Capstone Project - 3

Cardiovascular Risk Prediction

Presented by:

Mrugesh Patel

Content

1. Problem Statement
2. EDA and feature engineering
3. Data Transformation & preparation
4. SMOTE sampling
5. Applying different models
6. Model selection

Problem Statement

- A huge number of people experiences heart related problems in their life span, It can be very helpful if we can predict and prevent it from happening

- As the prediction made is whether a person is diagnosed with cardiovascular heart disease, it is very vital to predict the outcome accurately, especially predicting the class ‘YES (1)' accurately with as less false negatives as possible (predicting NO for actual YES) is very necessary

- Recall Score is a metric to identify the effect of false negatives while predicting.

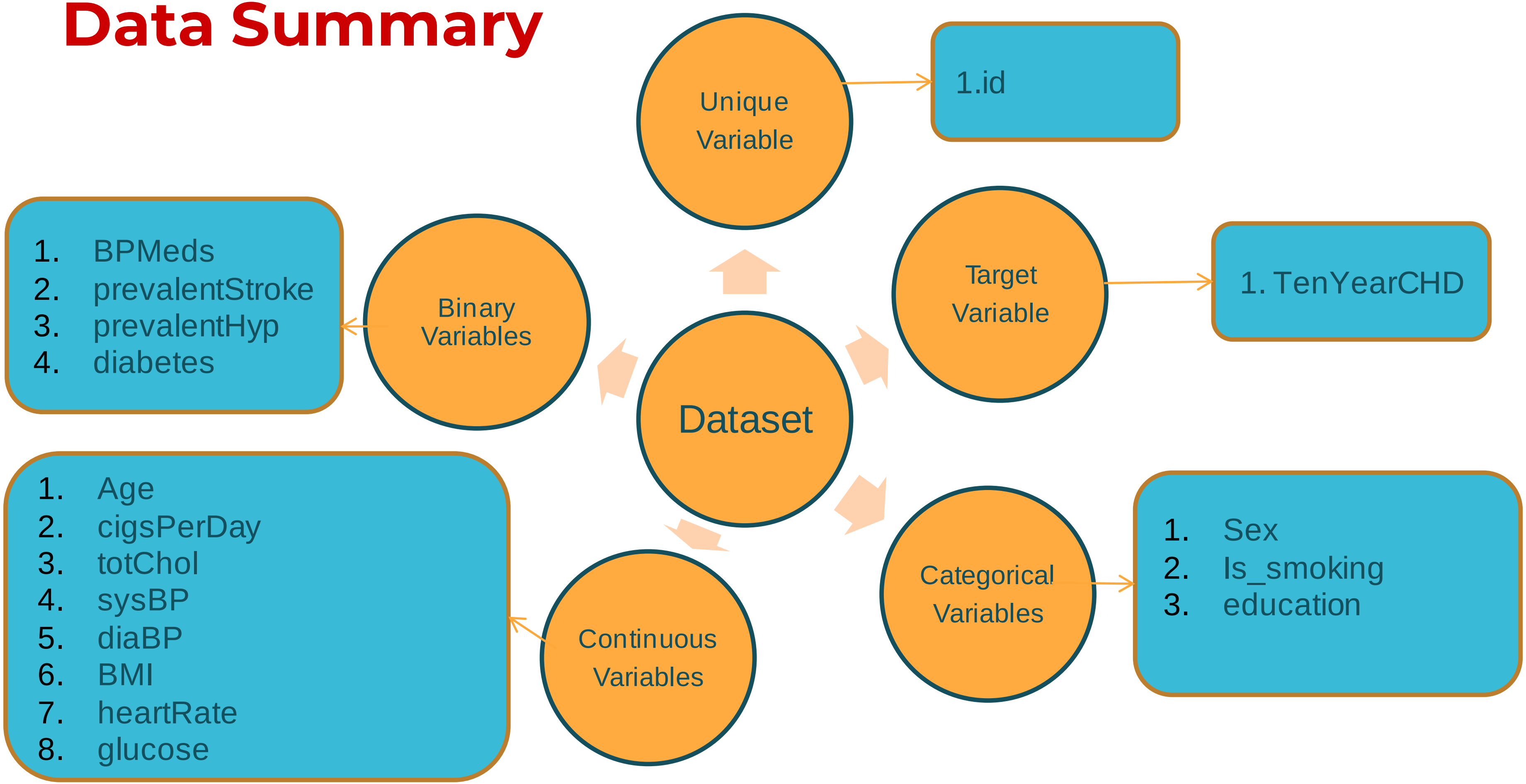
- So the problem statement for the project is that to get the recall score and f1 score as high as possible for test dataset, ideally 1.0.

