Capstone Project

Cardiovascular Risk Prediction

By - Shourya Chandra Sai

Project Agenda

- Segment 1: Data Description
- Segment 2: Defining the Problem Statement
- Segment 3: Data Preprocessing
- Segment 4: Exploratory Data Analysis(EDA)
- Segment 5: Logistic Regression
- Segment 6: K-nearest Neighbours
- Segment 7: Comparison Between the models
- Segment 8: Conclusion

Segment 1: Data Description

Below is the description of all the columns in the dataset.

- **Sex**: male or female("M" or "F")
- **Age**: Age of the patient
- **is_smoking**: whether or not the patient is a current smoker ("YES" or "NO")
- **Cigs Per Day**: the number of cigarettes that the person smoked on average in one day
- **BP Meds**: whether or not the patient was on blood pressure medication
- **Prevalent Stroke**: whether or not the patient had previously had a stroke
- **Prevalent Hyp**: whether or not the patient was hypertensive
- **Diabetes**: whether or not the patient had diabetes
- **Tot Chol**: total cholesterol level
- **Sys BP**: systolic blood pressure
- **Dia BP**: diastolic blood pressure
- **BMI**: Body Mass Index
- **Heart Rate**: heart rate

- **Glucose**: glucose level
- **education**: education level of a person.
- **10-year risk of coronary heart disease (CHD)**(binary: “1”, means “Yes”, “0” means “No”): Informs us whether that person has 10-year risk of CHD or not.

Segment 2: Defining the Problem Statement.

Dependent variable(Y): TenYearCHD is categorical in nature with 2 classes 0 & 1.

Independent variable(X):

Sex = It is categorical in nature with 2 classes Male & Female.
Age = It is continuous in nature.
Education = It is categorical in nature with 4 classes 1, 2, 3, & 4.
is_smoking = It is categorical in nature with 2 classes 1 & 0.
Cigs Per Day: It is continuous in nature as one can consume any number of cigs.
BP Meds: It is categorical in nature with 2 classes 1 & 0.
Prevalent Stroke: It is categorical in nature with 2 classes 1 & 0.
Prevalent Hyp: It is categorical in nature with 2 classes 1 & 0.
Diabetes: It is categorical in nature with 2 classes 1 & 0.
Tot Chol: It is continuous in nature.
Sys BP: It is continuous in nature.
Dia BP: It is continuous in nature.
BMI: It is continuous in nature.
Heart Rate: It is continuous in nature.

Segment 3: Data Preprocessing

- **Dealing with Null and Duplicate values -** There weren't any duplicate values in the dataset. There were a few columns(categorical and continuous) which had null values.

  Categorical variables - For these variables I had replaced the null values with their modes.

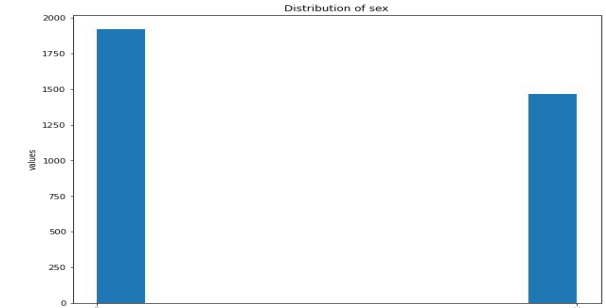
  Continuous variables - For these variables I had replaced the null values with median. There were outliers in these columns and since mean is sensitive to outliers I have used median for replacement of null values.

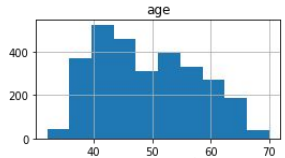
- **Dealing with outliers -** Removing outliers is important because logistic regression is sensitive to outliers. Using a for loop I have removed the outliers in the dataset and stored the data frame in no_outliers_df. This data frame, I have used as an input to logistic regression. I have used another data frame(with outliers) as an input to KNN because KNN is robust to outliers.

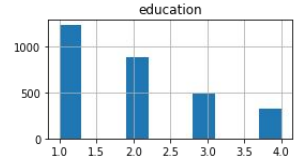
Segment 4: Exploratory Data Analysis.

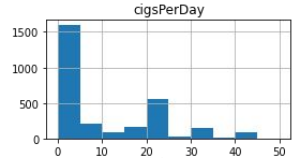
- **Uni-variate Analysis**

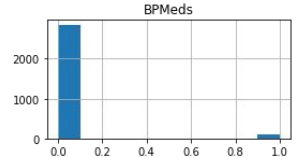
Objective: Helps us understand the distribution of a particular column.The most common way to understand distribution is through plotting histograms. Below are the histograms of all the variables.

