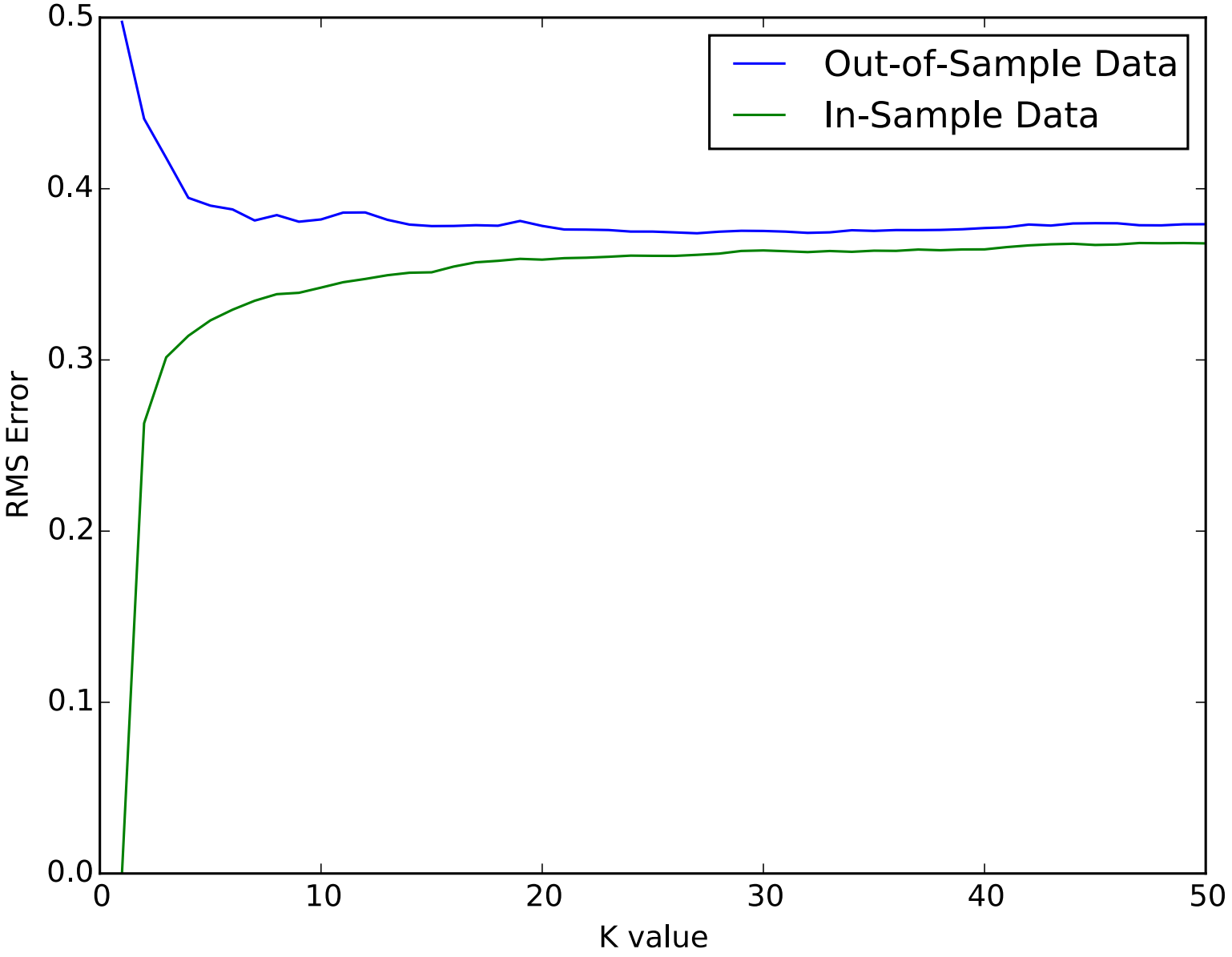
CS 4803/7646 – MLT (Machine Learning for Trading)

