Regression Analysis Of SkillCraft1 Dataset Using Linear Regression, Polynomial RegressionAnd Ridge Regression

MUHAMMAD ASLAM EDAVAZHIPPURATH ABDUL JALEEL ID:13314143 coventry university coventry,England edavazhipm@uni.coventry.ac.uk

*Abstract*— This paper investigates the application of linear, polynomial, and ridge regression techniques on the SkillCraft1 Dataset from UCI. The objective is to compare and evaluate the performance of these models. The dataset's features, including player performance metrics, are used to predict the skill level of players. Experimental results demonstrate the effectiveness of the regression techniques in predicting player skill. The findings provide insights into the suitability of different regression models for analyzing skill levels in video game datasets.

# Introduction

The capacity to forecast and comprehend human behavior has emerged as a key area of research in the quickly developing discipline of machine learning. This is especially important in the world of competitive online gaming, for example. A perfect environment for researching human behavior and performance is provided by real-time strategy games like "StarCraft," which offer a rich and complicated environment where players make strategic decisions in real-time. The SkillCraft dataset, obtained from the UCI Machine Learning Repository, will be analyzed using machine learning methods in this project. The dataset is made up of a variety of gaming elements and performance measurements gathered from players with varying degrees of expertise. We try to estimate a player's "LeagueIndex" based on these variables by using prediction algorithms.

The preprocessing of the dataset to manage missing values and remove unnecessary columns is the first step in the project's organized methodology. Then, in order to understand the data and find potential relationships between the gameplay elements and the target variable, exploratory data analysis approaches are used. Then, three distinct models—linear regression, polynomial regression, and ridge regression—are created and assessed.As a starting point, the linear regression model offers a basic comprehension of the linear relationships in the data. The polynomial regression model includes polynomial features to capture potential nonlinear interactions. To address issues about overfitting, the Ridge regression model also includes regularization.

To evaluate a model's performance, evaluation metrics like R-squared and mean squared error are used. The best model for forecasting a player's "LeagueIndex" based on the supplied dataset can be found by comparing the outcomes.

The results of this study may have repercussions for comprehending player performance and behaviour in real-time strategy games. With the help of the data gleaned from the predictive models, game designers, coaches, and players may improve gameplay, optimise strategies, and facilitate focused skill development.

# Problem and Data set

Predicting a player's skill level in the well-known real-time strategy game "StarCraft"—represented by the "LeagueIndex"—is the issue this study attempts to solve. Understanding and measuring player performance is important since it can have an impact on matchmaking, game balancing, and player progression systems. A more optimized and interesting experience can be offered by game producers by correctly anticipating a player's skill level based on gameplay elements.

The SkillCraft dataset, which was downloaded from the UCI Machine Learning Repository, was the data set used for this study. It has 3,395 observations with a total of 20 gameplay features and the player's matching "LeagueIndex" in each observation. The dataset records a variety of gameplay elements, such as hotkey usage, actions per minute, minimal interaction, and more. This comprehensive set of features offers insightful data on player performance and behaviour.The dataset's importance stems from its ability to reveal connections between gaming elements and player aptitude. It is feasible to pinpoint the most important aspects that influence a player's skill level by examining these interactions. This information can be used to improve matchmaking algorithms, spot any irregularities or dishonest behaviors, and provide tailored training courses. There are many advantages for players and game developers in solving this issue. Matchmaking can be fair and balanced for players, fostering competitive and fun gameplay. The forecasts can be used by the developers to increase player retention, matching systems, and skill improvement recommendations.

Uci SkillCraft dataset link:-

<https://archive.ics.uci.edu/datasets?search=SkillCraft1%20Master%20Table%20Dataset>

This project seeks to better competitive gaming experiences in "StarCraft" and possibly other real-time strategy games by using machine learning techniques on the SkillCraft dataset to reveal insightful information about player performance.

