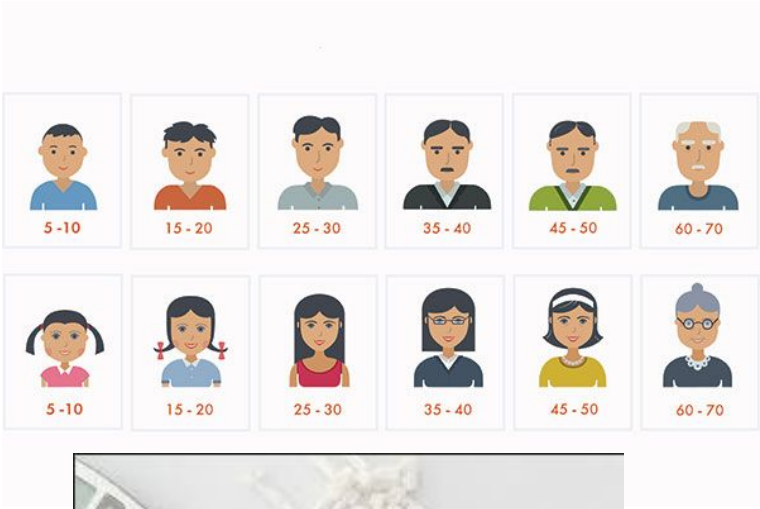
Capstone Project - II

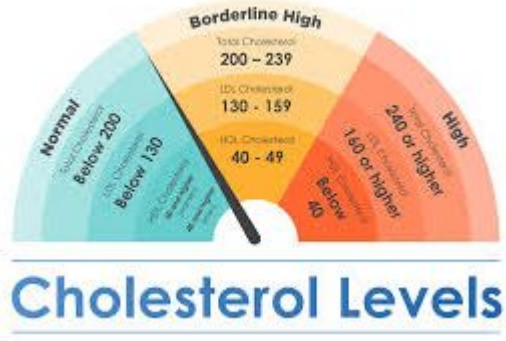
Cardiovascular Risk Prediction

Problem statement :

- Coronary heart disease is caused due to accumulation of plaque in major heart blood vessels leading to blockage of oxygen-rich blood to heart.
- It is the most common type of heart disease, killing about 300 K people in US alone every year.
- The goal of our project is to come up with a ML model that correctly predicts 10-year risk of a patient having coronary heart disease (CHD).
- The very important metric that we want to focus on is the **Recall** metric since we want to minimize false negatives i.e. person with 10-year CHD risk should be flagged positive by the model.

Features Present in Dataset :

