

Homework 3

IE 7275 Data Mining in Engineering

Before you start: Read the textbook Chapter 5: Evaluating Predictive Performance.

Submission Requirement: All steps and answers should be typed and organized in one answer sheet. The hand-written solution will not be accepted.

Note: Please make sure that you explain all the calculation steps and show the final answer separately. Also, ensure that you attach the cover letter while you submit this homework.


Problem 1

Roughly 21 million Covid cases were reported to the WHO over the last week, setting a new global record for weekly cases from the rapidly spreading omicron variant, Maria Van Kerkhove, WHO's Covid-19 technical lead, said during a live-streamed Q&A across the group's social media channels. The next Covid-19 variant (NV) that will rise to world attention will be more contagious than omicron, but the real question scientists need to answer is whether or not it will be more deadly, World Health Organization officials said on Jan 25, 2022. Recently, around 34 suspicious samples were collected by Johns Hopkins experts, and the results collected (categorized as New Variant: NV and Omicron: O) after performing tests in the laboratory are captured in the file [sub-variant.xlsx](#). As Omicron infection generally causes less severe disease than infection with prior variants, in this case, consider the Omicron variant as the class of interest.

Note: The dataset provides the *Probability of New Variant (NV) from the testing laboratory*, not the *Probability of Omicron (O) from the testing laboratory*.

- Calculate the sensitivity and specificity with 0, 0.2, 0.4, 0.5, 0.6, 0.8 and 1 as decision threshold probability (cutoff value) respectfully. Plot the result and use the plot to recommend the optimal cutoff value to balance sensitivity and specificity. (Show the axis titles and chart title)
- Calculate the Matthews correlation coefficient (MCC) and F-score with a default threshold of 0.5 and the optimal cutoff from the above question. Discuss your findings.
- When do you use MCC and F-score as measures for evaluation? Which measure do you think is advised to use here and why?

Problem 2

Diabetic retinopathy (DR), an eye condition that can cause vision loss to anyone who has diabetes, is the most frequent cause of new cases of blindness among adults aged 20–74 years. Nearly all patients with type 1 diabetes and >60% of patients with type 2 diabetes have retinopathy. Based on severity, it leads to a high risk for heart attack , stroke, and even death. The strange part is, that there is no cure for diabetic retinopathy. But treatment works very well to prevent, delay, or reduce vision loss. The sooner the condition is found, the easier it is to treat. Fluorescein angiography, and Optical coherence tomography scans, are the clinical methods for diagnosing DR which take around 48 hours right from exam room test to generating a report by the doctor to diagnosis. AI, being the need of the hour, not only is more accessible but can help identify those at risk of blindness and get them in front of an ophthalmologist for treatment before it is too late. Data scientists have built an EfficientNet augmented technique that allows doctors to identify Diabetic retinopathy from eye scanned radiology films easily and classify the severity level of DR. The model was trained on 100000 eye scanned radiology films, which shows 40000 Diabetic retinopathy conditions and 60000 normal conditions. The model classified 38950 Diabetic retinopathy conditions and 58500 other normal conditions correctly. On the validation set of 10000 radiology films (3750 Diabetic retinopathy conditions and 6250 normal conditions), the model classified 2500 Diabetic retinopathy conditions and 4975 other normal conditions correctly.

- a. Build confusion matrixes for training and validation set, respectively. Calculate error rate, sensitivity, and specificity for each.
- b. Comment on the model performance.

Problem 3

The Biolive pharmaceutical company wants to develop a predictive model to identify the Genetic Disorder Type-A. From domain knowledge, the prevalence of Genetic Disorder Type -A in the US population is 8%. The model was created on a dataset of 7000 samples. Among these samples, 2800 samples were diagnosed as positive. Their analytics team decided to partition the dataset into 70% training and 30% validation with a stratified sampling technique. The sensitivity and specificity achieved on the validation set are 60% and 80%, respectively.

- a. Explain and calculate the adjusted misclassification rate, precision, and recall on the validation set. Comment on the model performance.
- b. Recommend another scheme to deal with the unbalanced data for this data science team.