- So we need to formulate a model with at least 0.95 recall score and f1 score.

Data Pipeline

- **Data exploration and Reading:** Dataset summary and description, understanding features and target variable
- **EDA:** Uni and bi variate analysis for features
- **Feature Engineering:** Encoding of categorical variables and Data imputation of null values, correcting and changing unusable data, feature selection
- **Data Transformation preparation:** Transformation of features and target variable and preparing train and test datasets
- SMOTE Sampling : Generating data for balancing target variable classes
- **Creating a Model and Model Selection:** Creating different models with different machine learning algorithms, and selection of best model based on valid metric score using cross validation

Data Summary



Dataset

Unique Variable

1.id

Target Variable

1. TenYearCHD

Binary Variables

1. BPMeds
2. prevalentStroke
3. prevalentHyp
4. diabetes

Continuous Variables

1. Age
2. cigsPerDay
3. totChol
4. sysBP
5. diaBP
6. BMI
7. heartRate
8. glucose

Categorical Variables

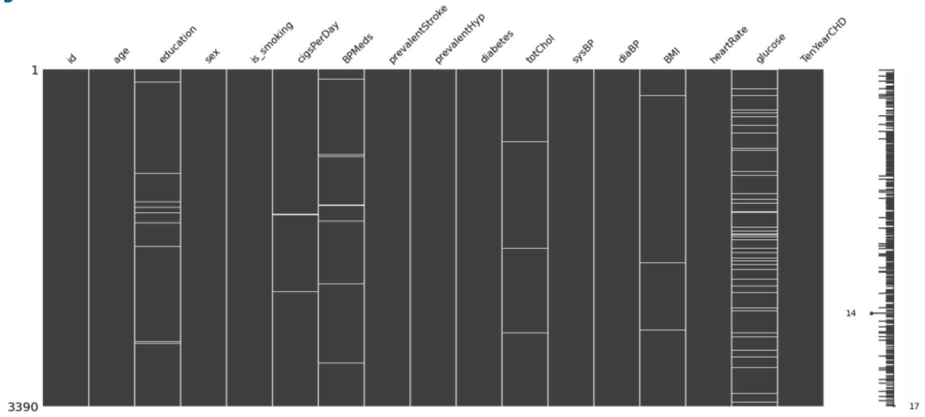
1. Sex
2. Is_smoking
3. education

Data Reading and Exploration

- 3390 patients records, with 15 different attributes for each record
- 4 types of attributes :

➢ Demographic, Behavioral, Medical (current and history)

- Out of 14 variables,; education cigsPerDay, BPmeds, totChol,BMI, heartrate, gluscose columns have some Null values.