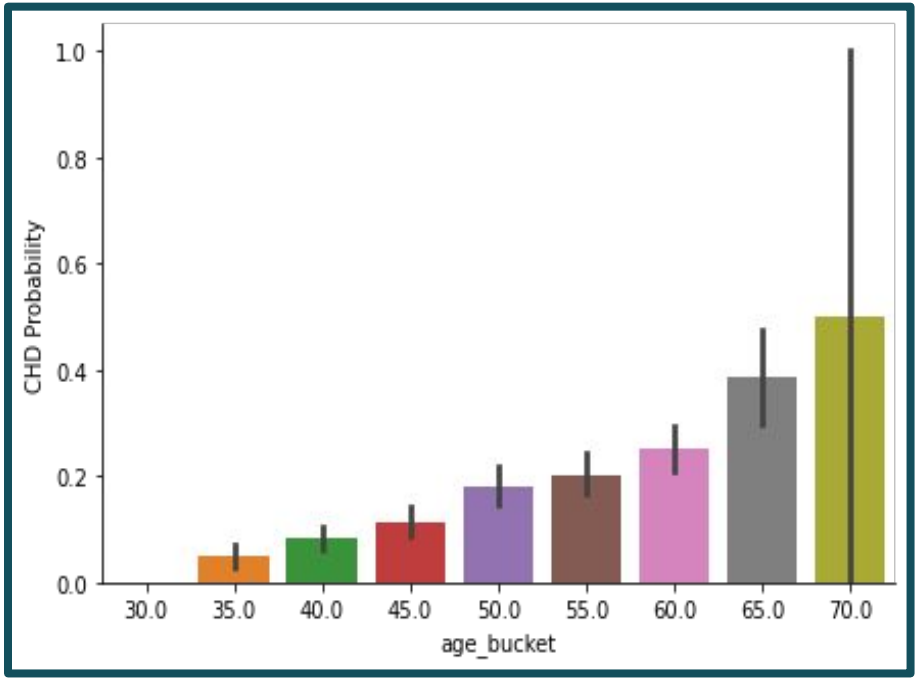




Exploratory Data Analysis

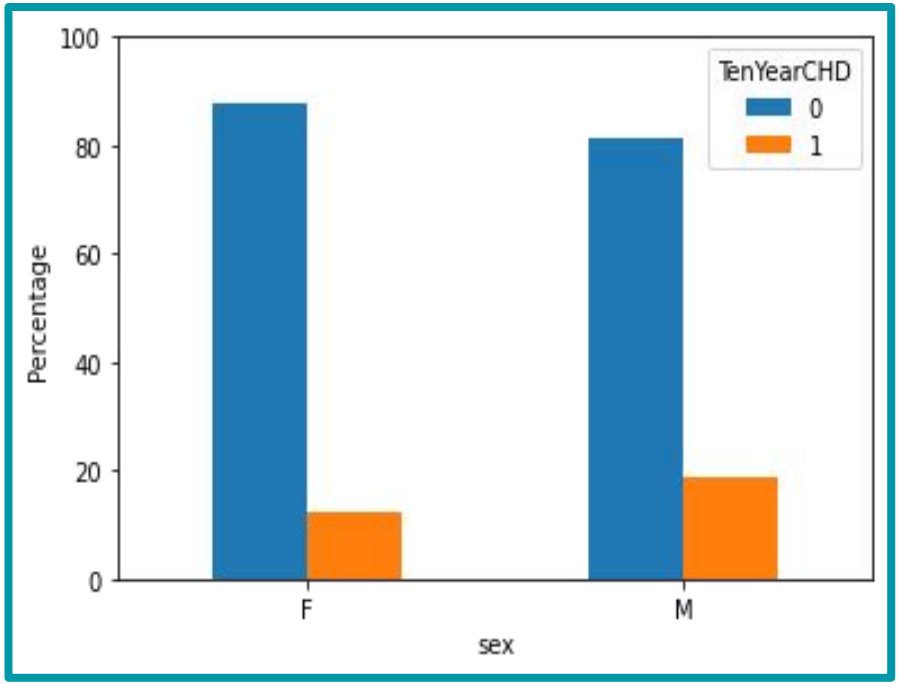
Does age play any role ?

Older people have a higher risk of Having coronary heart disease in next 10 years



Sex

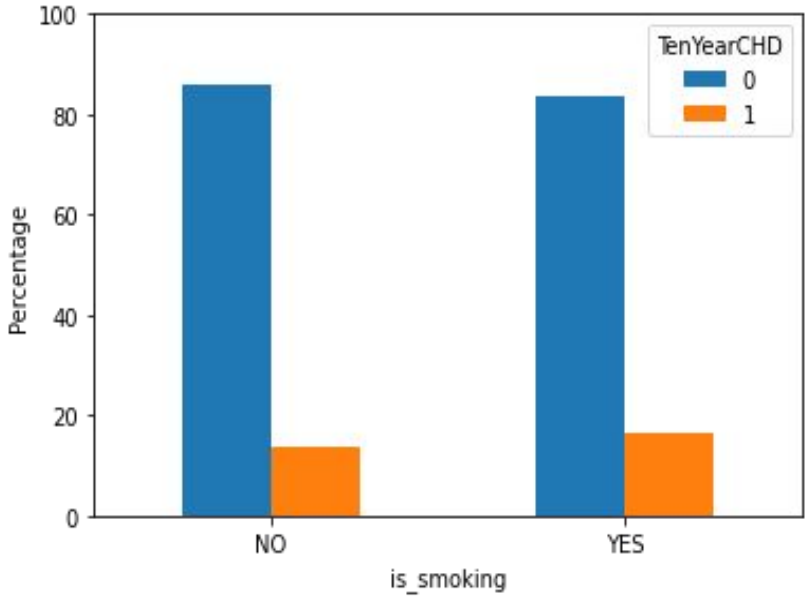
Men are generally at a higher risk of having coronary heart disease



Smoking ?

