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

Background
National Health Survey 2019 (PNS-2019) is an official program of the Brazilian Health Ministry. 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.

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

A previous tentative of developing a predictive model for the probability of the risk of someone to the heart disease based on an LGBM Classifier (https://github.com/embordin/PNS-2019-heart-diseases-prediction-LGBMClassifier), in which a large number of variables were used to describe it. For that, the performance evaluation on the test dataset shows *RoC-AUC: 0.7846, Recall: 0.6978 and Precision: 0.136.* The model underperformed on true-positive.

Before blaming the model, the PNS-2019 dataset comes from questionaries applied to select people who give spontaneous answers without medical accreditation. And also, the question is if someone has ever been diagnosed with a chronic disease. It does not clear if the person is still developing the disease or has already healed. Both points might impact the quality of prediction.

Anyhow, the purpose of that second tentative of developing a machine learning model through LGBM Classifier is 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).

For that current tentative, it proposes a dimensionality reduction based on the combination through SHAP values that show the most impact on the previous model with the risk factors to develop heart disease.

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

Modeling Development:
Raw 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 19 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
  - 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_GS_19, X=X_train_transformed, y=y_train, scoring='recall', cv=skf, n_jobs=-1)`
      - `Output: 0.7118`

  - LGBMClassifier ran on the training set
    - Parameters:
      - `Pipeline(steps=[('LGBMClassifier',LGBMClassifier(class_weight={0: 1, 1: 16}, learning_rate=0.025, max_depth=7, n_estimators=100, random_state=42))])`
  - Performance Evaluation
    - Training set
      - `ROC-AUC LGBM_train: 0.8038`
      - `Recall LGBM_train: 0.7561`
      - `Precision LGBM_train: 0.1401`
      - `F1 LGBM_train: 0.2364`

    - Test set
      - `ROC-AUC LGBM_test: 0.7810`
      - `Recall LGBM_test: 0.7210`
      - `Precision LGBM_test: 0.1293`
      - `F1 LGBM_test: 0.2193`

  - Confusion Matrix – test set:
    - The confusion matrix below shows that the True-Positive is underestimated, which also corroborates with the ROC-AUC curve for test set ~0.78

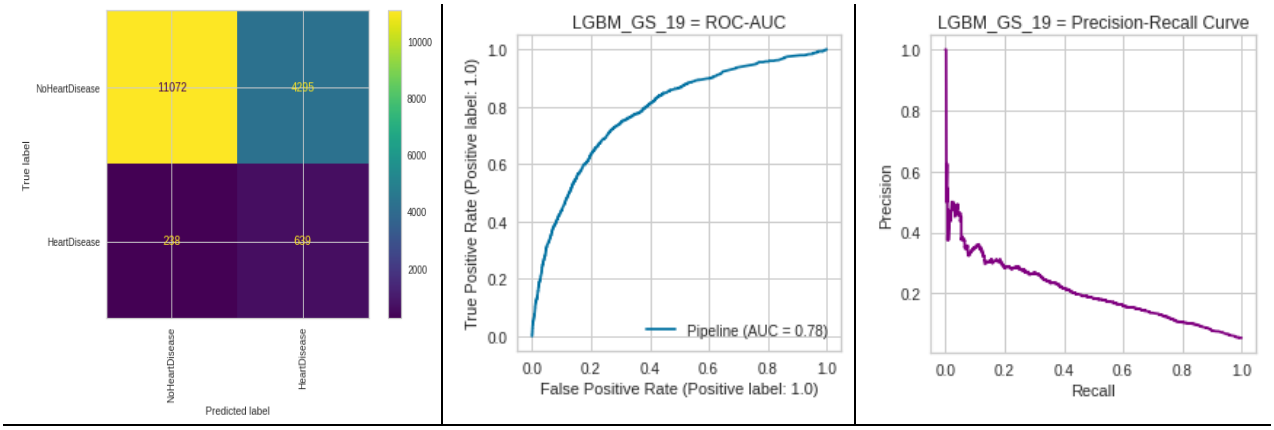

Fig 1 – LGBM performance evaluation on the test dataset

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. The model presents a weak prediction as the probability increases (higher than 50%).

