PNS-2019 (National Health Survey 2019), Prediction of Heart Disease by LGBM Modeling

Background

National Health Survey 2019 (PNS-2019) is an official program developed by Brazilian Health Minister. It has the main objective to take information about Brazilian residents regarding their health-related risk behaviors, chronic health conditions, use of the public healthcare system, and other information. PNS-2019 subsidizes public policies, either evaluating the current ones or promoting others to improve the quality of health for the entire population.

PNS-2019 has similarities with the Behavioral Risk Factor Surveillance System (BRFSS), developed by CDC (Center for Disease Control and Prevention, USA) https://www.cdc.gov/brfss/index.html

Dataset: raw data is available through the link: https://www.ibge.gov.br/en/statistics/social/health/16840-national-survey-of-health.html?=&t=resultados

A summary of the sampling program applied and the data collection are available at: https://www.scielo.br/j/ress/a/RdbtmCHjJGt8xDW6bV3Y6JB/?lang=en&format=pdf

The idea is to develop a machine learning model through LGBM Classifier to predict the probability of the risk of someone developing heart disease. For that, modeling data from PNS-2019 for the risk factors recommended by the CDC (Center for Disease Control and Prevention, USA) and WHO (World Health Organization).

LGBM Modeling Development (https://lightgbm.readthedocs.io/en/latest/Features.html)

Light GBM is a fast, distributed, high-performance gradient-boosting framework based on decision tree algorithm used for ranking, classification and many other machine learning tasks.

According to LGBM docs, the algorithm runs faster-training speed and higher efficiency, has lower memory usage, with better accuracy, supports parallel and GPU learning, and is capable of large-scale data handling.

Modeling Development:

Dataset:

- The file *df_pns_rawdata.csv* has data available for modeling regarding chronic diseases and health-related risk behaviors, such as food, physical activities, tobacco, alcohol use and other variables
- The dataset has 81,218 rows and 55 columns (numerical and categorical features)
- The *df_pns_rawdata_VariablesDictionary* shows labels and descriptions.

LGBM modeling:

- Feature Selection: from all features in the file df_pns_rawdata.csv, around 45 variables (numerical and categorical) were selected and split into (80/20) training, test dataset and the target feature (Heart Disease)
- Features Engineering:
    - Numerical variables transformed by applying *StandardScaler* converting to mean =0 and standard deviation =1
    - Boolean variables were concatenated to the numerical variables after transforming them
    - Categorical variables transformed by One Hot Encoder converting to dummy (0 and 1)

- Training Model_1
    - Cross-validation (k-fold cross-validation): it estimates the capability of the model to predict new data
        - Parameters:

        - `StratifiedKFold(n_splits=5, random_state=42, shuffle=True)`
          `cv_results = cross_val_score(estimator=lgbm, X=X_train_transformed, y=y_train, scoring='recall', cv=skf, n_jobs=-1)`
        - `Output: 0.6137`
    - LGBMClassifier ran on the training set
        - Parameters:
            - `Pipeline(steps=[('LGBMClassifier',LGBMClassifier(class_weight={0: 1, 1: 16}, random_state=42))])`
    - Performance Evaluation
        - Training set
            - `RoC-AUC LGBM_train: 0.8738`
            - `Recall LGBM_train: 0.7986`
            - `Precision LGBM_train: 0.1843`

        - Test set
            - `RoC-AUC LGBM_test: 0.7754`
            - `Recall LGBM_test: 0.6374`
            - `Precision LGBM_test: 0.1431`

The purpose is to maximize the "True-Positive" over "True-Negative." That comes from the fact that medical cases concern in to predict positive ones accurately, even overestimating the Falso-Positive. From that perspective, Recall and AUC-ROC curve should come first, and after the Precision

    - Confusion Matrix:
        - The confusion matrix below shows that the True-Positive is underestimated, which also corroborates with the AUC-ROC curve for test set ~0.77
        - Indeed, there is some room for optimization, and it is the next step