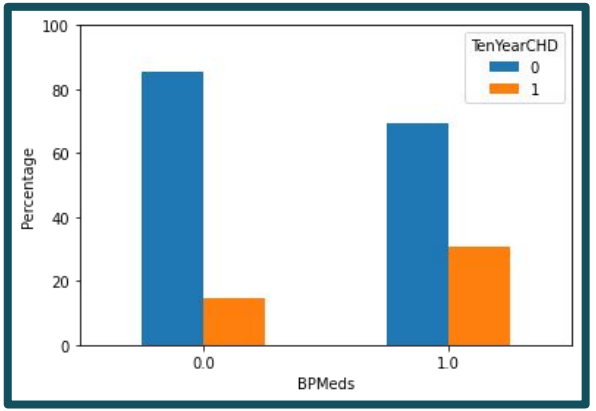
Contrary to what we might anticipate, Smoking has little to no role to play in affecting the risks of CHD.

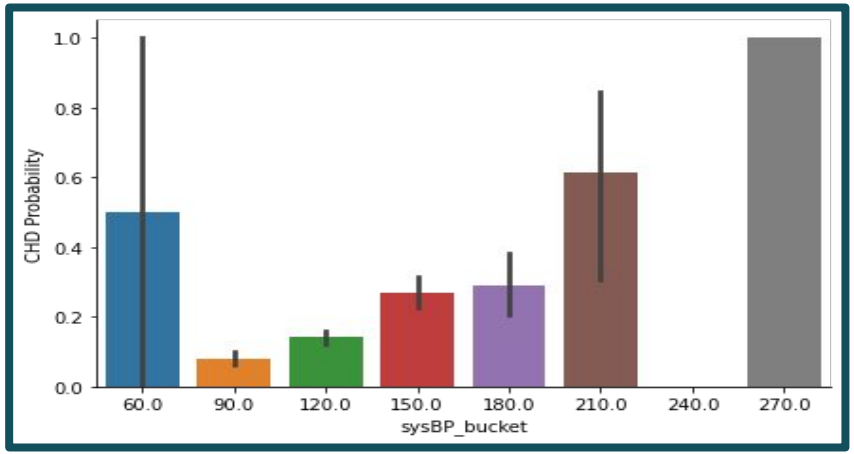
Statistically, 10-year risk of CHD is not dependent on smoking with a 95% confidence.

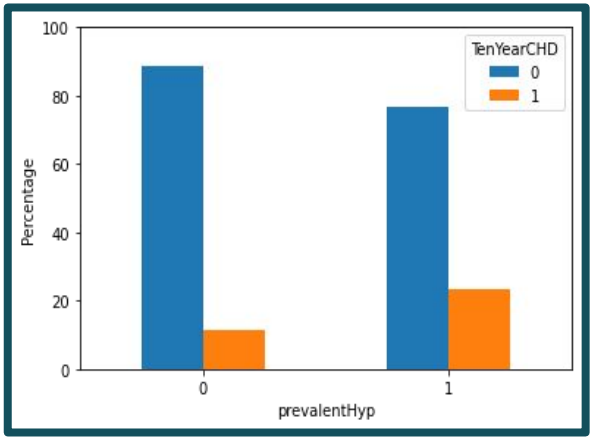


Other Notable Observations :

- Patients who have a high blood pressure, have a history of hypertension and have been taking BP medication have comparatively higher risk of CHD

