Statistical & Machine Learning

Prof. Minh Phan

Individual Project

Bank Subscription Prediction - Term Deposit

Ashwani Nitharwal

Ashwani.nitharwal@ieseg.fr

Contents

Introduction

Data is the key to majority of businesses these days, from customer behavior, sales, figures, demographics, etc.

Machine Learning plays a more advanced role in this whole process. It allows them to predict the coming patterns of them, estimate the sales, know the customer better and even the outputs/responses of coming campaigns.

Objective

We are provided by the bank dataset and we need to predict the customer's behavior based on different features that weather or not the customer will subscribe to the Bank's term deposit. We need to use different machine learning models and predict. We should apply benchmarking to those models like feature selection, cross-validation and validate our models by evaluating them by measuring the accuracies.

Dataset

The dataset contains the below columns:

Input variables
Group 1 - Bank client data:
1 - client_id : unique ID of the client (numeric)
2 - age : client age (numeric)
3 - job : type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
4 - marital : marital status (categorical: 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed)
5 - education (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
6 - default : has credit in default? (categorical: 'no', 'yes', 'unknown')
7 - housing : has housing loan? (categorical: 'no', 'yes', 'unknown')
8 - loan : has personal loan? (categorical: 'no', 'yes', 'unknown')

Group 2: Related with the last contact of the current campaign:
9 - contact : contact communication type (categorical: 'cellular', 'telephone')
10 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', …, 'nov', 'dec')
11 - dayofweek : last contact day of the week (categorical: 'mon', 'tue', 'wed', 'thu', 'fri')

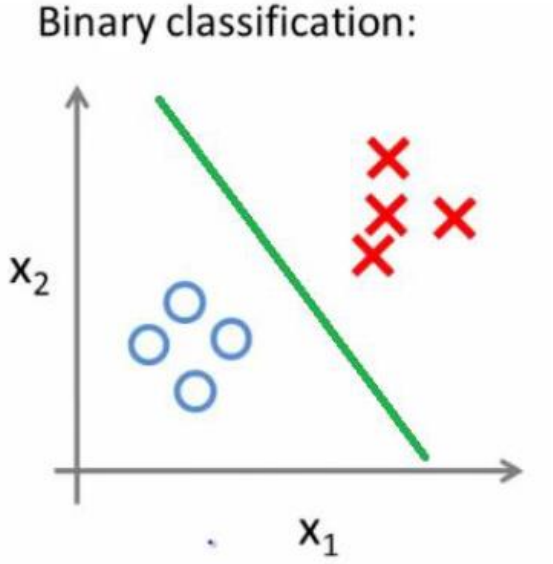
Group 3: Other attributes:
12 - campaign : number of contacts performed during this campaign and for this client (numeric, includes last contact)
13 - pdays : number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14 - previous : number of contacts performed before this campaign and for this client (numeric)
15 - poutcome : outcome of the previous marketing campaign (categorical: 'failure', 'nonexistent', 'success')

Group 4: Social and economic context attributes
16 - emp.var.rate : employment variation rate - quarterly indicator (numeric)
17 - cons.price.idx : consumer price index - monthly indicator (numeric)
18 - cons.conf.idx : consumer confidence index - monthly indicator (numeric)
19 - euribor3m : euribor 3 month rate - daily indicator (numeric)
20 - nr.employed : number of employees - quarterly indicator (numeric)

Target variable
21 - subscribe : has the client subscribed a term deposit? (numeric: 1='yes', 0='no')

