Capstone Project Submission

ML – Classification (Cardiovascular Risk Prediction)

Cardiovascular diseases (CVDs) are a group of disorders of the heart and blood vessels. Cardiovascular diseases are very common these days due to the change in lifestyle and poor eating habits. So, predicting the risk of cardiovascular disease for given person by analyzing his/her past and present medical data and habits can save them if they act early against the disease and take precautionary measures.

We have performed following steps while building the model:

- Exploratory data analysis
- Feature Engineering
- Training the model and Hyperparameter tuning
- Evaluating the ML model performance
- Conclusions

Exploratory data analysis:

In this step performed data preparation after that we have cleaned the data. We have performed Univariate, Bivariate, and multivariate analysis and we drove some conclusions.

Featuring Engineering:

In Feature engineering we have checked for multicollinearity are removed the features with high correlation. Some of the features are skewed so we performed LOG and Square root transformations. To convert categorical feature into numerical we performed one hot encoding on categorical feature. To select the features, we used Chi Squared test for categorical variables. Since the data is highly imbalanced, we used an oversampling technique known as SMOTE

Training the model and Hyperparameter tuning:

In this step, we applied different machine learning algorithms to train the model and predict 10-year risk of cardiovascular disease. Also, we performed hyperparameter tuning, to find out the best set parameters for the given algorithm.

Evaluating the ML model performance:

After training the different models with the data it is important to evaluate the model performance using different metrics but most of the times data is overfitting on training data Logistic regression and Random Forest are giving good F1 score compared to other models. We also observed that for low threshold values recall is good for Logistic regression and Random Forest

Conclusions:

After trying different models and with hyperparameter tuning we say that Random Forest and Logistics regression with low threshold values has highest Recall and F1 score, so we can Logistic regression and Random Forest for predictions.

Contributions:

1.Pavan Potnuru:

1. Exploratory Data Analysis
2. Feature Engineering
3. Training Models and Hyperparameter tuning
4. Evaluating models and conclusions

2. Mohammed Haseebuddin:

1. Exploratory Data Analysis
2. Feature Engineering
3. Training Models and Hyperparameter tuning
4. Evaluating models and conclusions

Git Hub link: https://github.com/Haseeb227/Cardiovascular-Risk-Prediction

Drive link: https://drive.google.com/drive/folders/17uObKw0CND9UnAX_XNLR1kqWmOxe1jaz?usp=sharing