DATA 11900 | Group 4
Patricia Ji, Jonathan Lai

Video Game Global Sales Prediction

Overview, Motivation, and Questions

Our dataset, sourced from Kaggle, contains detailed information about over 16,000 video games with more than 100,000 copies sold globally. It originates from a scrape of a database on the website VGChartz (up-to-date through October 2020), which compiles key attributes for each game, such as genre, publisher, platform, and year of release. The dataset includes sales data, in millions of copies sold, both globally and regionally. The regional sales data are broken into major markets like North America, Europe and Japan, providing valuable insights for cross-regional comparisons.

Our project is inspired by the rapidly growing and influential video game industry, which, according to Precedence Research, is projected to hit 583.69 billion USD by 2032 with a compound annual growth rate of 12.9% [1]. Bain & Company also estimates revenue to reach 257 billion USD by 2028 - surpassing other media sectors [2]. These figures highlight a market with tremendous potential and significant investment opportunities. Therefore, understanding video games sales trends is crucial for investors, developers, and marketers to identify emerging consumer demands and avenues for profit. Related work, such as a study by Zawar Ahmed on building predictive models using regression techniques on large video game datasets, demonstrates how these types of analyses can reveal important patterns and help predict future trends in the gaming industry [3]. In our project, we aim to apply a similar approach to explore the potential for predicting game sales and success based on various attributes, providing valuable insights that can help developers make data-driven decisions.

Initially, our analysis aimed to answer questions about the factors driving video game success, specifically, how genre, platform, publisher, and release year might influence market performance. Over time, our focus shifted toward understanding the predictive power of these features and the viability of model-based approaches (linear regression and classification) for predicting sales. Looking ahead, it would be meaningful to explore additional features and experiment different modeling techniques to enhance prediction accuracy.

Exploratory Data Analysis (EDA)

The dataset comprises key categorical features including platform, genre, and publisher, while numerical variables encompass year and na_sales, eu_sales, jp_sales, other_sales, and global_sales, which represent the sales figures across various regions. To understand trends in game releases and as well as both global and regional sales, a bar chart was created to visualize the number of video games released per year [Figure 1. Left] and a line graph was plotted to compare annual video game sales across different regions over time [Figure 1. Right]. The results show a steady increase in the number of games released from 1980 to 2010.. Regional sales follow a similar trend to global sales.

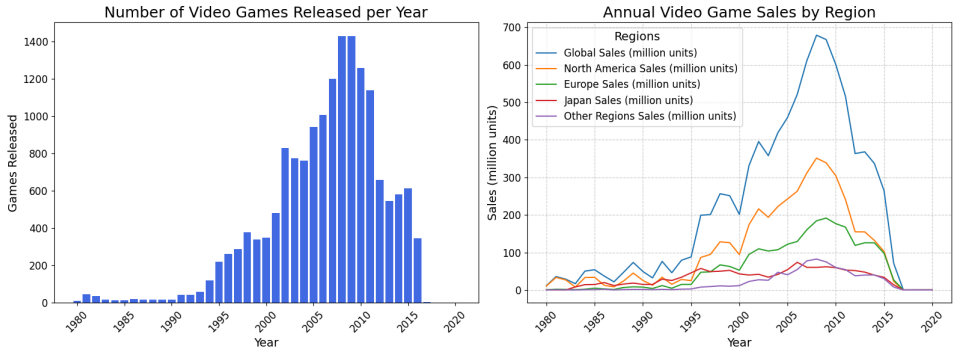

Figure 1. (**Left**) Number of Video Games Released per Year; (**Right**) Annual Video Game Sales by Region

A correlation heatmap was generated to examine relationships between the numerical variables, particularly regional and global sales. A strong correlation was observed between global sales and regional sales, especially in North America (NA) and Europe (EU) [Figure 2. Left]. Due to this collinearity, we decided not to use regional sales data as a predictor for global sales in our modeling. Our primary response variable, global_sales (in millions of copies), exhibits a highly right-skewed distribution [Figure 2. Right], with a mean of 0.54, a median of 0.17, and a variance of 2.46, indicating substantial variability in sales performance across different games. The interquartile range (IQR) is 0.42, with the 25th percentile (Q1) at 0.06 and the 75th percentile (Q3) at 0.48. The maximum value of 82.74 million copies significantly exceeds the median, highlighting the presence of extreme outliers. While this distribution can be explained by a few worldwide hit releases and the majority of games having mediocre popularity, the skewed distribution and the presence of outliers suggest that appropriate transformations or robust modeling techniques may be required to mitigate their impact on predictive performance.