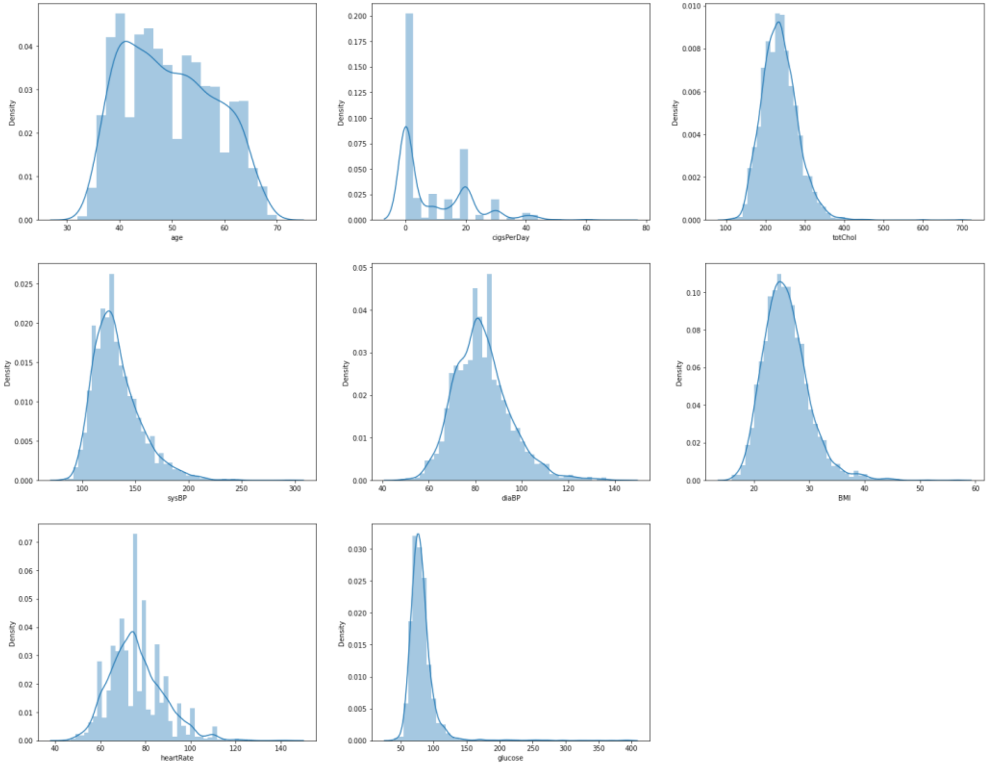


- **Target variable has imbalance of classes (0/1 : 2879/511)**

EDA

• Univariate Analysis

Distribution plots of continuous features



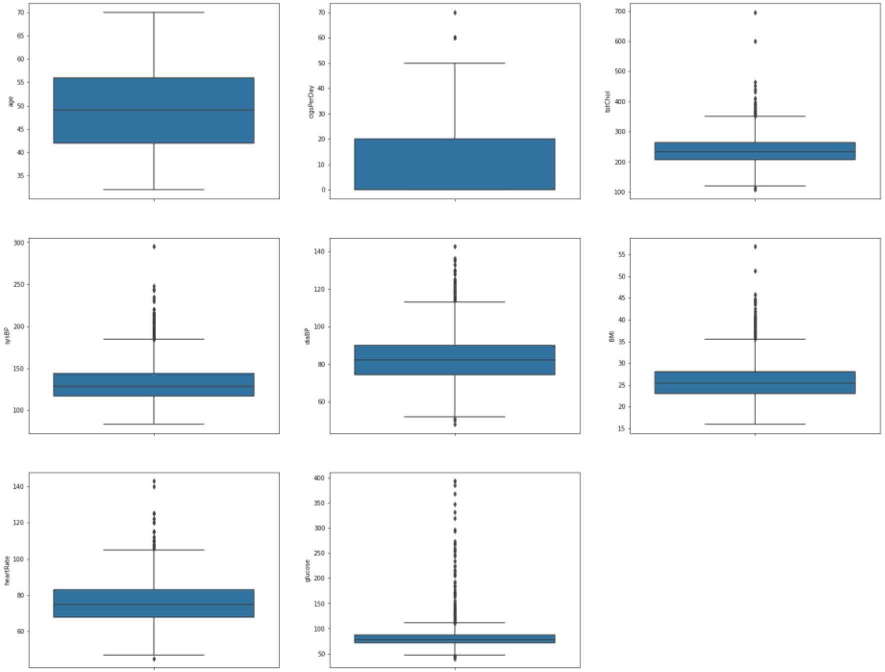
- Most distributions are positively skewed
- Some Transformation must be applied before data preparation.

EDA

• Univariate Analysis

Box plots of continuous features

- There are outliers in most of the numerical features, but as the dataset small we can not afford to drop any data. Also tree based algorithms and SVM , naive-bayes are **not** very sensitive to the outliers, so we can use those algorithms for the models.



Feature Engineering

- **Null Value Treatment**

➢It is safe to assume that the Null values in the cigsPerDay columns can be imputed as 0, as is_smoking values are NO.

➢totChol,BMI,glucose distribution is slightly positively skewed, so the null values can be imputed by the median of the rest of the values for respective columns.

➢heartrate values distribution is close to normal distribution, so mean value can be imputed for null values in the column.

➢ For BPMeds and education mode can be appropriate value for imputing null values for respective columns

- **Encoding**

➢Here there are two categorical variables with str type values, 'sex' and 'is_smoking'.

➢Now the column 'is_smoking' can be dropped, as 0 value in column cigsPerDay conveys same thing. So it would be redundant.

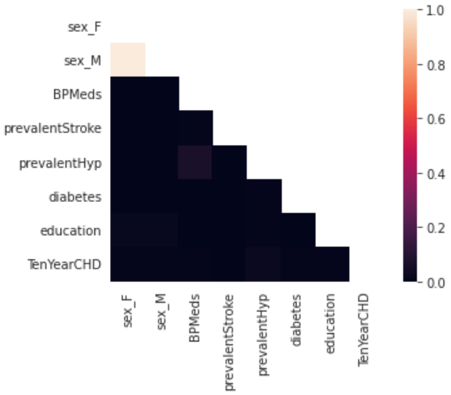
➢And for sex values, one hot encoding can be used.

