IEE 598: DATA SCIENCE FOR SYSTEM INFORMATICS:

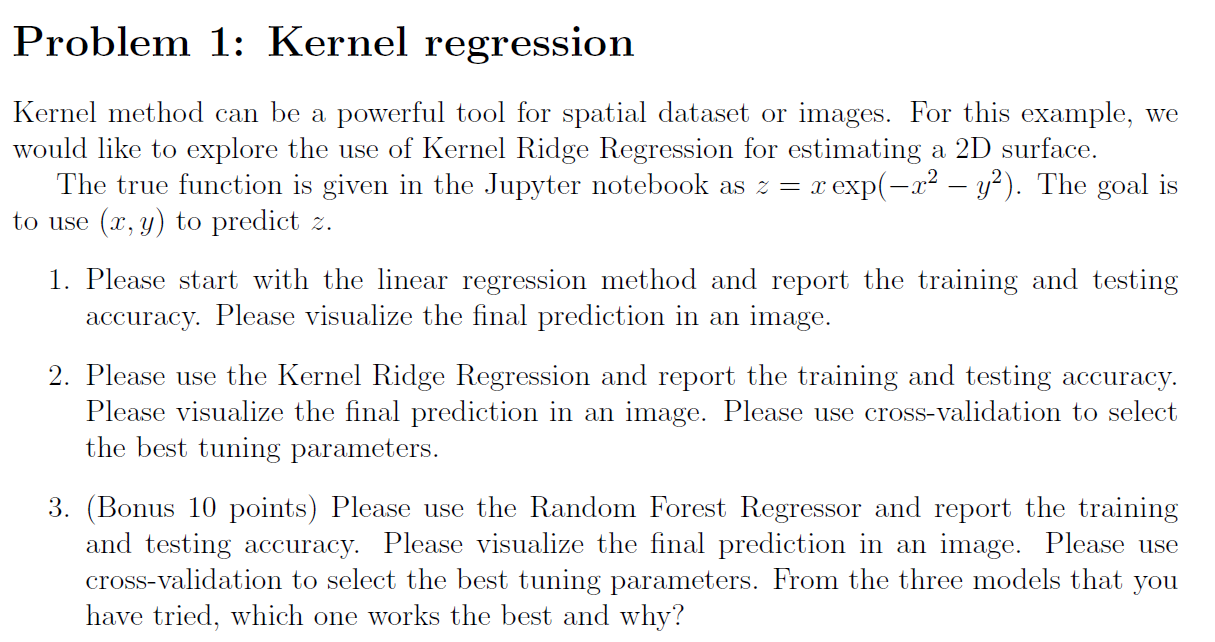
# ASSIGNMENT 3

HETUL VARAIYA

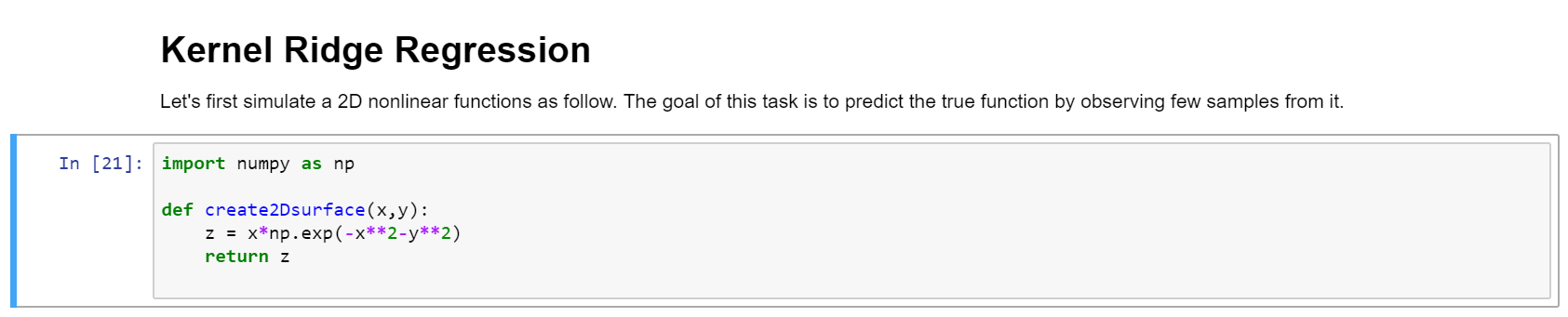
1211306106

Topics covered in this assignment:

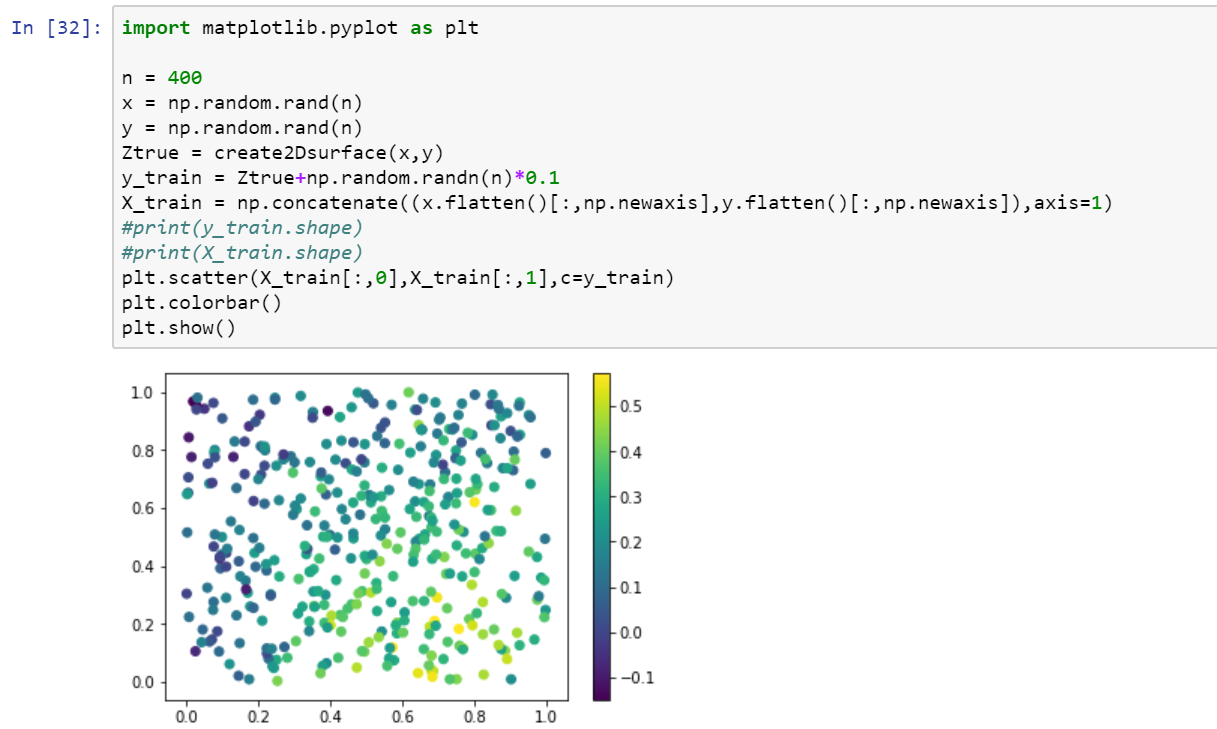
* Kernel Ridge Regression
* Linear Regression
* Random Forest Regression
* Training and Testing Split
* SVM
* Regularized SVM
* Logistic Regression
* precision/recall
* f1-score
* confusion matrix
* ROC Curve
* Area Under the curve
* Cross Validation using GridSearchCV



* Creating the function:



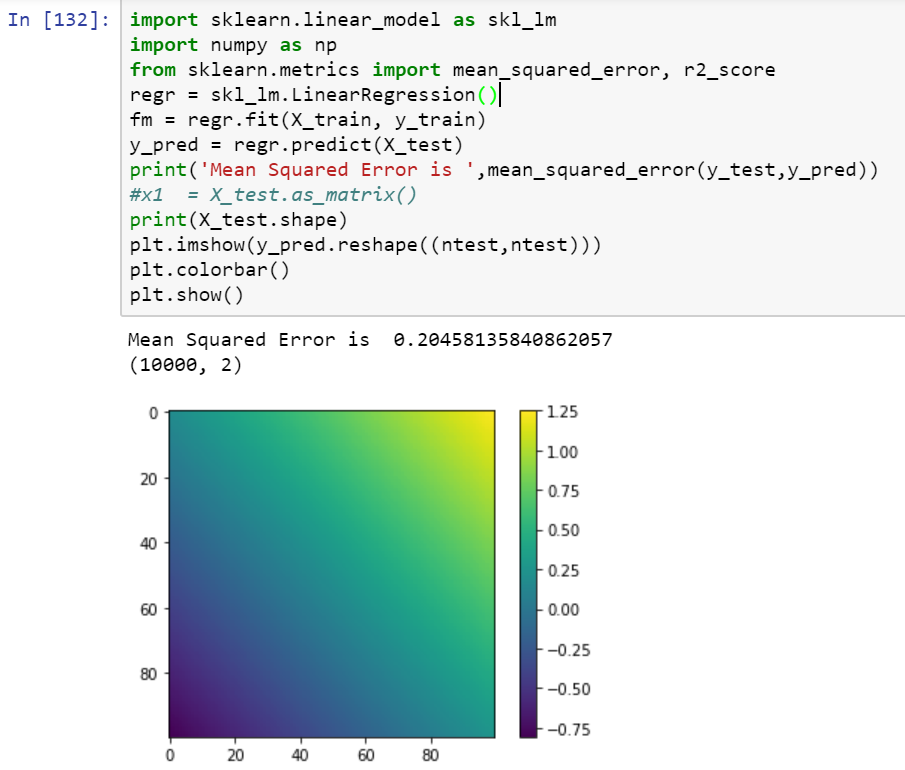
* Creating the Training Dataset and visualizing it:



* Create Testing Data:

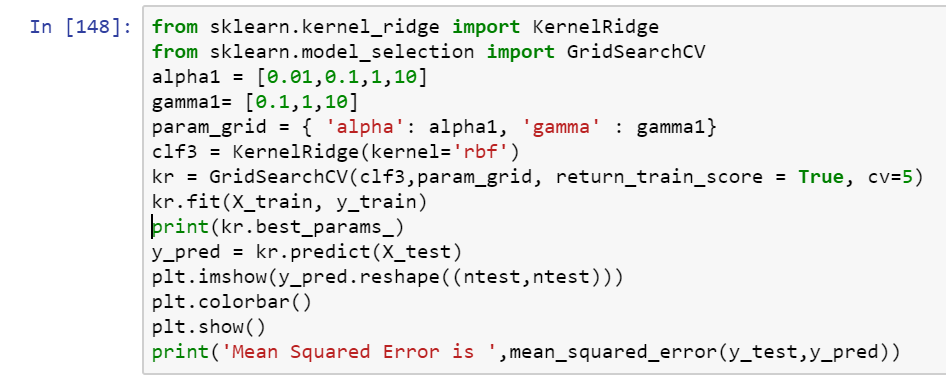


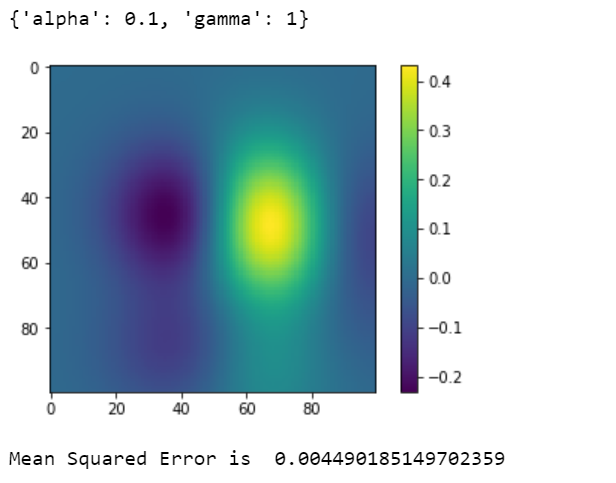
* Linear Regression on the Dataset with the Visualization and Mean Square Error:



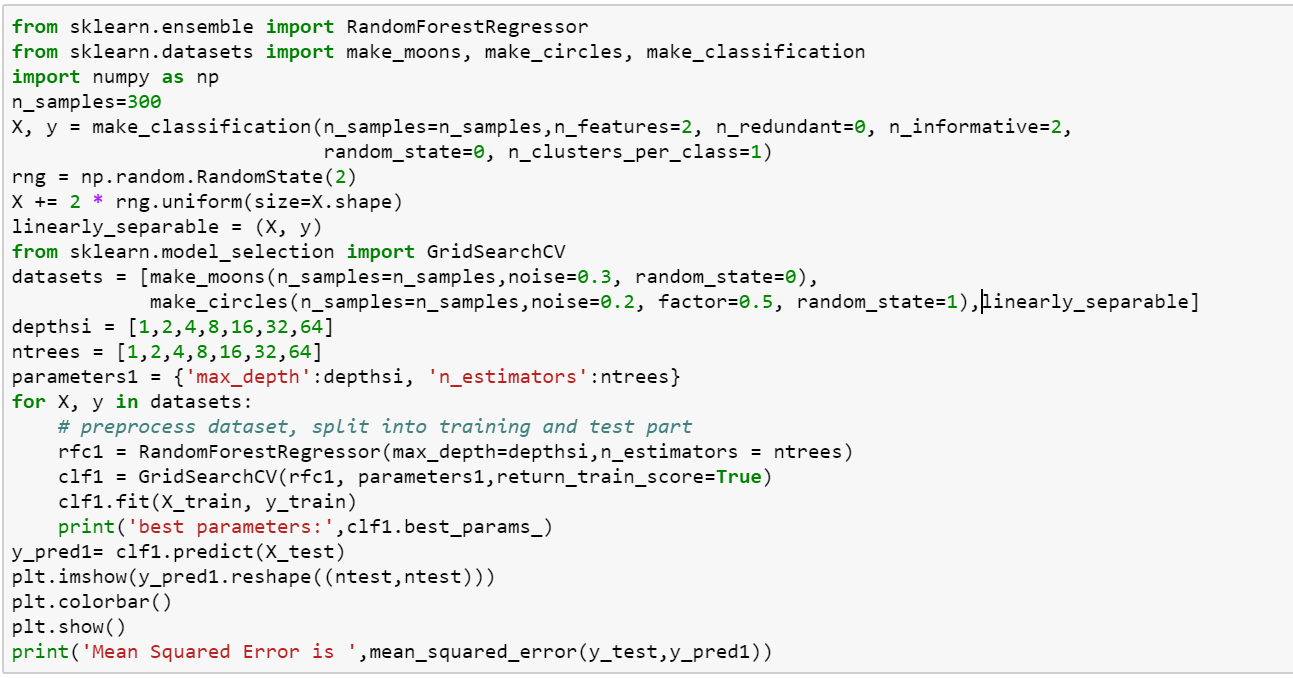
nfds

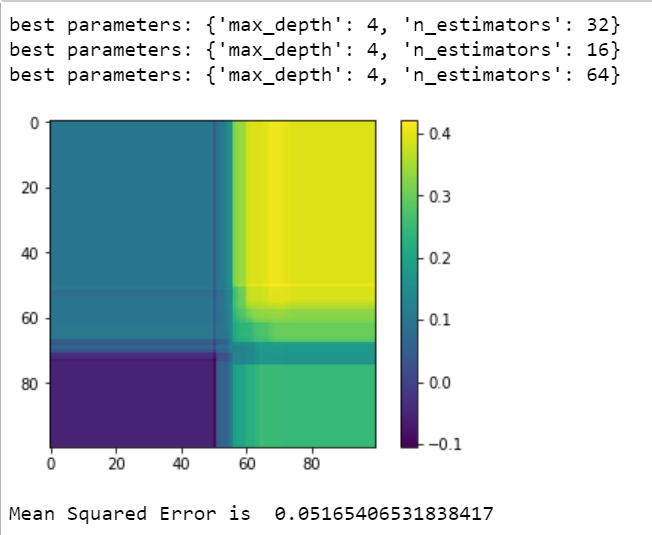
* Kernel Ridge Regression - Fitting the best parameters, visualizing the result, Mean Square Error:





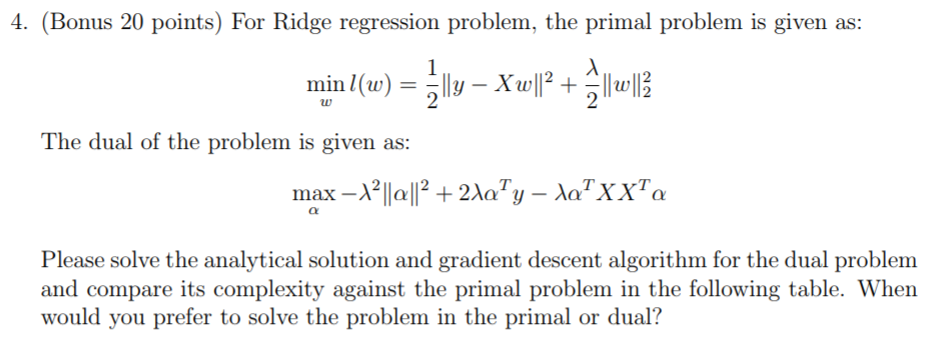
* Random Forest Regression:

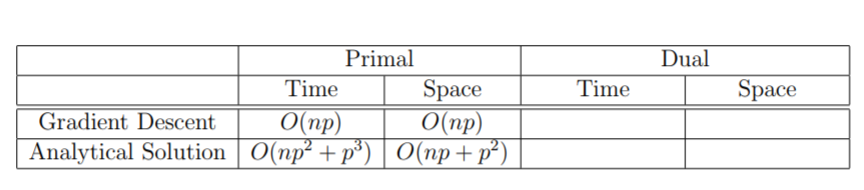


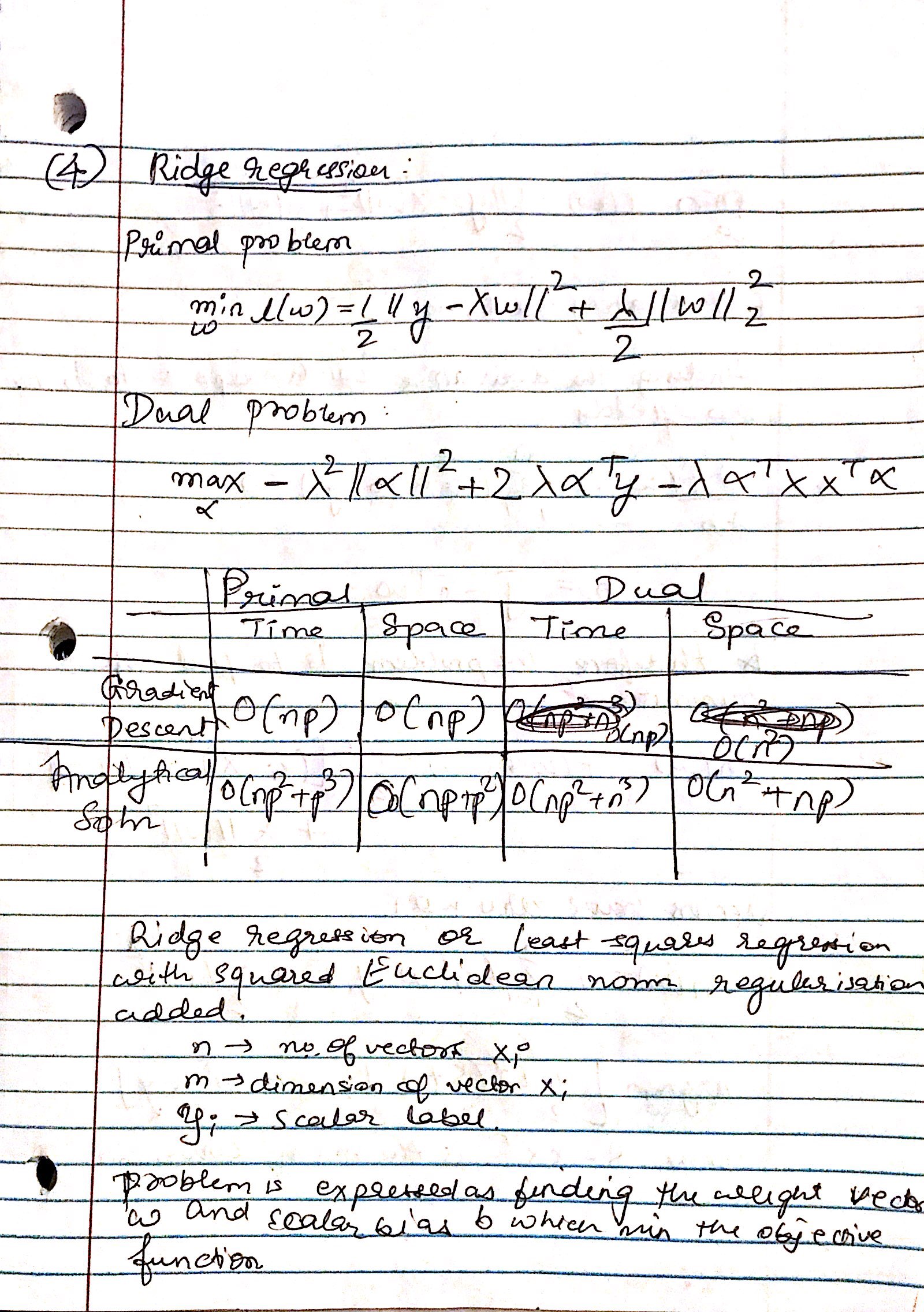


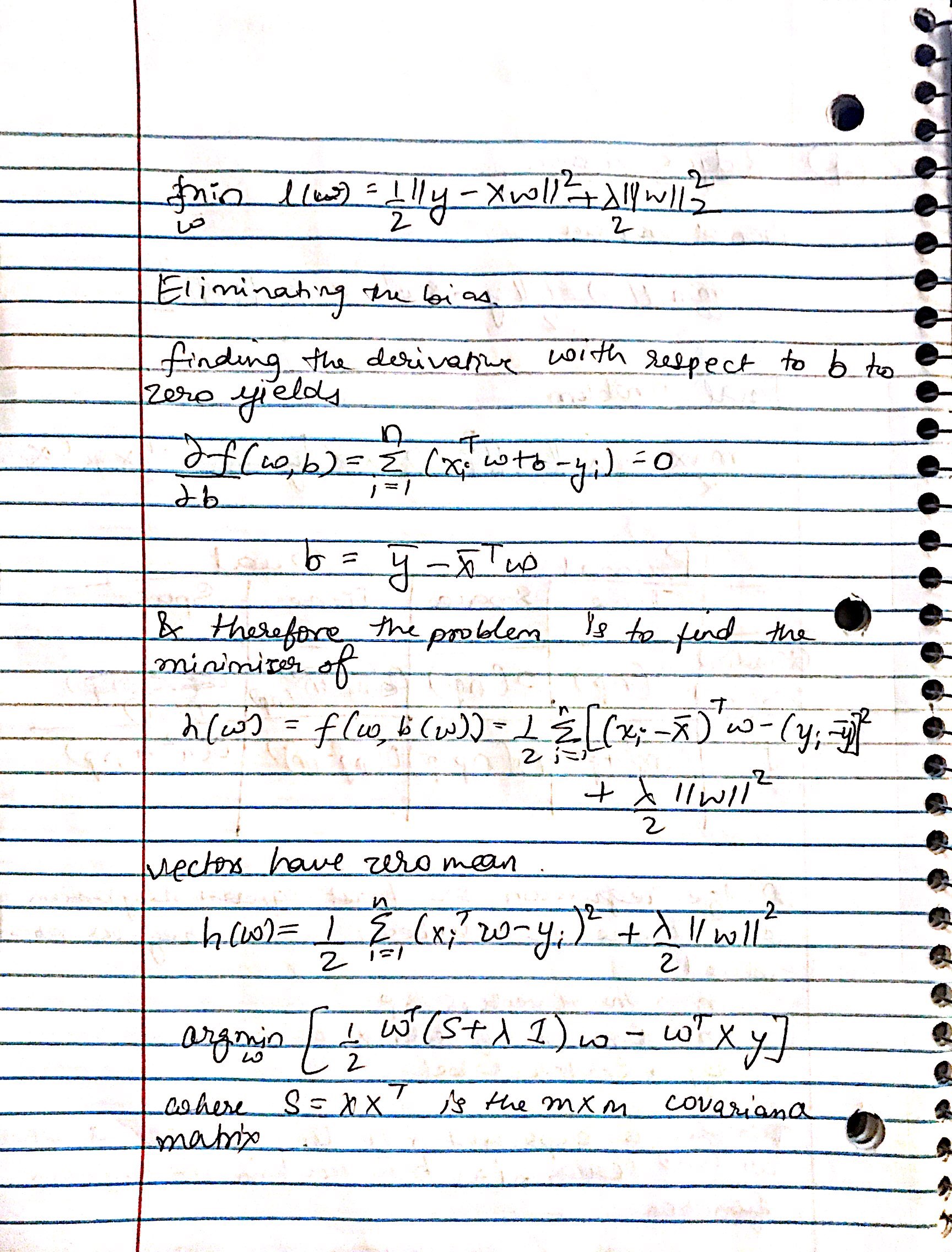
|  |  |  |
| --- | --- | --- |
| Model Name | Parameter Range | MSE |
| Linear Regression | None | 0.2045 |
| Random Forest | Depth: {1,2,4,8,16,32,64}  No of trees: {1,2,4,8,16,32,64} | 0.05165 |
