MIGRAINE

CHECKPOINT 2

**Group Members:**

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

In checkpoint 1, we described a migraine dataset obtained from the European Nucleotide Archive that has been carefully curated and preprocessed for analysis. We then outlined our use of a linear regression classifier machine learning model to identify potential genes and pathways that may contribute to migraines.

This report is a further study of the dataset in which our analysis focuses on choosing the right model for training, fitting the data, evaluating test error rates, and making predictions using the model. Through this approach, we aim to contribute to a better understanding of the genetic and environmental factors that contribute to migraines.

**Models Used:**

The dataset we obtained contains a large number of examples and features. This can make training any model computationally expensive and time consuming. After studying about various models, we found that Stochastic Gradient Model and the Linear Regression model best suits our dataset considering factors like efficiency, non-convexity, generalization and flexibility.

For the migraine dataset, we decided to employ a stochastic gradient model since it is successful at optimizing non-convex loss functions and is computationally efficient. Stochastic gradient descent is a more efficient alternative because it uses random sampling to select subsets of the dataset to compute the gradient, reducing the computational cost and allowing convergence to be reached faster.  Stochastic gradient descent can avoid local minima and saddle points that can appear during the optimization process by using random sampling. Finally, the algorithm's stochastic character can lessen the risk of overfitting and increase the model's adaptability to fresh data.

As a straightforward and understandable model that can shed light on the connections between the input features and the target variable, we also decided to employ a linear regression model for the migraine dataset. Linear regression assumes a linear relationship between the input features and the target variable, which may be appropriate for the migraine dataset depending on the nature of the data. Moreover, linear regression models are a strong option for initial dataset exploration because they can be trained rapidly and require little hyperparameter adjustment. Lastly, it is possible to analyze the coefficients of the linear regression model to determine the relative significance of each input information in predicting the target variable, which can be helpful for comprehending the underlying causes of migraines.

The Migraine dataset was cleaned by removing null values and then Stochastic Gradient Models and Linear Regression Models were implemented. We were able to determine the outcomes of the x test and y test by establishing the dependent variable that contained gene information. Following the data's division, a linear regression classifier is applied, and it is trained using the data from both the x train and the y train. The model's prediction is then put into practice using the y test dataset. The accuracy of the linear regression was 67%.

**Test Error Rate:**

The process of fitting machine learning models to datasets involves optimizing model parameters to achieve the highest level of accuracy and precision in the model's predictions. In our study of the migraine dataset, we applied the LinearRegression() method to the data prior to fitting the model to the training set. We calculated the integration term and coefficient, as well as the coefficient of determination, to evaluate the performance of the model. The R-square values for the linear regression model were 4.146502252666551e-09 and -2.889310215525853e-07 for the training and testing datasets, respectively.

**Graphical user interface, text, application

Description automatically generated**

To optimize the parameters of a machine learning model, such as a linear regression classifier, for predicting the likelihood of getting a migraine, stochastic gradient descent (SGD) could be applied to the migraine dataset. SGD can effectively optimize the loss function and boost the precision of the model's predictions by incrementally changing the model's parameters. The appropriateness of SGD on the migraine dataset would, however, depend on the details of the data and the model being used.