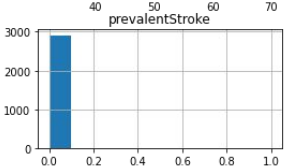


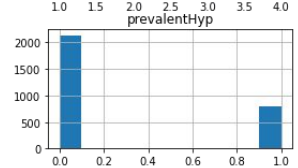


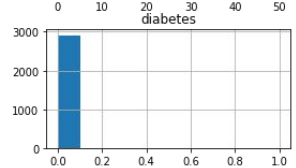


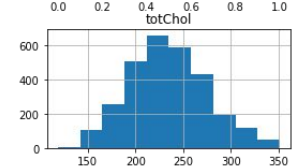


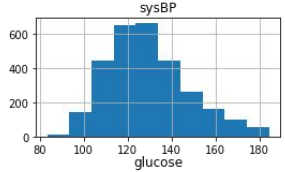


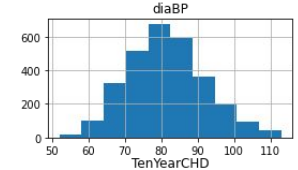


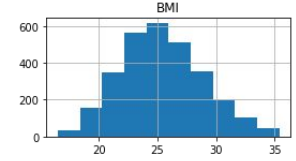


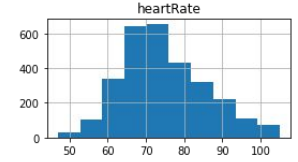


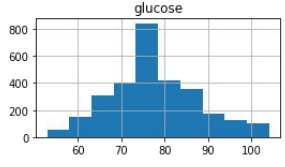


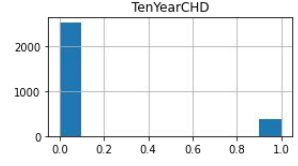








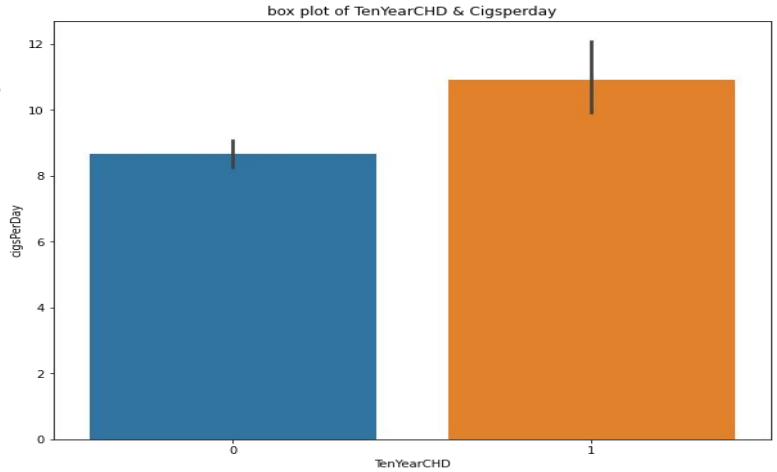


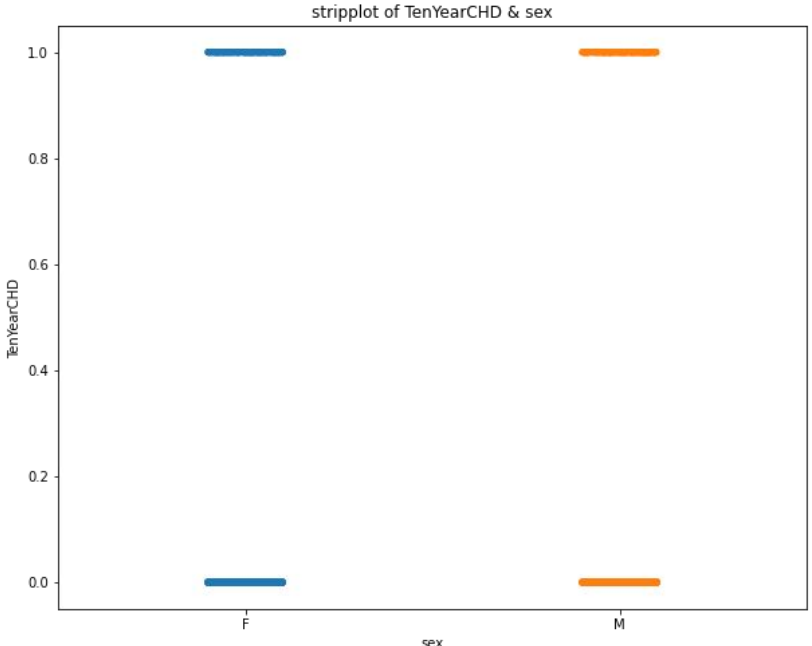


- **Bi-variate Analysis**

Objective: Helps us understand the relationship between the dependent and independent variables.