| Kernel Ridge Regression | Alpha: {0.01,0.1,1,10}  Gamma: {0.1,1,10} | 0.0045 |

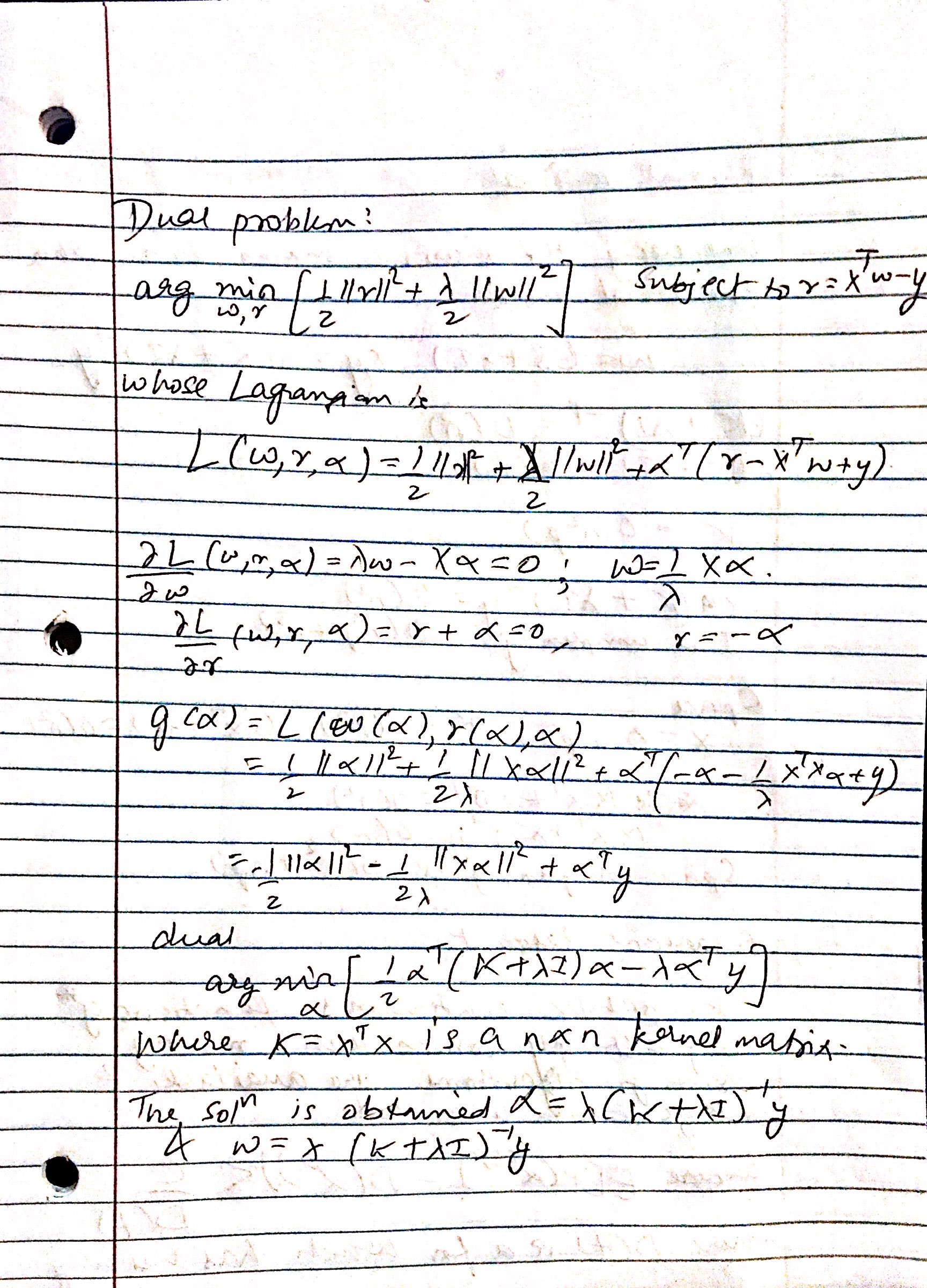
* From the above table is seen that Kernel Ridge regression produces the best model when the alpha value is 0.1 and the gamma value is 1 as it has the least mean square error.
* KRR uses squared error loss combined with l2 regularization.



