primarily focused on some major regions with a limited number of franchises and platforms. This might not account for niche markets or developing regions where sales patterns could differ significantly. Furthermore, the research left out external factors such as economic conditions, competition, and marketing techniques, all of which might have a significant impact on sales performance. The predictability of the model may be impacted if such parameters are left out.

Although the models used are robust, they are limited in their capacity to account for every possible features effecting video game sales. For example, while RF and XGB are good at handling non-linear relationships and interactions, they might not be able to capture very subtle or complex patters present in the data. Comparably, although ANN performed better, their dependency on large amounts of training data and computational resources might not always be feasible in practice.

## 5.7 Future Scope of the Study

Future studies might use a number of approaches to overcome these limitations and offer more comprehensive insights. One such possible avenue for improving the models’ predictive accuracy would be to explore the incorporation of other data sources, such as social media sentiment analysis, the impact of influencer marketing, or real-time data on consumer behaviour. Including these dynamic, real-time elements might improve forecasts and provide a more detailed view of sales trends.

Extending the geographic and categorical scope of the analysis might be another area that needs more investigation. A broader region, including developing economies, as well as diverse game genres and platforms, might give a complete picture of global sales trends. This broader scope may also help in identification of region-specific patterns and preferences that are not visible in the existing dataset.

In addition, it may be beneficial to investigate how external variables like changes in competition, marketing strategies, and economic conditions affect sales projections. Models that integrate these external variables might improve their robustness and applicability to real-world circumstances.

For both business and academia, future research should also focus on the creation of hybrid models that incorporate the capabilities of several forecasting methodologies. Integrating machine learning approaches with econometric models, for example, or including advanced feature engineering techniques, may result in more accurate and reliable forecasts. Furthermore, investigating Deep Learning and Artificial Intelligence developments may present new opportunities to enhance forecasting accuracy and identify complex patterns in sales data for the global video game industry.

# 6. Conclusion

This research offers a thorough analysis of forecasting video game sales using both traditional econometric models and advanced machine learning techniques. This paper provided important insights into how well these models predict video game sales by analysing the performance of LR, RF, XGB, and ANN across various regions globally. It also emphasises the implications on the video game industry along with possible future research opportunities.

The results show that although LR is a good baseline model, it is unable to adequately account for the complex non-linearities present in the historical video game sales data. Even though LR is straightforward to use and reliable, its average predictive accuracy and tendency to underfit adds to its shortcomings in this dynamic gaming industry. This discovery is consistent with existing research, which argues that in order to properly understand complex sales patterns, more sophisticated models are required.

On the other hand, machine learning models, particularly ANN, RF, and XGB excel in predicting video game sales. ANN stands out as the top performer, effectively modelling complex non-linear relationships and interactions, especially in North American and European markets. This shows ANN's ability to identify intricate patterns that traditional models might miss. RF also shows strong performance, particularly in Japan and Rest of the World, where it captures regional trends and interactions. XGB competes well at the global level, though it lags slightly behind ANN and RF in some aspects. The performance of these models highlights their flexibility and robustness, making them well-suited for forecasting sales in the global video game industry.

The video game industry is going to be significantly impacted by these findings. Overall, ANN provided the most accurate forecasts indicating that game developers and publishers should explore deep learning techniques to improve their forecasting models. This becomes crucial in a market with rapid technological improvements and ever-changing customer preferences. ANN’s ability to identify complex patterns can lead to better sales forecasts, helping companies make better informed decisions.

RF’s great performance in certain regions like Japan showcases the significance of including region-specific variables and interactions into forecasting techniques. This shows that a customised model selection strategy based on regional characteristics and data availability can enhance predictive accuracy. For markets with different trends and tastes, such as Japan and other regions, models like RF can offer significant insights into local sales dynamics.

Despite XGB’s performance falling short than that of ANN and RF, it shows its ability to process a wide range of datasets. Its efficacy on the global scale showcases its potential for extensive applications across various markets. Combining XGB with other algorithms might lead to a better predictive accuracy, resulting in a more comprehensive forecasting approach.

The report also highlights avenues for further research, such as incorporating new data sources and widening geographic and categorical scopes. Incorporating elements such as social media sentiment and real-time consumer behaviour could improve the models’ predictive accuracy. Extending the study to include new and emerging markets and other game genres can offer a more holistic view of global sales patterns.

In conclusion, this study demonstrates the superiority of advanced machine learning techniques over traditional econometric models in predicting video game sales. Particularly, ANN performs exceptionally well at capturing complex patterns and interactions, wile XGB and RF offer great alternatives with region-specific strengths. The implications for the video game industry are significant, emphasising the necessity for advanced forecasting methods to successfully navigate the dynamic market. Future research should expand on these findings by incorporating other data sources, expanding analytical scopes, and developing hybrid models to improve predictive accuracy and strategic decision-making in the video game industry.