[Ricky Gajeski](mailto:ricky.gajeski@valpo.edu)

Cheyann Odle

Data 151-A

Prof. Beagley

Video Game Sales

A classification model is used to try and predict whether a video game’s global sales will be “High” or “Low”. Google Collaboratory was used to create and produce four classification models for the data analysis. The data set is cleaned up by removing values with missing entry points, removing unnecessary columns for regional sales, and including a High/Low column to mark high and low selling video games. The high/low values were chosen by stating anything higher than the global sales median is High and equal to or below is Low. The data was then visualized in the form of bar graphs to get a better idea of how our results and predictions may look. From the preliminary data exploration some things found were Wii Sports sold the highest units and the year 2006 sold the most games. The variables chosen to predict High/Low sales from the video game sales data were platform, publisher, genre, and year. The variable names were encoded so the categorical data could be inputted into the models. From there the data was split into train and test, and four regression models were created in order to test and compare each of their results (Bayes, Decision Tree, Random Forest, and AdaBoost). From the models an array was predicted determining high or low. Once all the models were completed, it was found that the Random Forest model was the best model for predicting video games sales due to having the highest F1-score. It was also found that video game release year was the best predictor of a video game’s sale. The classification model was chosen to provide a general indicator that companies or game designers could use in determining how their game will do on market. In continuing or related studies include primary video game developer on the game dataset as a variable.