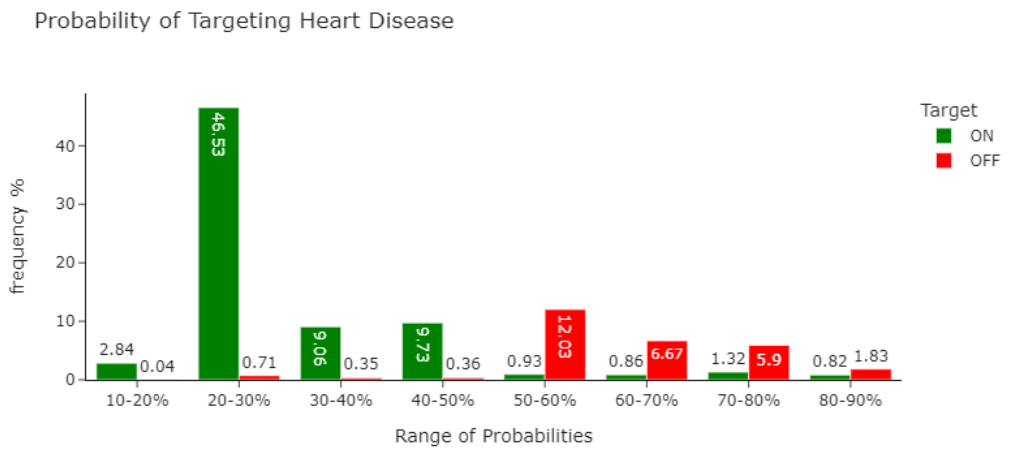

Fig 2 – Probability of the model targeting heart disease

The model classified many people as False-Negative or all off-target with a probability higher than 50%, as shown in the graphic below. In terms of risk factors, what makes them different from others who are classified as True-Positive?

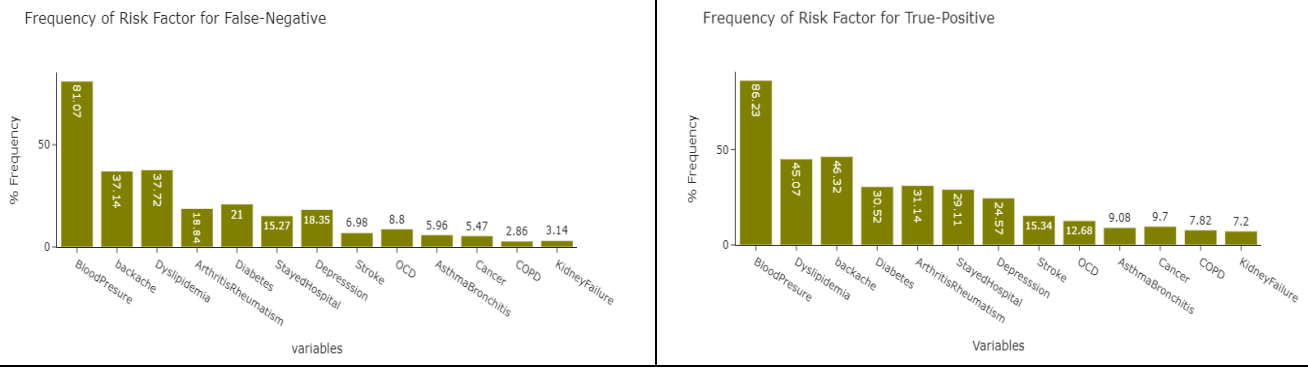

Fig 3 – Frequency of False-Negative and True-Positive for the risk factors to develop heart disease

Comparing the frequencies shown in the graphics above, it seems some tine difference between True-Positive and False-Negative. It would suggest that people spontaneously answer questions about their health conditions and chronic diseases. And it would also speculate that many people have sintomas and or have been developing risk factors for heart disease, but they don't have a medical diagnosis yet.

