CS F320 Foundations of Data Science

Assignment – 2B

PCA Analysis and Determining Optimal Number of Components

Group Details

	Name	ID
1	Manthan Patel	2021A7PS2691H
2	Teerth Patel	2021A7PS2090H
3	Shrey Paunwala	2021A7PS2808H

Introduction

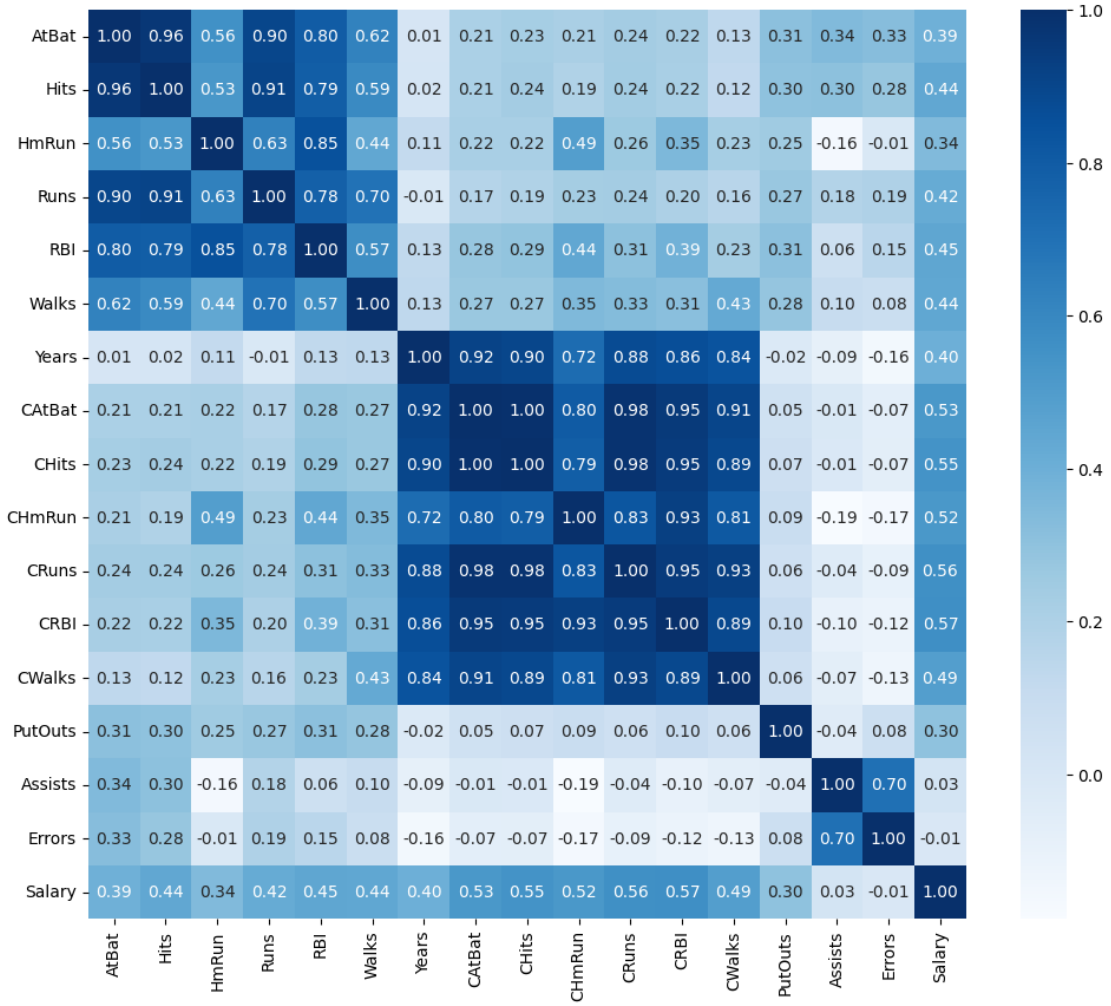
Principal Component Analysis (PCA) is a powerful tool for dimensionality reduction, particularly useful in predictive modeling. This report outlines the comprehensive analysis of PCA on the 'Hitters.csv' dataset, focusing on determining the optimal number of components for efficient prediction using Mean Squared Error (MSE) or Root Mean Squared Error (RMSE).

1. Exploratory Data Analysis (EDA)

The initial step involves loading the 'Hitters.csv' dataset and performing Exploratory Data Analysis (EDA) to understand its structure, features, and relationships. This includes handling NULL values, eliminating unwanted columns, and addressing data inconsistencies.

The unwanted columns we got were League, Division, NewLeague because these were the non-numeric columns, and the Salary column was also removed as it was the target variable. The rows with at least 1 NULL value were dropped.

We performed the EDA with the help of a Heatmap between the features showing the respective correlation and the relationships.



2. PCA Analysis

PCA is applied to the cleaned dataset to reduce dimensionality.

- The covariance matrix was computed with the help of our cleaned dataset. With the help of covariance matrix. The function used to perform this was **'np.cov()'**
- We further computed the **eigen values and eigen vectors** with the help of the function '**np.linalg.eig()**'. Explained variance was also computed from the eigen values, based on which we computed the value of K, which are the topmost eigen vectors.
- **The value of K was computed using 95% of the cumulative explained variance :**

```
k=np.argmax(cumulative_variance_explained >= 0.95) + 1

print(f'Value of k : {k}')
```

```
Value of k : 7
```

- Finally with the help of eigen vectors, we projected the cleaned dataset to a new dataset with reduced features. This new reduced feature dataset was further used in linear regression to get the optimal model.