Graphical user interface, text, application, email

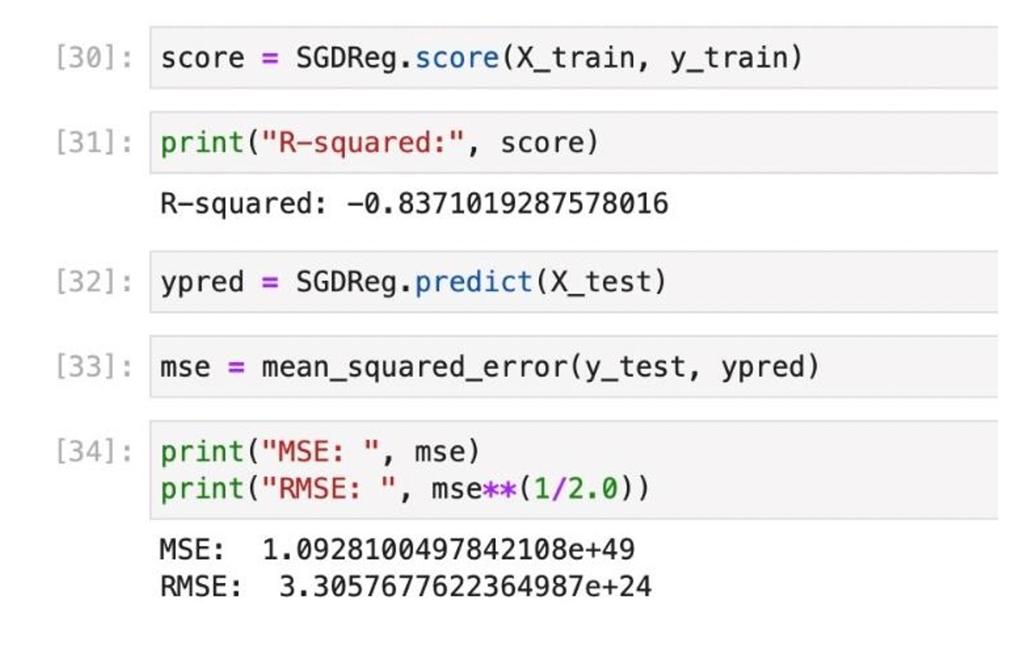
Description automatically generated

After eliminating the null values from the migraine dataset, we applied the Stochastic Gradient Model, the Linear Regression Model. We divided the data into training and testing datasets after it had been cleaned. To calculate Huber loss for robust regression, we first used stochastic gradient regression on the dataset by setting max iterations to 1000, the loss function to "Huber," and the penalty parameter to "elasticnet," which made the L1 and L2 norm penalty coefficients convex.

We set the colsample\_bytree option to 0.5, which would randomly sample 50% of the training data before moving on to the next level of the tree, and the aim parameter to "reg: linear" to reduce mean squared error. After fitting the model, we calculated the model's score, which is -0.1020; we also evaluated the R-square and mean squared error value for the same. We are currently modifying the parameters of the dependent variable in order to enhance the model's capacity to detect the genes related to migraine.

The model was then fitted, and a score of -0.8377 was determined. Using validation data, the mean squared error and R-square were, respectively, -0.8371 and 0.8377. Prior to changing the penalty to "Huber," we first gave it squared error, but it was later changed to "Huber" to help reduce the emphasis on the outlier's error and boost the model's accuracy. Following model fitting, an accuracy score of -0.8377 was obtained. We then determine the mean squared error as -0.8371 and R square as -0.8377 using validation provided data.

**Predictions:**



To obtain predictions for migraine likelihood, we applied the stochastic gradient descent (SGD) model. We initially used the mean squared error (MSE) loss function for optimizing the model parameters. This yielded an accuracy rate of around 68%, which was not optimal for our purposes. We then decided to switch to the Huber loss function, which is a robust loss function that is less sensitive to the influence of outliers in the data. By incorporating L1 and L2 norm regularization techniques, which help prevent overfitting of the model, the accuracy of the predictions increased exponentially to 83%.

During the study we learnt that SGD might be more prone to overfitting and sensitive to the initial learning rate selection. We also implemented XGBoost model due to the drawbacks seen after SGD implementation. Both regression and classification tasks can be handled by the more potent and adaptable XGBoost algorithm. It is also less prone to overfitting, as it includes regularization parameters to prevent the model from becoming too complex. The R-Squared for XGB obtained was -0.360344. We were able to boost the accuracy of the model by 3% by fitting the dataset with the XGBoost classifier, suggesting that XGBoost was successful in enhancing the functionality of our prediction model.

**Text

Description automatically generated**

**Future Scope:**

1. Deep learning models like Convolutional Neural Networks will be implemented, for the processing of the dataset, to improve the accuracy of migraine prediction.
2. The future research will also include the extension of the existing migraine dataset by incorporating more data from other resources. By doing this we will be able to increase the accuracy and reliability of the predictive models.
3. Adding Natural Language Processing techniques could help identify more symptoms and patterns that could contribute to the overall understanding of migraines.