Machine Learning Analysis

For the machine learning portion of this project, I attempted multiple different models until I found one that worked best on the natural language processing DataFrame. After trying out Lasso Regression, Support Vector Machines and more, Random Forest turned out to be the best overall. Since I was going to use that same model on each of the major types of sales (Global, North American, European and Japanese), one of the important preprocessing steps was removing the other types of sales when setting up the variables because they would have too big of an impact and take away from the importance of variables whose impact I wanted to observe. To dig deeper into this model, I also split each type of sales by their median and ran the model over the values equal to or above it. This would help when dealing with the different proportion of sales in each area. Lastly, new dataset from the University of Portsmouth with U.S. Sales info from 2004 to 2010 would be put through the same model (including sales equal to or above the median) to further test its accuracy as well as find out what other features could be important.

With each use of the model, the "X" value was set to most of the features except for things like types of sale, name, year of release and the categorical variables that either had been, or would be, turned into binary columns. The "y" value was set to whichever type of sale the model was meant to predict on. When it came to testing above the median, the data was filtered before the variables were created. Also, the data was split into training and testing sets with the test size being 20%. Each model used GridSearchCV with the parameters always having n_estimators from 10 to 100. When it came to parameters for the Random Forest Regressor, the random state was always set to 42 however, in certain situations, extra parameters were added to improve R^2 including min_samples_leaf of 3 and max_leaf_nodes of 1,000. For metrics in each, I got R^2 from the testing data for "X" and "y" and the Root Mean Squared Error by the way of the square root of the Mean Squared Error of the testing data of "y" as well as GridSearchCV predicting on the testing data of "X."

When it came to Global Sales, I received an R^2 of about 0.43 and a Root Mean Squared Error around 2.72. While not too high, this would prove to be one of the better uses of his model especially since it used the extra parameters. For features, the most important ones were Critic and User Scores plus User Count as well as words like "super," "mario" and "mw" (likely referring to "Modern Warfare" from "Call of Duty"). The lesser features tended to be different Developers and also certain consoles. Yet when it came to the biggest difference between test and prediction, User Count tended to have the biggest impact with one score under 40 and another over 800. After the data was split above the median, R^2 actually went up to around 0.45 and the RMSE raised to around 2.85. Because this is a smaller subset of the data, the model actually did better without the extra parameters. The feature importance was very similar with User Count and Critic Score leading the way but also included the word "kart," like in "Mario Kart," higher than before. And while Developers were still lower, certain ones like RockStar North and Polyphony Digital were fairly high. Also, once again, User Count had the biggest impact on the extreme differences between test and prediction.

For North American Sales, R^2 was also around 0.43 but had a lower RMSE, around 1.33 and this was using the extra parameters like for overall Global Sales. The important features were essentially the same as for Global Sales except for the rise in the word "madden."

This helps supplement all of the exploratory data analysis from earlier that showed the correlation between Global and North American Sales. And yet again, the extremes in User Count help explain the differences between test and prediction. Yet, when North American Sales were split above the median, R^2 dropped drastically to around 0.16 with an RMSE around 1.51. Done without the extra parameters, the reason R^2 may have decreased so much is because of how small the subset became. Much like the previous models though, the same features were near the top and bottom of the spectrum though the word “halo” was higher this time around. Extreme User Counts also had the biggest impact on difference between test and prediction especially “Grand Theft Auto V,” which had over 3,000 users. Between the first two types of sales, it is becoming apparent how important Critic Count and User Score are to sales as well as the words of certain major games like “Mario” and “Call of Duty.”

Things started to change when the model was used on European Sales. First off, the model for overall sales did better without the extra parameters with an R^2 around 0.38 and a Root Mean Squared Error around 0.93. This makes sense as it did have a very strong positive correlation with Global Sales but not quite as strong as North American Sales, which may be why R^2 was lower. Feature importance also differed a little bit with PC games as well as the words “fifa” and “engin” being higher though User Count still led the way, which still had the biggest impact on difference between test and prediction. After subsetting European Sales by the median, R^2 and the RMSE both dipped to around 0.33 and 0.85 respectively as well as did better with extra parameters. The most important feature ended up being word “kart” with the standard group of highly impactful features soon after. The reason “kart” may have become so important is because of how big Racing games are in the area including “Mario Kart.” And unlike the previous models, there was only one game with the extreme difference between test and prediction, “Blur.” The reason may be because of how frequently the words “mario” and “kart” were used in its reviews.

To wrap up this dataset, Japanese Sales saw the biggest difference when a Random Forest model was used on it. Without the extra parameters, it received an R^2 around 0.37 and a RMSE around 0.46. But while the word “super” was the most important feature, Japanese Sales was the only one that had Publishers or Developers very high with Nintendo and SquareSoft having big impacts. For the other models, both types tended to be near the bottom of the spectrum. Much like what was shown in the plots from the exploratory data analysis, Nintendo is very popular in Japan. And this was further exemplified in the difference between test and prediction as the common trait for those extreme games was that Nintendo or SquareSoft did not make them. But this model gets even stranger when split above the median; everything stayed the exact same. And not just R^2 and the Root Mean Squared Error but feature importance and the difference between test and prediction did not change at all. This may be due to Japanese Sales being lower than the other types so the median could be close to, if not exactly zero so nothing really changed with the model. Either way, Japanese Sales definitely tells a different story when run through the Random Forest model.

After using this model on the four major types of sales, it is interesting how R^2 overall got weaker in the different regions, signifying weaker models, while the RMSE also got lower, signifying stronger models. Also, despite what I may have thought in the exploratory data analysis, User Count and Critic Score had big impacts on these models. For the natural language processing, some of the words of major franchises tended to have big impacts including “Mario,” “Call of Duty,” “FIFA” and “Halo.” And except for Japan, the companies that made the games were not that important, nor were the genres and consoles. While this data was not used in the exploratory data analysis, it does say a lot about what features are truly important in the sales of video games when used with a predictive model.

To finish off this section, I used the exact same model and process on the data from University of Portsmouth on U.S. Sales from 2004 to 2010. Although this data does not have any natural language processing, it does include binary columns of when in the year a game was released. When the Random Forest model was used (with the extra parameters), it received an R^2 around 0.86 with a Root Mean Squared Error around 0.41; this is a very strong model especially compared to the previous ones. The features that were most important were the time of year the games came out especially Block 4, the holidays. This was further evident in the difference between test and prediction when none of the extreme games came out in any of the major blocks of the year. When the data was filtered around the median, both R^2 and RMSE increased to around 0.92 and 0.42 respectively with a very similar spectrum of feature importance. Although Block 4 was still the most important (by a lot), some of the others changed position including Block 2 and Block 0.5, which dropped, while Block 1 increased. Along with the drop in the importance of the Nintendo Wii console and its possible that many of the games below the median came out in Blocks 2 and 0.5 as well as were played on the Wii. And once again, the blocks of the year had the biggest impact on the difference between test and prediction.

What this new data shows is how important the time of the year is for the sale of video games. Holiday time is as important as expected and, in general, these blocks of the year have a bigger impact than any genre or console. Couple this with the earlier dataset and it becomes more apparent about what is critical for video game sales. Name recognition for major series is important along with the amount of users who rate a game. Things like ESRB rating as well as Publishers and Developers are not as key (except in Japan). And while some games are hindered by rushed deadlines, when a game comes out during the year is very important when it comes to successful sales. Instead of focusing on the hot type of game, like an online Shooter, or a specific console, companies should look more at the most popular brands in the medium.