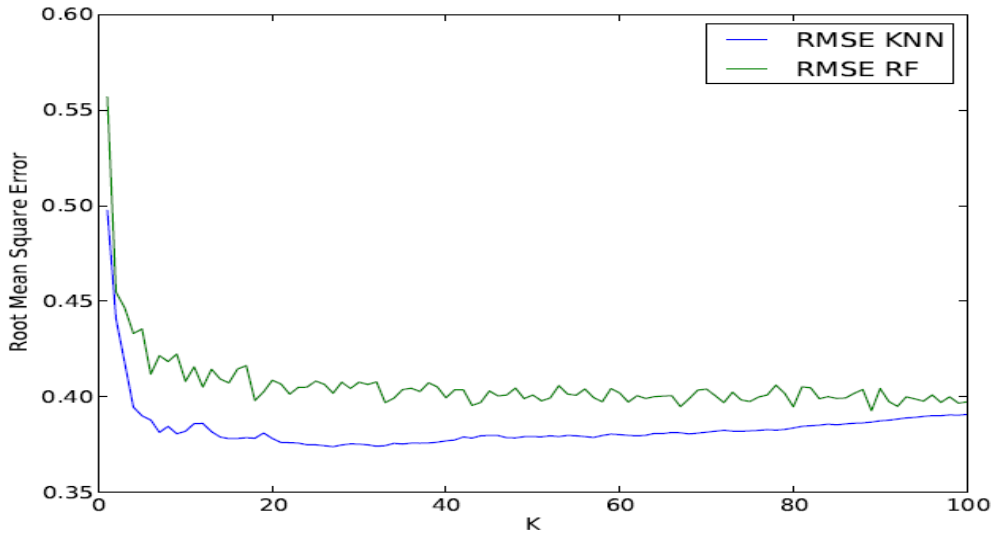
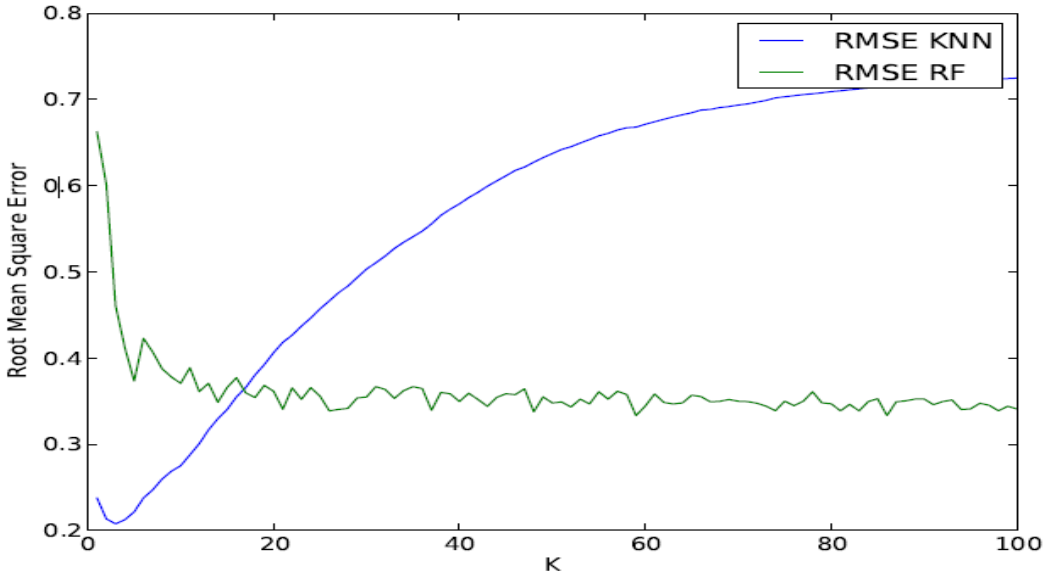
Student Name: Aishvarya Krishnan
Course Name: CS7646- Machine Learning for Trading
Homework Name: 2014Fall7646 Project 3

Charts:

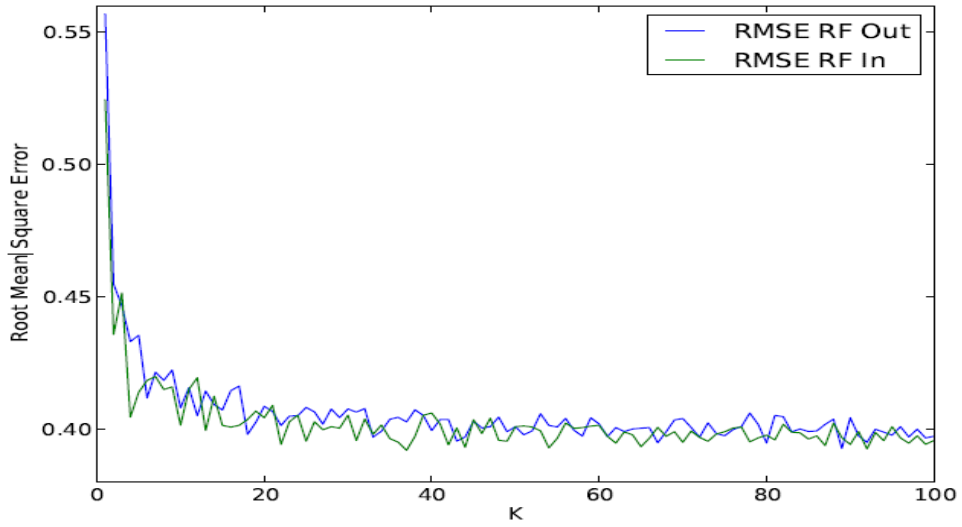
1) RMS Error – KNN and Random Forests – data-classification-prob



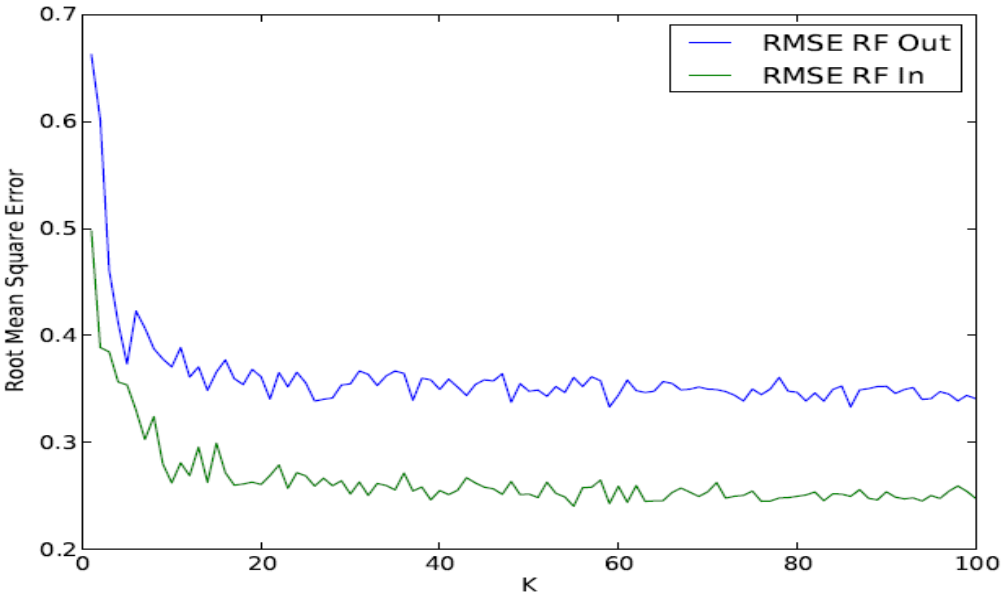
2) RMS Error – KNN and Random Forests – data-ripple-prob



3) Random Forest – In Sample vs Out of Sample – data-classification-prob



4) Random Forest – In Sample vs Out of Sample – data-ripple-prob



KNN vs Random Forest – Which approach is more subject to overfitting?

KNN is more subject to overfitting at lower values of 'k' since it takes into consideration and hence affected more by noise and hence leads to overfitting. This is evident from the varying values of RMS errors for in sample and out of sample in the plot. Whereas in case of Random Forest algorithm, comparitively we have lower chances of overfitting and chances of overfitting decreases with increasing number of trees.

Improvement in performance using more trees? Which data set and why?

Yes, I noticed an improvement in the performance of Random Forest Learner with increasing number of trees in the data-ripple-prob data which is possibly because ripple data has more fluctuating values and Random Forest Learner averages out the values with increased number of trees.

KNN vs Random Forest vs Linear Regression – Which is best and why? Dependant on dataset? And why?

Among the three approaches, Random Forest is comparatively the best approach since it gives lower RMS error especially while using more number of trees. KNN works well with larger values of 'k' for less fluctuating data sets. Random Forest works well with more number of trees and for especially for ripple data. Linear Regression is used to show linear dependence of the data and hence works well only for such datasets thus, its applications are limited.