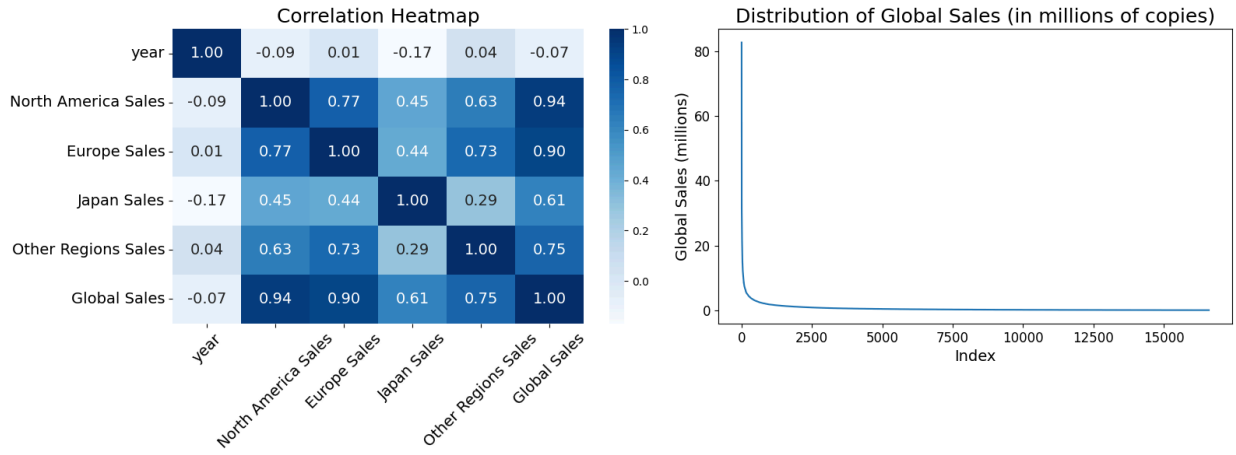

Figure 2. (**Left**) Correlation Heatmap; (**Right**) Distribution of Global Sales

To address the right-skewed distribution and the presence of outliers in global sales, we applied a log transformation to the variable. We used the log1p function (ln(1+x)), which is particularly useful for handling values close to zero, as well as logarithms with bases 2, e, and 10 [Figure 3]. However, the dataset still exhibits significant right-skewness after transformation. This could be due to the extreme high outliers and values close to zero. While log1p helps by adding 1 to all values before taking the logarithm, it does not fully eliminate the skewness, particularly when a large proportion of the data is concentrated near zero, making the transformation less effective.

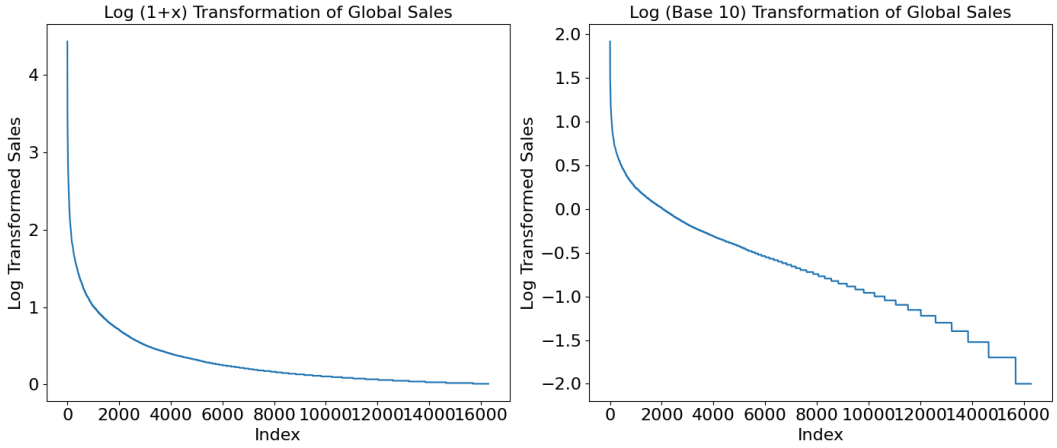

Figure 3. (**Left**) Log1p Transformation of Global Sales; (**Right**) Log10 Transformation of Global Sales

