Fundamentals of Data Science

Project Proposal 1

Tanner Patrom

My first proposal would be to predict the sales of a video game based on its genre, publisher, or platform. The dataset can be found at <https://www.kaggle.com/datasets/gregorut/videogamesales>. It contains information from video games that sold more than 100,000 copies across North America, Europe, Japan, and other countries. This would be a regression problem and I could use a supervised learning model, as we have prior knowledge that we are using to predict the future sales of a video game. I could start by trying to predict the number of sales for a video game in North America and then move to predict the sales in Europe, Japan, and globally to get a better idea of the different tastes a population may have to a genre or affinity to a particular publisher. Some of the models I could use are linear regression, decision trees, and random forests. A linear regression may tell me if there is a linear relationship between certain metrics provided. I would like to explore what metrics might have a strong linear relationship in the dataset though to figure out whether a linear regression model would be a good fit. A decision tree or random forest however may offer more if the data has much variability or if the none of the metrics have strong linear relationships.

You can use a mix of the categorical variables as well as one of the regions of sales to predict its sales performance globally