Capstone\_Final\_Report

**Introduction**

HypotheticalVideo Games (HVG) is looking to create a new game for the new year. They are a video game company looking to recoup the loss of last year’s game and increase revenue and be known as great video game company to the audience. The problem that we are attempting to solve is for HVG to choose a publishing company to publish their game within the next year based on video games’ sales and genres that can help raise revenue by 5% within the next two years and have a sale price above 3% the market within the next year. The data sets that were used are from Kaggle: <https://www.kaggle.com/datasets/gregorut/videogamesales?resource=download>

<https://www.kaggle.com/datasets/xtyscut/video-games-sales-as-at-22-dec-2016csv>

These data sets contain a list of top 10000 video games with their rating, sales, publisher, and genre.

**Data Cleaning and Exploration**

The approach I took to cleaning the data was to first make sure there were no NaN values nor duplicate values either. I dropped any duplicate columns that I deemed unnecessary and then renamed all of the columns to make them readable. The NaN values were mainly in Sales and Year, which I replaced those values with the average from that publisher or I removed the row entirely. Once I did that, I needed to adjust columns User\_Score and Global\_Sales to be float values instead of object64 values. Finally, I also had to remove the ‘TBD’ values in my dataset.

The first thing wanted to check was which was the best-selling platform, I did this by creating a horizontal bar chart and seeing which platform had the best numbers. This happened to be the PS2 with the PS3 nearby, therefore Sony had the best-selling platforms. From there I checked what genre had the greatest number of sales via a bar chart. The genre for this was ‘Action’ followed by ‘Shooter’ and ‘Sports’. When I looked at the overall sales by Publisher, Nintendo had sold far more than any of its competitors. Finally, I wanted to see the correlation of sales by region, so I conducted a correlation heat map, which showed that North America and Europe had a correlation of positive while others were going negative.

**Model Selection and Evaluation**

I explored multiple regression algorithms to predict the target variable, including Linear Regression, Support Vector Regression, Random Forest, Gradient Boost, and Decision Tree Regressor. To evaluate the performance of each model, I calculated the Mean Squared Error (MSE) and R-squared (R2) metrics.

Based on this analysis, the models produced these results:

Linear Regression produced the Mean Squared Error of 0.05304434995923028 and R-squared of 0.9868073650691855, Support Vector Regression produced 2.466909808860618 and 0.3864560402652205, Random Forest produced 0.002039394212546326 and 0.9994927832399337, Gradient Boost produced 0.0019149379576947712 and 0.9995237366956058, and Decision Tree Regressor produced 0.00966420664206642 and 0.997596419784148.

The R-squared scores were similar for Linear Regression, Random Forest, Gradient Boost, and Decision Tree Regressor, while Support Vector Regression had a lower R-squared score. For the cross-validation R-squared scores, Linear Regression and Decision Tree Regressor had positive scores, while the others had negative scores.

Considering the evaluation metrics and the simplicity of the model, we chose Linear Regression as the final model for our problem statement. Linear Regression provides a good balance between model performance and interpretability.

In the next steps, I trained the Linear Regression model on the entire dataset and made predictions for the new data. The final model metrics are as follows:

- Mean Squared Error: 1.8304

- R-squared: 0.093745559842249085

**Model Interpretation**

In this project, we developed a Linear Regression model to predict the target variable based on the given input features. The model's performance was evaluated using Mean Squared Error (MSE) and R-squared. Our model achieved an MSE of 1.8304 and an R-squared of 0.937455, indicating that it is a good fit for the data and can make accurate predictions.

The high R-squared value (0.937455) signifies that approximately 93.75% of the variance in the target variable is explained by the input features. This suggests that the features we selected for our model are indeed relevant and have a strong relationship with the target variable. As a result, our Linear Regression model can make accurate predictions.

Upon analyzing the data and the model's performance, several interesting findings and patterns emerged:

1. **Linear Regression as the best mode**l: One of the most interesting findings was that Linear Regression emerged as the best model for our problem statement. Despite its simplicity, Linear Regression outperformed other, more complex models such as Decision Trees and Gradient Boosting. This can be attributed to its strong performance metrics and consistent results in cross-validation. By choosing the Linear Regression model, we can minimize the likelihood of error and ensure accurate predictions for our target variable.