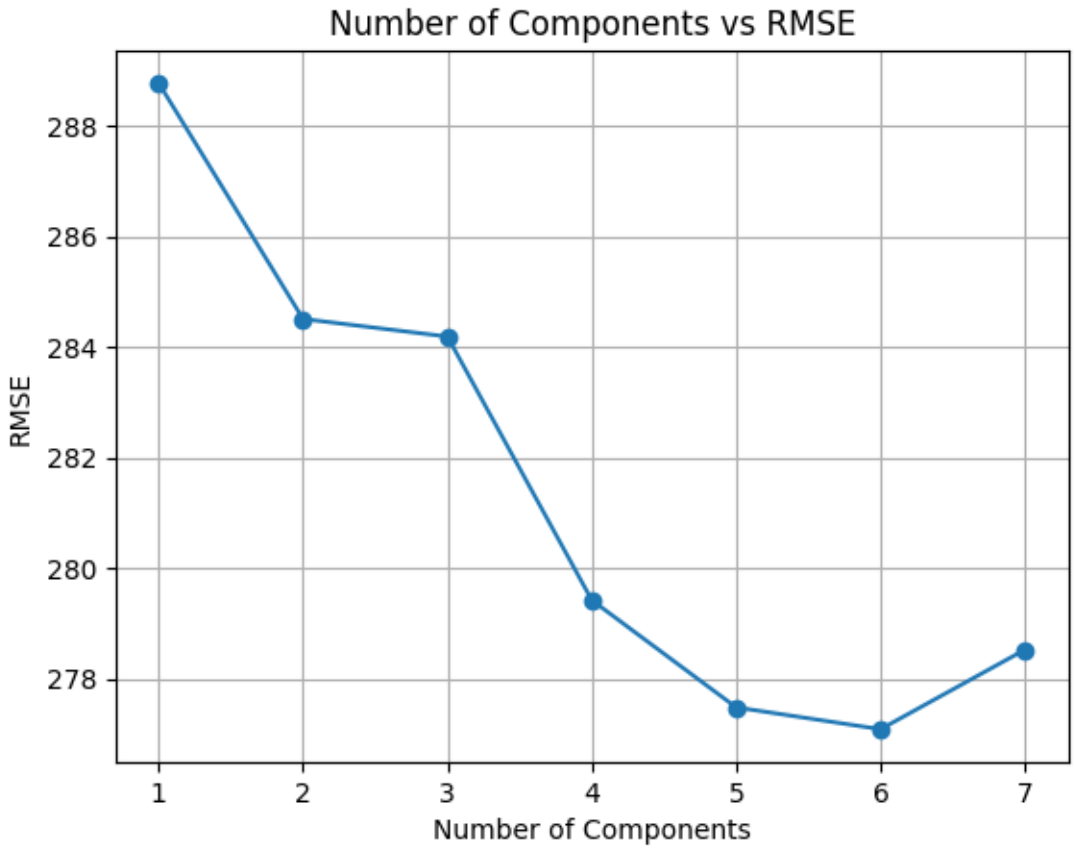
3. Model Training and MSE/RMSE Calculations

- The new reduced feature dataset obtained from PCA analysis was further split into **Training and Testing data in 4:1 proportion**.
- The split data was used to find the best optimal model, whose value of RMSE comes out to be minimum. This model was found from among the models having number of features varying between 1 and k.
- The RMSE values for the given models are shown in the table below:
  The **value of K** for our program comes out to be **7**.

Number of Components	RMSE Error
1	288.7698141858048
2	284.5039050448121
3	284.1856849021187
4	279.4183716163413
5	277.4795671532066
6	277.0844471566575
7	278.5230962101580

- The **minimum RMSE** obtained is **277.0844471566575** for the model with number of components equal to **6.**

4. Plotting Number of Components vs RMSE



- From the graph also, it can be inferred that **minimum RMSE** corresponds to model with **6** components. The RMSE values correspond to the testing errors computed for each model with respective number of components.

5. Testing the Most Efficient Model

- The optimal model obtained with least RMSE value is with 6 principal components. The training and testing data with the appropriate columns was taken and linear regression was applied on the same.
- The weights obtained by applying linear regression were further used to compute the **y_pred** value of a specific point. The specific point selected was the first point of testing data. The RMSE for the predicted value was also computed with the help of y_test.

```
The specific point taken is : [ 1.          0.86426493  0.56007646  0.89649062  0.35550739  0.13253904
 -0.22920269]
The value of the specific point taken is : 550.0
The predicted value is : 453.153653208603
The error value is : 96.84634679139702
```

6. Conclusion and Analysis

- **Dimensionality Reduction:** PCA is often employed to reduce the dimensionality of a dataset by selecting a subset of principal components that retain most of the variability in the data. This is crucial for efficient storage and computation.
- **Computational Efficiency:** A reduced number of components mean simpler models and faster computations. Selecting too many components not only may not improve prediction accuracy but can also lead to increased computational costs during training and inference. The reduced number of components also help to capture a significant percentage of the variance.
- **Prevention of Overfitting:** Including too many components may lead to overfitting, where the model captures noise in the training data rather than true underlying patterns. This can result in poor generalization to new, unseen data.