Project Name: Project 2

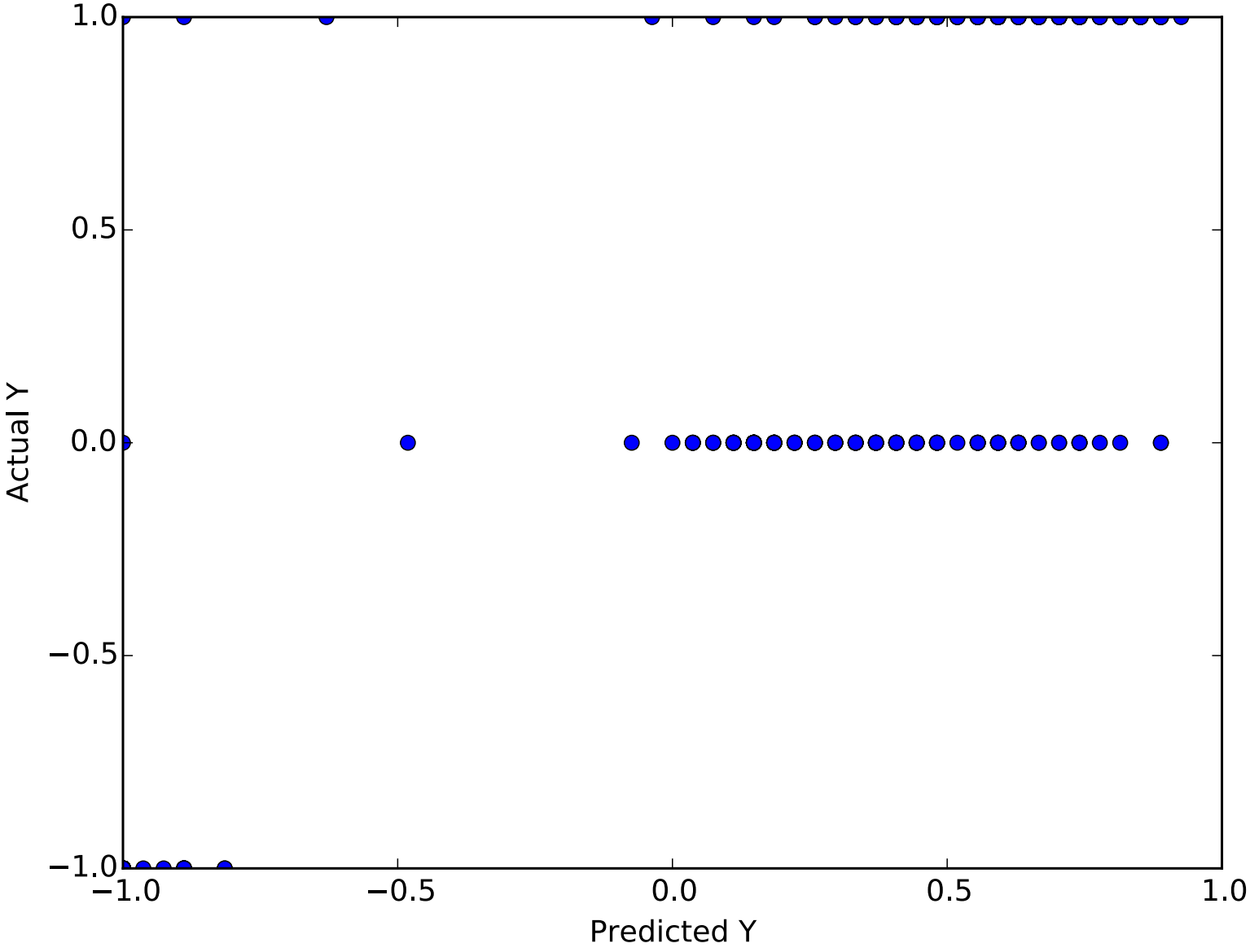
Student Name: Utkarsh Garg

