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

**Assignment IV**

**<Insert NAME>**

Directions:

1. Fill out below information (tables and methods)
2. 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. GLM model : This model utilizes patient characteristics and most-recent lab recordings to predict 30-day admissions in a logistic regression framework.
3. ANN model: This model utilizes patient characteristics and most-recent lab recordings to predict 30-day admissions using an artificial neural network. Feature engineering steps include balancing classes using SMOTE as well as data normalization/standardization of continuous variables.
4. RF Model : This model utilizes patient characteristics and most-recent lab recordings to predict 30-day admissions using a random forest.
5. GBM Model: This model utilizes patient characteristics and most-recent lab recordings to predict 30-day admissions using a gradient boosted machine.

Utilize a five-fold cross-validation technique to build your model.

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

|  |  |
| --- | --- |
|  | AUC (95% CI) |
| Naïve Model |  |
| Logistic Regression |  |
| ANN |  |
| Random Forest |  |
| GBM |  |

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

Methods:

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.
3. ANN model: This model utilizes all features to predict 30-day mortality using an artificial neural network. Feature engineering steps should use normalization/standardization of continuous variables.
4. Random Forest model: This model utilizes all features to predict 30-day mortality using a random forest.
5. GBM model: This model utilizes all features to predict 30-day mortality using a gradient boosted machine.

Utilize a five-fold cross-validation technique to build your model.

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

|  |  |
| --- | --- |
|  | AUC (95% CI) |
| Logistic Regression |  |
| Ridge Regression |  |
| ANN |  |
| Random Forest |  |
| GBM Model |  |

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.

1. Explain briefly the intuition behind using information gain when building decision trees.’

Ans.

1. What is difference between test error as measured by cross-validation and test error measured through bootstrap?

Ans.

1. Why do we need de-correlated trees when developing random forests?

Ans.

1. What does a varImp() on a random forest model indicate?

Ans.

1. True or False: Boosting involves combining independent decision trees as learners.

Ans.

1. What is the interpretation of learning rate in a gradient boosting model?

Ans.

1. Suppose I used the top 10 features as indicated by varImp() from my random forest model to create a multiple linear regression model. Is this a valid method? Why or why not?

Ans.