Name Guo, Xinyu Wang, Yuqin Date: 12/10/2022

(last name, first name)

NYU ID: xg693, yw3875

Course Section Number: DS-GA 2433-001

Project 3

Total in points (100 points total): \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Professor’s Comments:

2433 Project3 Report

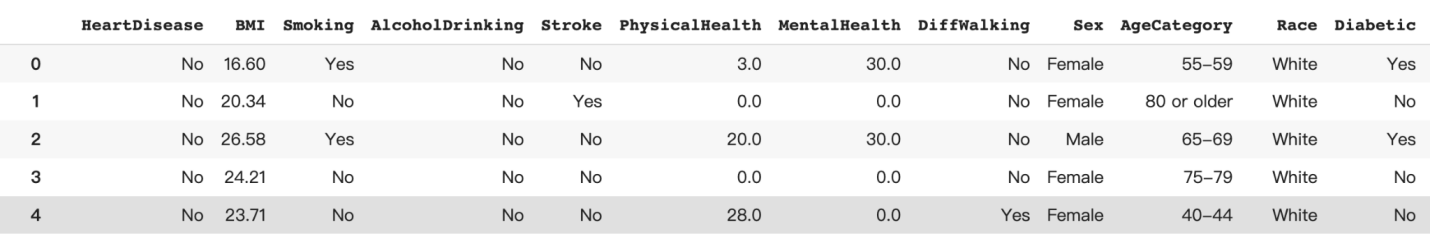
*Xinyu Guo, Yuqin Wang*

***Design the database system:***

Initially, our database would be built by loading the cleaned datasets into the cloud storage, and then turn the data into SQL format using SQL commands. In the future, if we have more data, we can add them into our existing database using similar methods.

***Dataset:***

We use the data collected in Project 2 (heart\_2020\_cleaned.csv).





***Workflow-based application:***

To implement our database in practice, such as a business case that deals with insurance quotes with patients enduring heart disease, approaches that can analyze each insurance quote to verify the reliability are also required. Thus, we deploy some machine learning models to achieve this goal.

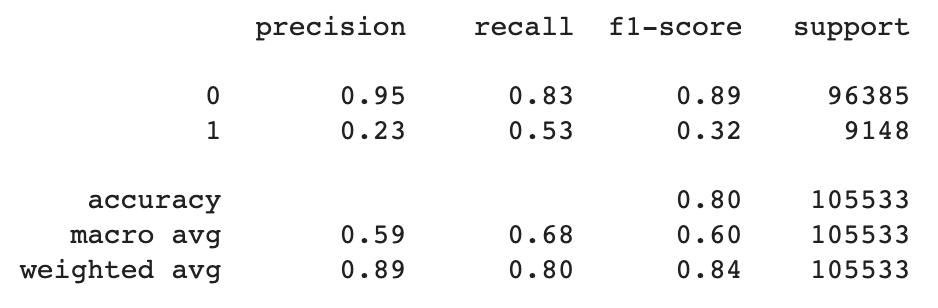
***ML models:***

We implement the following models to perform the analysis over the data.

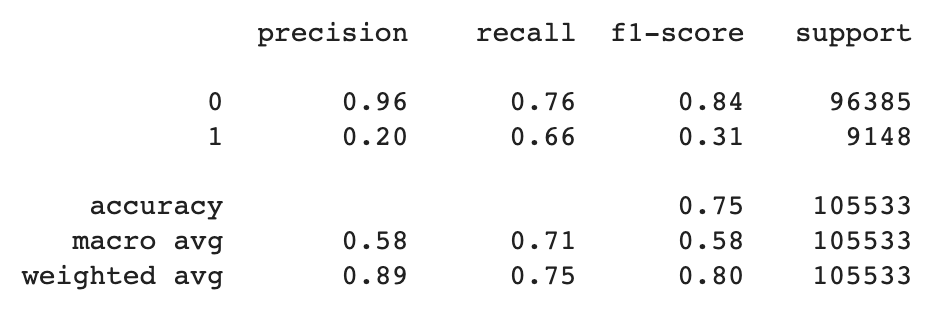
* Logistic Regression is a popular method for classification in machine learning. It is a type of regression algorithm that predicts the probability of an outcome, rather than predicting the outcome itself. This makes it well-suited for binary classification problems.
* Random Forest is an ensemble learning method that is based on decision trees. It is an efficient and powerful method for both regression and classification tasks. Random forest effectively avoids overfitting, which makes it a good choice for datasets embracing high variance.
* XGBoost is an efficient implementation of gradient boosting. It is designed to be extremely flexible and powerful with robust performance. It always performs fast and accurately, and is able to handle large datasets with ease.
* CatBoost is specifically designed to automatically detect and handle categorical data, making it a popular choice for datasets with a large number of categorical features. It is also an efficient implementation and is easy to use and customize.