The original dataset contains 576 unique publishers, making it impractical to use them all as one-hot encoded categorical variables. To address this, we encoded the 9 most renowned and successful developers/publishers, grouping the remaining publishers as “Others.” Other relevant categorical features, including platform, genre, and publisher, were one-hot encoded. Regarding missing values, there were 271 missing entries in the "Year" column (1.6% of the dataset) and 58 missing entries in the "Publisher" column (0.34%). These missing values appeared to be missing at random and contained specific information about the respective video games, which could not be reliably imputed. Given their small and negligible proportion of the dataset, they were dropped. Additionally, the columns representing sales ranking and game name were removed, as they were deemed irrelevant for predictive analysis.

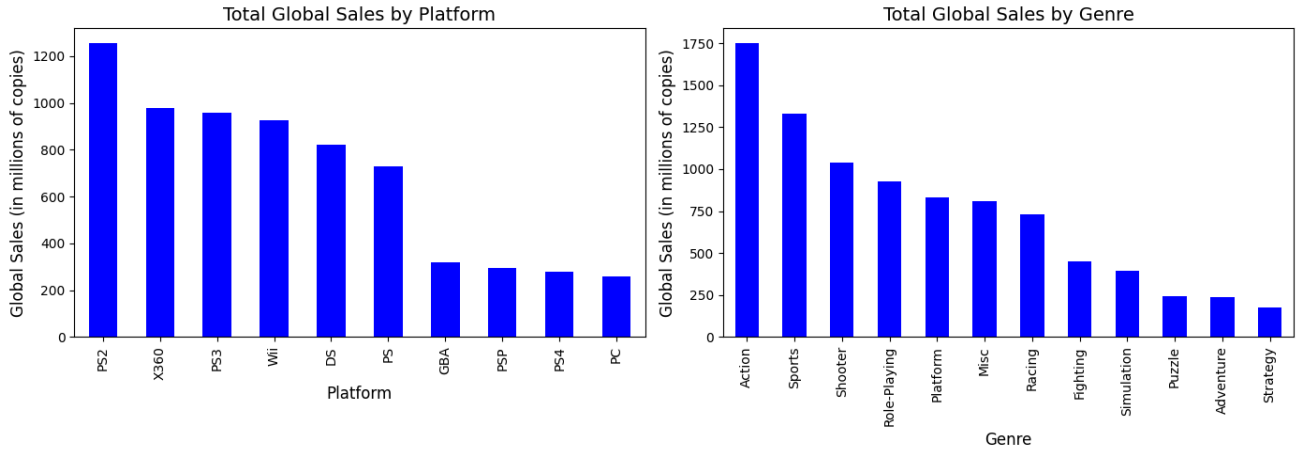

Figure 4. (**Left**) Relationship between Platform and Global Sales; (**Right**) Relationship between Genre and Global Sales

Model Fitting and Parameter Tuning

We fit a baseline multiple linear regression model, as well as Ridge and LASSO regression models, to our data, systematically evaluating model performance based on R^2, Mean Squared Error (MSE), and Mean Absolute Error (MAE). For Ridge and LASSO, we tested various exponents of 10 as alpha values using 5-fold cross-validation to achieve a balanced Bias-Variance tradeoff. Ultimately, our analysis showed that Ridge regression with $\alpha = 0.1$ performed best in terms of minimizing MSE while maintaining reasonable values for R^2 (0.2128) and MAE (0.2321). While both $\alpha = 0.01$ and $\alpha = 0.1$ produced the same cross-validation scores, we chose $\alpha = 0.1$ because it provided a slightly less aggressive regularization, making the model more interpretable and offering a better balance between model complexity and generalization.

