Video Games Sales Project

Aadyant Khatri

10/03/2022

Overview

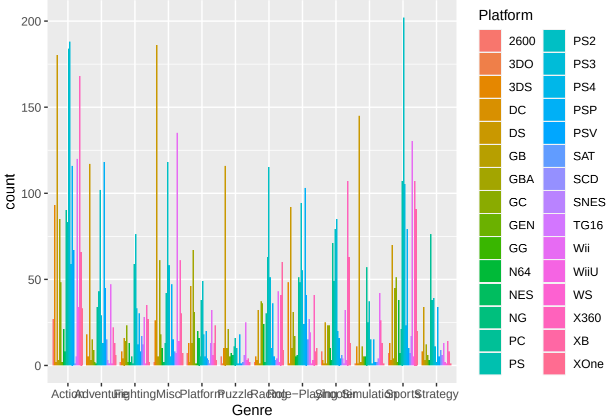
This project is aimed at predicting the genre of video game given the year of its release, its platform, its publisher and the total sales it had globally. If I am able to do so successfully, I would be able to conclude that their is a significant relation between the predictors and the genre.

For this project, **Video Game Sales** data set available on Kaggle was used (https://www.kaggle.com/gregorut/videogamesales). This data set contains sales data for more than 16,500 games. For each of the games, the year of its release, its platform, its publisher and its sales in North America, Europe, Japan and globally are provided.

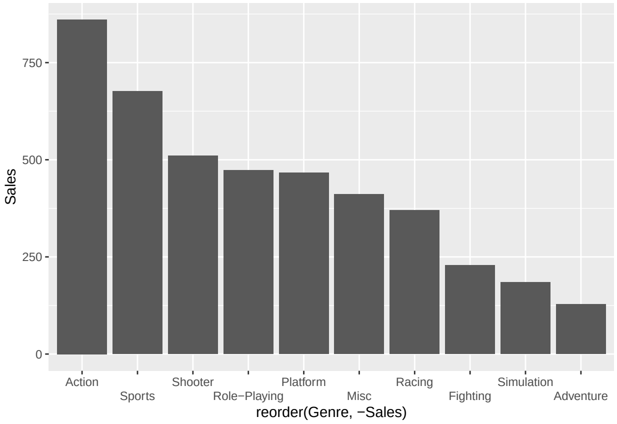
I have tried tackling the problem using several machine learning techniques such as **k-nearest neighbors**, **decision trees** and **random forest**.

Visualizing Data

In this figure, we can see the number of sales by different publishers for each of the genres -



In this figure, we can see the global sales in each of the genres -



Methodology

Model - 1

K-nearest neighbors (KNN) tries to predict the correct class for the test data by calculating the distance between the test data and all the training points. Then select the K number of points which is closet to the test data.
In this method, tune grid was used to find the optimum value of the k.

```
fit_knn <- train(Genre ~ .,  method = "knn", tuneGrid = data.frame(k = seq(39, 42, 1)), data = train_se

prediction_knn <- predict(fit_knn, newdata = test_set)

accuracy_knn <- mean(prediction_knn == test_set$Genre)
```

This model gave an accuracy = 0.2657343

Model - 2

Decision Trees are a type of supervised machine learning where the data is continuously split according to a certain parameter.
In this method, tune grid was used to find the optimum value of the complexity hyper parameter.

```
fit_tree <- train(Genre ~ ., method = "rpart", tuneGrid = data.frame(cp = seq(0, 0.01, len = 30)), data

prediction_tree <- predict(fit_tree, newdata = test_set)

accuracy_tree <- mean(prediction_tree == test_set$Genre)
```

This method gave accuracy = 0.267791

Model - 3

Random forest is a machine learning algorithm that builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

```
fit_rf <- train(Genre ~ .,
                    method = "Rborist",
                    tuneGrid = data.frame(predFixed = 2, minNode = c(2, 10)),
                    data = train_set)

prediction_rf <- predict(fit_rf, newdata = test_set)

accuracy_rf <- mean(prediction_rf == test_set$Genre)
```

Model 3 resulted in accuracy = 0.2344714

Results

The accuracies obtained with each model have been compiled in the table below.

Method	Accuracy	RMSE
kNN Model	0.2657343	NA
Decision Tree Model	NA	0.2677910
Random Forest Model	NA	0.2344714

Conclusion

I was able to create a model which could predict the genre of video games using several sales parameters with a decent accuracy.

This project helps in concluding that video games of certain genres, by certain publishers and on certain platforms are preferred and enjoyed more by users than others.

However, a key factor that limited the accuracies of my models is low computation power. On a machine having higher computational capabilities, much more robust models with even better hyper-parameters can be created.