Predicting Video Games Sales Based on Genre, Platform and Critic Reviews

# Research

Applying machine learning (ML) when solving issues like predicting sales is a method that has been used before since 1980 through the use of artificial neural networks (ANN). These methods have delivered results and have proved their potential however research has been lacking due to issues in validation or implementation.

Machine learning now has gained a very good reputation due to the overcoming of previously mentioned issues and very sizeable advancements in technologies in both hardware and software. Current studies done show that ML can be used in both demand and sales forecasting. While methods other than ANN exist, they are still attractive because of performance and generalization, being able to also represent non-linear relations.

Studies also show that proper data preprocessing techniques result in much simpler models, good accuracy and lower computational requirements than ever before while also not sacrificing accuracy.

In machine learning algorithm, data can’t be used in its normal form as it is the as the way it is obtained, so the data needs to be devised before employing it in machine learning models. This technique is used to solve problems that are not yet known by the knowledge extractor. Proper formatted and cleaned data is essential for preprocessing.

Machine learning has been proved to outperform traditional models by mitigation of the Bullwhip effect. Even if Data Preprocessing techniques could substantially reduce computing time, they are still complex and therefore expensive to establish in a company, as they require qualified people with proper equipment.

Dealing with a high volume of data is one of the more common challenges imposed by modern business trends. At this stage, it can be noticed that no study using data mining techniques on D&SF was found, however techniques such as rule mining may allow to analyze large amounts of data and to identify correlations between items that can provide explicit knowledge, able to improve the skills of the decision makers.

# Development

In the development of this prediction model several training algorithms were used. Following is a short description of each trained model and then the results will be presented and the best model is saved based on accuracy.

1. Linear Regression

Linear regression is a fundamental machine learning algorithm used for predicting numerical values based on input features. It assumes a linear relationship between the features and the target variable. The model learns the coefficients that best fit the data and can make predictions for new inputs.

1. Ridge Regression

Ridge regression is a linear regression method that adds a bias to reduce overfitting and improve prediction accuracy. Unlike ordinary least squares, ridge regression includes a penalty on the magnitude of coefficients to reduce model complexity.

1. Least Absolute Shrinkage and Selection Operator

Least absolute shrinkage and selection operator, also Lasso or LASSO, is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model. The lasso method assumes that the coefficients of the linear model are sparse, meaning that few of them are non-zero.

1. Decision Tree Regression

Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.

1. Random Forest Regressor

Random forest regression is a supervised learning algorithm and bagging technique that uses an ensemble learning method for regression in machine learning. The trees in random forests run in parallel, meaning there is no interaction between these trees while building the trees.

1. Support Vector Regression

Support Vector Regression (SVR) is a type of machine learning algorithm used for regression analysis. The goal of SVR is to find a function that approximates the relationship between the input variables and a continuous target variable, while minimizing the prediction error.

1. Neural Networks

A neural network is a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain. It is a type of machine learning process, called deep learning, that uses interconnected nodes or neurons in a layered structure that resembles the human brain.

1. XGBoost Regressor

XGBoost is a type of supervised learning algorithm that can be used to make predictions on continuous numerical data. It is an implementation of the gradient boosting machine learning algorithm, which is a type of ensemble learning method that combines the predictions of multiple weaker models to create a stronger, more accurate model.

1. LightGBM

LightGBM or ‘Light Gradient Boosting Machine’, is an open source, high-performance gradient boosting framework designed for efficient and scalable machine learning tasks. It is specially tailored for speed and accuracy, making it a popular choice for both structured and unstructured data in diverse domains.

