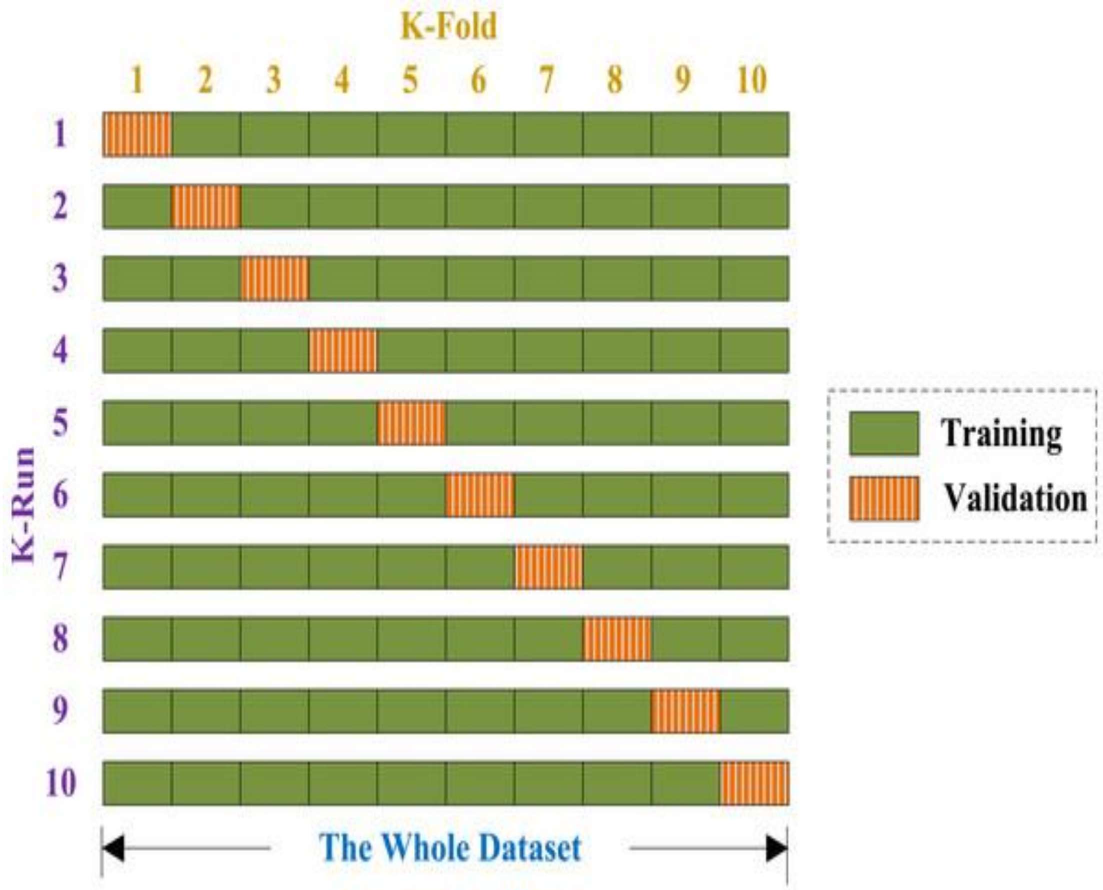
Multiple Linear Regression

Cross Validation - II

Contents

1. K-Fold Cross Validation
2. Repeated K-Fold Cross Validation
3. Leave One Out Cross Validation

K-Fold Cross Validation



- In k-fold cross-validation the data is first partitioned into k equally (or nearly equally) sized segments or folds
- Then k iterations of training and testing are performed such that each time one fold is kept aside for testing and model is developed using k-1 folds

- Model performance measure is aggregate measure based on above iterations

K-Fold Cross Validation in Python

```
# Import necessary libraries
```

```
import pandas as pd
import numpy as np
from sklearn.model_selection import cross_val_score
from sklearn import linear_model
```

```
#Splitting data into X variables and Y variable
```

```
motor=pd.read_csv('Motor_Claims.csv’)

X=motor.drop(['claimamt'], axis = 1)
y=motor.claimamt
```

```
#Regression Model object is lm_reg
```

```
lm_reg = linear_model.LinearRegression()
```

K-Fold Cross Validation in Python

```
# Perform K fold cross validation with K=4 and report R squared values
cv_r2_scores_lm = cross_val_score(lm_reg, X, y, cv=4,scoring='r2')
print(cv_r2_scores_lm)
[0.75432031 0.72743804 0.69157553 0.7365273 ]
print("Mean 4-Fold R Squared: {}".format(np.mean(cv_r2_scores_lm)))
Mean 4-Fold R Squared: 0.7274652945788245

cv_rmse_scores= cross_val_score(lm_reg, X, y, cv=4,
scoring='neg_mean_squared_error')
np.sqrt(-(np.mean(cv_rmse_scores)))
11461.73251088066
```

