3.

Network Backup Dataset

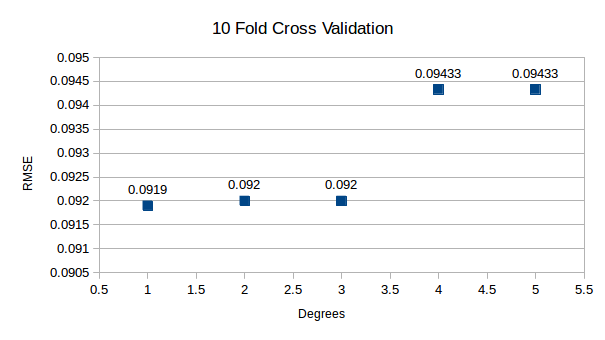
Piecewise Linear Regression Model

The dataset is splitted into training dataset(90%) and testing dataset(10%).Then separate the instances of the training dataset into five workflows and train 5 different Linear Regression models corresponding to each of the workflows separately. So, we are training a piece-wise linear regression model in this case. After training the different models, we then predict the labels on testing dataset from these trained models and calculate the RMSE. The RMSE on the testing dataset for this model comes out to be 0.05033662.

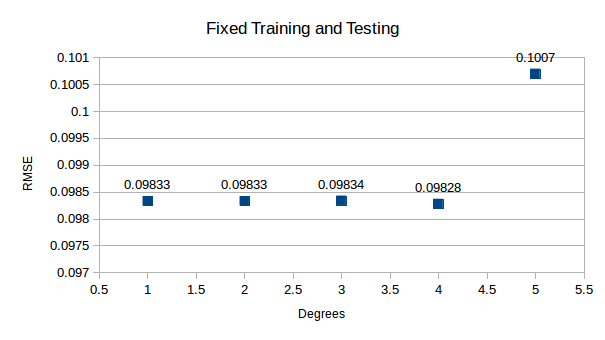
The RMSE in this case is less as compared to the linear regression model. So the fit is improved in this case.

Polynomial Regression Model

We used one-hot encoding to featurize the dataset.This increased the feature vector length to 64. So, for doing the polynomial regression, we have taken only the significant variables obtained from the linear regression model. Initially, the length of the feature vector was 64(after doing one-hot encoding), then we took 16 significant variables out of them and trained polynomial regression model for different degrees.

a) Graph plotted for average RMSE of the 10 Fold Cross Validation values

b) Graph plotted for a RMSE values for a fixed training and testing dataset



In the graph we can see that, for increasing degrees of polynomial function we are getting nearly same RMSE values.(This is after we have used one-hot encoding and took 16 significant variables from that.) Computational time for higher degrees was very high, so that was not performed. So according to the results using degree 5 will not give poor generalization. So we can fit our data on a polynomial function of degree 5.

Complexity of a model increases as it tends to overfit the training data. When model overfits the data, the RMSE value will be large on the testing dataset. When we do cross-validation we get to know that the model is overfitted, as we get high RMSE value. So, we can control the complexity of the model.

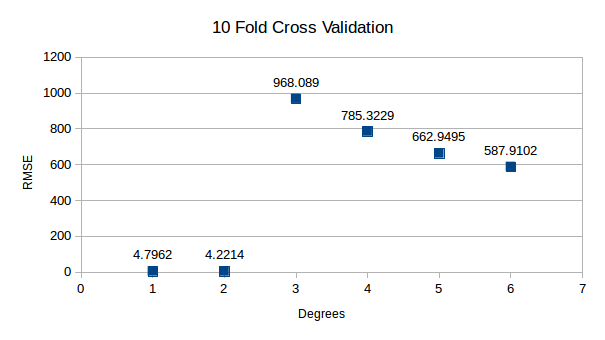
So cross-validation helps us to know whether the model is overfitted or not. An overfitted model will have more complexity and poor generalization.

Housing Boston dataset

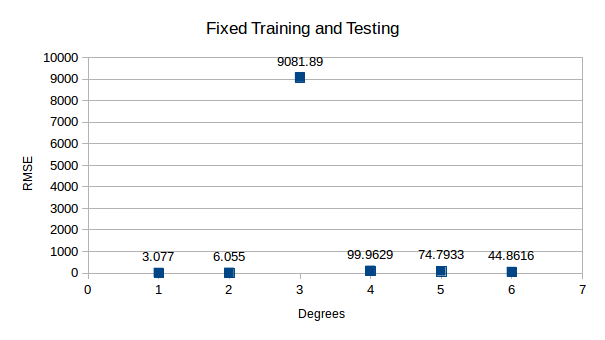
Polynomial Regression Model

This dataset was also used to fit the polynomial regression model for various degrees.

a) Graph plotted for average RMSE of the 10 Fold Cross Validation values



b) Graph plotted for a RMSE values for a fixed training and testing dataset



After degree 2, the RMSE values are large and the generalization of the model gets worse. So the optimal degree of fit is 2.