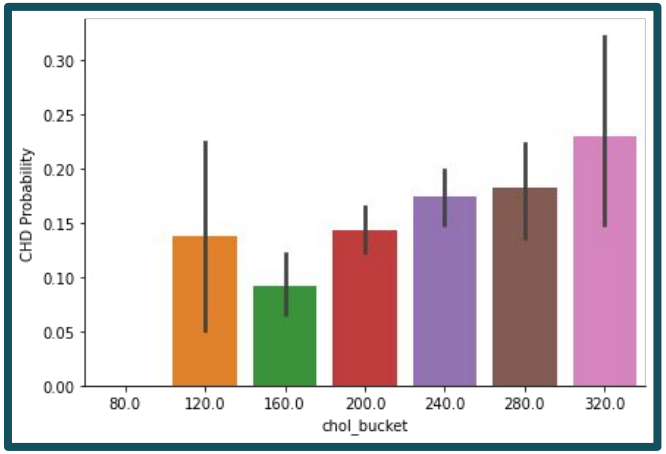


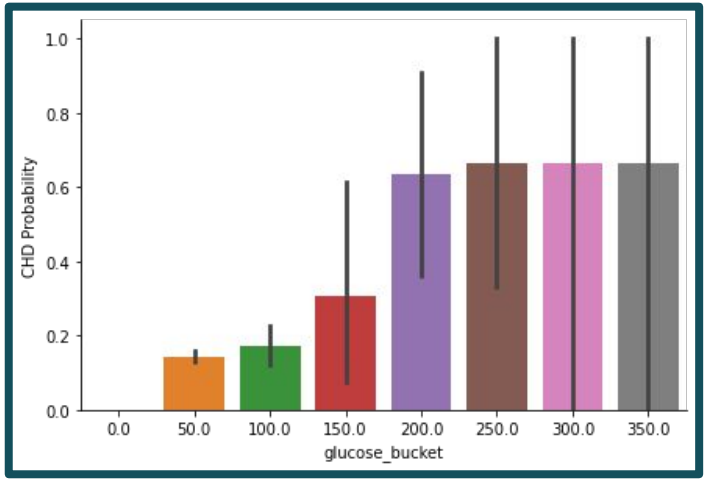


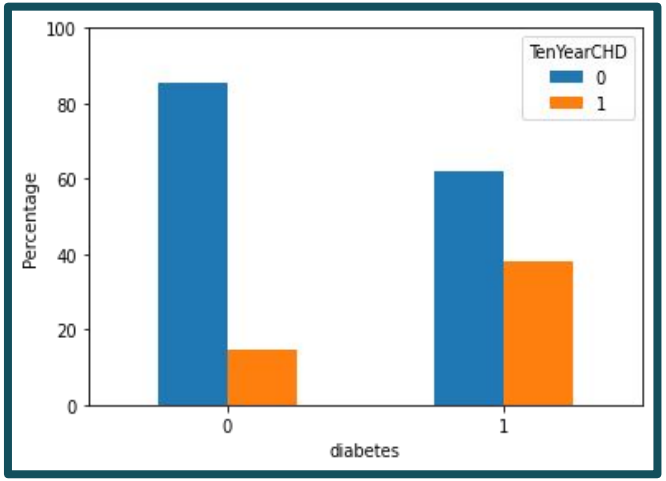


Other Notable Observations :

- Similarly, patients with high cholesterol and glucose levels (with diabetes) have higher risk of having CHD.







