Predicting the success of NBA players is a challenging task due to the complex nature of the sport and the multitude of factors that contribute to player performance. This program focuses on feature selection, a critical step in building a predictive model, to identify the most relevant features from a pool of statistics. By selecting the most influential features, the model aims to improve prediction accuracy and provide insights into the factors that contribute to player success.

The program begins by cleaning the college basketball player statistics dataset, filtering out missing values and mapping the year of the players' college careers. Next, the NBA draft data is cleaned, ensuring only common players with college statistics are included. A weighted formula is applied to calculate a performance metric for each player in the draft dataset. The datasets are then merged based on player names, resulting in a dataset that contains both college and NBA performance information.

The program utilizes the SelectKBest algorithm from the scikit-learn library to perform feature selection. The algorithm evaluates the statistical significance of each feature using the f_regression scoring function and selects the top K features. A linear regression model is trained using the selected features to predict the performance metric. The model's performance is evaluated using mean squared error (MSE) and R-squared metrics.

The program presents the results of the feature selection process, providing a list of the top selected features. The importance of these features in predicting player success is highlighted, shedding light on the key factors that contribute to performance. The analysis also discusses the skewness of the target variable and presents descriptive statistics for the performance metric.

Furthermore, the program evaluates the model's predictive accuracy by comparing MSE scores for the full feature set and the selected feature set. The R-squared scores are also reported to assess the model's overall goodness of fit. The findings demonstrate the impact of feature selection on model performance and highlight the potential for improved predictions by focusing on the most relevant features.