**Heart Disease Prediction Analysis**

**Machine Learning Model Proposal**

**Objective:**

Heart Disease is the leading cause of death for men and women in the United States. It is said that women are at a greater risk for Heart Disease than men. About 1 in every 5 female deaths is due to heart disease([www.cdc.gov](http://www.cdc.gov)). The diagnosis of heart disease can be challenging especially in women since they present atypical symptoms, as opposed to men. According to <https://pubmed.ncbi.nlm.nih.gov/26210899/>, “this, in association with a consistent underestimation of their risk for ischemic heart disease, leads to underdiagnosis and undertreatment in women.”

An automated system based on Machine Learning Models, that predicts the probability of heart disease in patients, based on medical and lifestyle parameters could be effective in predicting and preventing the disease and increased accuracy of diagnosis. This system would be beneficial to patients at risk, to hospitals and health systems in providing improved quality of care as well as to insurance companies.

**Data:**

The data extracted from the database is a join of 4 tables – Patient Demographics, Patient Habits, Pre-existing conditions and Parameters. The columns of the data frame created from the extracted data are :

* Female Patient ID – Unique ID of each patient
* Age
* Education
* Whether the patient is a smoker currently
* How many cigarettes per day
* Whether they take BP Meds
* Whether they have a stroke
* Whether there is prevalent Hypertension
* Whether the patient has Diabetes
* The Total Cholesterol level
* Systolic Blood Pressure reading
* Diastolic BP reading
* BMI
* Heart rate
* Glucose level
* Ten year CHD: Whether they are likely to develop heart disease in the next 10 years.

**Data Preprocessing:**

The following actions were performed as part of data preprocessing:

1. The raw data from our data source had data for both male and female patients. This was then cleaned as part of the ETL process to only retain data for female patients in order to address the needs of this project.
2. Null Values were identified and dropped.
3. The number of rows and columns in the data were determined
4. The datatypes were determined to ensure that they were valid and there were no “object” datatypes.
5. Data was checked for existence of duplicate rows.
6. The number of unique values under each column was determined.
7. The Categorical and Numeric features were identified as follows:

|  |  |
| --- | --- |
| Categorical Features | Numeric Features |
| Education | Age |
| Current Smoker | Cigarettes per Day |
| BP Medication | Total Cholestrol |
| Prevalent Stroke | Systolic BP |
| Prevalent Hypertension | Diastolic BP |
| Prevalent Diabetes | BMI |
|  | Heart rate |
|  | Glucose |

Categorical features predominantly indicated whether a condition was present or not (except for Education). Numeric features were mostly actual parameters (except for Age).

1. The Value counts for each unique value under the categorical columns were determined.

**Feature Engineering:**

As part of preliminary feature engineering, data distribution and relationships were examined for each feature.

1. Boxplots were used to determine outliers. Extreme outliers in the Total Cholesterol and Systolic BP were removed.
2. A Statistical analysis was performed on the data frame using the describe() function.
3. Data Distribution Analysis was performed on each of the categorical features using Count Plots. It was found that the data for BP\_Meds, Diabetes and Prevalent\_Stroke were highly imbalanced.
4. There was a large imbalance in the distribution of the target value - Ten\_year\_CHD.
5. Data Distribution for Numeric Values was analyzed using distplots. It was found that the distribution for Cigarettes per Day was very uneven. Total Cholestrol, Systolic BP, Diastolic BP,Heart Rate, BMI and Glucose were more or less normally distributed.
6. In order to analyze the relationships between the different features, a Correlation heat map was generated. It was observed that both Education and Current\_Smoker had negative correlation with the target variable.

**Train- Test Split:**

1. Since the numeric variables were at a higher scale than the categorical variables, feature scaling was done using the Standard Scaler method.
2. Data was split into training and testing sets using the train\_test\_split method in a 3:1 ratio. The split was stratified.

**Feature Selection:**

1. Patient\_female\_ID was removed from the Dataset since it was a unique ID and would not have an impact on the target variable.
2. Feature Scores were calculated using f\_Classif, Mutual\_info\_Classif and Chi2 tests.
3. SelectKBest method was used to help select the best features.
4. Based on the scores calculated and EDA, the features with consistent low scores/ importances were removed, namely – Prevalent Stroke, Heart Rate, Education, Cigarettes per day, Current Smoker.

**Feature Scaling:**

1. The data across the different features were on varying scales with some features being categorical 0s and 1s, while others were numerical parameters.
2. The StandardScaler, MinMax Scaler and RobustScaler methods were attempted to scale and normalize the data. The scores of the Machine learning models did not show a great difference based on the Scaling method selected.
3. Finally, the RobustScaler method was selected since the method was robust to outliers.

**Machine Learning Model:**

From the dataset, it is evident that the Ten year CHD is the target column which represents if a patient is likely to develop Coronary Heart Disease within the next 10 years or not. Most of the other columns are features that determine what the result would be in the Target column.