From the public/private health system, false-negative patients are the most important patients to come into a preventive and/or informative program to minimize or slow down the development of heart diseases that would bring an important impact on the health cost in the medium and long-terms.

- **SHAP VALUE:**

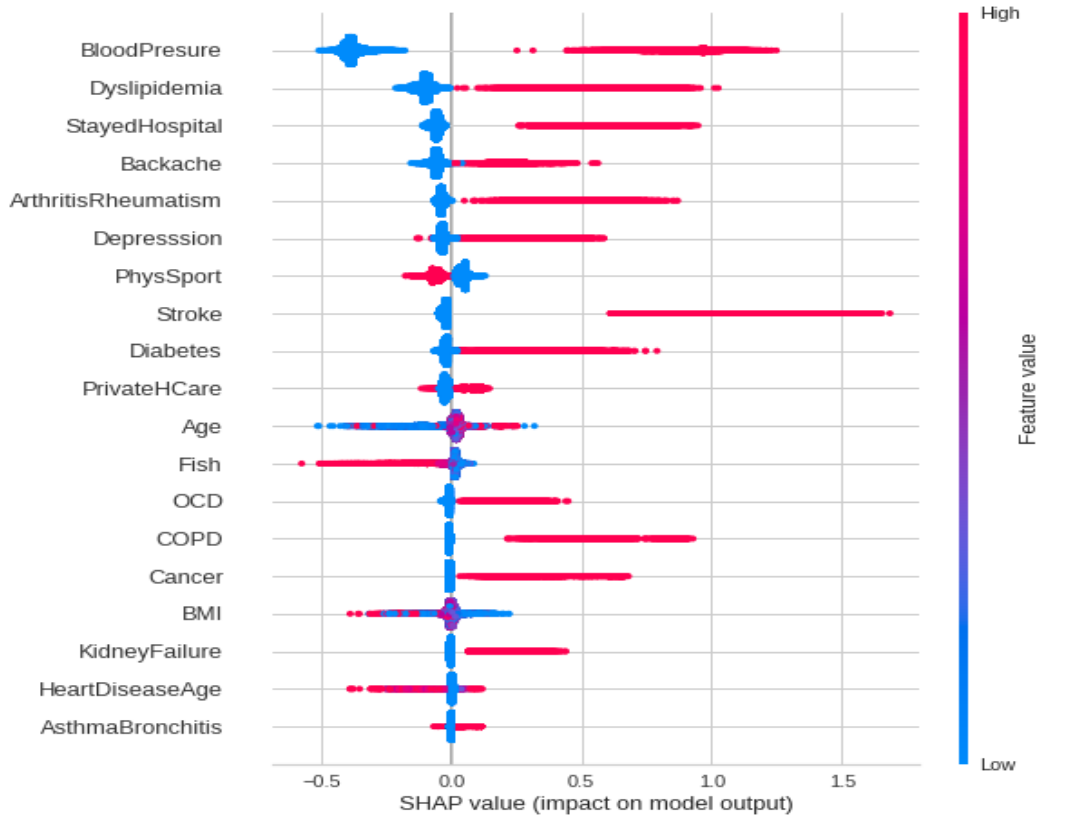

Fig 4 – SHAP value in the training dataset

  - Features Importance through SHAP values are in figure 4 below. Generally, the most important variables 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.

Conclusion:

The model proposed didn't perform as expected to figure out the True-Positive. However, it would be helpful to identify potential heart disease patients that are classified as False-Negative.

A health preventive/awareness program would be helpful to minimize the risk of those False-Negative developing heart disease in the medium or long term and, in this way, bringing intense demand pressure on both public and private health systems. And also, a program like that would give those patients a better quality of life and, at the end of the day: minimize an economic impact.