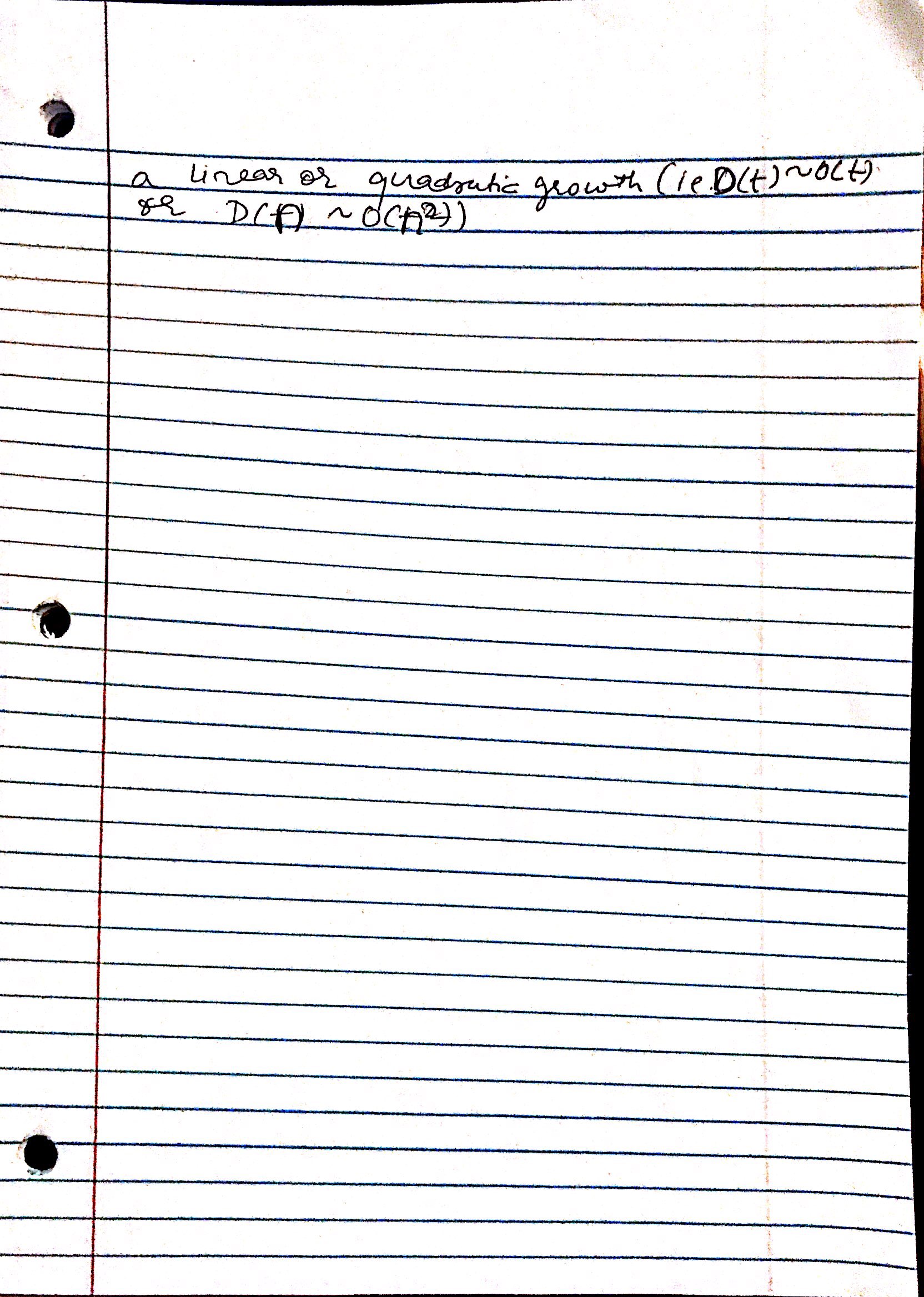


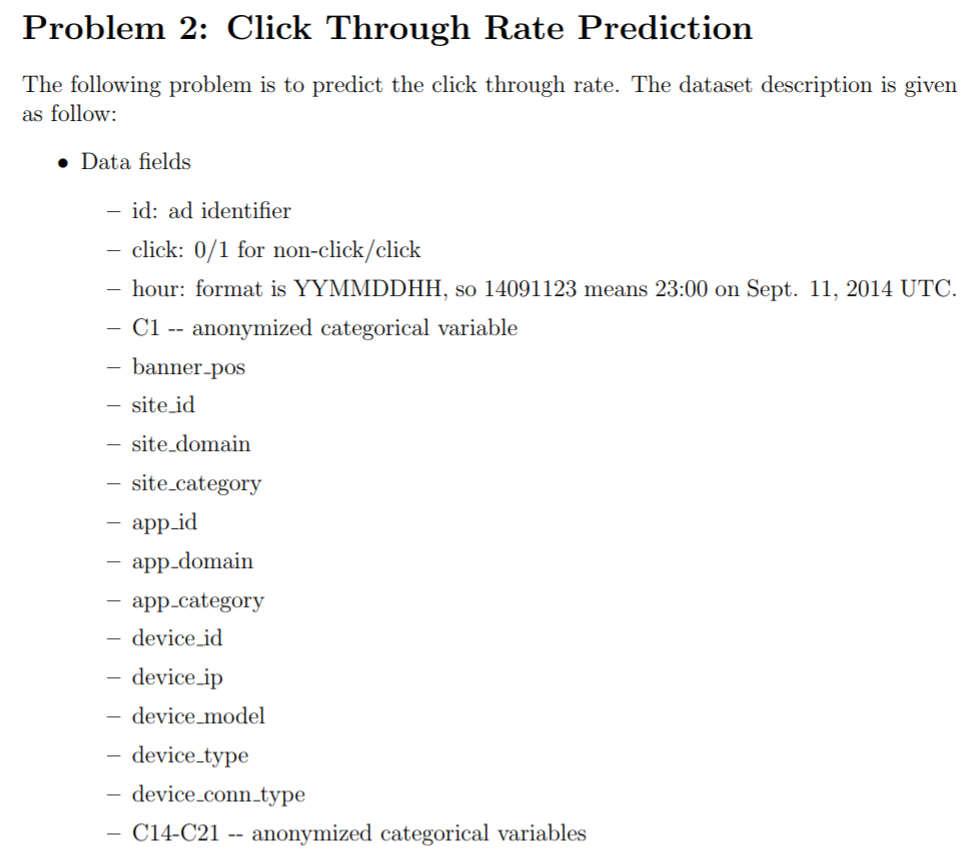


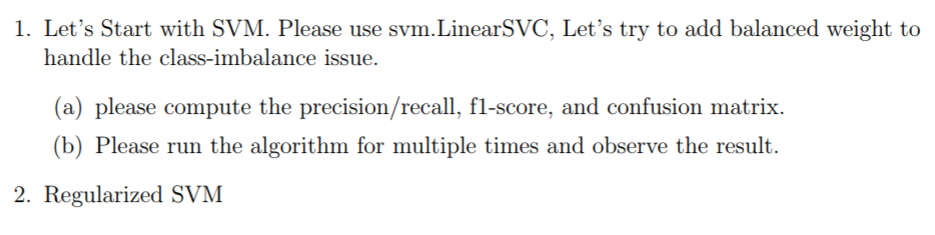


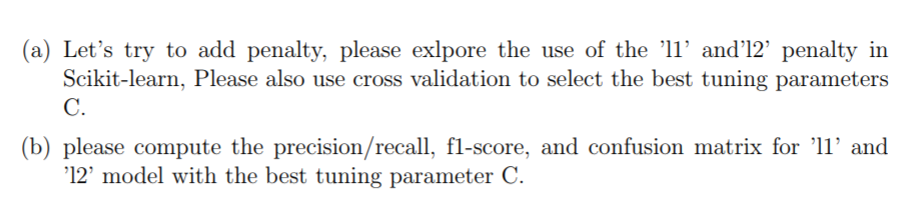


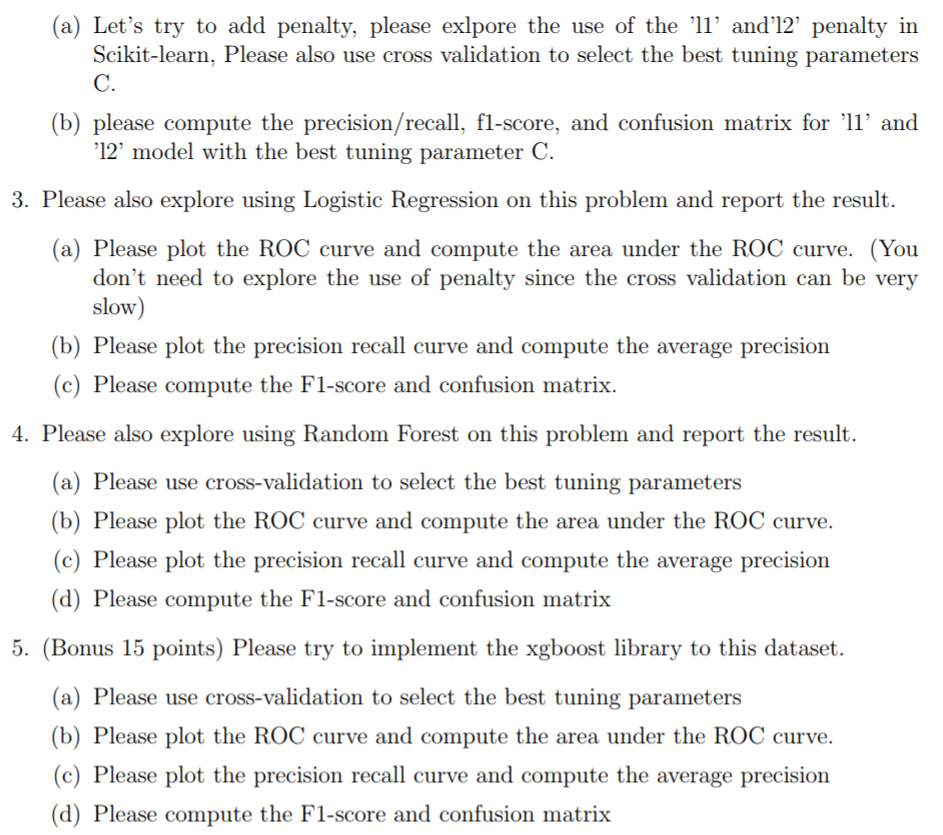