R squared value

RMSE value

❑ *cross_val_score computes cross validation score*

❑ *cv=4 sets number of K-folds.*

❑ *scoring=‘r2’ gives r squared value ; 'neg_mean_squared_error’ gives negative mean squared error*

Interpretation :

- *R^2 of the original model is 73.19%*
- *RMSE for the original is model is 11444.51*
- *Comparing the RMSE values, we can say that the model is stable*

Repeated K-Fold Cross Validation

- As the name suggests, repeated k-fold cross validation technique **undertakes cross validation and repeats the process m-number of times**
- This ensures that more robust measure of model performance is generated
- **K-fold** is repeated m times with different randomization in each repetition

For instance,

- Five repeats of 10-fold cross validation will generate 50 total resamples.
- These results are again averaged to produce a single estimate
- This is not the same as 50-fold cross validation

Repeated K-Fold Cross Validation in Python

```
#Creating ‘5’ Folds and ‘5’ repeats
```

```
from sklearn.model_selection import RepeatedKFold
rkfold = RepeatedKFold(n_splits=5,n_repeats=5)
```

- *RepeatedKFold() is used to prepare the cross-validation procedure for implementation of repeated k-fold cross-validation.*
- *n_splits = specifies the number of folds.*
- *n_repeats= specifies the number of repeats.*

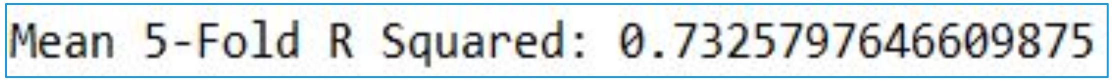
```
# Finding R squared value & RMSE value
```

```
cv_r2_repeated = cross_val_score(lm_reg, X, y, cv=rkfold)
print("Mean 5-Fold R Squared: {}".format(np.mean(cv_r2_repeated)))


cv_rmse_repeated= cross_val_score(lm_reg, X, y, cv=rkfold,
scoring='neg_mean_squared_error')
np.sqrt(-(np.mean(cv_rmse_repeated)))
```

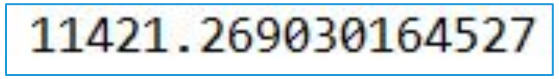
Repeated K-Fold Cross Validation in Python

```
# Output
```



```
Mean 5-Fold R Squared: 0.7325797646609875
```

R squared value



```
11421.269030164527
```

RMSE value

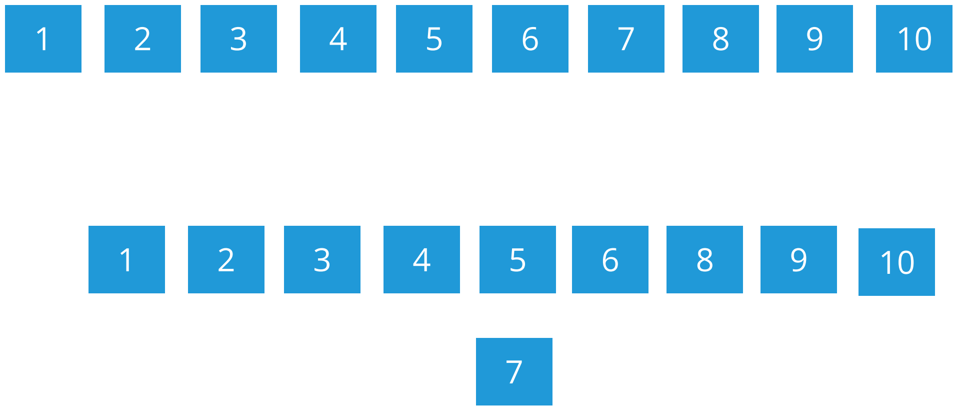
Interpretation :

- *R^2 of the original model is 73.19%*
- *RMSE for the original is 11444.51*
- *RMSE values of the cross validated model indicates stability.*
- *As the observations are selected randomly, output may vary slightly.*

Leave One Out Cross Validation (LOOCV)

- **LOOCV** is a special case of k-fold cross validation where **k equals the sample size (n)**
- Each time one observation is kept aside and the model is developed on the remaining data.
- The left out observation is predicted using the model.
- This process is repeated n times
- RMSE is computed based on these predicted residuals

- Sample size is 10 and one observations (say 7) is chosen to be kept aside
- The model is developed on the new sample with n=9 and observation 7 is predicted



Quick Recap

K-Fold Cross Validation

- Data is first partitioned into k equally (or nearly equally) sized segments or folds
- Then k iterations of training and testing are performed such that each time one fold is kept aside for testing and model is developed using k-1 folds

Repeated K-Fold Cross Validation

- This is an extension of k-fold method wherein the process is repeated m number of times