Based on the Ridge coefficient graph at $\alpha = 0.1$ [Figure 5], PS3 and PS4 had the highest coefficients among platforms, indicating that they are the strongest predictors of high global sales. This aligns with historical industry trends, as both consoles had extremely successful game libraries and strong global appeal. Among publishers, Nintendo and Electronic Arts (EA) had the highest coefficients, reflecting their dominance in the industry. This is consistent with real-world observations, as Nintendo has produced numerous best-selling franchises, while EA excels with its sports and action game titles. Genre, however, had a relatively weak impact on sales - 8 out of 12 genres showed negative coefficients, with the majority close to 0, especially in comparison to publisher and platform. This suggests that while certain genres may perform well in niche markets, they are not strong predictors of sales at the global scale. Since genre is a broader category that does not account for factors like game quality, franchise recognizability, or developer reputation, it may not be worth revisiting in future analyses unless more granular attributes within each genre are considered. For instance, RPGs and Strategy games may be regionally popular but lack universal appeal. The presence of negative coefficients for some genres suggests that they may be associated with lower commercial performance, either due to their niche nature or lack of mainstream adoption. This highlights the importance of

considering factors such as more specific sub-genres, player demographics, and evolving regional consumer preferences when evaluating the role of genre in forecasting sales.

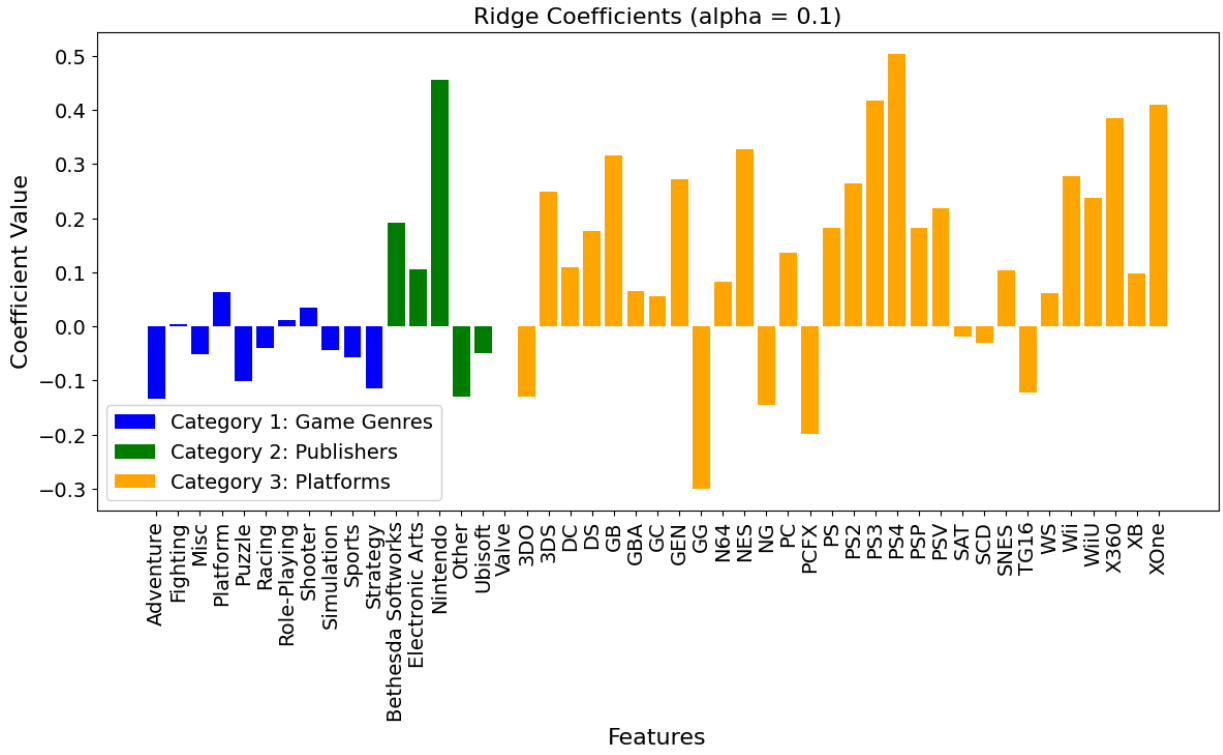

Figure 5. Ridge Coefficients of Genre, Publishers, and Platforms

Both the residual plot and the scatter plot showing actual vs. predicted values [Figure 6] exhibit significant heteroscedasticity, with widely dispersed residuals and non-random clusters forming in both graphs. This suggests that even the best linear model we fit does not fully capture the underlying patterns in the data, indicating potential non-linearity in the relationships between our features and the response variable. This aligns with our earlier discovery of the non-normal distribution of global sales.

