**DS 501 STATISTICAL & MATHEMATICAL METHODS FOR DATA SCIENCE**

**ASSIGNMENT 4**

Name: Ali Nauman Roll Number: 18L1863

**QUESTION 1:** How did you map the values of predictions from the regression model to labels? Give an exact mathematical expression or pseudocode.

I have taken the absolute difference of prediction value from 2 and 5. Label is selected based on smallest distance from both of the differences:

**if abs(val-2) < abs(val-5):**

**label = 2**

**else:**

**label = 5**

**QUESTION 2:** (ASSUME 2 IS THE POSITIVE CLASS. MARKS DEDUCTED IF YOU DO NOT ASSUME THIS)

**RESULTS**

|  |  |  |
| --- | --- | --- |
| **RESULTS FOR TRAINING DATA** | **RESULTS FOR TEST DATA** | **Predicted value of y for test data** |
| Lambda = 0.001  Actual 2 Actual 5 Total  Predicted 2TP = **150** FP =**0** **150**  Predicted 5FN = **0** TN =**125** **125**  Total -> **150** **125** **275** | Lambda = 0.001  Actual 2 Actual 5 Total  Predicted 2TP = **150** FP =**0** **150**  Predicted 5FN = **0** TN =**125** **125**  Total -> **150** **125** **275** | Row 1 – 1.94 – Label = 2  Row 10 – 4.94 – Label = 5  Row 15 – 5.08 – Label = 5  Row 100 – 5.13 – Label = 5  Row 120 – 4.97 – Label = 5  Row 200 – 5.14 – Label = 5 |
| Lambda = 1  Actual 2 Actual 5 Total  Predicted 2TP =**150** FP =**0** **150**  Predicted 5FN =**0** TN =**125** **125**  Total -> **150** **125** **275** | Lambda = 1  Actual 2 Actual 5 Total  Predicted 2TP =**150** FP =**0** **150**  Predicted 5FN = **0** TN =**125** **125**  Total -> **150** **125** **275** | Row 1 – 2.12 – Label = 2  Row 10 – 4.77 – Label = 5  Row 15 – 5.20 – Label = 5  Row 100 – 4.86 – Label = 5  Row 120 – 4.79 – Label = 5  Row 200 – 5.34 – Label = 5 |
| Lambda = 1000  Actual 2 Actual 5 Total  Predicted 2TP =**150** FP =**108** **258**  Predicted 5FN = **0** TN =**17** **17**  Total -> **150** **125** **275** | Lambda = 1000  Actual 2 Actual 5 Total  Predicted 2TP = **150** FP =**94** **244**  Predicted 5FN = **0** TN =**31** **31**  Total -> **150** **125** **275** | Row 1 – 2.68 – Label = 2  Row 10 – 3.16 – Label = 2  Row 15 – 2.69 – Label = 2  Row 100 – 3.59 – Label = 5  Row 120 – 2.36 – Label = 2  Row 200 – 3.88 – Label = 5 |
| Lambda = 0.2 (should have a good result)  Actual 2 Actual 5 Total  Predicted 2TP =**150** FP =**0** **150**  Predicted 5FN = **0** TN =**125** **125**  Total -> **150** **125** **275** | Lambda = 0.2  Actual 2 Actual 5 Total  Predicted 2TP = **150** FP =**0** **150**  Predicted 5FN = **0** TN =**125** **125**  Total -> **150** **125** **275** | Row 1 – 2.07 – Label = 2  Row 10 – 4.84 – Label = 5  Row 15 – 5.13 – Label = 5  Row 100 – 4.98 – Label = 5  Row 120 – 4.87 – Label = 5  Row 200 – 5.25 – Label = 5 |

**QUESTION 3:** Give YOUR opinion or conclusion about the results

From the results we can clearly see that there is no error on the training set for all the values of **lambda**, except for **lambda=1000**. In this scenario, the count of **False Positives** is way too high. For the rest of the **lambda** values, as our model is trained for the same data and our model will learn the data well. When applying on the test data, the values generated for smaller **lambda** values, the model is highly efficient. However, as the value of **lambda** increases to **1000**, the count of **False Positives** increase. From these results, we can see that a smaller value of **lambda**, produces a very excellent result. As the value is increased to a higher value of **1000**, the error appears. Moreover, we have taken a value of **0.2**, which also gives us a very excellent and to the point result. This also further satisfies the notion that smaller values of **lambda**, are the way on this dataset.