CSCE822- Homework 2 Report

Problem – 1:

1. Here are the accuracy, precision and recall, AUC score, true positive rate, false positive rate, specificity, sensitivity (for each class) calculated manually:

	True Positive	False Negative	False Positive	True Negative
politics	140	1	5	589
business	160	7	5	563
tech	128	5	3	599
entertainment	127	1	0	607
sport	165	1	2	567

Precision:

Precision (politics) = 140 / (140 + 5) = 0.9655

Precision (business) = 160 / (160 + 5) = 0.9697

Precision (tech) = 128 / (128 + 3) = 0.9771

Precision (entertainment) = 127 / (127 + 0) = 1

Precision (sport) = 165 / (165 + 2) = 0.9880

Recall / Sensitivity / True Positive Rate:

Recall (politics) = 140 / (140 + 1) = 0.9929

Recall (business) = 160 / (160 + 7) = 0.9581

Recall (tech) = 128 / (128 + 5) = 0.9624

Recall (entertainment) = 127 / (127 + 1) = 0.9922

Recall (sport) = 165 / (165 + 1) = 0.9940

Accuracy:

Accuracy (politics) = (140 + 589) / 735 = 0.9918

Accuracy (business) = (160 + 563) / 735 = 0.9837

Accuracy (tech) = (128 + 599) / 735 = 0.9891

Accuracy (entertainment) = (127 + 607) / 735 = 0.9986

Accuracy (sport) = (165 + 567) / 735 = 0.9959

False Positive Rate (FPR):

FPR (politics) = 5 / (5 + 589) = 0.0084

FPR (business) = 5 / (5 + 563) = 0.0088

FPR (tech) = 3 / (3 + 599) = 0.005

FPR (entertainment) = 0 / (0 + 607) = 0

FPR (sport) = 2 / (2 + 567) = 0.0035

AUC Score:

AUC (politics) = 0.99306

AUC (business) = 0.99390

AUC (tech) = 0.99231

AUC (entertainment) = 1

AUC (sport) = 0.99398

Specificity:

Specificity (politics) = 589 / (589 + 5) = 0.9916

Specificity (business) = 563 / (563 + 5) = 0.9912

Specificity (tech) = 599 / (599 + 3) = 0.995

Specificity (entertainment) = 607 / (607 + 0) = 1

Specificity (sport) = 567 / (567 + 2) = 0.9965

2. Understanding test: I answered all the questions correctly.

3. Properties of the ROC curves:

a. It has an AUC of 1. Because it ranks all positives above all negatives. So, theoretically this is the best a model can achieve. But in practice, it is almost impossible to achieve, so the model might have overfitted the training samples.
b. It has an AUC approximately ≥ 0.5 and ≤ 1. So, for more than almost half of the time it ranks a random positive example higher than a random negative example. Most of the models for binary classification have AUC value in this range.
c. It has an AUC approximately ≥ 0 and ≤ 0.5. So, for less than almost half of the time it ranks a random positive example higher than a random negative example. It indicates the model is not a good model.
d. It has an AUC value 0, indicating that this is the worst a model can achieve.

Problem – 2:

1. Handling Missing Values: There were no missing values in the dataset.

2. Attribute/Feature selection: I used the "Variance Threshold" method for selecting features among the 334 features of the dataset. It selects features using the variance of each column. The threshold was set to 0.01 and it reduced the number of features to 165. I used Scikit learn package for it.

3. Data Balancing: As the dataset was highly unbalanced, I also balanced the data. I stratified the data by grouping them using the target values and I selected data such that the ratio of all the different target values remains very close.

4. Normalization: I normalized the data using the StandardScalar package of Scikit learn.

Algorithm	Precision	Recall	MCC	ROC Area
SVM (10-fold-CV)	0.6516	0.5922	0.2747	0.6367
Result on test data	0.1403	0.6251	0.1348	0.6140

Problem – 3:

Results	MSE	RMSE	MAE	R2 Score
10-fold-CV (using y-exp as target)	72.0947	8.4909	0.9226	0.99245
Comparison with y-theory	2561.7206	50.6134	49.6441	-3.2395