This is a box plot showing the relationship between the dependent variable and cigs per day. As you can see those who smoke more cigs are prone to 10yearCHD.

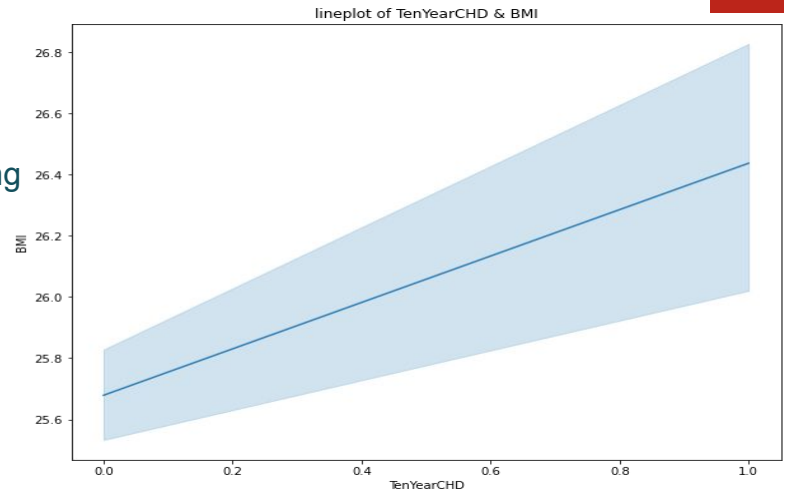




This is a stripplot showing the relationship between sex and 10YearCHD. There isn't any noticeable relationship here. Looks like equal number of male and female have 10YearCHD.

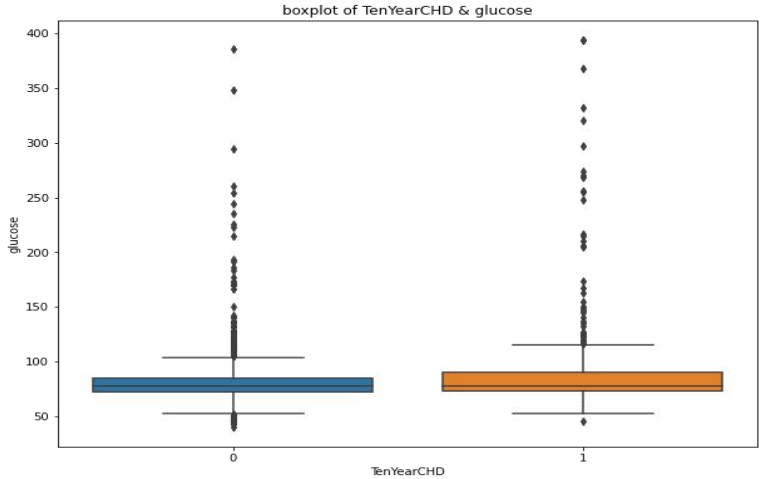
This is a line plot showing the relationship between BMI variable and 10YearCHD variable.

The ideal BMI is between 18.5 & 24.9. As you can see from the plot anything above 24.9 BMI has a positive relationship with 10YearCHD.



This is a box plot showing the miniscule relationship between glucose and 10YearCHD

The median value of glucose for those who smoke is slightly larger than the median value of glucose for those who do not smoke.



Segment 5: Logistic Regression

Logistic regression is a classification algorithm that predicts the probability of an outcome that can only have two values (i.e. a dichotomy). A logistic regression produces a logistic curve, which is limited to values between 0 and 1.

Objective: In this segment I had made sure that all the assumptions of Logistic regression have been met, In order to get the best out of Logistic Regression. And also I had implemented the Model.

Assumptions of Linear Regression:

- Binary logistic regression requires the dependent variable to be binary and ordinal logistic regression requires the dependent variable to be ordinal.
- Logistic regression requires the observations to be independent of each other. In other words, the observations should not come from repeated measurements or matched data.
- Logistic regression requires there to be little or no multicollinearity among the independent variables. This means that the independent variables should not be too highly correlated with each other.

Encoding the variables.

Encoding the categorical variables is important because machine learning models can only interpret numbers.

Categorical Variables are classified into 3 types:

- Dichotomous variables = These are those categorical variables which have only **2 classes** of categories. We can perform binary encoding for these kind of variables.