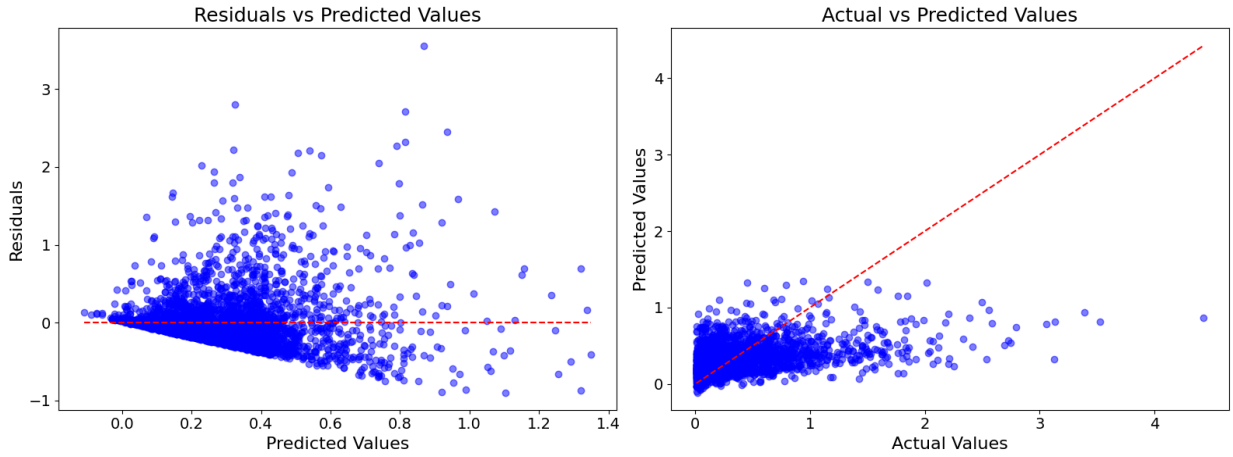

Figure 6. (**Left**) Residual Plots; (**Right**) Actual vs Predicted Values Plot

Classification: Logistic Regression, Random Forest, Neural Network, and Gamma

Given that our linear regression model was a poor fit for predicting global sales due to the non-linearity and heteroscedasticity we observed, we shifted our focus to classification. We defined "high sales" as games with global sales above 0.48 million copies, corresponding to the 75th percentile from our exploratory data analysis. We selected the 75th percentile as the threshold because it is close to 500,000 copies, a natural demarcation for a well-performing game that effectively identifies games that perform significantly better than the majority. The percentile is also a more resistant metric against the extreme outliers in the global sales data. In addition to a baseline logistic regression model, we also fitted a Random Forest and a Neural Network.

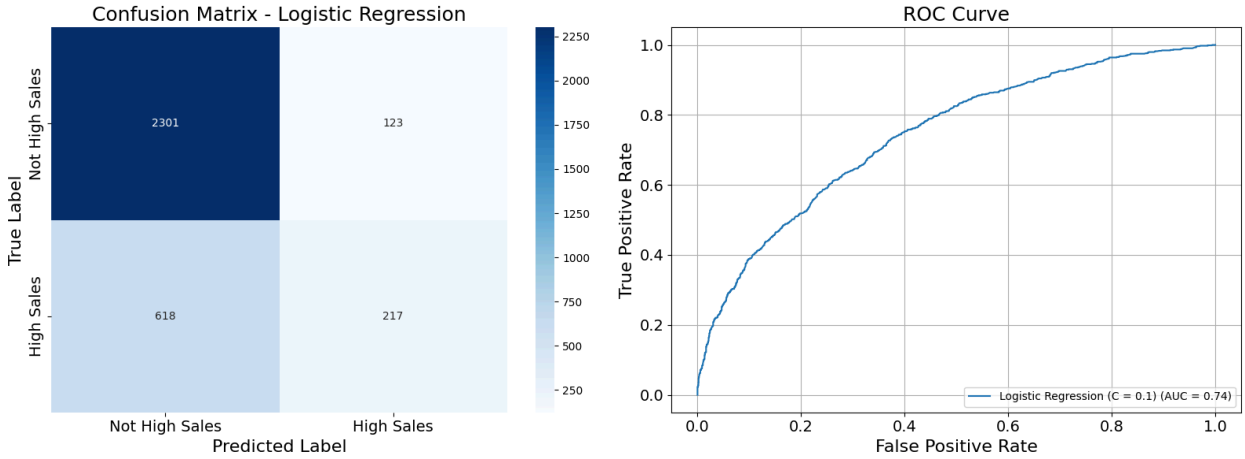

Figure 7. (**Left**) Logistic Regression Confusion Matrix; (**Right**) ROC for Logistic Regression