***Results and Evaluations:***

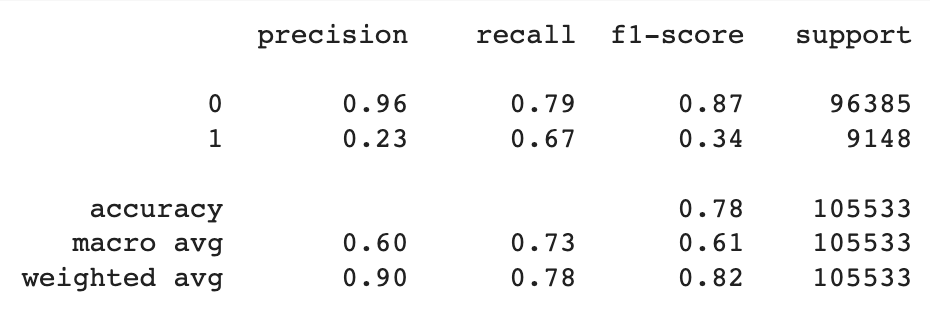
Logistic Regression:



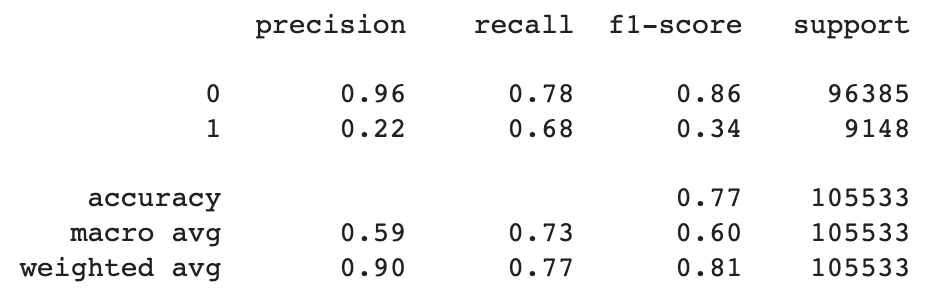
Random Forest:



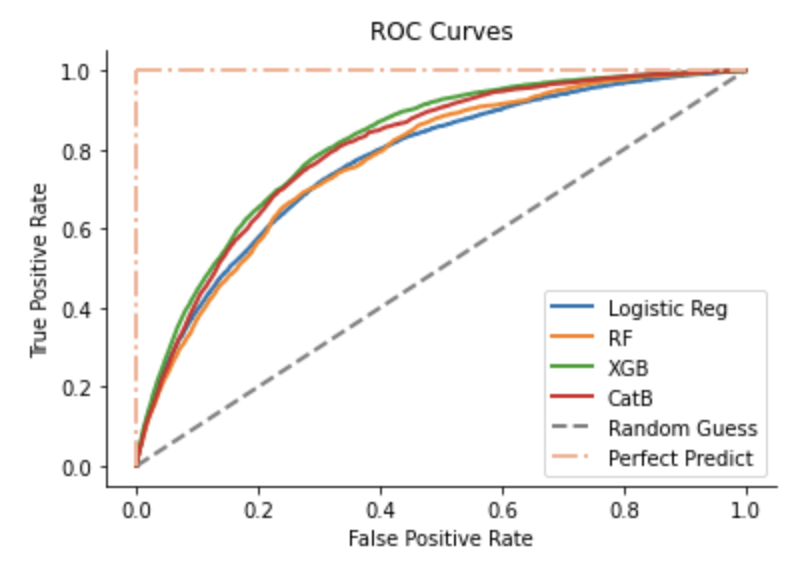
XGBoost:



CatBoost:



ROC Curves:



***Conclusion:***

Overall, the performance of random forest, XGBoost, and CatBoost are quite similar,

with a much higher recall score for 1 compared to the logistic regression. Also, implied by the ROC curves, boosting algorithms tend to perform better than the others. The reason for this might be that our data set is quite small and relatively simple. If we can have more data in the future, the current models might show better results.