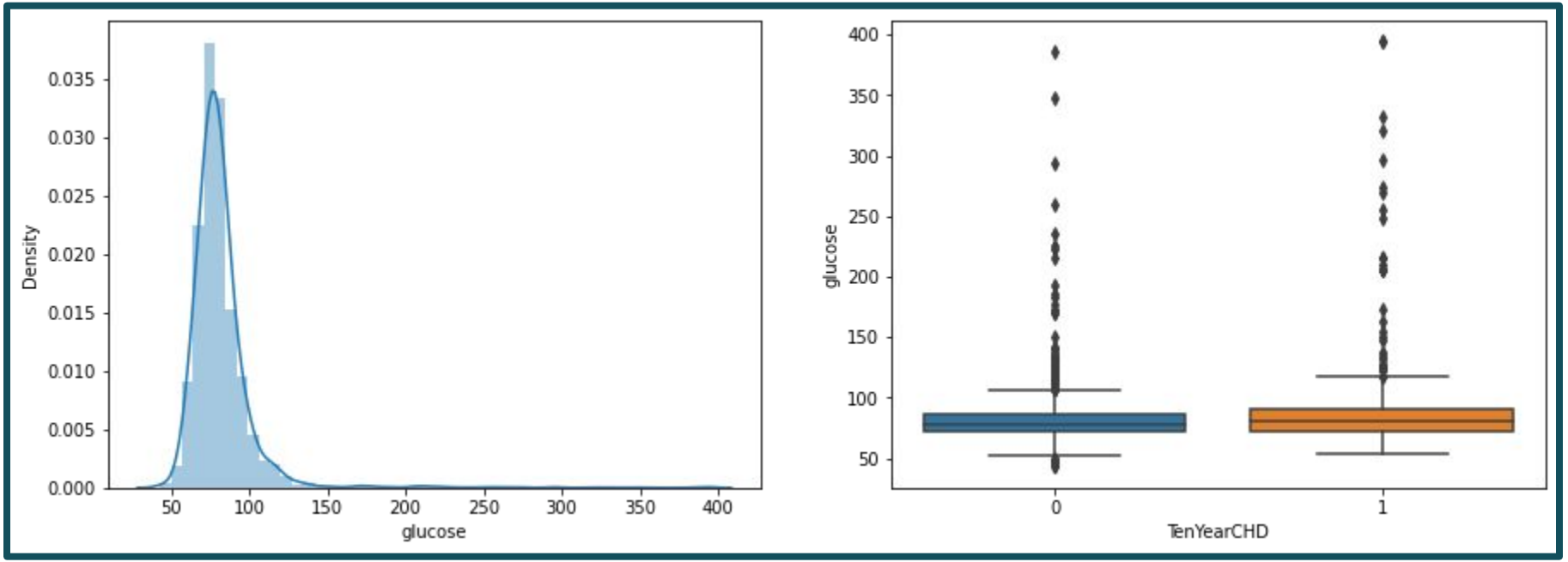
Data Cleaning & Feature Selection

Dealing with Nulls

- **Categorical Variables :** To fill up the absence of data in our categorical variables we have used simple imputer that imputes the null values with feature label that is most frequent in the feature column.

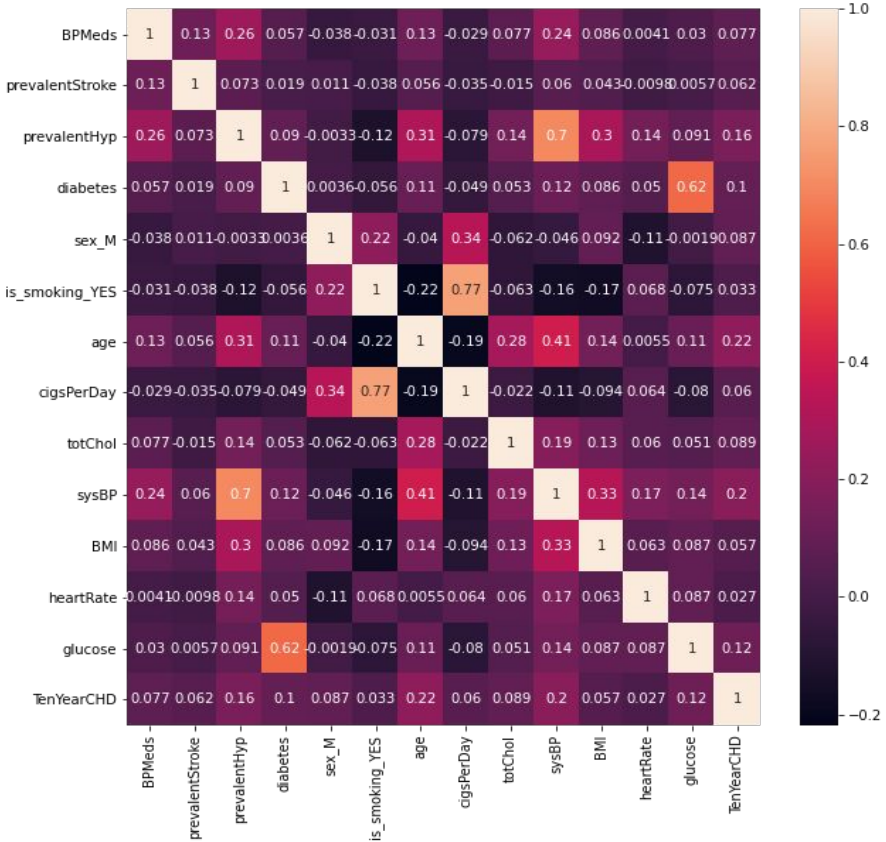
- **Continuous Variables :** To treat the null values in continuous variables, we use KNN imputer which uses a unsupervised clustering algorithm to come up with values of the features.

Dealing with outliers



Feature Selection

- There is significant correlation between systolic BP and prevalent hypertension.
- Also features like is smoking and cigarettes per day are correlated.
- Similarly glucose level and diabetes are correlated.



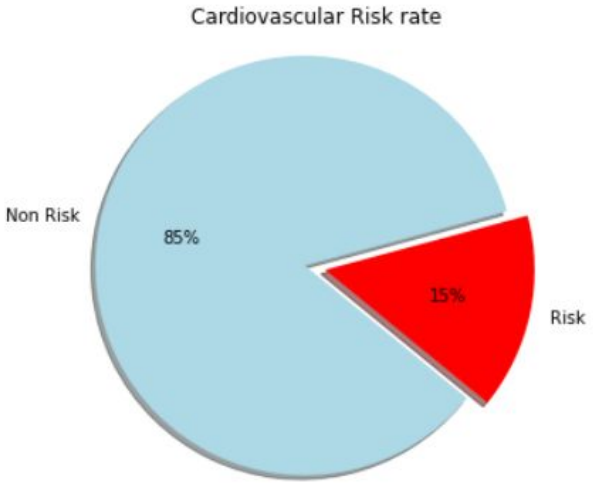
Final set of features :