GTID: 902904045

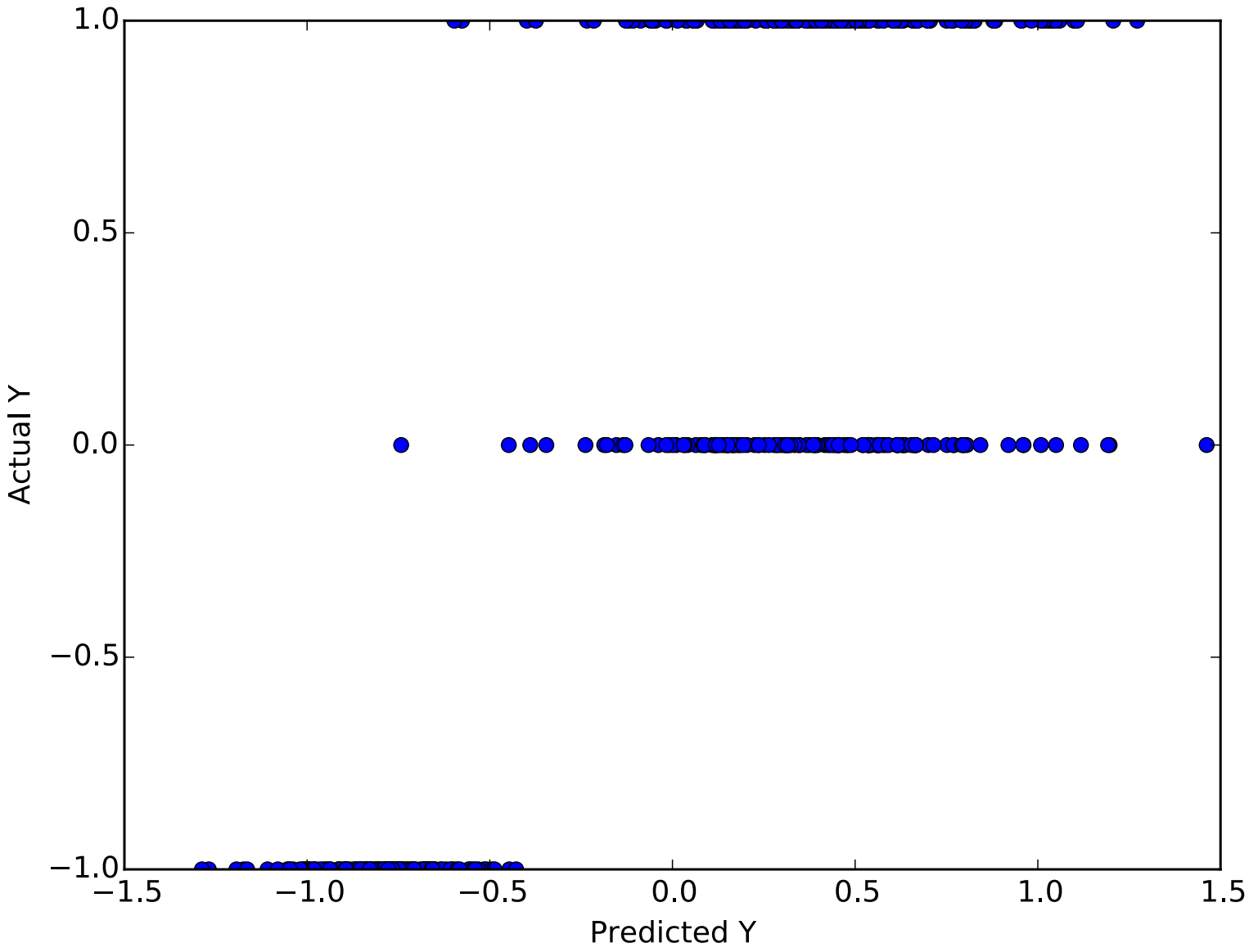
Graphs for ‘data-classification-prob.csv’



