SEP 767: Multivariate Statistical Methods for Big Data Analysis and Process Improvement

Yash Dham
400279698

9.12.19
Monday

Project Report

Index

SEP 767: Multivariate Statistical Methods for Big Data Analysis and Process Improvement

Yash Dham 9.12.19

400279698 Monday

Project Report

Abstract

The Basic Idea Behind this Paper is to Find a Dataset that has a data that is relevant and is required to find answers to some question by applying PCA and PLS. The Dataset Used in this Paper is About model that Describes the salary of different players and their respective run scored in different leagues in different years along their career. Now, depending on that we'll Predict whether the model is accurate or not in displaying their salaries. The Process that we'll go through will be depend on Principal Component Regression using PCA and doing the scaling of data finally to do Partial Least Squares to Find the Mean Square Error in our Model. Depending on which we'll decide how accurate is our dataset. The Accuracy is called the best when the Value of R-squared is almost equal to 1.

R-squared = 1 – (First Sum of Errors / Second Sum of Errors)

We'll Scale the Data for PCA to Apply the Regression that will help us understand the data more and find mean square error if required. Upon concluding we'll check if the mean square error calculated is accurate enough for the dataset to be considered as a good model.

Once we have done PCA we will move towards doing the partial least square method (PLS) In which baby try to do 10-fold cross validation where we will shuffle the data and Find the mean square error. if it is same for both PCA and PLS then we can say that our data set is quite accurate.

SEP 767: Multivariate Statistical Methods for Big Data Analysis and Process Improvement

Yash Dham
400279698

9.12.19
Monday

Project Report

Introduction

The Dataset that I Selected was given up online as a challenge by a Prof. of Smith College. Now, This Dataset made me excited because of the number of input variables it had. There are total 18 input variables to which output of the 19th variable depends. Also, I am always interested in solving problems regarding Prediction Models. Now, the reason why I selected to do this project in Python and not in ProMV are:

- Python is an Open source language which is available online free of cost and is a standard base for any software tool for Big Data Analysis.
- My Background is from IT, which is why I am more comfortable with Programming Languages.

The total process went on from Creating PCA to PLS and finding out the results before and after Training and Testing of the Data.

The Dataset was one opportunity to work with multiple variables and reducing them down to only those that were important for my work and required me to use them with Principal Component Analysis and Partial Least Squares to Finally Compare the Mean Squared Errors Between Them.

SEP 767: Multivariate Statistical Methods for Big Data Analysis and Process Improvement

Yash Dham
400279698

9.12.19
Monday

Project Report

Dataset

As Stated Earlier the dataset selected was Hitters Dataset with following Variables:

- AtBat
- Hits
- HmRun
- Runs
- RBI Walks
- Years
- CAtBa
- Chits
- CHmRun
- CRuns
- CRBI
- CWalks
- League
- Division
- PutOuts
- Assists
- Error

And **Salary** as the Independent variable. The dataset used required us to take inputs and predict the accuracy of the salary model.

SEP 767: Multivariate Statistical Methods for Big Data Analysis and Process Improvement

Yash Dham 9.12.19
400279698 Monday

Project Report

Methodology

So, To Analyze the Prediction Model, I Used Two Methods, PCA and PLS:

PCA

So, for Principal Component Analysis I opted for the method of Regression applying 10-fold Cross Validation to find out mean Square Error. The Data was Initially Used to Find the PCA where only the necessary data was used after reduction and all the unnecessary data was left out. Then, we tried to perform 10-fold cross validation in which we divided data into 10 random segments and did PCA on it. The Results were then added to the mean square error algorithm to find the error matrix which we will store and move to PLS to compare results from its Performance. This PCA was found to be accurate enough for a Considerable model with a good R-square and the Mean square error to be approximately around 14%. This was then passed on to the training and testing dataset phase where we split the data into two equal parts with random distribution where the first half was used for training and the second half was used for testing. The Final Results were good enough for a PCA model to be considered as Ideal. Still PLS was needed to cross verify the Results.

SEP 767: Multivariate Statistical Methods for Big Data Analysis and Process Improvement

Yash Dham 9.12.19
400279698 Monday

Project Report

Methodology

So, To Analyze the Prediction Model, I Used Two Methods, PCA and PLS:

PLS

For Partial Least Squares We Used the predefined Libraries in sklearn to find out the PLS on the given dataset. Again, the target variable was Salary which was an independent variable and because of data we do not needed to do the training and testing of dataset for PLS. The Process was similar to PCA just that the mean square error was found to be near about 25% for the first time. But then the glitch was found to be in the formula and was figured out how it should work. Once the error was solved, we got the mean square Error for PLS to be around 14% again which was similar to that of PCA which showed that our dataset used was correct and up to the mark and no more changes were required.

Results

PCA

Mean Square Error: 14.27%

PLS

Mean Square Error: 14.58%