Model

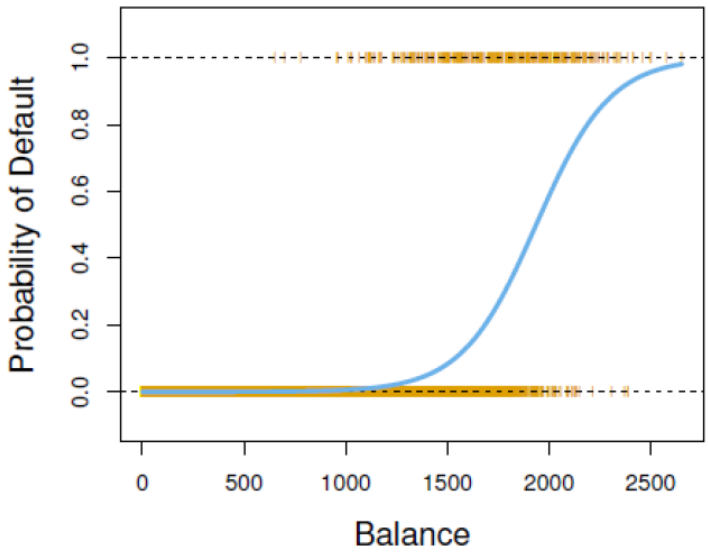
- Logistic Regression



Logistic Regression is a classification model which works on the probability and gives the class to which a particular data point belongs. It takes parameters and return the probability of it to belong to the particular class. It is usually used for binary classification.

It introduces non-linearity in the form of Sigmoid function:

$$f(x) = \frac{e^x}{1 + e^x}$$



The formula on which it works and finds the prediction and classification is:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

- e ≈ 2.718 Euler's number
- p(X) = Pr(Y=1|X): probability that observation X is in class 1, return value either 0 or 1
- β0, β1 : coefficient of logistic regression model

The likelihood of the data points are calculated based on the below formula and the curve for which we get the maximum likelihood gives the best selected curve for logistic regression for that data.

Advantages of Logistic Regression:

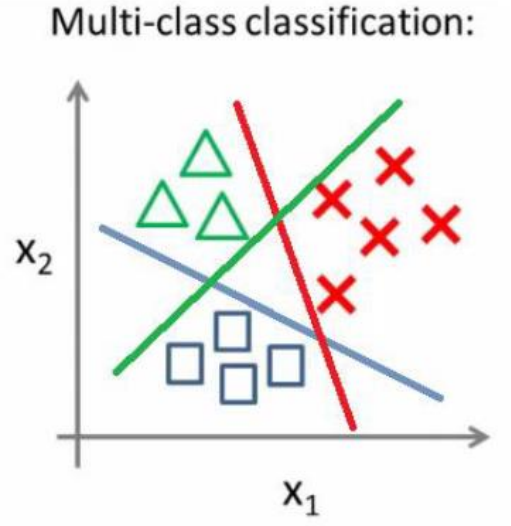
- Logistic regression is easier to implement, interpret, and very efficient to train.
- It can easily extend to multiple classes(multinomial regression).
- It not only provides a measure of how appropriate a predictor(coefficient size)is, but also its direction of association (positive or negative).
- It is very fast at classifying unknown records.
- Good accuracy for many simple data sets and it performs well when the dataset is linearly separable.
- It can interpret model coefficients as indicators of feature importance.

Disadvantages of Logistic Regression:

- If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to overfitting.

- The major limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables.
- It can only be used to predict discrete functions. Hence, the dependent variable of Logistic Regression is bound to the discrete number set.
- Logistic Regression requires average or no multicollinearity between independent variables.
- It is tough to obtain complex relationships using logistic regression. More powerful and compact algorithms such as Neural Networks can easily outperform this algorithm.