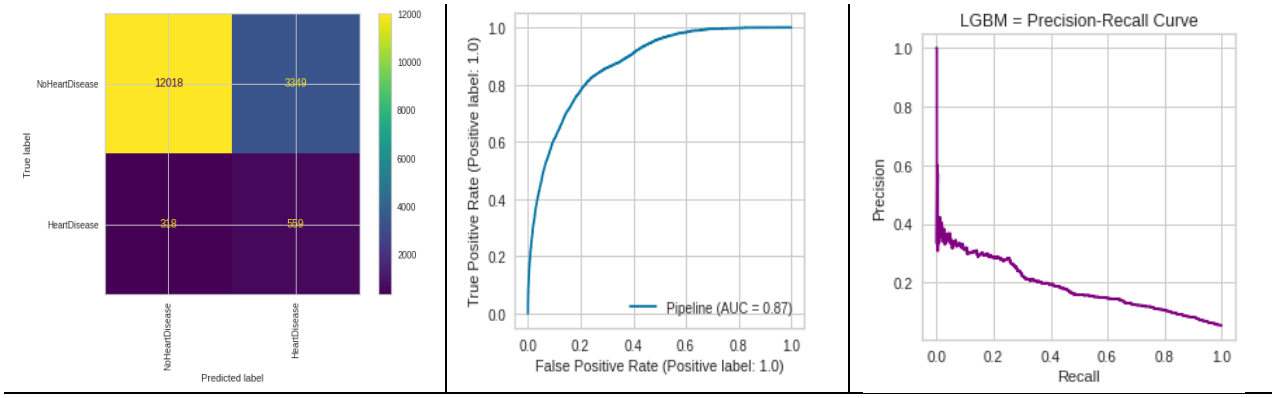

Fig 1 – LGBM performance evaluation

- Training Model_2 (Parameters Optimization)
    - Cross-validation (k-fold cross-validation): it estimates the capability of the model to predict new data
        - Parameters:
            - `skf=StratifiedKFold(n_splits=5, random_state=42, shuffle=True)`

```
cv_results = cross_val_score(estimator=lgbm_GS, X=X_train_
transformed, y=y_train, scoring='recall', cv=skf,n_jobs=-
1)
```

- Output: 0.6815

- LGBMClassifier ran on the training set
  - Parameters:
    - ```
      lgbm_GS = Pipeline(steps=[
            ('LGBMClassifier', LGBMClassifier(random_state=42,
      class_weight={0:1,1:16}, learning_rate=0.02, max_depth=9,
      n_estimators=200))])
      ```
- Performance Evaluation
  - Training set
    - RoC-AUC LGBM_train: 0.8298
    - Recall LGBM_train: 0.7642
    - Precision LGBM_train: 0.1551
  - Test set
    - RoC-AUC LGBM_test: 0.7846
    - Recall LGBM_test: 0.6978
    - Precision LGBM_test: 0.136
- Confusion Matrix:
  - True-positive is better than the model_1 (before optimization). That reflects in the Recall shows some improvement. On the other hand, AUC-ROC and Precision turned worse. It demonstrates the effect of the "trade-off" between Recall and Precision. The challenge is to balance it properly.
  -

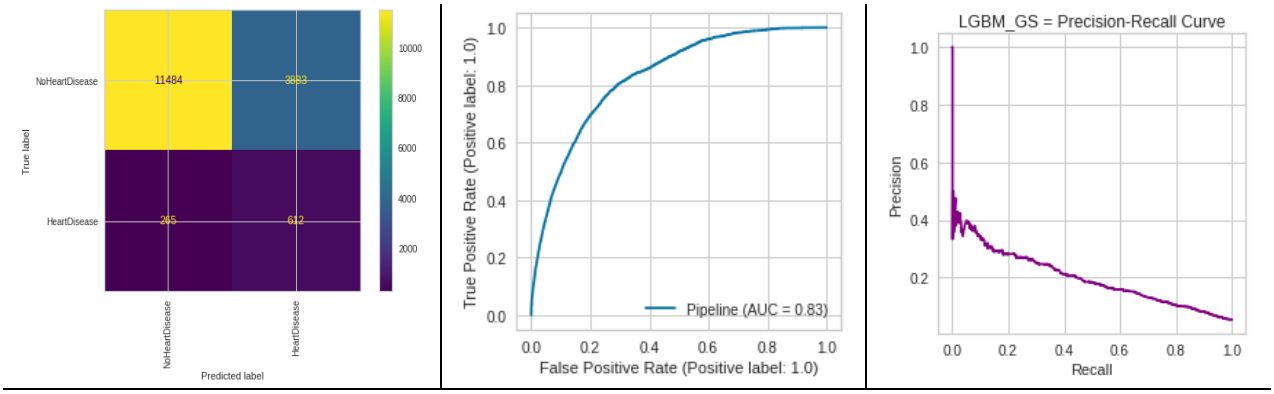

Fig 2 – LGBM optimized, performance evaluation

The current model aims to predict the probability of someone developing heart disease. So, the graphic below shows how the model performs in a range of possibilities using the test dataset—the graphic compared heart disease diagnosis vs. heart disease predicted. On target means the model predicted accurately; on the other hand, Off target means inaccurate.