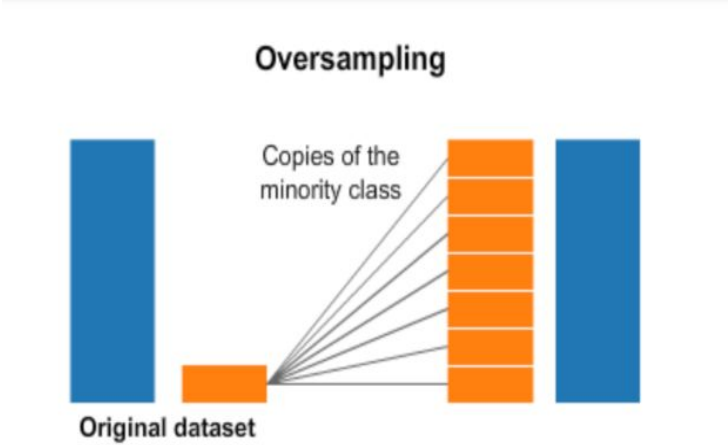
- Age
- Sex
- BP Meds
- Prevalent Stroke
- Systolic BP
- Glucose
- Total Cholesterol
- Body Mass index
- Heart Rate

Train-Test Split

- Train dataset has 2712 samples while test dataset has 678 samples.
- The split is such that the target variables classes are equally stratified over train and test dataset
- Out of 2712 samples, 2303 samples are of class 0 i.e. patients with no risk of CHD, while 409 samples belong to class 1 i.e. patients with a risk of CHD.

