PowerUp: Unlocking Video Game Sales Predictions

Anbar Saleem, aas27
Daniel Sternberg, ds44
Nishant Joshi, nj26
Raphael Roxas, rcr24

Keywords: Video Game Sales, Linear Regression, Machine Learning, Neural Networks, Ensemble Methods, Random Forest, Gradient Boosting, Predictive Modeling, Sales Forecasting, OneHot Encoding

Abstract: This paper presents our findings of our project which was to predict global video game sales using a linear regression model trained on the dataset below. The dataset contains an extensive amount of information on video game sales within America, Europe, Japan, and Other (as specified by the dataset). Through preprocessing and our different approaches to achieve this goal, we establish a strong correlation between regional sales figures and global sales, with the models effectively capturing this relationship.

Dataset: https://www.kaggle.com/datasets/gregorut/videogamesales/data

Links: https://github.com/nishantjoshi-007/PowerUp

I. INTRODUCTION

In this paper, we address the need for accurate sales predictions in the video game industry. While not a huge problem, these predictions can help companies produce accurate budgets for their teams to develop games on time and be able to earn a profit. By creating these three models, we have begun to analyze what forms of machine learning work well for this issue and which forms may fall behind others.

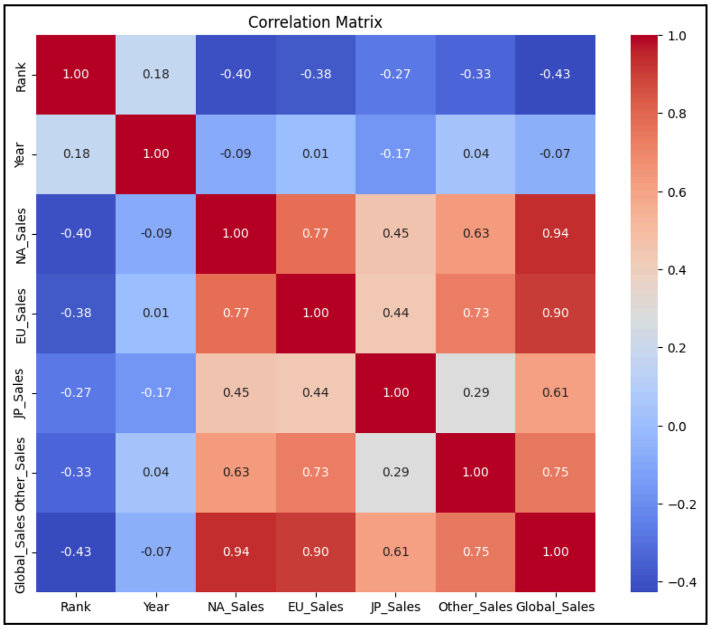
II. RELATED WORKS

Our project goes beyond the linear regression analysis often found on platforms, like Kaggle by using a multifaceted strategy to forecast video game sales. We improve the strength of our models by conducting evaluations employing cross validation to gauge performance across data segments. By incorporating advanced machine learning techniques such as Random Forests and Gradient Boosting into our analysis we are better equipped to capture linear relationships compared to traditional linear methods. Our utilization of OneHotEncoder and MaxAbsScaler ensures that categorical data is accurately transformed and all features are standardized to prevent differences in scale from impacting model performance. This thorough preprocessing process along with our enhancements in managing outliers and missing data establishes a groundwork for our predictive models. Additionally we showcase visualizations to depict the connections within the data providing insights that go beyond regression analyses. Our approach not only improves accuracy but also enhances understanding of the intricate dynamics in global video game sales.

One of the specific models on kaggle we looked into was a very well documented linear regression model program made by Omer Senol. He had his code preprocessed to predict global sales given EU sales. Working off of his various data visualizations included in his notebook, we were able to learn how to properly adjust and add to his work in order to approach our own goal properly.