Feature Engineering

- **Correlation**

❑ **Categorical-categorical correlation**

cramers_V values as a Correlation measure for Categorical-Categorical pairs

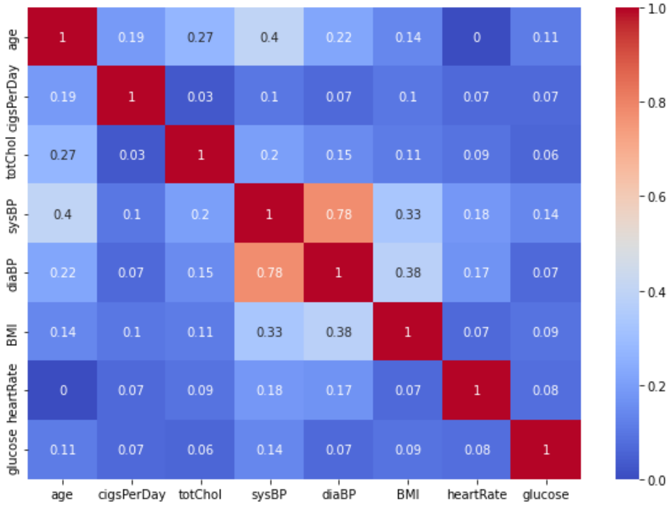


❑ **Continuous-Continuous Correlation**

➢ high correlation for sysBP and diaBP pair

➢Pulse pressure , difference b/w sysBP and diaBP

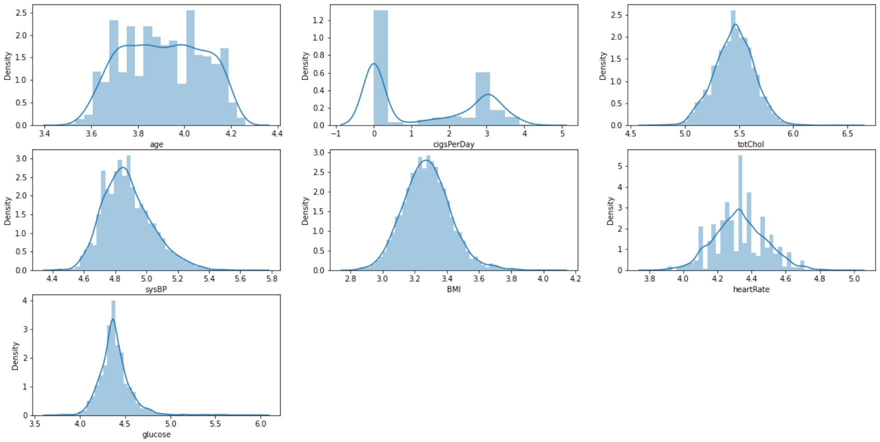
➢ So diaBP column is dropped and Pulse pressure is calculated and added in the dataset to move forward.



Data Preparation

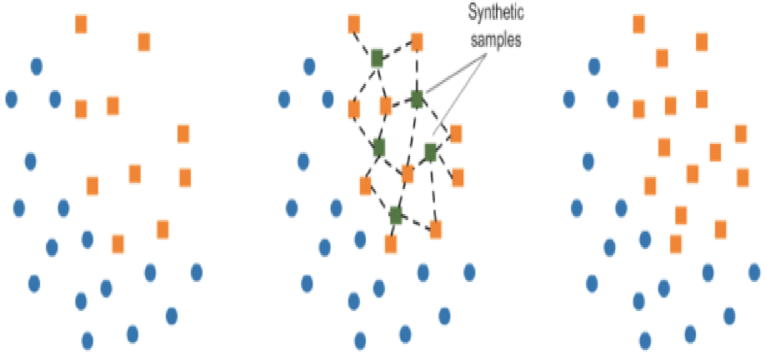
❖ Transformation and Standardization

- log1p transformation is applied to continuous variables
- All continuous feature values were standardized before sending the dataset for training, because standardization is very important to achieve uniformity in the dataset in order to have the better model.



SMOTE sampling

- SMOTE (Synthetic Minority Oversampling Technique) works by randomly picking a point from the minority class and computing the k-nearest neighbors for this point. The synthetic points are added between the chosen point and its neighbors.



Models experimentation

- Base line model : - Logistic Regression
- Random Forest, XGBoost, Adaboost
- Guassian Naïve Bayes
- KNeighborsRegressor
- SVC

➢All models were fitted on SMOTE sampled data and the results for both SMOTE sampled and original test datasets were generated with those models.