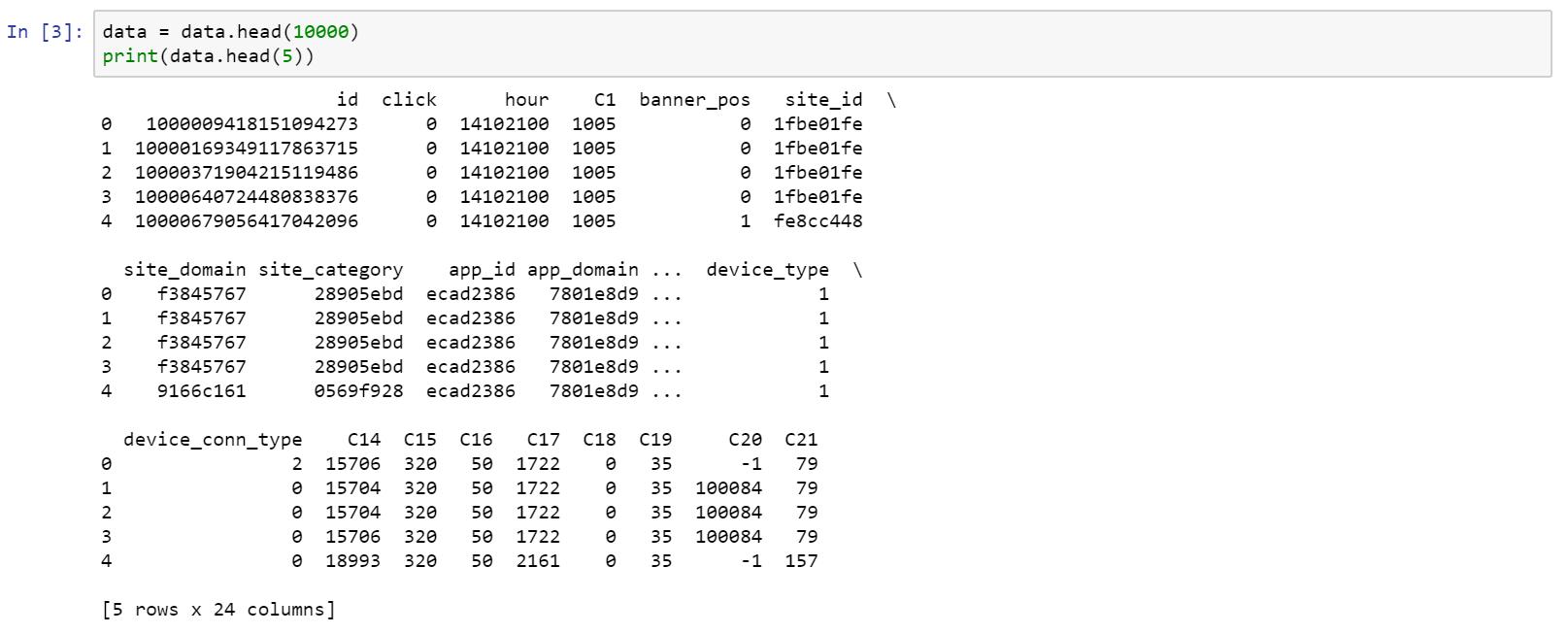




* Read the Data file:



* Overview of how the data looks like:

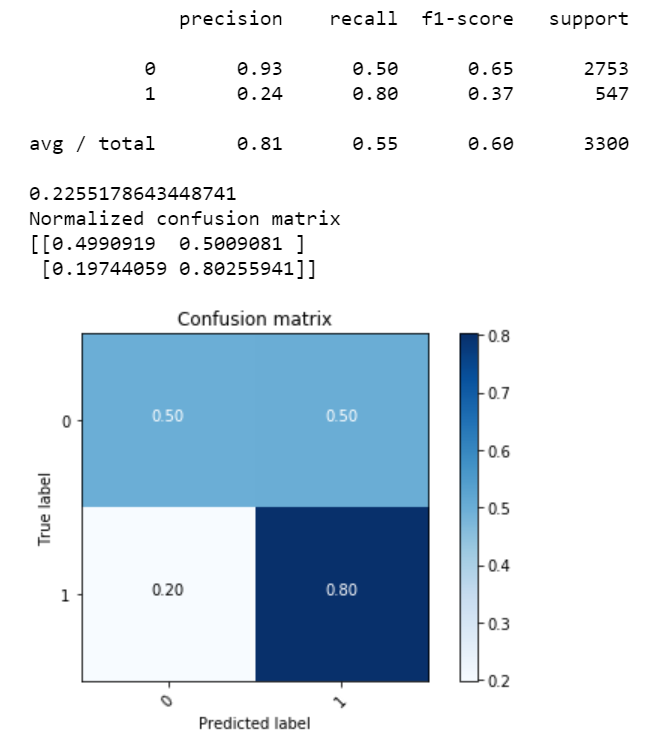


* Splitting the training and the testing data:



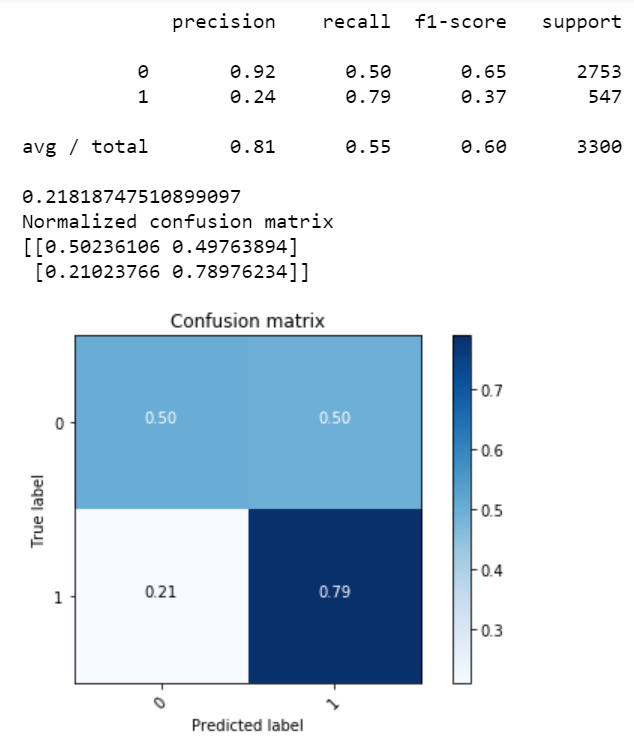
* Using SVM(svm.LinearSVC) with balanced weights to handle the class – imbalance issue:
* precision/recall, f1-score, and confusion matrix:

Running 1st Time:

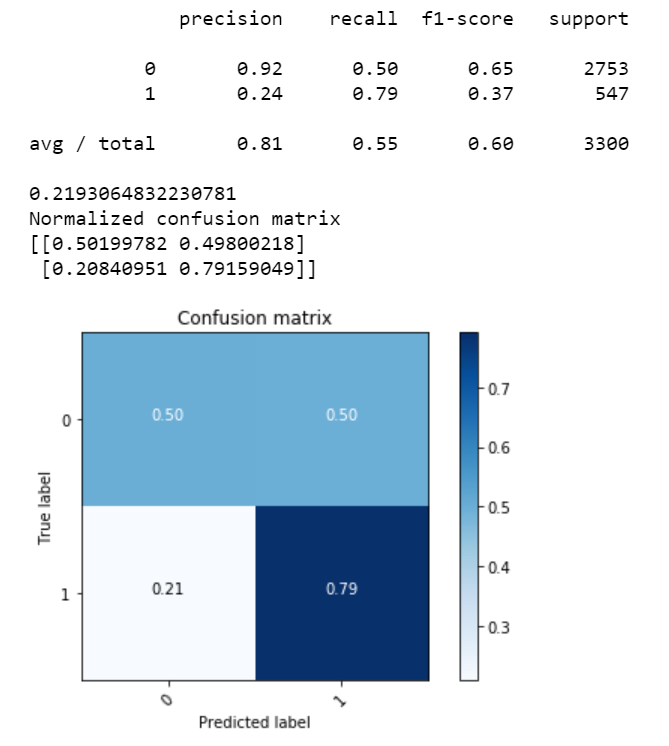


Running the algorithm multiple times has no effect on the accuracy, F-1 score and the confusion matrix. The average of all the values has little or no variance after multiple runs.

Running 2nd Time:



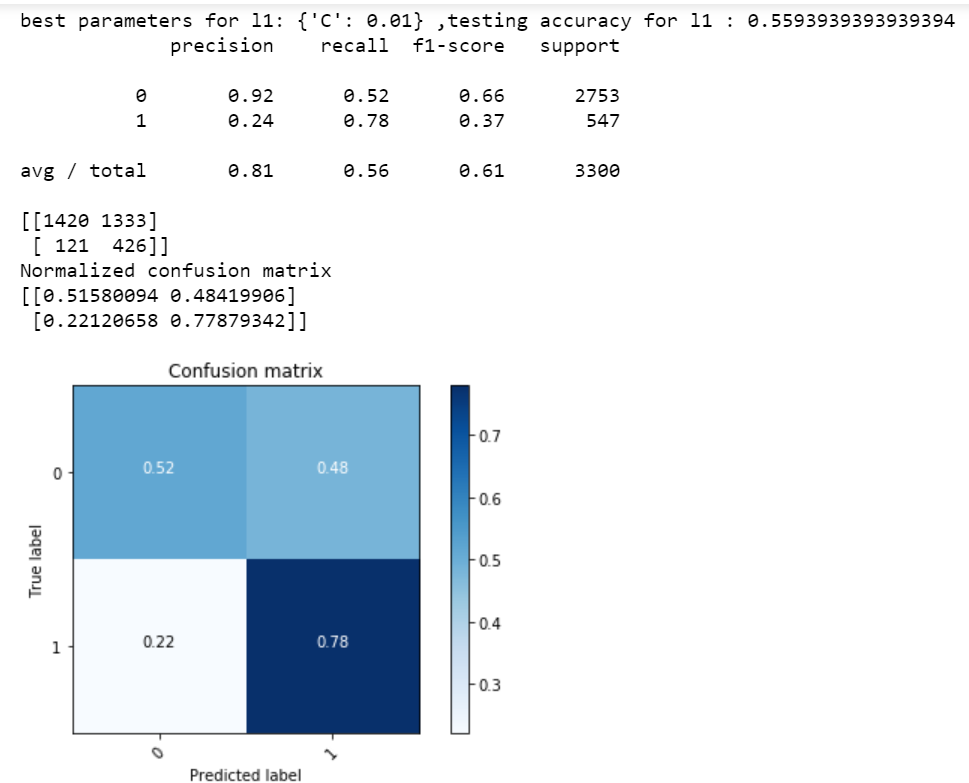
Running 3rd Time:



Ajefjdkjfbsdkjbfdbjdkfb

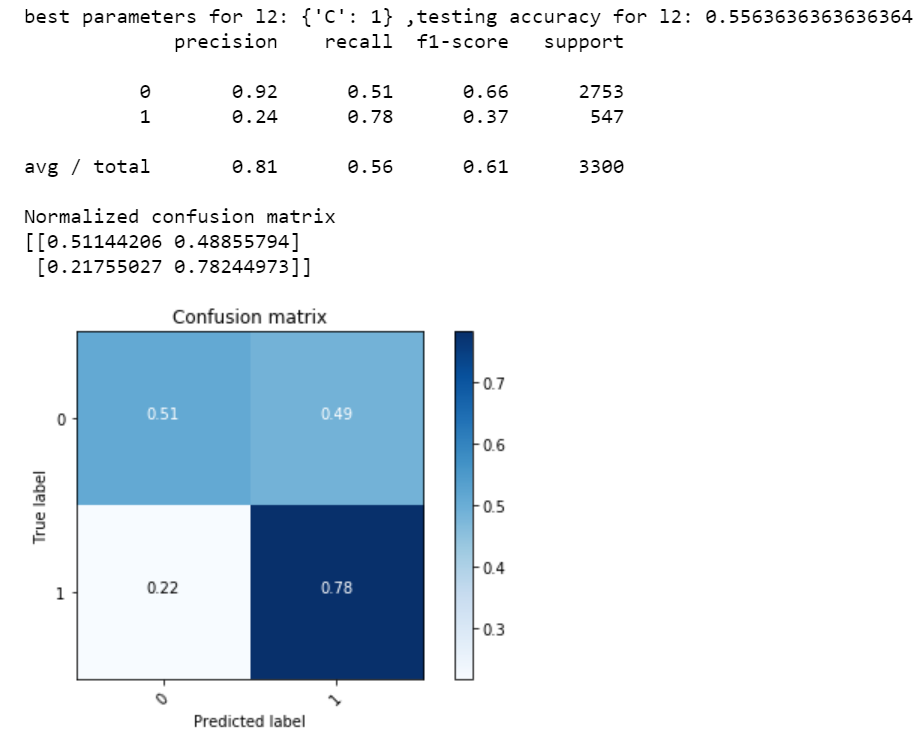
* Regularized SVM:

L1 as the penalty for regularization

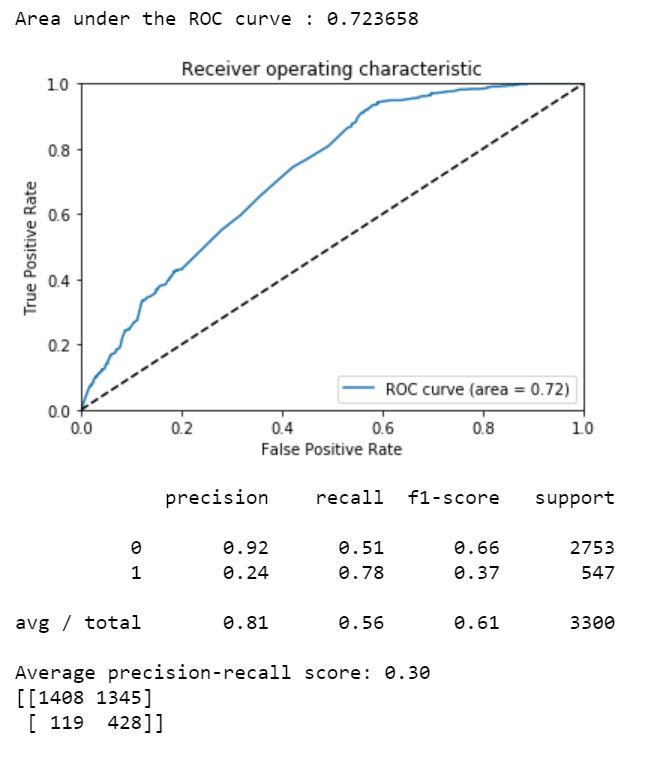


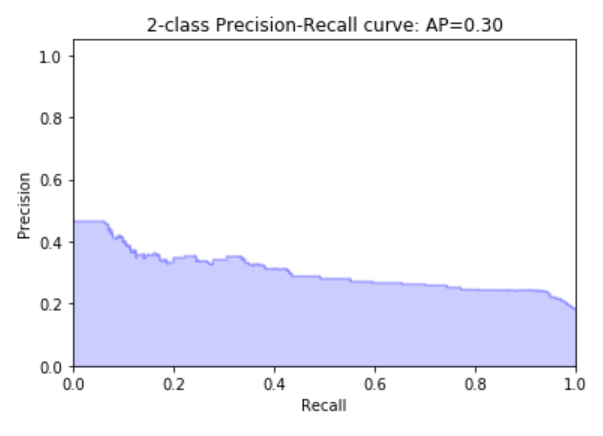
* Regularized SVM:

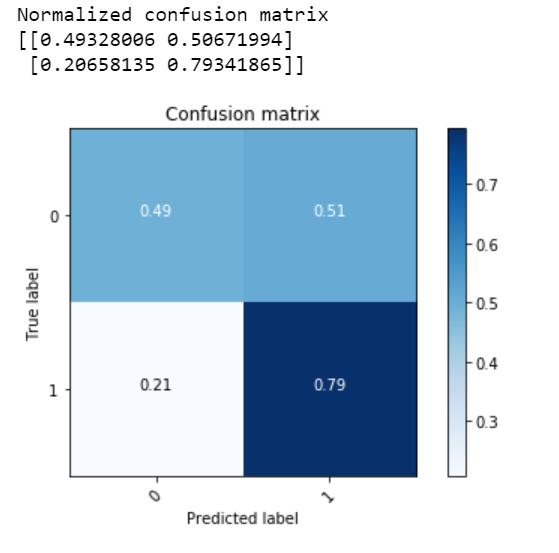
L2 as the penalty for regularization



* Logistic Regression:

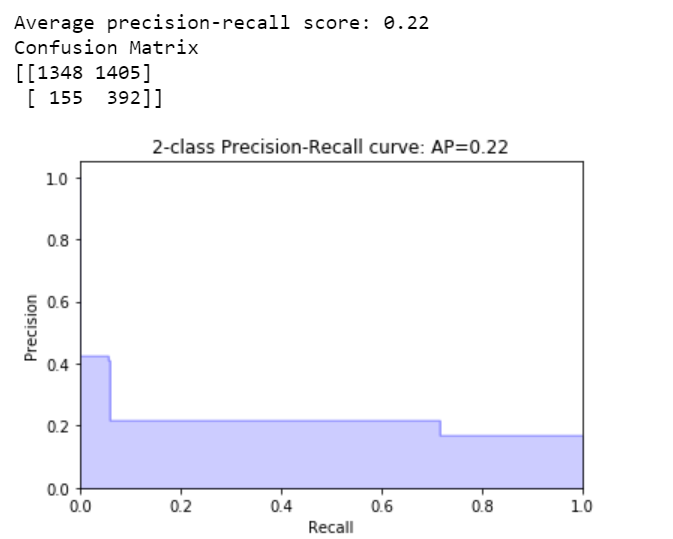




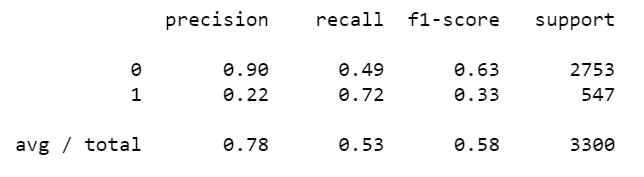


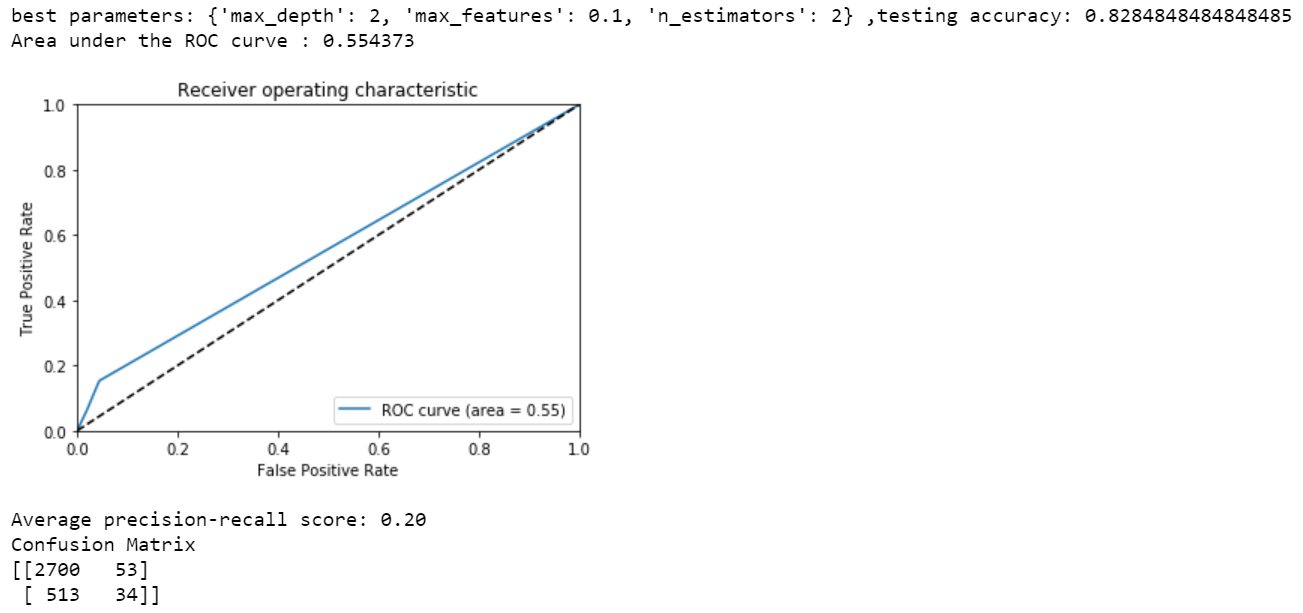
* Random Forest:





* Gradient Boosting Tree:







=