Addressing Class imbalance



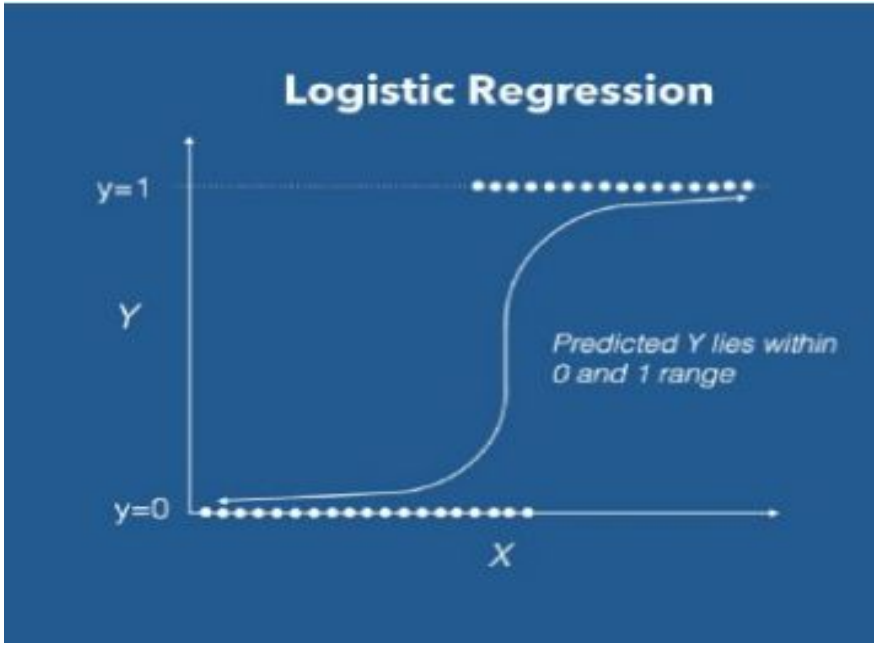


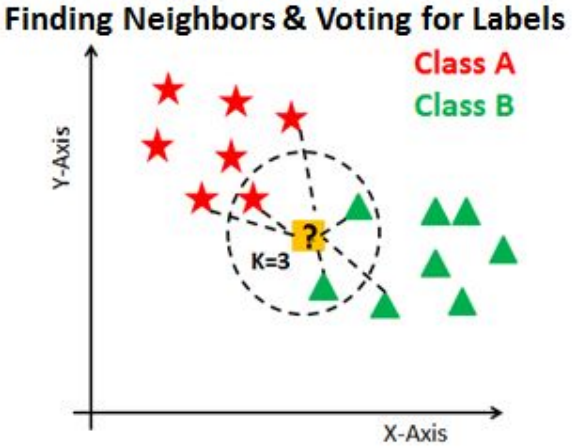
Random over sampler on train dataset

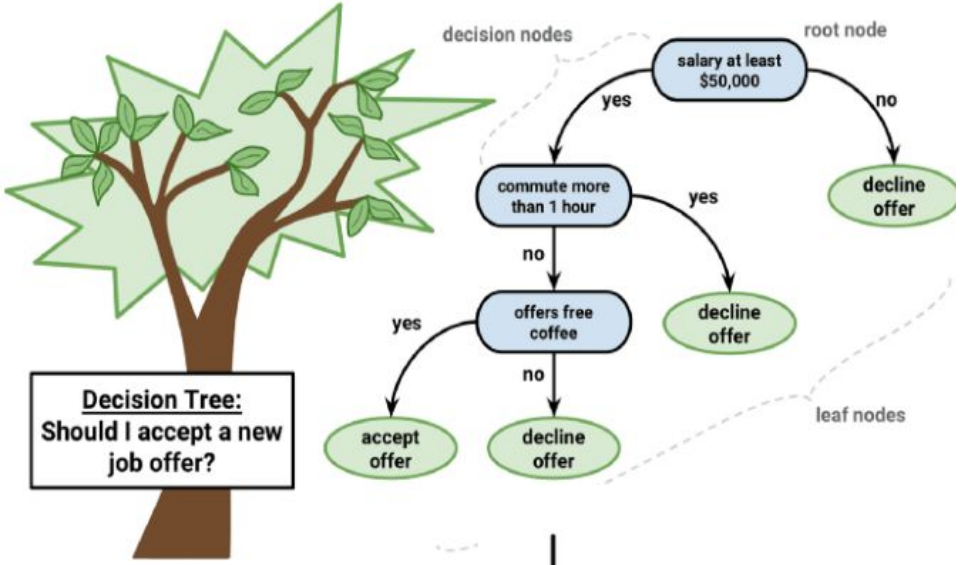
After over sampling we have train set of size 4606 with 2303 samples of each of the class. Our dataset is now ready for training.

Modeling and Results

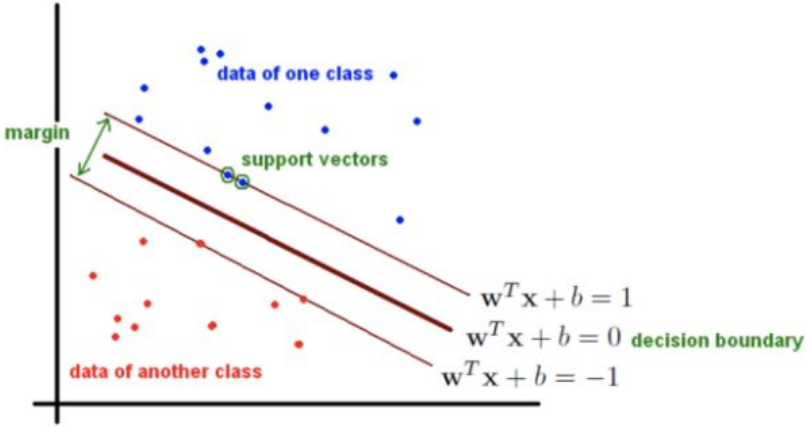
Using the training set, We trained five classifiers, i.e.,:



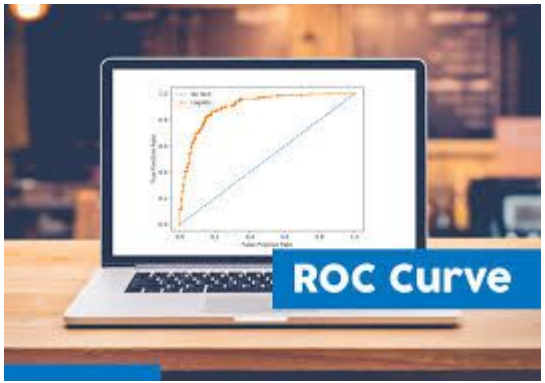



decision nodes
root node
salary at least $50,000
yes
no
commute more than 1 hour
yes
decline offer
no
offers free coffee
yes
decline offer
yes
no
accept offer
decline offer
leaf nodes
Decision Tree:
Should I accept a new job offer?

