Mohammad Zana Majidi

Final Project

course: CE 610

December 2023

1. Introduction

The project revolves around forecasting and approximating the salaries of baseball players for a team in the American baseball league during the 1986-1987 season. During this season, the team faced financial challenges, prompting the need to predict salary amounts for individual players based on their abilities, disregarding their names. So, the main purpose is creating the best model that helps the club manager to estimate the salary of players by using related variables (variables are in the CSV file).

Comprehensive insights into the project can be gleaned from either my Python code or its accompanying HTML file. Through these resources, I have visualized diverse graphs and tables to provide a clear overview of my work. In this report, my aim is to succinctly encapsulate the key aspects of the project using simple language.

To achieve the club's objectives, I employed four distinct models. In the subsequent sections, I will delve into the results produced by each model and engage in a comparative analysis to determine which one aligns best with the club's goals. The models compared are Model 1: Linear Regression (LM_t-test), Model 2: Linear Regression Using the Best Subset Selection (BestSubset), Model 3: Forward and Backward Stepwise Selection (Forward Stepwise), Model 4: Using K-fold Cross-Validation Approach (Forward Stepwise CV).

2. Model summary

The result of four models is showed in the following table:

Table 1 table of result

Models Factors	Mean of AbsErrors	Median of AbsErrors	SD of AbsErrors	IQR of AbsErrors	Min of AbsErrors	Max of AbsErrors
LM_t-test	257.94559	212.04905	221.12737	321.24472	3.643909	973.384257
BestSubset	227.88187	132.66203	277.19107	180.65888	0.677151	1743.01428
Forward Stepwise	227.89325	132.72754	277.42331	180.73521	0.716825	1745.20379
Forward Stepwise CV	234.48339	149.64758	300.28404	180.86951	9.620813	1905.55317

The Factors considered are:

Mean of AbsErrors: The average of the absolute values of the errors.

Median of AbsErrors: The median of the absolute values of the errors.

SD of AbsErrors: The standard deviation of the absolute values of the errors.

IQR of AbsErrors: The interquartile range of the absolute values of the errors.

Min of AbsErrors: The minimum of the absolute values of the errors.

Max of AbsErrors: The maximum of the absolute values of the errors.

Here is a conclusion and report based on the results: Comparative Performance Analysis of Predictive Models. The assessment of four predictive algorithms, namely LM_t-test, BestSubset, and Forward Stepwise (with and without cross-validation), presents a detailed overview of their error distributions when applied to a specific dataset.

The LM_t-test exhibits the highest mean absolute error (Mean of AbsErrors) at 257.945587 and a relatively high median absolute error (Median of AbsErrors) of 212.049049, indicating that on average, its predictions deviate from the actual values significantly. Additionally, the largest error recorded (Max of AbsErrors) is 973.384257, which suggests potential outliers or extreme cases where the model performed poorly.

In contrast, the BestSubset model, listed twice possibly due to a replicated row, shows improved performance with a lower mean (227.881868) and median (132.662027) of absolute errors. Interestingly, the standard deviation (SD of AbsErrors) is quite high (277.191072), suggesting variability in the prediction errors. The interquartile range (IQR of AbsErrors) is comparatively lower (180.65888), which could indicate that most errors are not too far from the median. The maximum error (Max of AbsErrors) is significantly larger (1743.014276) than the LM_t-test, but this may be due to a few extreme cases as the minimum error is quite small (0.677151).

The Forward Stepwise models demonstrate similar error metrics to the BestSubset model. The Forward Stepwise without cross-validation has a mean absolute error marginally higher than the BestSubset model but a median absolute error slightly better, which can suggest it has a more consistent performance in the middle 50% of predictions. However, its maximum error is slightly higher, indicating potential overfitting to the training data or sensitivity to outliers.

The Forward Stepwise with cross-validation (Forward Stepwise CV) has a mean and median absolute error slightly higher than its counterpart without cross-validation. This could be due to the model being more robust and generalizable, as cross-validation tends to prevent overfitting by validating the model on multiple subsets of the data. The maximum error, however, is significantly higher (1905.55317), which is concerning and should be investigated further.

3. Conclusion

In conclusion, the BestSubset and Forward Stepwise models show better overall performance compared to the LM_t-test model. The use of cross-validation in the Forward Stepwise CV model does not unequivocally translate into better performance according to these metrics and may suggest a need for further parameter tuning or investigation into the nature of the data and outliers. It is essential to consider the context of the application and the cost of errors when choosing between these models.