2. **Significance of Genre**: Our analysis showed that the genre of a video game has a strong influence on its success. Action games, in particular, demonstrated a strong correlation with the target variable, indicating that they are more likely to achieve success in the market. This finding underscores the importance of selecting the right genre for the video game company's upcoming release, as it can significantly impact its performance and revenue.

3. **Trends in the data**: We observed certain trends in the data, such as the increasing popularity of certain platforms (e.g., Sony's PS5) and the dominance of specific regions (e.g., North America and Europe) in terms of market share. These trends provide valuable insights into the gaming industry and can help the video game company make informed decisions about which platforms to target and which markets to prioritize in their marketing and development efforts.

1. **Outdated data**: One significant limitation of our current model is that the data we used for our analysis is not current to 2023. This may affect the accuracy and relevance of our predictions, as the gaming industry is continually evolving, and market dynamics can change rapidly. To address this limitation, the company could consider obtaining more recent data and updating the model to ensure its findings remain relevant and applicable.
2. **Limited contribution of Rating**: In our analysis, we found that the ESRB rating did not contribute significantly to the model's performance. This suggests that other factors, such as genre, platform, and region, have a more substantial impact on a game's success. While rating might still play a role in some contexts or market segments, it may be worth considering whether its inclusion in the model is necessary or if it could be replaced with a more relevant feature.
3. **Potential bias in User Scores**: Another area of concern is the potential for bias in user scores, as these ratings can be influenced by various factors, such as personal preferences, marketing campaigns, or even organized efforts to manipulate scores. This bias could impact the model's predictions and lead to less accurate insights. One potential solution to mitigate this issue would be to obtain user score data from multiple sources or apply techniques to detect and filter out potential bias in the data.

**Recommendations**

Based on the analysis and findings from our Linear Regression model, we propose the following recommendations for the video game company to achieve success in their upcoming game release:

1. **Choose Sony's PS5 as the platform**: Our analysis indicates that games released on the PlayStation 5 platform have a strong potential for success. By choosing this platform, the video game company can leverage the popularity and user base of the PS5 to maximize the reach of their new game.

2**. Develop an Action Game**: The data suggests that action games perform well in the market, and our model indicates a strong relationship between this genre and the target variable. By developing an action game, the company can cater to the preferences of a large audience and increase the chances of their game's success.

3. **Focus on North America and Europe**: Our analysis shows that the North American and European markets are significant contributors to the target variable. By focusing on these markets, the video game company can ensure that their game reaches a wide audience with a high likelihood of interest and engagement.

4. **Adopt the Linear Regression model**: Our Linear Regression model has demonstrated strong predictive capabilities and provides valuable insights into the relationships between the input features and the target variable. By adopting this model and using it to guide decision-making, the company can make informed choices throughout the game development process and optimize their strategies for success.

**Future Work**

For future work, I recommend the following steps to improve the model and further explore the video game industry dynamics:

1. Update the dataset: As the gaming industry is continually evolving, it is essential to keep the dataset up to date. This would involve acquiring more recent data and updating the model accordingly to ensure that the predictions remain relevant and accurate.

2. Explore additional features: There may be other features not included in the current dataset that could impact a game's success, such as marketing budget, game development time, or online features. Incorporating these additional features into the model could provide more comprehensive insights and improve the model's predictive capabilities.

3. Assess other modeling techniques: While Linear Regression has proven effective in our analysis, it may be worth exploring other modeling techniques, such as deep learning or ensemble methods, to further enhance the model's performance and uncover additional insights.

4. Investigate regional preferences: Our analysis showed that North America and Europe are significant markets for video games. It may be valuable to explore regional preferences in more detail, such as the differences in preferred genres, platforms, or game mechanics, to tailor the game development and marketing strategies for these markets.

**Conclusion**

In conclusion, we developed a Linear Regression model to predict the success of a video game based on various input features such as genre, platform, and region. Our model achieved an R-squared of 0.937455 and a Mean Squared Error of 1.8304, indicating its strong performance and predictive capabilities.

Through our analysis, we identified several key factors that contribute to a game's success, such as the choice of platform (Sony's PS5), genre (Action), and market focus (North America and Europe). Based on these findings, we provided recommendations for the video game company to increase their chances of success in their upcoming game release.