# Results

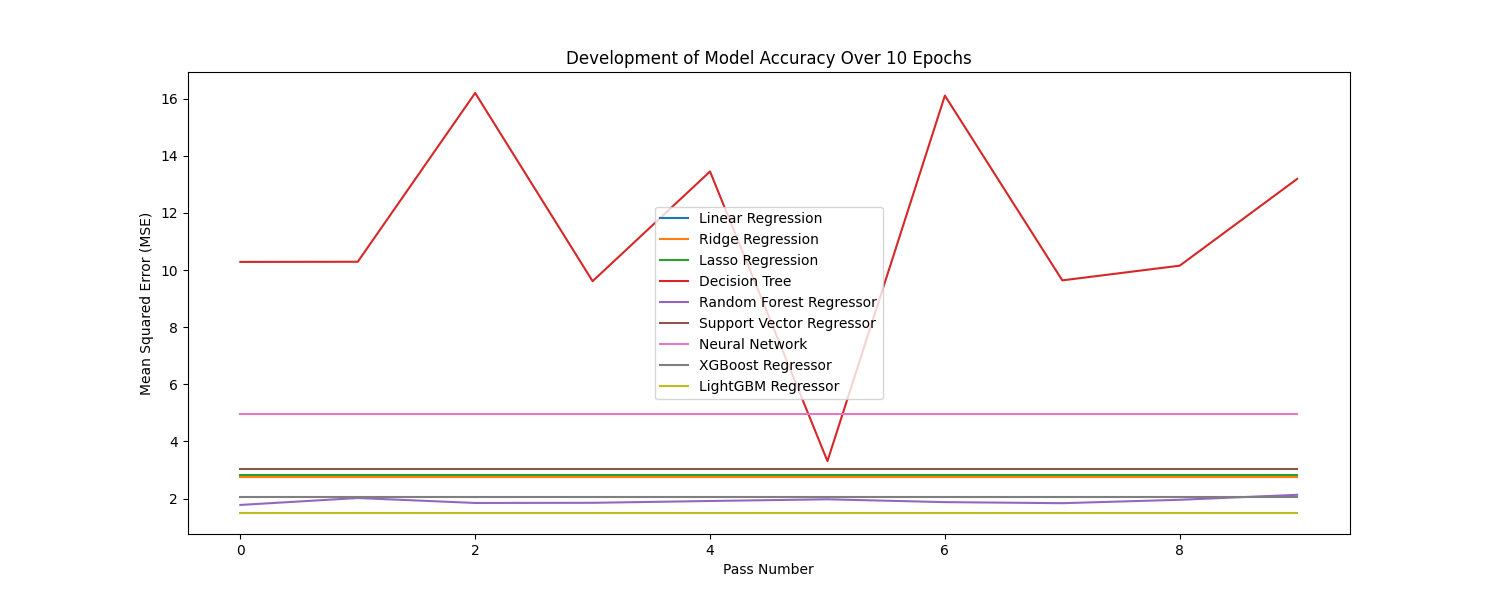
The dataset used for training can be found at here: <https://www.kaggle.com/datasets/sidtwr/videogames-sales-dataset?select=Video_Games_Sales_as_at_22_Dec_2016.csv>

The dataset includes the following fields:

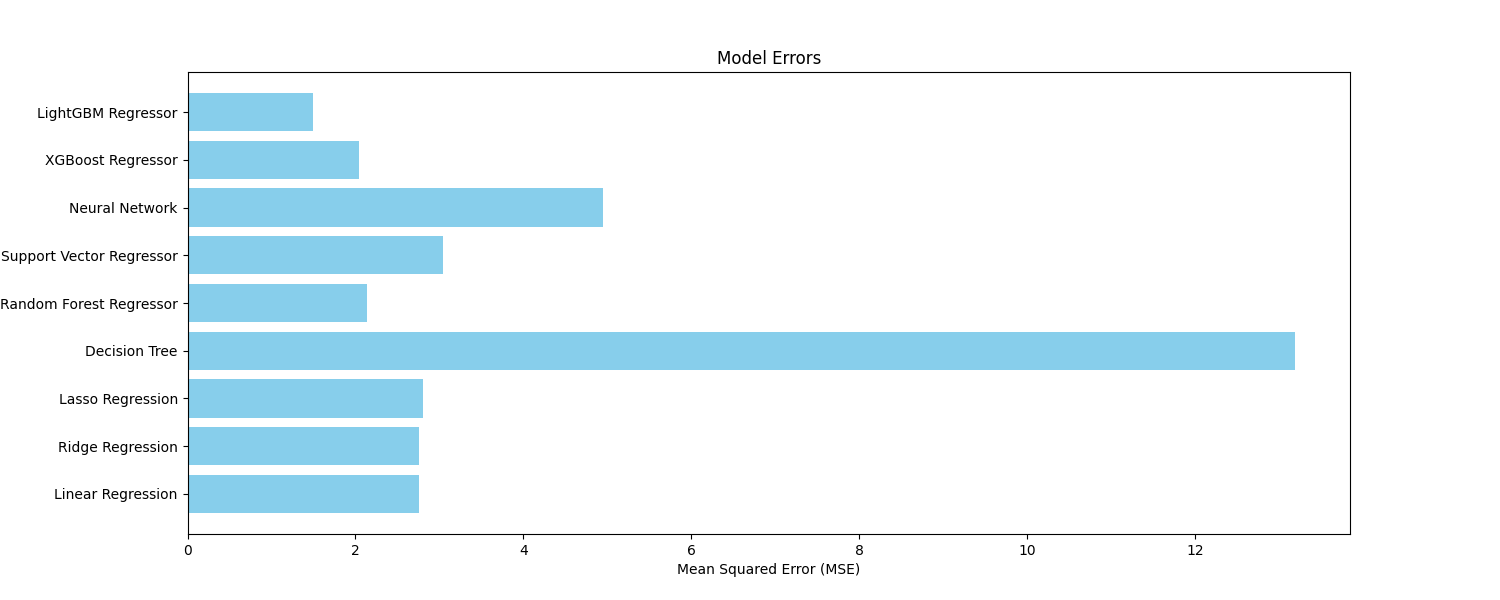
* Rank - Ranking of overall sales
* Name - The games name
* Platform - Platform of the games release (i.e. PC, PS4, etc.)
* Year\_of\_Release - Year of the game's release
* Genre - Genre of the game
* Publisher - Publisher of the game
* NA\_Sales - Sales in North America (in millions)
* EU\_Sales - Sales in Europe (in millions)
* JP\_Sales - Sales in Japan (in millions)
* Other\_Sales - Sales in the rest of the world (in millions)
* Global\_Sales - Total worldwide sales.
* Critic\_Score – The score resulting from Game Critic Reviews
* Critic\_Count – The number of critics who reviewed the game
* User\_Score – The score resulting from Game Critic Reviews
* User\_Count – The number of critics who reviewed the game
* Developer – The Developer of the game
* Rating – Age rating of the Game (i.e. E, M, T, etc.)

The fields used for training are the following: Platform, Year\_of\_Release, Genre, Publisher, Critic\_Score, Critic\_Count, User\_Score, User\_Count, Developer and Rating. Using these fields, the training had the Global\_Sales field as the target.

The graph below shows how the mean square error (MSE) evolves of 10 epochs of training on the dataset. Chaotic results can be seen for the Decision Tree Regressor. That is because Decision-tree algorithms are known to be unstable: small variations in the training set can result in different trees and different predictions for the same validation examples.

The other models provide expected results with the Light Gradient Boosting Machine providing the lowest MSE out of all the trained models in testing.

The graph below shows the MSE of each model after being trained.



Further analyzing the graph shows that the Light Gradient Boosting Machine is actually the only model that provides a MSE lower than 2 on the validation data. While the error of all models is reasonably large, that might be caused by the small number of training epochs since it has been proven that increasing this number will almost always generate a higher accuracy for almost all ML algorithms.

# References

1. Cadavid, Juan Pablo Usuga, Samir Lamouri, and Bernard Grabot. "Trends in machine learning applied to demand & sales forecasting: A review." *International conference on information systems, logistics and supply chain*. 2018.
2. Bajaj, Purvika, et al. "Sales prediction using machine learning algorithms." *International Research Journal of Engineering and Technology (IRJET)* 7.6 (2020): 3619-3625