III. EXPLORATORY DATA ANALYSIS

In the initial phase of our analysis on the video game sales dataset, comprehensive data preprocessing was undertaken to ensure the data quality and usability for modeling purposes through the use of the Pandas Python Data Analysis library. The dataset contained several key features, including game title, platform, release year, genre, publisher, and regional sales figures. We then proceeded to handle any missing values in the original dataset through dropping any of the null values in order to maintain data integrity. To further prune the dataset for values that were necessary for our analysis, we utilized the following correlation matrix heatmap (see Fig. 1) in order to decide what values we maintain throughout the running of our models on the data.



	Rank	Year	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
Rank	1.00	0.18	-0.40	-0.38	-0.27	-0.33	-0.43
Year	0.18	1.00	-0.09	0.01	-0.17	0.04	-0.07
NA_Sales	-0.40	-0.09	1.00	0.77	0.45	0.63	0.94
EU_Sales	-0.38	0.01	0.77	1.00	0.44	0.73	0.90
JP_Sales	-0.27	-0.17	0.45	0.44	1.00	0.29	0.61
Other_Sales	-0.33	0.04	0.63	0.73	0.29	1.00	0.75
Global_Sales	-0.43	-0.07	0.94	0.90	0.61	0.75	1.00

Fig. 1: Heatmap for Correlations

Furthermore, categorical variables such as 'Platform' and 'Genre' underwent One-Hot Encoding to convert these nominal data into a numerical format, allowing them to be processed by our machine learning algorithms. This step was crucial for ensuring that the algorithms could interpret the categorical data accurately. Additionally, the sales figures underwent normalization using the Min-Max scaler, adjusting the data to a standard scale without distorting the differences in the ranges of values.

Finally, the prepared dataset was split into training and testing sets in a 75-25 ratio. This setup provided a robust basis for training the models and subsequently validating their performance on unseen data. These meticulous preprocessing steps were crucial in preparing the dataset for the subsequent modeling phase, ensuring that the data was clean, relevant, and structured appropriately for effective analysis.

IV. METHODS

A. Group 1: Traditional ML Model

In this section we constructed a pipeline that integrated categorical transformation, feature scaling using MaxAbsScaler, and linear regression. This pipeline streamlined the preprocessing and modeling steps, maintaining consistency across both training and testing phases.We fitted this pipeline to the training data, then proceeded to evaluate the model using the testing subset. For evaluation, we used Root Mean Squared Error (RMSE) and the Coefficient of Determination (R^2). Our linear model achieved an incredibly low RMSE of 0.0053 and an R^2 of approximately 0.99999 on the test set, suggesting that it could explain nearly all the variance in global sales based on the input features which are Platform, Genre, NA_Sales, EU_Sales, JP_Sales, and Other_Sales.

To ensure the reliability of our model's performance, we implemented cross-validation using K-Fold splitting. The results showed consistently low RMSE values, reinforcing the model's effectiveness. Despite the extremely promising performance metrics, the exceptionally high R^2 prompted us to consider potential over-simplifications or data leakage issues. This could affect the model's ability to generalize to new data, indicating a need for further scrutiny and possibly external validation.

Our linear regression model proved to be a strong baseline, illustrating potential for relatively straightforward predictions of video game sales from available data.