Problem 4

Boston Consultancy is an award-winning AI-first digital engineering company driven by the desire to solve transformational problems at the heart of the business. They provide System integrations service to clients across the globe to solve their problems related to data storage, cloud solutions, and Business intelligence. In order to conduct a discovery call with a new client, they require Data Engineer, Business Intelligence (BI) engineer, and a solutions architect on the call along with a Sales Engineer and a Project Manager. The cost for each designated person during the discovery call is mentioned below:

| Sn | Designation | Cost/call |
|----|---------------------|-----------|
| 1 | Data engineer | 60 \$ |
| 2 | BI engineer | 75 \$ |
| 3 | Solutions architect | 100 \$ |
| 4 | Sales engineer | 50 \$ |
| 5 | Project manager | 50 \$ |

If Boston Consultancy wins an opportunity, each client pays respective service charges as per their end-to-end solutions requirement mentioned in the excel workbook. The profit of Boston Consultancy for each opportunity shall be service fees paid by each client post deducting the charges attained by each of the team members. The cloud consultants of Boston Consultancy then sought help from the data science team to maximize the profit.

The data science team built a predictive model to classify a client as a potential opportunity or non-potential opportunity based on its cloud data size, demographic information, and market value. The file [bostonconsultanct.xlsx](#) contains the model output on the validation set. Note that 1 indicates the case of winning the opportunity.

- a. Build a lift chart, a cumulative gain chart, and plot the net profit.
- b. Assume only 10 discovery calls can be conducted by the data science team during a month. Based on the data, which are the 10 clients that the cloud consultants should reach out to (Name them all)? Why?
- c. Is there any pattern or common trend you see in the top 10 potential customers with respect to their high probability and the datawarehouse/region/BI tool/Cloud storage space they are using?

Problem 5:

Dataset:

abalone.csv

Attribute Information:

Given is the attribute name, attribute type, the measurement unit, and a brief description.

| Name | Data Type | Measurement Unit | Description |
|----------------|------------|------------------|-----------------------------|
| Sex | nominal | -- | M, F, and I (infant) |
| Length | continuous | mm | Longest shell measurement |
| Diameter | continuous | mm | perpendicular to length |
| Height | continuous | mm | with meat in shell |
| Whole weight | continuous | grams | whole abalone |
| Shucked weight | continuous | grams | weight of meat |
| Viscera weight | continuous | grams | gut weight (after bleeding) |
| Shell weight | continuous | grams | after being dried |
| Rings | integer | -- | +1.5 gives the age in years |

Considering the abalone dataset, a mutual friend of you and your project mate tried building a Linear Regression Model to predict the number of rings using the sklearn library from python. Since he/she made pre-bookings to the Hans Zimmer orchestra this weekend in Boston. They had no option but to ask you both to compute the performance metrics for 5 test data points in each model. They left you some details for you to proceed further.

Problem 5:

Linear Regression Model

The intercept and coefficient values produced by a Linear regression model using sklearn library.

Considering only “Height” and “Whole weight” as features for building the model, the weights calculated were as below.

Intercept: 5.108486109050556

Coefficients: [22.6777052, 1.98547094]

Linear regression Generalized Objective function:

$$f(\mathbf{x}) = b_0 + b_1x_1 + \dots + b_rx_r.$$

The variable b_0 is the intercept. The variables b_1, \dots, b_r are the estimators of the regression coefficients, which are also called the predicted weights or just coefficients.

Test Data:

| Height (X1) | Whole weight (X2) | Rings (Y) |
|-------------|-------------------|-----------|
| 0.125 | 0.5225 | 7 |
| 0.2 | 1.463 | 11 |
| 0.16 | 0.644 | 9 |
| 0.155 | 0.8715 | 10 |
| 0.1 | 0.378 | 7 |

- Using the above given information predict the values for the test data and compute R2 score, MAE and RMSE. Comment on the model performance.

Note: All the calculation must be done manually using excel and please include the step-by-step process and formulas you’ve used to arrive at the solution.