K versus RMS Error representing the In-Sample and Out-of-Sample data



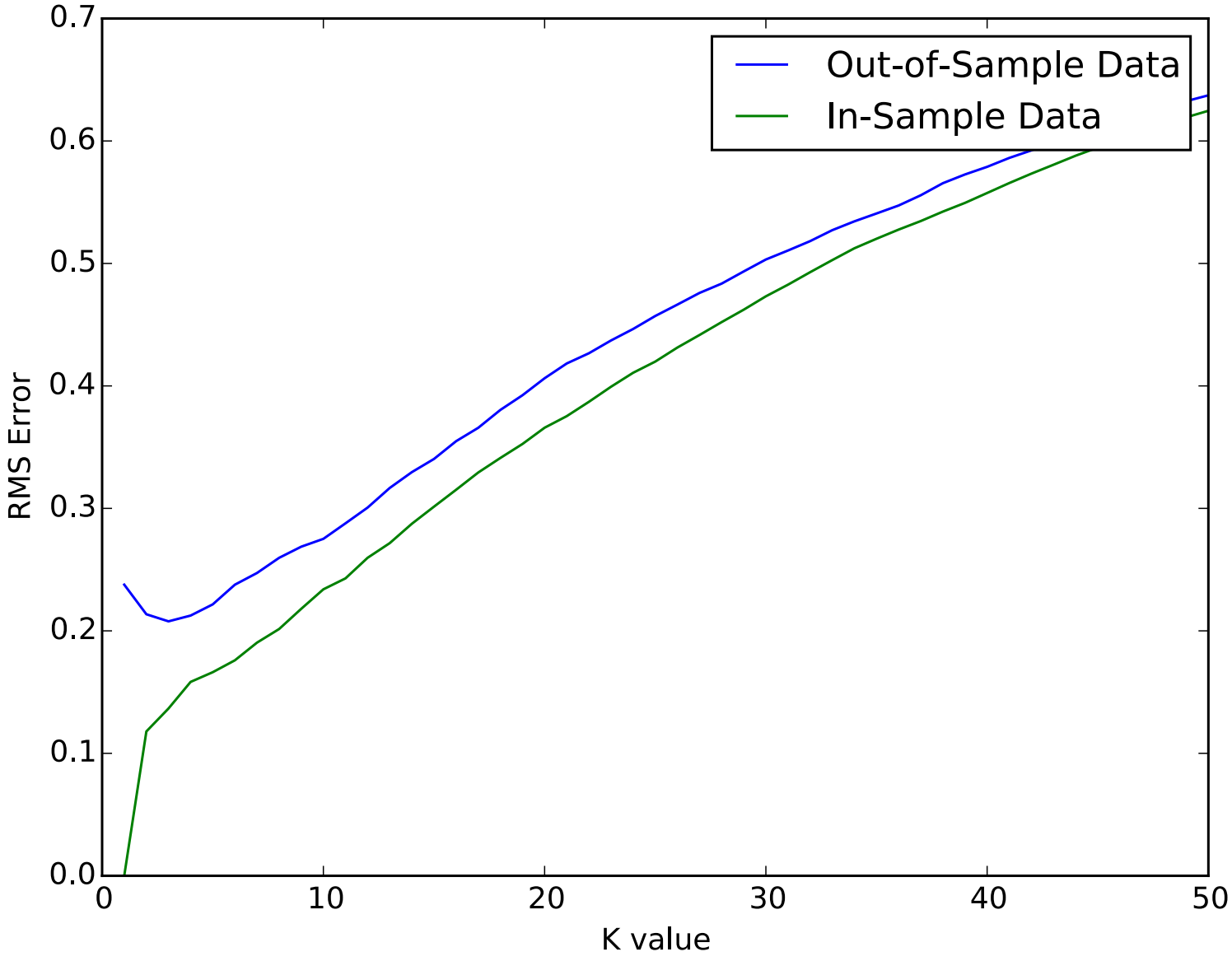
Scatter Plot of the Predicted Y versus the
Actual Y values for KNN Learner



