Final Project Proposal

Nan Chen
Cameron Calder
Kevin Gwinn

February 2023

1 Problem Statement

This project will use a data set of video game sales from 1980 to 2020 to predict regional sales based on popularity. We will use 7 different supervised learning algorithms for training and predicting in this project: linear regression, logistic regression, decision tree, random forest, support vector machine, Bayesian learning and ANN. These machine learning methods will then be compared for accuracy and efficiency based on testing data.

2 Description of Data Set

For our machine learning project, we will be using a data set of video game sales found on kaggle.com. This data set has a large enough set of detailed data to allow us to train algorithms, and test them.

There are some identical datasets on Kaggle that have many code and discussion forums, but the one we're looking into has more than half new features comparing to the popular one. Also, we will be digging into our own new target problem that has not been used with this data set before.

There are 16 features on the dataset: Name, Platform, Year of Release, Genre, Publisher, North American Sales, European Sales, Japan Sales, Other Sales, Global Sales, Critic Score, Critic Count, User Score, User Count, Developer, and Rating.

Our goal is to predict sales of video games in 4 regions(NA, EU, JP and others) based on other 12 features provided.

URL of data set:
`https://www.kaggle.com/datasets/rishidamarla/video-game-sales`

3 Implementation Plan

We have planned three phases for our implementation plan:

First phase (Mar 1 to Mar 15): cleaning up the data, which would be used for every method.

Second phase (Mar 16 to Apr 26): each person in the group picks 2-3 different methods to implement individually.

Third phase (Apr 27 to May 8): compare the results we get together using ensemble learning. Generate a report along with presentation slides.

4 Team Members and Task Allocation

Together: Cleaning up the data; making init() functions to generate dataframes and vectors in Python. Will also clearly define the libraries to work with and platforms to work on, as well as how to share files (Github, Drive, .py or .ipynb, etc.). Use ensemble learning to compare the models created by all the group members.

Nan Chen: Linear regression, logistic regression, and ANN to help determine the complexity of model needed.

Cameron Calder: Decision tree, Bayesian learning classifier models to compare to linear/logistic regression. See differences in variance, look for any indicators of bias in the data.

Kevin Gwinn: Random forest, support vector machine models; see if there is evidence of over or underfitting in the other models, and consider feature selection based on performance of SVM.