Syed Qavi

EXECUTIVE SUMMARY

**GOALS**

The dataset looks at video games sales around the world along with video game information and score and count of the users and critics. The dataset contains around approximately 6900 complete cases. The goal of the dataset is to whether the dataset variables can predict the target variable and to determine the relationship between variables.

**METHODS**

Different types of regression including standard linear regression, Lasso regression, Stochastic Gradient regression, Ridge regression and Random Forestation Regressor.

Clustering and PCA to determine number of variables that can be grouped together.

Using Pandas, Matplotlib, Numpy and sci kit learn.

**CONCLUSION**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dependent  Variable:  **NA\_Sales** | Lasso regression | Ridge regression | Stochastic Gradient regression | Linear Regression  (sklearn) |
| RMSE on Test | 0.0117 | 0.0083 | 0.0083 | NA |
| MAE on Test | 0.0076 | 0.0046 | 0.0045 | 0.004575 |

|  |  |
| --- | --- |
| RMSE on Training (Linear Regression function) | 0.3949 |
| RMSE on 10-fold CV (Linear Regression function) | 0.0740 |
| Optimal number of features (feature selection) | 46 |
| Optimal percentile of features (feature selection) | 9 |

SGD has lowest MAE value.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dependent  Variable:  **User\_Score** | Lasso regression | Ridge regression | Stochastic Gradient regression | Linear Regression  (sklearn) |
| RMSE on Test | 0.1425 | 0.1217 | 0.1222 | NA |
| MAE on Test | 0.1099 | 0.0923 | 0.0928 | 0.092198 |

|  |  |
| --- | --- |
| RMSE on Training (Linear Regression function) | 20.6288 |
| RMSE on 10-fold CV (Linear Regression function) | 9.1301 |
| Optimal number of features (feature selection) | 46 |
| Optimal percentile of features (feature selection) | 9 |

Linear Regression using sklearn has lowest MAE value. Ensemble Method, Random Forest Tree showed after splitting tree, one of the leaf had MAE = 0.071 which is lower than Linear Regression. Explanatory variables are better at predicting NA\_Sales compared to User\_Score.

PCA resulted in 12 components explaining at least 95% of variance.