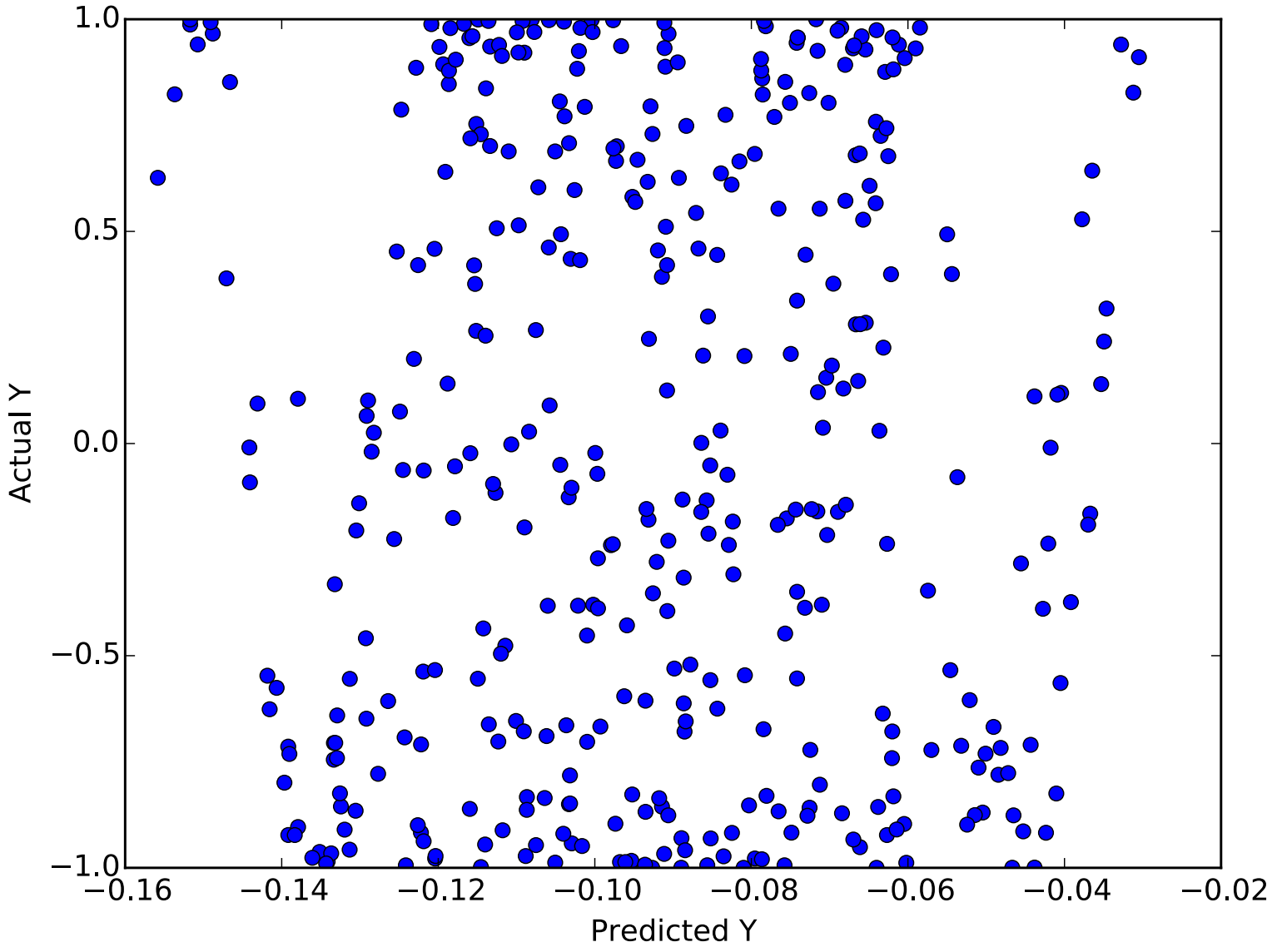
Scatter Plot for the Predicted Y versus the
Actual Y for the Linear Regresser