Cache awareness and out-of-core computing
Regularization for avoiding overfitting
Tree pruning using depth-first approach
Efficient heandling of missing data
XGBoost
Parallelized tree building
In-built cross-validation capability


data of one class
margin
support vectors
$w^T x + b = 1$
$w^T x + b = 0$ decision boundary
data of another class
$w^T x + b = -1$

After training each model and tuning their Hyper-parameters using Grid Search, We evaluated and compared their performance using the following metrics:



Models Trained

SL NO	MODEL_NAME	Train Accuracy	Train Recall	Train Precision	Test Accuracy	Test Recall	Test Precision
1	Logistic Regression	0.66	0.74	0.66	0.68	0.65	0.26
2	KNearest Neighbors	0.89	0.99	0.83	0.63	0.49	0.20
2	Random Forest Classifier	1.0	1.0	1.0	0.85	0.17	0.49
3	Support Vector Machine	0.68	0.74	0.66	0.64	0.74	0.26
4	XGBoost Classifier	0.78	0.81	0.76	0.68	0.56	0.24

- **Best Performing Model : Support Vector Machines (SVC)**
- **Since the recall on test set is 74% for SVC. However the precision is low ~ 26%, although precision on train set is 66%**

Confusion Matrix and Classification report of SVC

- **Train Set :**

1144	699
471	1372

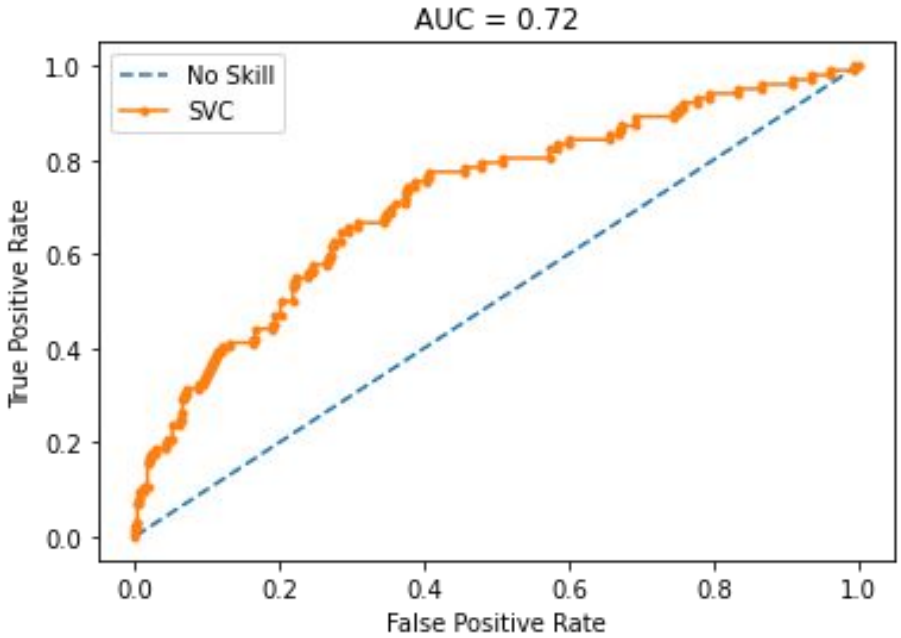
	precision	recall	f1-score	support
0	0.71	0.62	0.66	1843
1	0.66	0.74	0.70	1843
accuracy			0.68	3686
macro avg	0.69	0.68	0.68	3686
weighted avg	0.69	0.68	0.68	3686

- **Test Set :**

359	217
27	75

	precision	recall	f1-score	support
0	0.93	0.62	0.75	576
1	0.26	0.74	0.38	102
accuracy			0.64	678
macro avg	0.59	0.68	0.56	678
weighted avg	0.83	0.64	0.69	678

ROC Curve


AUC = 0.72
No Skill
SVC
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

Conclusion

This model can then be used as a simple screening tool and all that we need to do is to input ones: age, BMI, systolic and diastolic blood pressures, heart rate and blood glucose levels after which the model can be run and it outputs a prediction.

However, as a sanity check, most of the data on the positive cases were artificially created using ROS and as such they may not be a true representation of the actual population data thus more data, especially on the positive cases, is needed to build better models and much more potent screening tools.

Challenges and future work

- Although the oversampled training data show higher recall and precision for minority class, precision of minority class in test data still remains a concern. However overall precision is good.

- Although we have done feature selection based on their relevance to the target variable, it was challenging to come up with new engineered features that could explain hidden patterns in the data and classify our target variable better.

- We might need to work more on feature engineering and improve our precision. We might as well expect data samples with positive risk of CHD to be available in future.

Thank You