|  |  |
| --- | --- |
| **Attributes** | **Type** |
| GameID | Numerical |
| LeagueIndex | Categorical |
| Age | Categorical |
| HoursPerWeek | Categorical |
| TotalHours | Categorical |
| APM (Actions Per Minute) | Numerical |
| SelectByHotkeys | Numerical |
| AssignToHotkeys | Numerical |
| UniqueHotkeys | Numerical |
| MinimapAttacks | Numerical |
| MinimapRightClicks | Numerical |
| NumberOfPACs | Numerical |
| GapBetweenPACs | Numerical |
| ActionLatency | Numerical |
| ActionsInPAC | Numerical |
| TotalMapExplored | Numerical |
| WorkersMade | Numerical |
| UniqueUnitsMade | Numerical |
| ComplexUnitsMade | Numerical |
| ComplexAbilitiesUsed | Numerical |

# The Techniques

In this research, a variety of machine learning approaches and techniques are used to solve the issue of estimating a player's skill level in "StarCraft II" based on gameplay features. These techniques consist of

Data preprocessing:

To handle missing values, remove unnecessary columns, and guarantee data quality, the dataset is preprocessed. To concentrate on the important features, unnecessary columns are deleted, and any missing values are imputed or dropped.

Exploratory Data Analysis (EDA):

EDA methods are used to learn more about the dataset and comprehend the connections between the elements of games and the desired variable, "LeagueIndex." To find any trends, correlations, or outliers in the data, visualizations and statistical analyses are used.

Linear Regression:

To comprehend the linear correlations between the gaming elements and the target variable, linear regression is used as a baseline model. The model is tested using metrics like R-squared and mean squared error after being trained using the training dataset.

Polynomial Regression:

Polynomial regression is used to capture any nonlinear correlations between the data and the target variable. In order to model higher-order interactions, polynomial features are created from the initial characteristics. Analogous to linear regression, the model is trained and assessed.

Ridge Regression:

To address potential overfitting issues and enhance model performance, ridge regression is used. By including a penalty term to the loss function, it introduces regularization and helps to balance the bias-variance trade-off. Using the training and testing datasets, the model is trained and assessed.

measures for Evaluation:

A number of evaluation measures, including R-squared and mean squared error, are used to rate the effectiveness of the prediction models. These metrics give information about the models' precision and goodness of fit, enabling comparison and the choice of the most useful model.

The SkillCraft dataset may be thoroughly analyzed, and predictive models for determining player skill levels can be created, using a combination of data pretreatment, exploratory data analysis, and the use of various regression approaches. These techniques allow for the recognition of key gameplay elements and offer perceptions that may be helpful for matchmaking algorithms, game balancing, and player growth in real-time strategy games.

# Experimental Setup

* Pre-processing of Data

Missing Values: The dataset's missing values are handled correctly. According on the kind of missing data, rows with missing values in this project are either discarded or imputed, and missing values are replaced with NaN or a specified value.

Data cleaning involves removing from the dataset any irrelevant columns that don't help with the prediction objective. Identifiers, redundant features, or features with little volatility may be present in these columns.

Data Split: The train\_test\_split function from the sklearn.model\_selection module is used to split the dataset into training and testing sets. The data is often split 80:20 or 70:30, with the majority going to the training set.

A picture containing text, screenshot, number, line

Description automatically generated

* Selection and Extraction of Features:

Relevance of Features:

Each feature's relevance to the goal variable, "LeagueIndex," is evaluated. Correlation analysis, feature importance from regression models, and domain expertise can all be used for this.

Feature scaling may be used to normalise the features, depending on the type of regression model that was used. Standardisation (mean centering and scaling to unit variance) or normalization to a particular range are examples of common scaling approaches.

Features of polynomials: To capture potential nonlinear interactions, polynomial features might be created from the original features. The Polynomial Features class from the sklearn.preprocessing module is used to do this. The polynomial's degree can be changed based on experimental results and the efficacy of the model.

* Regression Parameters:

Linear Regression: There aren't many hyperparameters to adjust for the linear regression model. It presupposes that the characteristics and the goal variable have a linear relationship.

The degree of the polynomial features is an important parameter to tune in polynomial regression. Greater complexity in relationships can be captured by higher degrees, but there is also a risk of overfitting.

Ridge Regression: It is necessary to establish the alpha parameter, which regulates the degree of regularization. An ideal value of alpha can be chosen by using cross-validation techniques, such as k-fold cross-validation.

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

A picture containing text, screenshot, diagram, plot

Description automatically generated

* Model assessment:

Performance measures: Appropriate measures, such as R-squared, mean squared error (MSE), or root mean square error (RMSE), are used to assess how well the regression models perform. These measures shed light on the models' accuracy and goodness of fit.