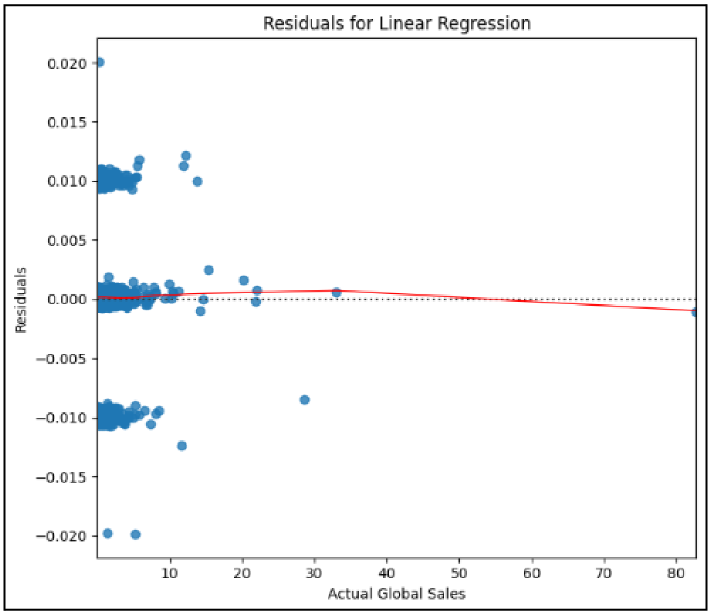


Fig. 2: Residual Plot for Linear Regression

B. Group 2: Neural Network-Based Model

In this section we used a feed forward neural network (FNN) model in an attempt to evaluate its effectiveness in achieving our specific goal. Through numerous tests and optimizations, we determined our model achieved the best results with a three-layer model. We used ReLU activation for the first 2 layers. Additionally we compiled the model using the Adam optimizer

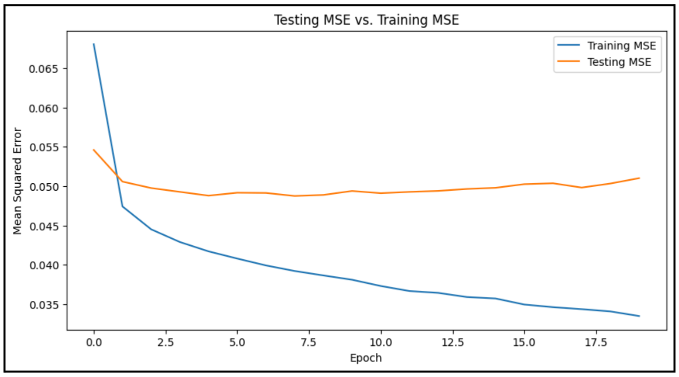


Fig. 3: (FNN) Training vs Testing MSE

	MSE	R-squared
Training	0.03188	
Testing	0.05101	0.2118

Table.1 (FNN) Results

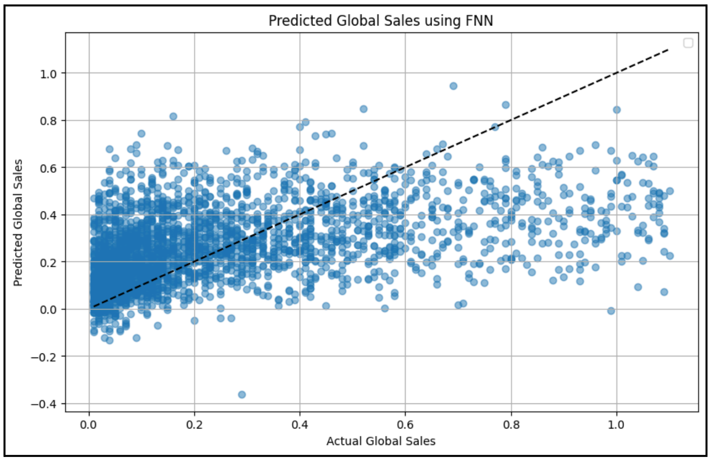


Fig. 4: (FNN) Predicted Global Sales Results

After optimizing and testing, it was determined that 20 epochs yielded the best results for our goal. Anything more than that allowed for the MSE of the training dataset to continue decreasing while the test MSE began to increase or become erratic at times.

C. Group 3: Ensemble Based Models

For our ensemble methods, we chose to include complex models like Random Forest and Gradient boosting, aiming to enhance our understanding and predictive accuracy of global video game sales. Both models are known for their robustness and ability to handle nonlinear relationships, making them good candidates for our dataset which include both numerical sales data and categorical variables such as Platform and Genre.

Random Forest Model: This model leverages the power of multiple decision trees to create a more stable and accurate prediction by averaging the results of individual trees, which reduces the risk of overfitting associated with single decision trees. For our implementation, we built a pipeline similar to that of the linear model, incorporating preprocessing steps like one-hot encoding for categorical variables and feature scaling. We then trained the Random Forest regressor using 100 estimators, optimizing its ability to generalize by introducing randomness in the selection of features and splits in each tree. In terms of performance, the Random Forest model yielded an RMSE of 0.785 on the test data and an R^2 of approximately 0.830, demonstrating substantial predictive power but not as high as the linear regression model.