RMS Error: 0.515389956744
Correlation Coefficient: 0.774943176781

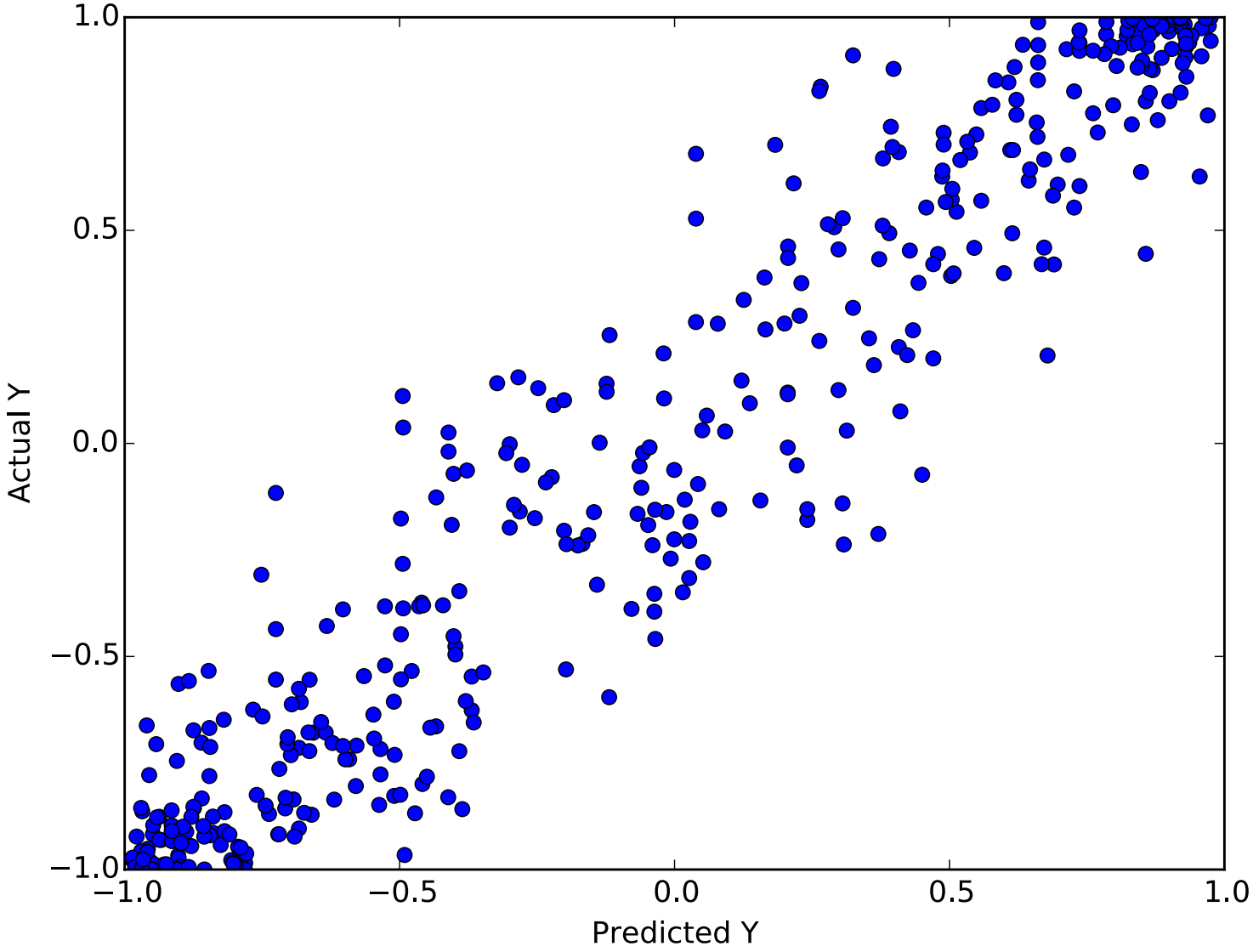
Graphs for ‘data-ripple-prob.csv’



K versus RMS Error representing the In-Sample and Out-of-Sample data



Scatter Plot of the Predicted Y versus the
Actual Y values for KNN Learner



Scatter Plot for the Predicted Y versus the
Actual Y for the Linear Regresser

```
RMS Error: 0.70409341685
Correlation Coefficient: 0.0162663457175
```

‘best’ k:

For both data sets, I plotted graphs for the rms error versus the K values and also a graph for the correlation coefficient versus the K values.

data-classification-prob: In this case, the rms error dips between 25-30 and the value of the correlation coefficient peaks in the same range. I chose 27, but I suspect that 26 or 28 would give almost the same results.

data-ripple-prob: In this case, k=3 clearly has the lowest rms error value and the highest correlation coefficient, thus being the optimal k.

Overfitting:

Yes, in both cases, overfitting occurs as k decreases. In the first data set, classification, overfitting starts occurring somewhere 25<k<27 and occurs as k decreases and in the ripple dataset, it occurs for k<3. I detected these k values by plotting the in sample and out-of-sample data and saw that around these values, the in-sample error started dipping while the out-of-sample error started peaking.

Overfitting occurs because for these data sets, below the identified k values, the number of neighbors being compared to becomes to small and thus the data is overfitted.

Code Disclosure: The code for the KNN Learner was inspired from the one on the Quant github link provided on the wiki.