Models experimentation

❖ Observations:

- Baseline Model produced 0.65 recall score for SMOTE data and 0.51 for original, for class 1 of target variable
- Random and Adaboost performed well for SMOTE data but generated poor results for original dataset
- XGBoost produced 0.88 recall score for both sets of datasets after tuning
- Knn produced excellent recall score for both datasets, but precision was very poor
- SVC model produced 0.93 recall score initially, after tuning it reached 0.96 for SMOTE sampled data, and for original data generated recall score of 1.0 with 0.95 precision

Results

SMOTE sampled data

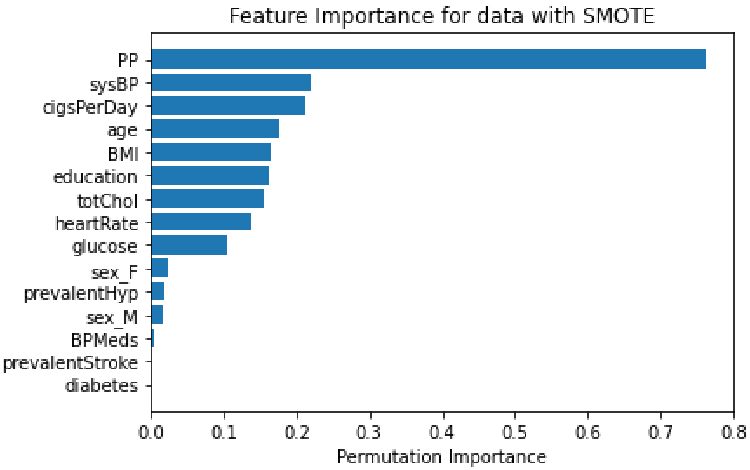
Model	Precision	Recall	f1-score
Logistic Regression	0.75	0.65	0.7
Random Forest	0.81	0.81	0.81
XGBoost	0.93	0.88	0.9
Gaussian NB	0.7	0.5	0.58
ADA Boost	0.91	0.81	0.85
KNN	0.78	0.96	0.86
Support Vector Machine	0.96	0.96	0.96

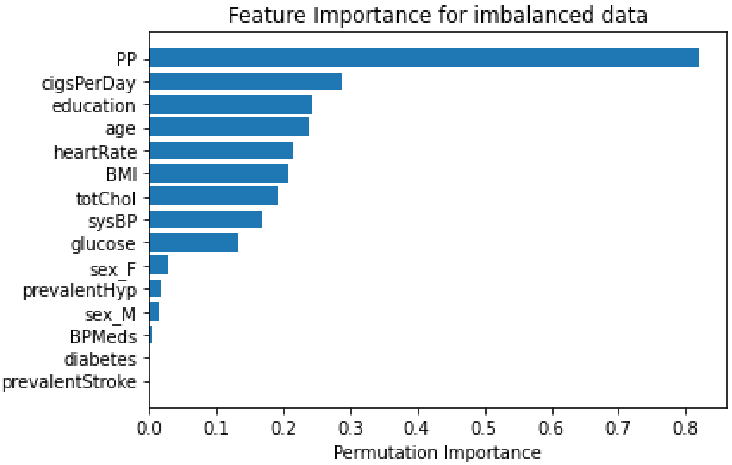
Original data

Model	Precision	Recall	f1 score
Logistic Regression	0.24	0.51	0.33
Random Forest	0.36	0.62	0.46
XGBoost	0.9	0.88	0.89
Gaussian NB	0.25	0.53	0.34
ADA Boost	0.25	0.53	0.34
KNN	0.52	0.98	0.68
Support Vector Machine	0.95	1.0	0.97

The Parameters of the finally selected model are,
Kernel='**rbf**', class_weight='**balanced**’, Probability=True, C = 10, gamma= 0.5.

Feature Importance





Future Scope

- As the prediction made by study is very essential, more data for class YES can be collected to get better validation.
- XGB model after cross validation improved much better can be further improved to match svc model's result.

Conclusion

- SMOTE sampling as class imbalance in the target variable, experimentation with 7 different models
  - Random and Adaboost performed well for SMOTE data but generated poor results for original dataset
  - XGBoost produced 0.88 recall score for both sets of datasets after tuning
  - Knn produced excellent recall score for both datasets, but precision was very poor
- The model performance parameters for SVC,

➢ For SMOTE sampled Data,

Precision	**Recall**	**f1 score**
0.96	**0.96**	**0.96**

➢ For Original Data

Precision	**Recall**	**f1 score**
0.95	**1.0**	**0.97**

➢ As we can see the recall value of 1.0 was achieved for test dataset of original data (without smote sampling). So we can confidently classify the cardio vascular heart disease with svc model.

➢ Pulse pressure and cigsPerDay are most important features for predicting the target variable.

Thank You