Given the pattern of the dataset, it has been proposed that a Binary Classification model be used as our Machine Learning Model in order to predict the likelihood of Heart Disease in a patient.

**Logistic Regression Method:**

Initially the logistic regression method was applied to build the classification model.

**Benefits:**

* Generated a good Accuracy score of 0.8779.
* Good precision, recall and f1 scores for 0.
* Good Precision score for 1.

**Limitations:**

* The recall score for the model predicting ‘1’ is low which means that the model has low sensitivity. This means that if a person does have heart disease, the probability of our model predicting it is low. In other words, there is a high probability of False Negatives.

The dataset has a high degree of imbalance between the target values 0 and 1, where 0 indicates that the probability of heart disease is False while 1 indicates that the probability of heart disease is true.

To address this, the following methods were attempted.

**Logistic Regression with Balanced Class Weights:**

The Logistic Regression method was applied with the parameter: Class\_Weight = “balanced”. According to the documentation, The “balanced” mode uses the values of y to automatically adjust weights inversely proportional to class frequencies in the input data as n\_samples / (n\_classes \* np.bincount(y)).

**Benefits:**

* Generated Good precision, recall and f1 scores for 0 and 1.

**Limitations:**

* Accuracy is low at 0.6830

**Logistic Regression after Moving Threshold:**

Source: https://machinelearningmastery.com/threshold-moving-for-imbalanced-classification/

When predicting whether a value belongs to a certain class ie.0 and 1, the machine learning algorithms use a threshold such as 0.5 where all values equal or greater than the threshold are mapped to one class and the rest are mapped to the other class. Imbalanced classification can be addressed by moving the threshold used to map probabilities to class labels.

The y values were predicted using this approach and the Logistic regression method was applied.

**Benefits:**

* Generated Good precision, recall and f1 scores for 0 and 1.

**Limitations:**

* Moderate Accuracy Score at low at 0.7775

**Other Models:**

The following methods were also used to build the model.

* Support Vector Machine Model
* Decision Tree Model
* Decision Tree Model with Class Weights
* Random Forest Classifier Model
* Gradient Boosting Method

**Benefits:**

* Almost all the models produced good Accuracy Scores as can be seen below:

|  |  |
| --- | --- |
| Model Name | Accuracy Score |
| Support Vector Model | 0.8779 |
| Decision Tree Model | 0.7637 |
| Decision Tree Model w/Class Weights | 0.807 |
| Random Forest method | 0.875 |
| Gradient Boosting Model | 0.8602 |

* All models generated good Precision, Recall and F1 scores for 0 and 1.
* The Decision Tree Classifier with Class weights generated a decent accuracy score at .807 and a better recall score for 1, than other models.

**Limitations:**

* The Decision Tree Classifier had a lower accuracy score but had a better recall score for 1.

**Challenges:**

1. The primary challenge faced was the Class Imbalance in the target values. There were a disproportionately large number of 0 as compared to 1s.
2. Resampling – both over and under sampling as well as Combination Sampling were attempted but they led to very low accuracy scores.
3. Finding an optimal balance between the accuracy score as well as precision and recall scores was a challenge. In the several models attempted, with class imbalance adjustments, either the precision and recall scores were impacted or the accuracy score was impacted. A tradeoff had to be made to arrive at the optimal sample which had both a decent accuracy score as well as precision and recall scores.

**Model Recommendation:**

Based on the several models attempted and the results, the first model of choice is the Decision Tree Classifier with class weights. The model has an accuracy of 0.807. Even though this is lower than the other models, this model is recommended for its better recall score for 1, since a good recall is important in Heart Disease Diagnosis.

A picture containing text, receipt

Description automatically generated

The above model’s confusion Matrix suggests the following:

|  |  |  |
| --- | --- | --- |
|  | Predicted True | Predicted False |
| Actually True | True Positive 397 | False negative 49 |
| Actually False | False Positive 49 | True negative 13 |

* There are 397 true Positives(for 0). This means that if a person is predicted to not have heart disease(0), then it is really true.
* There are 49 False Positives(for 0). This means that if a person is predicted to not have heart disease, but actually has heart disease .
* There are 49 False Negatives (for 0). This means that if a person is predicted to have heart disease but does not have it.
* There are 13 True Negatives (for 0). This means a person is predicted to have heart disease and has it.

Based on Accuracy score, the Gradient Boosting Model is recommended. It has an accuracy score of 0.8602. While other classifier models have better accuracy scores, this one has a better recall score.

A picture containing text, receipt

Description automatically generated

The best model in terms of Recall score is the Logistic regression Model after moving thresholds. Even though the accuracy score is at 0.77 which is nearly 0.80, it has a very good recall score for 1 (predicting risk of heart disease is true).

A picture containing text, receipt, screenshot

Description automatically generated