- Ordinal Variables = These are those categorical variables which have **more than 2 classes** of categories and they signify **some order**. We can perform Ordinal encoding.

- Nominal Variables = These are those categorical variables which have **more than 2 classes** of categories and they do not signify **any order**. We can perform One hot encoding.

From the above dataframe we can see that only is_smoking column and sex column has to be encoded. I will be performing a binary encoding on these columns. All the other categorical variables have already been encoded.

Correlation and Multicollinearity.

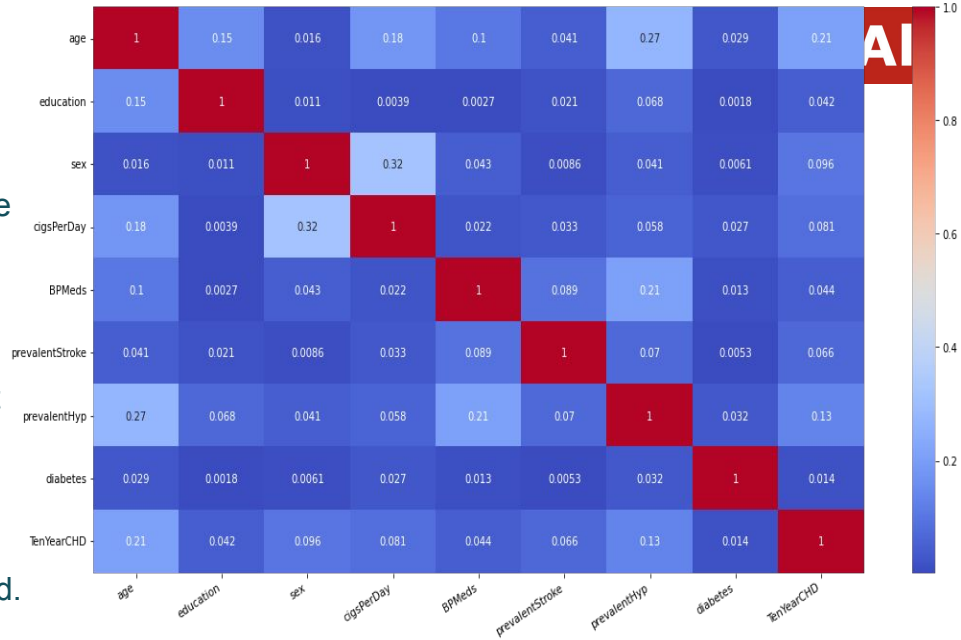
Correlation: It tells us how strong the relation is between the variables. The main aim here is to drop those variables that have a low correlation with the dependent variable. Also drop those variables whose correlation with the variables (excluding the dependent variable) is high.

Multicollinearity: Logistic regression assumes that there is no correlation between the independent variables. One way to deduct multicollinearity is through VIF(variance inflation factor).

As a rule of thumb if VIF of a particular variable is below 2 we can say that the variable in subject is not related to other independent variables. If the VIF is between 2 and 5 they are moderately related. and if the VIF is above 5 we say that they are highly correlated.

Approach used to reduce the VIF- We will check what other variable has a VIF value of the same range.

We will check both the variables correlation to confirm our finding that they indeed are correlated and remove the feature that has the lowest correlation with **TenYearCHD**



	variables	VIF
0	age	5.275794
1	education	4.034014
2	sex	2.011139
3	cigsPerDay	1.756689
4	prevalentHyp	1.521079
5	BPMeds	1.092781
6	prevalentStroke	1.018130
7	diabetes	1.007193

Dealing with Imbalanced data

Meaning of an Imbalanced dataset: When the classes of the dependent variable are not equally distributed then we say that the dataset is imbalanced.

Why is Imbalance dataset bad?

Imbalance Dataset is not good because the model might predict skewed results towards the majority class.

Approach that I will be using to solve this problem: Since the number of observations is small I can't choose undersampling techniques. From oversampling techniques I have decided to go for SMOTE(Synthetic Minority Oversampling Technique).
What is SMOTE?

This technique generates synthetic data for the minority class.

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.

SMOTE algorithm works in 4 simple steps:

* Choose a minority class as the input vector
* Find its k nearest neighbors (k_neighbors is specified as an argument in the SMOTE() function)
* Choose one of these neighbors and place a synthetic point anywhere on the line joining the point under consideration and its chosen neighbor
* Repeat the steps until data is balanced

Evaluating the results of logistic regression

I have used 4 metrics to evaluate the performance of the logistic regression model. They are confusion matrix, classification report, accuracy score and roc_auc_score.

