Report

12 October 2021

Regression

1 . Data: our data set is of SpaceKind company who focused on finding and propagating life on other planets.

Only Two rows are given X_inv and Y_inv and we have to focus on the value of Force

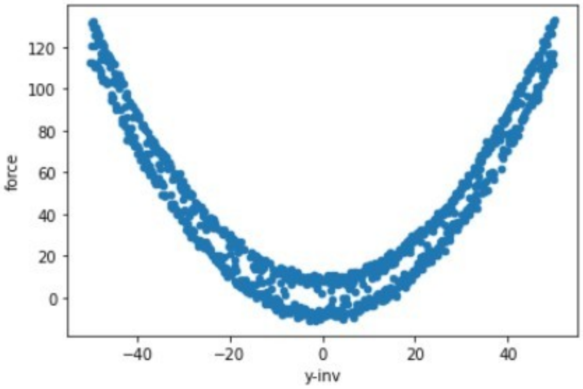
2. Job 4: Predict Force

Data Preprocessing and Cleaning :

1. Null values were replaced with median value .
2. Data were standardized such that to decrease the influence of the columns having a high range of data in the output. Every column now had data in the range of 0 to 1.
3. Used box plots to identify outliers, removed data points having values greater than 3 rd quartile.
4. Splitting dataset .

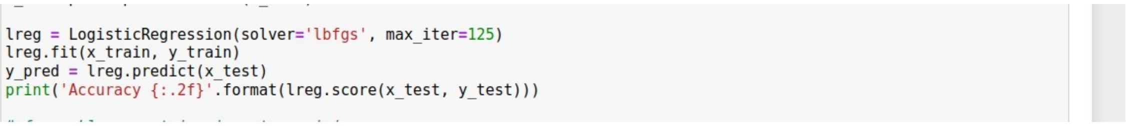
Using linear regression:

1. First we have tried linear regression but the data hyperbolic in nature
2. So the accuracy is coming in -ve



Using logistic regression:

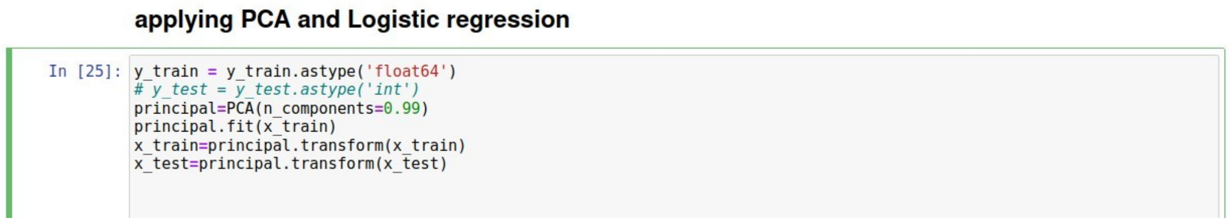
1.Use sklearn to apply the Logistic Regression model and get an accuracy score of 96% but the roc_auc score is very low, which suggests that there are many false positives in our model.



```
lreg = LogisticRegression(solver='lbfgs', max_iter=125)
lreg.fit(x_train, y_train)
y_pred = lreg.predict(x_test)
print('Accuracy {:.2f}'.format(lreg.score(x_test, y_test)))
```

Using PCA:

1. All the steps are the same as above. The only different thing that we tried now is to apply PCA in our dataset. As our dataset is pretty large we tried to reduce it using PCA and by that the precision score increases to 0.5



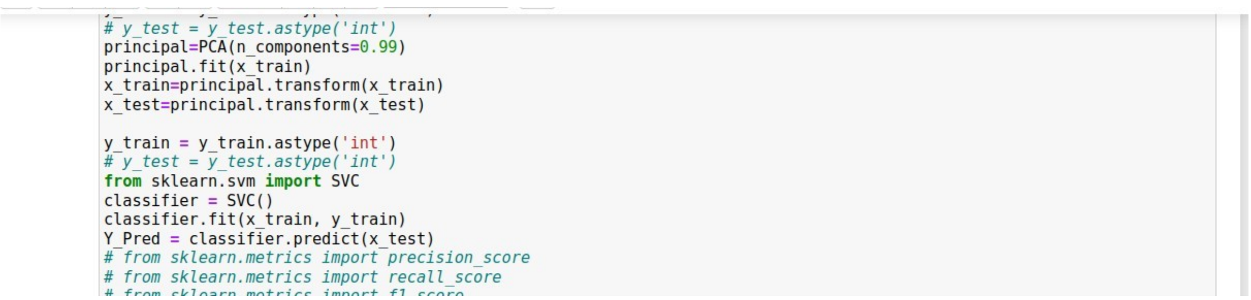
applying PCA and Logistic regression

```
In [25]: y_train = y_train.astype('float64')
         # y_test = y_test.astype('int')
         principal=PCA(n_components=0.99)
         principal.fit(x_train)
         x_train=principal.transform(x_train)
         x_test=principal.transform(x_test)
```

But still the accuracy was not good .so we have used SVM.

Using PCA: and SVM :-

1. In this to improve accuracy, I did one extra step. I remove outliers as data is too big and there are outliers also so I remove them and clean the data a bit. After that when I apply SVM.



```
# y_test = y_test.astype('int')
principal=PCA(n_components=0.99)
principal.fit(x_train)
x_train=principal.transform(x_train)
x_test=principal.transform(x_test)

y_train = y_train.astype('int')
# y_test = y_test.astype('int')
from sklearn.svm import SVC
classifier = SVC()
classifier.fit(x_train, y_train)
Y_Pred = classifier.predict(x_test)
# from sklearn.metrics import precision_score
# from sklearn.metrics import recall_score
# from sklearn.metrics import f1_score
```

We have applied different different method but the accuracy is 0.92