In the trainhelp & test.py for assignment submitted on
19 Feb; all the lines were written by me &
in the post was written by me. I understand that
cases of copying lines that deviate from the
signed declaration as liable to be awarded F
grade

B Sharma

EE 769: Introduction to Machine Learning
Assignment#1

Name: Balraj Parmar
Roll No: 14D070001

Date: 19 February 2018

Answer 1: Data Preprocessing

The training data contained lot of NA values in certain columns namely MasVnrArea and GaragaYrBlt, those had to bbe replaced by 0.

Id column was removed as it does not have any important information associated with SaleStatus

Data was then normalized (mean made to be zero and variance = 1)

Data was then divided into 60% training set, 20% cross validation test and 20% pseudo test set

For the test data, the NA values from numerical features was replaced by zero. It was also ensured that the number and order of columns in both the training and test set is same by adding new columns if necessary

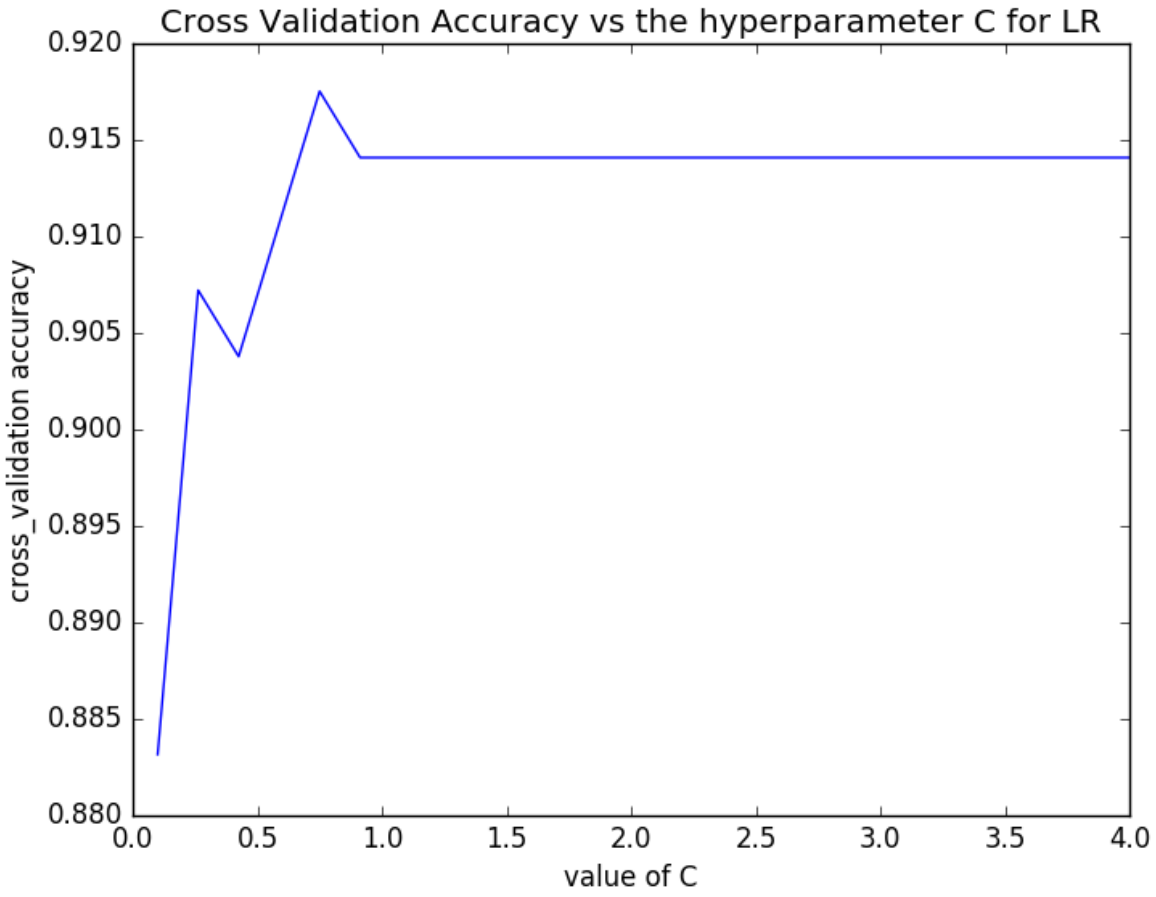
In my experience the data preprocessing part was the most clumsy and time consuming part as there were lot of things to be taken care of like NA values, order of the features and so on.

Answer 2: Random Forest worked best for me. I tried using SVM (with RBF kernel), logistic regression and Random Forest.
A lot of features were categorical in nature, RF is an ensemble of decision trees which works well with such data. As it is at the end of a day a decision maker at each node. One hot encoding of categorical features helped this nature of Rfs.

Logistic Regression is in general too simple an algorithm to give best accuracy.

Answer 3:

Hyperparameter tuninng for Logistic Regression



Data set was divided into 60% training, 20% CV and 20% pseudo test set.
The algorithm was trained on the training set, hyperparameters were tuned on CV set and the final testing to determine performance was done on the pseudo test so that the final result does not have effect of overfitting from both the training and the hyperparameter tuning.

1. Logistic Regression;
The parameter 'C' which is inverse of lambda was varied linearly and the value which performs best on CV set was chosen. You can observe the graph above.
Accuracy obtained for the pseudo test test is in the range of 86-90%

2. SVM:
The parameters C and gamma were varied in the exponential range and the C, gamma pair which gives the best result on CV was chosen. Pseudo test set accuracy in range 89-94%

3. Random Forest:

n_estimators, max_depth, min_samples were varied in suitable ranges and a random sample of all the possible values was used for hyperparamter tuning with the help of RandomizedSearchCV

function of scikit-learn
Accuracy on pseudo test set is 93-96%