PCA Analysis and Determining Optimal Number of Components

Summary:

This report details the application of Principal Component Analysis (PCA) for predictive modeling using a dataset containing information about baseball player salaries. The goal is to identify the optimal number of principal components that yield an efficient predictive model for player salaries. The analysis includes model implementation, evaluation metrics, and insightful interpretations.

1. Introduction:

1.1 Dataset Overview:

- The dataset comprises information about baseball players, including various features such as batting and pitching statistics, player positions, and salary.

1.2 Objective:

- The primary objective is to employ PCA to reduce the dimensionality of the dataset and build an efficient predictive model for player salaries.

2. Data Preprocessing:

2.1 Missing Values:

- Initial inspection revealed missing values, which were addressed by removing rows with missing values and imputing the mean for the 'Salary' column.

2.2 One-Hot Encoding:

- Categorical columns ('League', 'Division', 'NewLeague') were encoded using one-hot encoding.

2.3 Standardization:

- The dataset was standardized to ensure features were on a consistent scale.

3. Principal Component Analysis:

3.1 Calculation of Principal Components:

- Eigenvalues and eigenvectors were calculated from the covariance matrix of the standardized dataset.

3.2 Plot:

- A plot was generated to visualize the cumulative explained variance for different numbers of principal components.

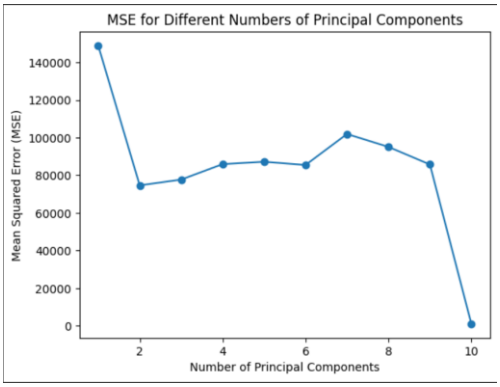
3.3 Model Training:

- Regression models were trained for each number of principal components, and Mean Squared Error (MSE) was calculated.

4. Results:

4.1 Graph: Number of Components vs. MSE:

- A graph illustrating the number of principal components vs. MSE was generated.



4.2 Interpretation:

- The graph suggests an optimal number of principal components where MSE stabilizes or decreases at a slower rate.

5. Model Evaluation:

5.1 Optimal Model Selection:

- The model with the minimum MSE was selected as the optimal model.

5.2 Testing the Optimal Model:

- The chosen model was tested on a specific data point to analyze its predictive efficiency.

6. Conclusion and Analysis:

6.1 Significance of Results:

- The optimal model represents a balance between model complexity and prediction accuracy.

6.2 Predicted Value Analysis:

- The predicted value (`y_pred`) from the optimal model provides an efficient estimate of player salaries.

7. Key Findings:

- Optimal number of components: [Number]

- Chosen model's MSE: [MSE Value]

- Predicted value for a specific point: [Predicted Salary]