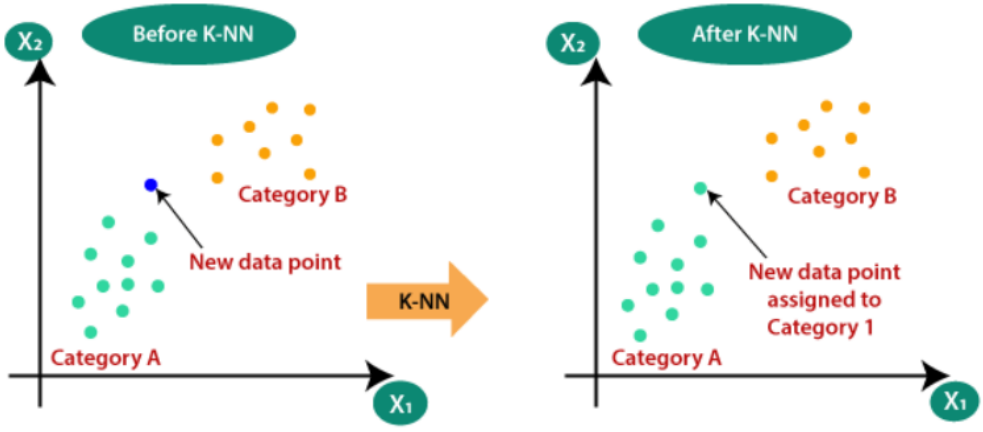
It is also possible to use logistic regression for multiple class output.



It can be done so by doing it by taking 1 class at a time and training the model and then predicting the values for all.

- KNN

KNN is one of the simplest algorithm. It works on the principle that we place a data point and then decide the value of k (number of neighbors) to look around that point.



The distance decides the neighbors of the new data point. The most used method to measure distance to find neighbor is Euclidean Distance which can be calculated by

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Suppose, if the nearest k neighbors are 7 of green class and 3 of yellow class then the data point is predicted to be of the majority class, here, green class. This way, it puts the data into pre-existing categories.

This algorithm is easy to interpret and is also called lazy learner algorithm. The model does not learn from the available data rather, it stores the dataset and at time of classification, it performs an action on dataset based on previous observation.

The selection of k value is the key to the model to perform well. We need to try multiple value so as to get favorable results. A very low value for K such as K=1 or K=2, can be noisy and lead to the effects of outliers in the model. Large values for K are good, but it may find some difficulties or may cover more than one categories which may distort results.

Advantages of KNN algorithm:

- It is simple to implement.
- It is robust to the noisy training data.
- It can be more effective if the training data is large.
- It can be used for prediction as well as classification processes.

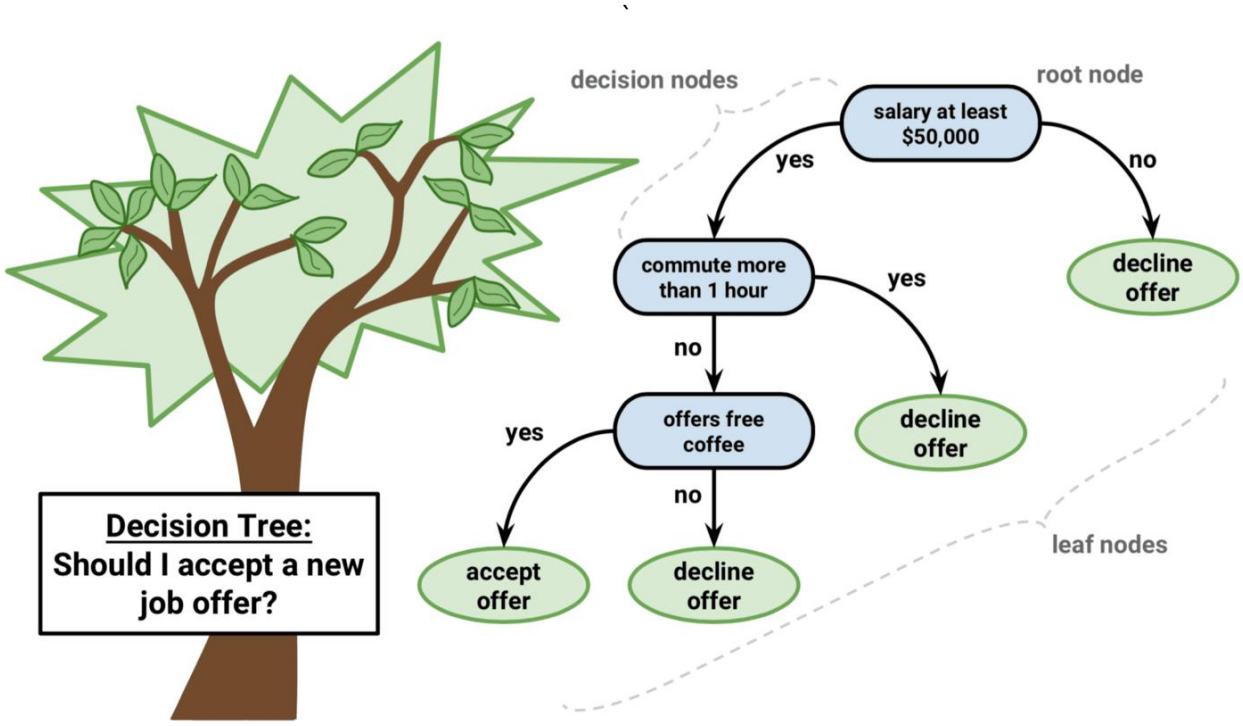
Disadvantages of KNN algorithm:

- Always needs to determine the value of K which may be complex some time.
- The computation cost is high because of calculating the distance between the data points to find neighbors.

- Decision Tree

Decision tree is one the machine learning algorithm which is highly interpretable and looks very similar to a tree structure having ROOT NODE, NODEs and LEAVES or LEAF NODEs. Decision tree is most powerful and popular tool for classification problems.

Decision tree is constructed based on the recursive partitioning based on the test conditions of the train dataset values. This tree can then be utilized for predicting the classes of test data. It can high dimensional data as it splits the node into two whenever a new value is encountered in a column.



Advantages of Decision Tree:

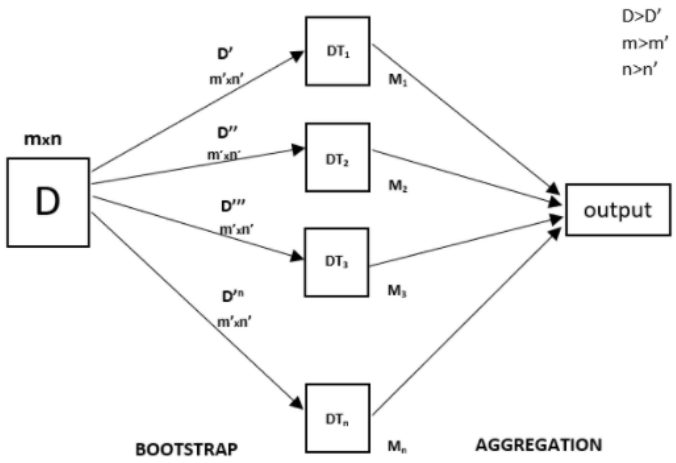
- Decision trees can generate understandable rules.
- Decision trees perform classification without requiring much computation.
- Decision trees can handle both continuous and categorical variables.
- Decision trees provide a clear indication of which fields are most important for prediction or classification.

Disadvantages of Decision Tree:

- Decision trees are prone to errors in classification problems with many class and relatively small number of training examples.
- Decision tree can be computationally expensive to train. The process of growing a decision tree is computationally expensive. At each node, each candidate splitting field must be sorted before its best split can be found.
- Pruning algorithms can also be expensive since many candidate sub-trees must be formed and compared.
- Decision trees can more easily lead to overfitting unless pruned.

- Random Forest

Random Forest as the name suggests, is a combination of decision trees used for both classification as well as prediction. It is complex than a decision tree model, obviously. Random datasets are created from the given data and decision trees are made based on that data, this process is called bootstrap. While predicting the result values of all these are aggregated for the final output. The whole process of Bootstrap and aggregation is called Bagging. The result is not relied on the basis of single decision tree. Random Forest provides much better accuracy and efficiency than a decision tree, this comes at a cost of storage and computational power.