Gradient Boosting Model: This approach builds on the concept of boosting, where weak learners (typically decision trees) are sequentially improved by focusing on the errors of previous learners. Our Gradient Boosting model was also embedded within a pipeline that ensured consistent data preprocessing. We also used 100 estimators, carefully tuning the interaction depth to manage the trade-off between model and complexity and overfitting. On evaluation, the Gradient Boosting model presented an RMSE of 0.701 and an R^2 of 0.864, indicating effective learning and predictive capability, surpassing that of the Random Forest model but still below the linear model in terms of RMSE and R^2.

For both models, we implemented cross-validation using K-Fold splitting as well. The cross-validation results were encouraging, showing lower RMSE values compared to single-split evaluations, which underscores their robustness and reliability in predictive settings.

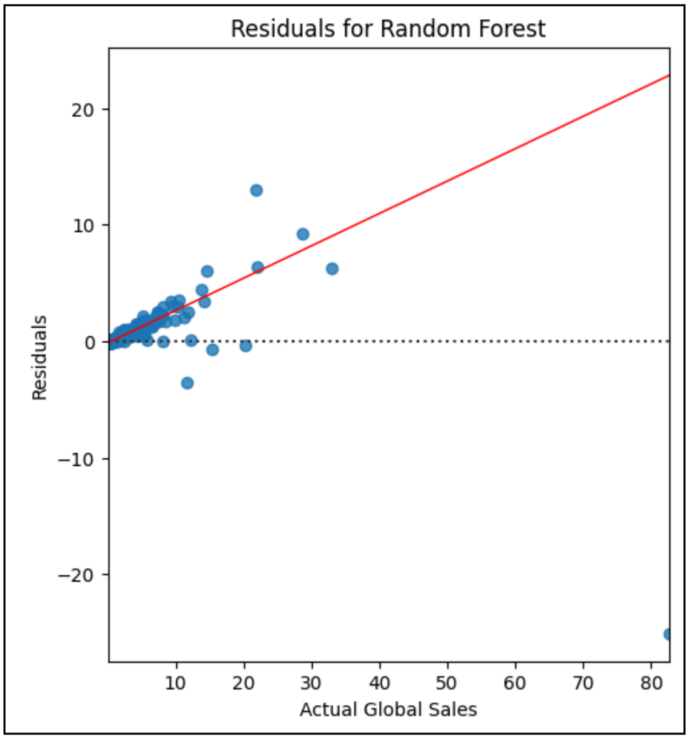


Fig. 5: Residual Plot for Random Forest

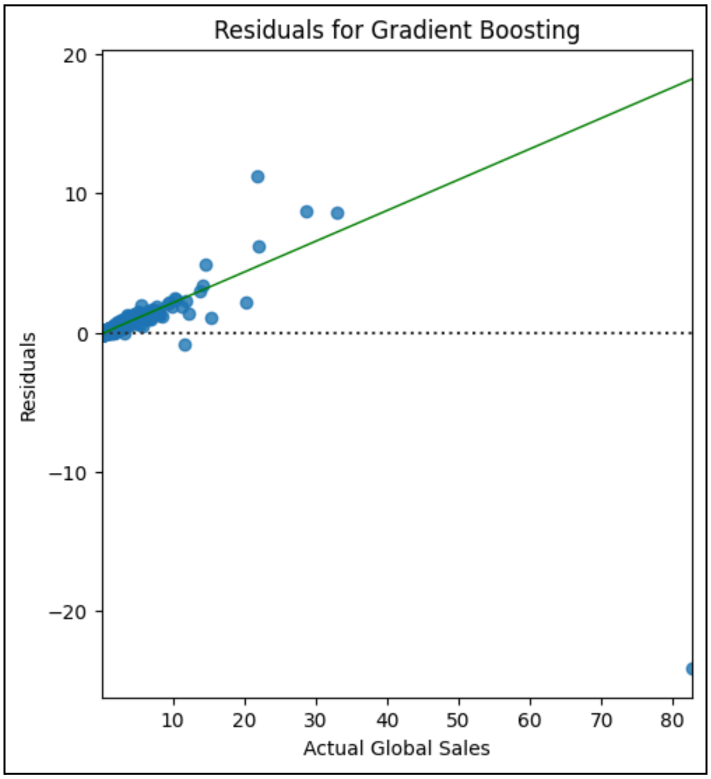


Fig. 6: Residual Plot for Gradient Boosting

V. DISCUSSIONS

Despite all of our optimization the R-squared value for our FNN was lower than we had expected. In comparison to the Random Forest, Gradient Boosting, and Linear Regression models sections there is a significant difference between those R-squared values and that of our FNN. This could be due to the fact that FNN is not a typical solution to this kind of problem, or possibly the amount of epochs ran to minimize overfitting.When first implementing this model, the end results included MSE values that were 100 times more than what we have gathered in later iterations. This was due to our initial preprocessing methods. After implementing a function that takes the interquartile range of all the data points and removes any outliers, our results were in a much more reasonable range.