According to the confusion matrix [Figure 7. Left] and results of the logistic regression, the mean accuracy after 5-fold cross validation was 77.57%. However, this was largely driven by its high recall (95%) for the negative class (not “high-sales”). In contrast, the recall for the positive class (“high-sales”) was only 26%. The high specificity and low sensitivity suggests that the model is heavily biased toward classifying games as not high-selling and struggles with correctly classifying games as high-selling. As observed through the ROC Curve [Figure 7. Right], the AUC = 0.74, which indicates that the logistic regression model is definitely better than randomly guessing but its bias toward predicting the negative class limits is a significant limiting factor.

The Random Forest model achieved similar results, with an accuracy of 77.54%. Again, this was largely driven by the specificity, and the sensitivity of this model (23%) was even lower than that of the logistic regression model. The Neural Network model performed similarly, with an accuracy of 77.23%, but with a sensitivity of 32% - the highest of the three although still quite low. All of this points to a need for improvement, potentially through dataset rebalancing or model adjustment. The low sensitivity observed across all three models may be attributed to the fact that some predictors (e.g. genre, publisher) may not contain enough discriminative power to differentiate high-sellers from low-sellers. This can cause the models to rely on weak signals and fail to generalize effectively. Potential future work includes rebalancing the dataset using oversampling (e.g., SMOTE) or undersampling to ensure that models receive enough examples of high-sales games (which are few by nature) during training. Additionally, introducing more predictive features, such as game ratings, might help improve model performance.

Conclusions and Future Work

Since the distribution of our response variable is highly skewed and has extreme outliers, we used log transformations to attempt to mitigate this issue. To predict sales, we tested multiple linear regression as well as Ridge and LASSO regression, with Ridge regression ($\alpha = 0.1$) displaying the most desirable scores after 5-fold cross validation. The coefficients upon fitting this model revealed that platform and publisher

significantly impact global sales, with PS3, and PS4 leading in platform influence and Nintendo and Electronic Arts dominating publishers, whereas genre has a minimal effect. For classification, we set 0.48 million sales as the “high-sales” threshold (75th percentile) and tested Logistic Regression, Random Forest, and Neural Network models. Logistic Regression achieved 77.57% accuracy but had poor recall (26%) for high-sales games, which persisted across Random Forest (23%) and Neural Network (32%). This suggests that existing features lack sufficient predictive power for high-sales classification, likely due to class imbalance and the lack of more critical determining features like marketing impact or player engagement.

In terms of future work, the priority lies in addressing the skew in our response variable - global sales. Given the limited effectiveness of the log transformation, the Gamma regression model can be explored as a solution. Although we tried to use this model, we received uninterpretable results due to time constraints and our unfamiliarity with this method. Additionally, working with larger datasets with more quantitative features, such as user ratings and other player engagement metrics, could be key to improving predictive models. The sensitivity of our classification models could also be improved through oversampling (SMOTE). Finally, detailed cross-regional analysis could provide insight into how features and coefficients might behave differently in different regions, which could uncover a more nuanced understanding of global sales makeup.

References

[1] Zoting, Shivani. “Video Games Market Size to Hit USD 721.77 Billion by 2034.” Precedence Research, January 31, 2025. https://www.precedenceresearch.com/video-game-market.

[2] Bain & Company, "Global Video Game Revenue to Reach $257 Billion by 2028, Outpacing Combined Revenues of Other Media Types, Bain & Company," PR Newswire, August 28, 2024, https://www.prnewswire.com/news-releases/global-video-game-revenue-to-reach-257-billion-by-2028-outpacing-combined-revenues-of-other-media-types-finds-bain--company-302232174.html.

[3] Zawar Ahmed, "Learning About Video Games Through Regression," Medium, September 6, 2020, https://zawar-ahmed.medium.com/learning-about-video-games-through-regression-c83bc6bbd228

Group Member Contribution

Patricia Ji (EDA and Data Cleaning, Coefficient Interpretations for Linear Models, Neural Network and Random Forest)

Jonathan Lai (Project Overview, Preprocessing + Fitting Linear Models, CV for Ridge and LASSO, Logistic Regression, Future Work)