**Migraine**

*Checkpoint 3*

**Group Members:**

* Anjan Kumar Depuru
* Ruthwik Reddy Bommana
* Sandali Nemmaniwar
* Manasa Thatipamula

**Insights gained from Checkpoint 2:**

The report shed light on the use of stochastic gradient descent (SGD) and linear regression models to analyze a migraine dataset. SGD was chosen due to its success in optimizing non-convex loss functions, computational efficiency, and ability to avoid local minima and saddle points. Linear regression was chosen for its simplicity and ability to shed light on the connections between input features and the target variable. The models were optimized using parameters such as the Huber loss function and L1 and L2 norm regularization techniques. The accuracy of the models was evaluated through metrics such as R-square values, mean squared error, and coefficient of determination. The use of these models yielded an accuracy rate of around 83% in predicting the likelihood of getting a migraine.

Q1. Justify your model choice based on how your response is measured and any observations you may have made in your EDA and the first model you have estimated.

**Observations made in EDA:**

We have noticed that each sequence had a special sequence ID and was made up of the letters C, A, G, and T. The essential components of the DNA molecule, the nucleotides Cytosine (C), Adenine (A), Guanine (G), and Thymine (T), are represented by these letters. The genetic code of an organism is determined by the specific arrangement or sequence of these nucleotides. The primary finding is that every sequence ID has a distinctive RNA sequence, which distinguishes each data point. To help with the prediction of quality scores using training and testing data, we have currently estimated the mean value for each quality score point.

**Graphical user interface, text, application, email

Description automatically generated**

We chose the dataset named ERR4796172.fastq.gz which contained 19318921 rows of data.

Table

Description automatically generated with medium confidence

**Choice of the Model:**

After an initial attempt to predict outcomes in a large dataset using SGD and linear regression models, the accuracy obtained (83%) was suboptimal.

Further research led to the adoption of the **Random Forest and Xgboost** models. Random Forest is a popular algorithm that can handle large datasets with many features and observations, while XGBoost is a distributed gradient boosting library that is optimized for efficiency, flexibility, and portability. Both algorithms use a divide-and-conquer approach, building decision trees on subsets of data and combining them for final predictions. Random Forest handles missing data and outliers well, while XGBoost further prevents overfitting with a regularized model. Overall, both algorithms are efficient, accurate, and robust for analyzing large real-world datasets like the migrane dataset.

**Random Forest Model:**

Graphical user interface, text, application, Word

Description automatically generated

**Graphical user interface, text, application

Description automatically generated**

**Table

Description automatically generated**

**XGBoost Model:**

**Graphical user interface, text, application

Description automatically generated**

After evaluating both Random Forest and XGBoost models on the migraine dataset, we found that XGBoost performed better with an **accuracy of 89%,** compared to Random Forest with an **accuracy of 87%.** This implies that XGBoost is more efficient and accurate for analyzing the given dataset than Random Forest.

Q2. Report the model’s test error rate using one of the techniques we discussed in lecture. Justify your choice.

**R-squared and Mean Squared error:**

Common metrics for assessing the effectiveness of regression models for test error rate analysis include mean squared error (MSE) and R-squared. We determined our model’s test error rate using these techniques. These measures are helpful for evaluating the model's precision and capacity for generalization to new data. We can identify whether the model is overfitting or underfitting and make appropriate corrections to enhance performance by analyzing the MSE and R-squared on test data.

**Test error rates for Random Forest:**

Text

Description automatically generated

Text

Description automatically generated

**Test error rates for XGBoost:**

Text, letter

Description automatically generated

Text

Description automatically generated

Q3. Based on the estimated test error rate, discuss how well the model fits the data.

Mean Squared Error (MSE) is a metric that measures the average squared difference between the actual and predicted values of a model, with lower values indicating better performance. In the case of the **random forest model**, the MSE values for the train and test datasets are **2.293 and 2.159**, respectively. On the other hand, the MSE values for the train and test datasets for the **XGBoost model are 1.189 and 1.028**, respectively. These results suggest that the XGBoost model is a better fit for the data, as it has lower MSE values for both the train and test datasets. Therefore, we can say that the XGBoost model outperforms the random forest model for this dataset.

A higher R-squared value indicates a better model performance. For the **random forest model,** the R-squared values for the train and test sets are **0.658 and 0.614**, respectively. In contrast, for the **XGBoost model**, the R-squared values for the train and test sets are **0.776 and 0.735,** respectively. Therefore, based on the R-squared values, we can say that the XGBoost model is better than the random forest model for this dataset.

In conclusion, it is safe to say that XGBoost performs better than random forest for the migraine dataset considering both the metrics.

Q4. Use the models to make predictions for at least three cases of interest and compare between models.

**Prediction 1:**

The AUC-ROC curve will be a plot of the TPR versus the FPR for all possible threshold values, with the area under the curve (AUC) representing the overall performance of the model. A model with an AUC of 1.0 would have perfect performance, while a model with an AUC of 0.5 would perform no better than random guessing. To interpret the AUC-ROC curve for our XGBoost model predicting migraine description, we looked for a high AUC value and a curve that is as close to the top-left corner of the plot as possible which indicated that model can correctly identify higher proportion of migraine-related description while minimizing the false positives.

Chart, line chart

Description automatically generated

**Prediction 2:**

We created a boxplot graph using the gene sequence data that was provided. By looking at the graph the gene sequence having the value maximum which in our case is 34 would make an estimated prediction of having presence of migraine in the required patient.

The 30 is the average quality score for this specific prediction according to the trained model.

**Chart

Description automatically generated**

**Prediction 3:**

Based on CGRPmAbs value and patients having migraine more than 4 times in a month,  our model can help predict presence of migraine by comparing responders and non-responder age in which age is a positive predictive factors for detection.

Accuracy: 76%