In summary, the model predicted quite accurately to lower risk of someone developing heart disease. As the probability increases (higher than 50%), the model presents a weak prediction
      ```

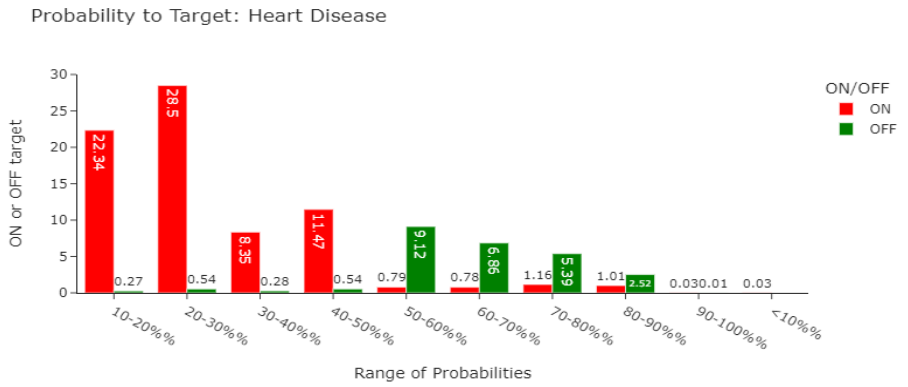

Fig 3 – Probability of the model targeting heart disease

- Features Importance through SHAP values are in figure 4 below. Generally, the most important variables pointed out here are associated with risk factors preconized by international health organisms, such as blood pressure, dyslipidemia, diabetes, physical inactivity, and unhealthy foods. Those have a high impact on the model to predict heart disease.
- It is also interesting to note that other chronic diseases also highly impact the current prediction model, such as backache, depression, stroke, and arthritis/Rheumatism. It might relate to the fact that people tend to be physically inactive, increasing the risk of heart disease development.
  -

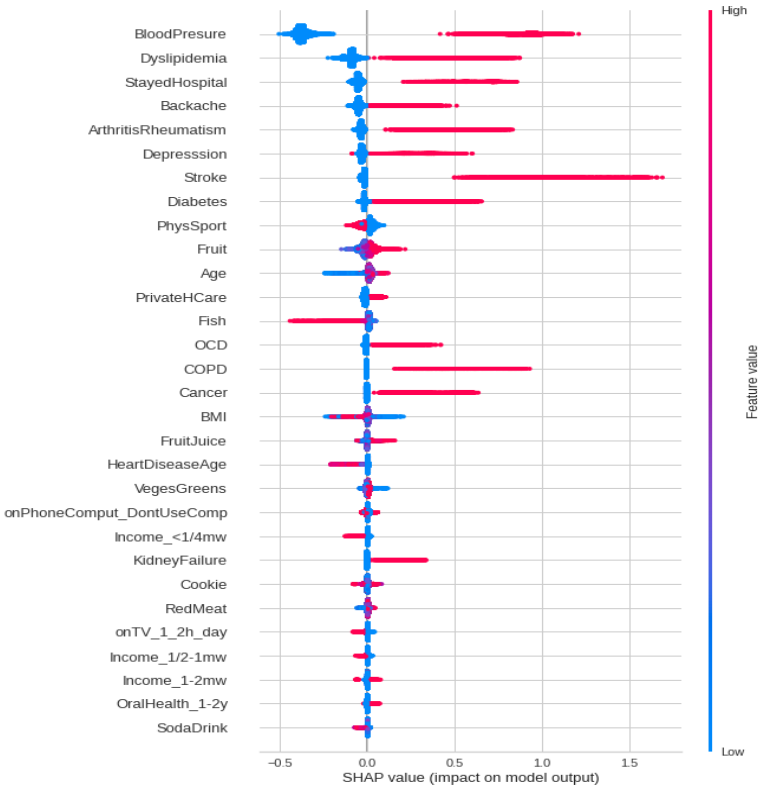

Fig 3 – SHAP value in the training dataset

Conclusion:

- More optimization is necessary to accurately target people associated with a higher risk of developing heart disease
    - Next step:
        - Feature Selection – reduce dimensionality working with variables that impact the model response
        - Balance the database – bring the minority class to the same level as the majority class
        - Look for another classifier algorithm