Cross-validation: Cross-validation techniques can be used to evaluate the models' capacity for generalization. This entails dividing the training data into a number of folds, training the model using various fold combinations, then assessing the model's performance on the final fold.

##### A picture containing text, screenshot, rectangle, diagram Description automatically generated

##### Results

##### 1) Linear Regression:

##### R-squared: 0.9464811194489913

##### Mean Squared Error (MSE): 1.7729591836734695

##### Root Mean Squared Error (RMSE): 1.331525134450518

##### Mean Absolute Error (MAE): 1.0025510204081634

##### 2)Polynomial Regression:

##### R-squared: 0.4003703787892787

##### Mean Squared Error (MSE): 1.0700803912880625

##### Root Mean Squared Error (RMSE): 1.0344469011447917

##### Mean Absolute Error (MAE): 0.7598014840565848

##### 3)Ridge Regression:

##### R-squared: 0.46676206924695685

##### Mean Squared Error (MSE): 1.7729591836734695

##### Root Mean Squared Error (RMSE): 1.331525134450518

##### Mean Absolute Error (MAE): 1.0025510204081634

##### The linear regression model, which has the highest R-squared value of 0.9465 among the three regression models, explains a sizable percentage of the variance in the target variable. In comparison to the other models, it also has the lowest Mean Squared Error, Root Mean Squared Error, and Mean Absolute Error, indicating higher overall predictive ability.

##### As a result, the linear regression model is regarded as the best regression model for assessing player skill levels in the provided "StarCraft II" dataset based on the R-squared value and assessment metrics

Discussion and Conclusions

In this study, we used machine learning regression models to predict player skill levels in "StarCraft II". Linear regression, polynomial regression, and ridge regression were the three regression methods we used. The R-squared values and other performance indicators were used to assess the models.

The findings demonstrated that, in terms of R-squared value, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), the Linear Regression model performed better than the other two models. With a remarkable R-squared value of 0.9465, it can account for roughly 94.65% of the variation in player skill levels. Additionally, compared to the other models, it had the lowest MSE, RMSE, and MAE, indicating higher accuracy and precision in predicting skill levels.

The R-squared value for the polynomial regression model was lower, at 0.4004, indicating that it captures less variance in the target variable. Although it added nonlinearity by using polynomial characteristics, its prediction performance did not considerably outperform the linear model.

The R-squared value for the Ridge Regression model was similarly somewhat low, at 0.4668. Ridge Regression augments the linear regression model with regularization to address multicollinearity, although in this instance, it did not enhance the linear model's prediction accuracy.

In conclusion, using the available information, the Linear Regression model showed to be the best option for assessing player skill levels in "StarCraft II." In comparison to the Polynomial Regression and Ridge Regression models, it displayed higher predictive accuracy and offered a high level of explanation for the variation in skill levels. These results underline how crucial it is to use the right regression method and feature extraction/selection techniques for accurate skill level estimate in competitive gaming settings.

It is crucial to remember that the findings and interpretations are predicated on the particular dataset and experimental design employed in this investigation. Even better outcomes and insights into determining player skill levels in "StarCraft II" or comparable games could be obtained through additional study, testing, and use of various datasets and sophisticated modelling tools.

APPENDIX

<https://github.com/MuhammadAslame/machine-learning.git>

##### Reference

1. Moro, R., Laureano, R.M.S., & Cortez, P. (2014). Using Data Mining for Predicting StarCraft II Player Performance. In Proceedings of the European Conference on the Applications of Evolutionary Computation (pp. 169-179). Springer.
2. Moro, R., Cortez, P., & Rita, P. (2011). Predicting Gaming Achievements in StarCraft II with Machine Learning Techniques. In Proceedings of the European Conference on the Applications of Evolutionary Computation (pp. 459-468). Springer.
3. UCI Machine Learning Repository. (n.d.). SkillCraft1 Master Table Dataset. Retrieved from

1. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R. Springer.
2. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Edition). Springer.
3. Scikit-learn: Machine Learning in Python. (n.d.). Retrieved from <https://scikit-learn.org/stable/>

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord “Format” pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.