**MScBMI 33200 – Machine Learning for Biomedical Informatics**

**Assignment I**

**<Insert NAME>**

**Directions:**

1. Fill out below information (tables and methods)
2. Write as much detail as possible for the methods section for each problem. The idea is for you to have a record of all the steps you took for building each model.
3. Submit this document along with your code in an HTML/PDF format

**Section 1: EMR Bots 30-day Readmission study**

Using the training datasets, create the following models:

1. Naïve model: This model utilizes only patient characteristics (age, gender and race) to predict 30-day readmission in a logistic regression framework

2. Logistic Regression model : This model utilizes patient characteristics and most-recent lab recordings to predict 30-day admissions in a logistic regression framework.

Calculate AUC on the test dataset. Fill out the following Table.

|  |  |
| --- | --- |
|  | AUC (95% CI) |
| Naïve Model | 0.4909 |
| Logistic Regression | 0.4863 |

Insert details (packages, parameter selection, etc.) on the models that were developed in the space given below.

Methods:

I used Python for modeling. I imported GaussianNB for the Naïve Model and LogisticRegression for the Logistic Rregression model from sklearn package.

from sklearn.naive\_bayes import GaussianNB

I fitted the training set and predicted testing set

Finally, I used predict\_proba to calculate the probability of my predictions.

**Section 2: Gusto Study**

Using the training datasets, create the following models:

1. GLM model : This model utilizes all features to predict 30-day mortality in a logistic regression framework.
2. Ridge Regression model : This model utilizes all features to predict 30-day mortality in a logistic regression framework with regularization. Utilize a 5 fold cross validation to build the parameters for your model.

Calculate AUC on the test dataset. Fill out the following Table.

|  |  |
| --- | --- |
|  | AUC (95% CI) |
| Logistic Regression | 0.8285 |
| Ridge Regression | 0.8280 |

Insert details on the models that were developed in the space given below.

Methods:

**Section 3: Short Answer Questions**

Please answer the following questions briefly.

a. What are the assumptions of a linear prediction model?

Ans. The regression has five key **assumptions**: **Linear** relationship. Multivariate normality. No or little multicollinearity.

b. Explain briefly what the interpretation of R2 is.

Ans. R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. Whereas correlation explains the strength of the relationship between an independent and dependent variable, R-squared explains to what extent the variance of one variable explains the variance of the second variable. So, if the R2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs.

c. What is learning rate in the gradient descent algorithm? Explain what happens when it is too high. Explain what happened when it is too low.

The learning rate is the step size on how much the data point move toward to the global minimum. If the learning rate is too high, it will possibly pass over the minimum point. If the learning rate is too low, it will take too long to reach to the minimum point.

d. Consider the following scenario. You developed two models for predicting mortality in the hospital using the same dataset. Model 1 had an AUC of 0.78 (95%CI 0.76-0.80). Model 2 had an AUC of 0.72 (95% CI 0.70-0.73). When you took a closer look at the sensitivity and specificity measures for various thresholds (predicted probabilities x 1000), you saw the following:

|  |  |  |
| --- | --- | --- |
| Model Cutoff | Sensitivity (%, 95%CI) | Specificity (%, 95% CI) |
| Model 1 (Predicted Probability x1000) | | |
| ≥ 17 | 57 (55-59) | 88 (87-88) |
| ≥ 20 | 53 (51-55) | 90 (90-91) |
| ≥ 23 | 49 (47-51) | 92 (92-92) |
| ≥ 27 | 40 (38-41) | 93 (93-93) |
| ≥ 67 | 19 (18-21) | 99 (99-99) |
| Model 2 (Predicted Probability x 1000) | | |
| ≥ 16 | 57 (55-59) | 90 (90-91) |
| ≥ 20 | 40 (38-41) | 97 (97-97) |
| ≥ 30 | 21 (19-22) | 99 (99-99) |

Which is a better model to operationalize? And why?

Ans.

e. Consider the following calibration plot. What can you infer from this plot?



Ans. The y-axis is the relative frequency of what was observed, and the x-axis is the predicted probability frequency. Usually the closer the points appeared along the main diagonal from bottom left to upper right, the better calibrated or more reliable a forecast. The line usually over-forecasts low probabilities and under-forecasts high probabilitiesAs the predicted probability of outcome increases, the outcome prevalence also increases. The plot line is always above the diagonal means the model has under-forecast; the probabilities are too small.

Section 4 (Optional): Watch the following talk: Yann Le Cunn: Who’s afraid of non-convex loss functions? <https://www.youtube.com/watch?v=8zdo6cnCW2w>