The results were not fruitful. The accuracies & roc_auc_scores were low. The precision, recall, f-1 score from classification report were low. The FN(false negatives) & FP(false positives) were also also high from the confusion matrix. This means that the model has done a lot of misclassifications.

Hyperparameter Tuning

Even after tuning the hyperparameters like C, Regularization penalty(L1, L2) the models performance has not improved at all.

Segment 6: K-nearest Neighbor

K-nearest neighbors (kNN) is a supervised machine learning algorithm that can be used to solve both classification and regression tasks. kNN as an algorithm seems to be inspired from real life. People tend to be affected by the people around them. Our behaviour is guided by the friends we grew up with. Our parents also shape our personality in some ways. If you grow up with people who love sports, it is highly likely that you will end up loving sports. There are of course exceptions. kNN works in a similar fashion.

Below are the steps involved in KNN:

Step-1: Select the number K of the neighbors

Step-2: Calculate the Euclidean distance of K number of neighbors

Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.

Step-4: Among these k neighbors, count the number of the data points in each category.

Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.

Step-6: Our model is ready.

Evaluating the results of KNN.

Like in logistic regression, In KNN also I have used the same evaluation metrics. The results from KNN were encouraging. The accuracy and roc_auc_score were high. The precision, recall, and F-1 score were high. There were fewer misclassifications by the model.

Hyperparameter Tuning

After tuning the k hyperparameter, the models performed improved. The best K parameter was 1.

Segment 7: Model Comparison

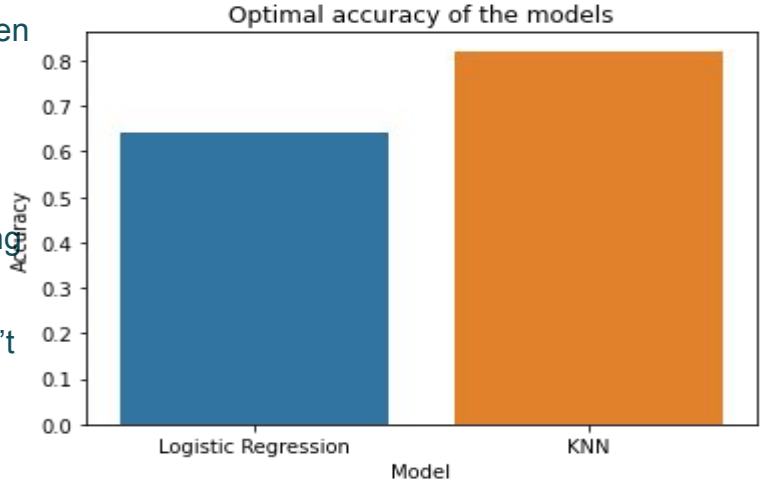
I have taken accuracy as a metric for comparison because the dataset has been balanced. If the dataset had not been balanced I would have taken recall metric(for this problem) for comparison.

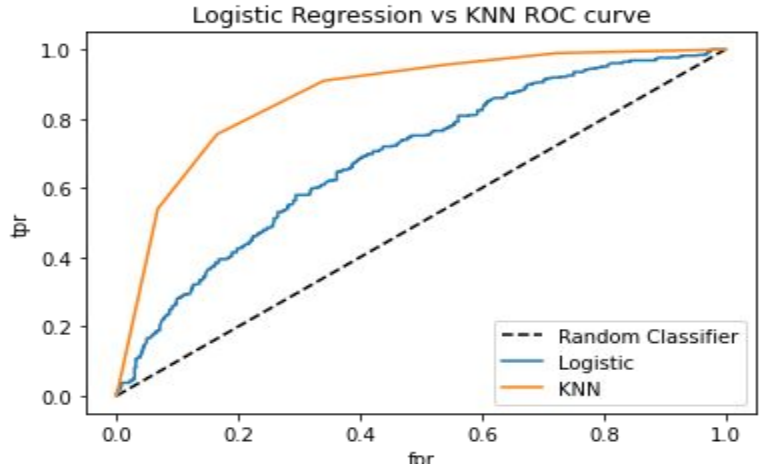
Why recall?

For this problem,reducing FN(actual:1, predicted:0) is important. Since reducing FN is important I would have selected recall(TP / FN + TP) for optimisation. Reducing FN is important because if the model predicts that the person doesn't have 10YearCHD and he actually does have it then his life might be in danger.

From the bar plot, you can see that the accuracy of KNN is way more than Logistic regression.

From the roc_curve also we can come to the same conclusion. KNN model outshines logistic regression.





Segment 8: Conclusion.

After implementing two models for the classification task I came to the conclusion that using KNN would provide fruitful results. Sometimes simple models too can get the work done.

Thank you.