The Random Forest and Gradient Boosting models offered deeper insights into the data, especially through the examination of feature importance, highlighting the significant drivers of global video game sales and provided valuable complexity and learning dynamics that could be crucial for handling more diverse or complex datasets. Despite that, the Random Forest and Gradient Boosting models did not achieve the near-perfect R^2 of the linear regression mode.

VI. CONCLUSION

In our project we utilized machine learning techniques to forecast video game sales by analyzing a dataset containing detailed regional sales information. Out of the models we experimented with, the linear regression model stood out as the one showing remarkable predictive accuracy, with an R squared value close to 1. This indicates that the linear regression model effectively captured the connection between global sales. However its high accuracy raised concerns about overfitting. To address this issue we incorporated cross validation which confirmed the models performance across data segments and provided reassurance regarding its reliability.

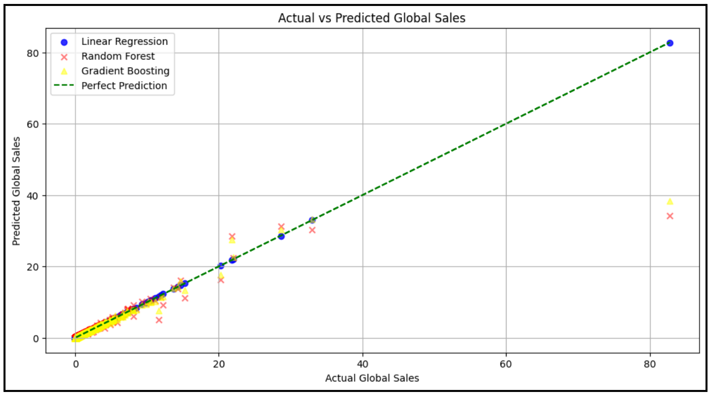


Fig. 7: Actual Vs Predicted Plot for 3 ML Models

Our thorough exploration of data and preprocessing steps such as removing outliers and scaling features played a role in ensuring that our dataset was well prepared for modeling. While ensemble methods, like Random Forest and Gradient Boosting as neural network models offered valuable insights and effectively handled non linear data patterns they did not surpass the linear regression model in terms of overall accuracy. These results highlight the practicality of using the regression model in this scenario. In the future we aim to delve into machine learning methods and fine tune our models to improve their abilities leveraging the solid groundwork laid by the exceptional performance of the linear regression model, in our research.

REFERENCES

[1] Ö. Şenol, "Linear Regression," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/code/omersenol/linear-regression. [Accessed: 02, April, 2024]

[2] A. Saleem, D. Sternberg, N. Joshi, R. Roxas "VGsalesRegression," GitHub repository, 2024. [Online]. Available: https://github.com/nishantjoshi-007/PowerUp. [Accessed: 19, April, 2024].