Each box in the middle represents a different decision tree with the subset dataset.

Advantages of Random Forest:

- It reduces overfitting in decision trees and helps to improve the accuracy
- It is flexible to both classification and regression problems
- It works well with both categorical and continuous values
- It automates missing values present in the data
- Normalizing of data is not required as it uses a rule-based approach.

Disadvantages of Random Forest:

- It requires much computational power as well as resources as it builds numerous trees to combine their outputs.
- It also requires much time for training as it combines a lot of decision trees to determine the class.
- Due to the ensemble of decision trees, it also suffers interpretability and fails to determine the significance of each variable.

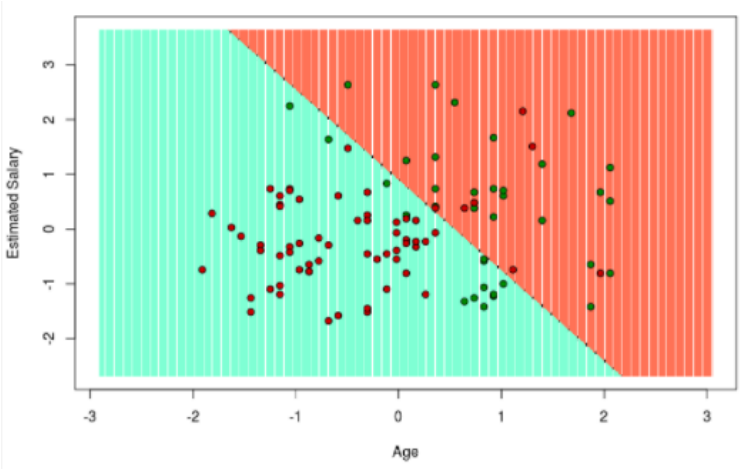
- Support Vector Machine

SVM is a machine learning algorithm used for classification and regression. It gives a hyper-plane that separates 2 classes. The data points which are on 1 side of the plane are of one class and the ones on the other side of the plane belongs to the other.

The equation of separating hyperplane is :

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

The separating plane looks like a line if a data is in 2D plane and if the data is projected in a 3D plane, then the separating plane looks like a 2D plane which separates the two classes.



The above figure shows the implemented SVM model. The data points are divided into 2 categories by a single line.

Advantages of SVM:

- SVM works relatively well when there is a clear margin of separation between classes.
- SVM is more effective in high dimensional spaces. SVM is effective in cases where the number of dimensions is greater than the number of samples.
- SVM is relatively memory efficient

Disadvantages of SVM:

- SVM algorithm is not suitable for large data sets.
- SVM does not perform very well when the data set has more noise i.e. target classes are overlapping.
- In cases where the number of features for each data point exceeds the number of training data samples, the SVM will underperform.
- As the support vector classifier works by putting data points, above and below the classifying hyperplane there is no probabilistic explanation for the classification.

Variable Selection

To reduce the complexity of model and to increase the interpretability as well as dependency of the independent variables on dependent variable, variable selection or feature reduction or dimensionality reduction is done.

We have used multiple ways to find the most important features of the given data. We finally, took the intersection of selected features by different algorithms and use it for further models. Methods used are:

- **Linear Regression Model**

We ran data through the linear regression model and observed the dependency of each variable on the target variable. The variables which are significant are selected.

- **Boruta method**

From the package, Boruta, we used Boruta method to find the importance of each variable and extract the important ones.

- **Caret train by rpart method**

From caret library, we trained dataset using method rpart and extracted important variables.

- **XG Boost**

From caret library, we trained an xgb model to carry out important features.

- **Step-wise forward and backward Selection**

We used step-wise algorithm to find the important features, with both backward and forward selection.

Finally, taking the important and common features from above methods, we decided to move by reducing features and proceed only with the important ones.

Cross-Validation Methods

We used k-fold CV for all the models. Cross-Validation is a resampling procedure to split the data into multiple folds to evaluate the machine learning models and allow us to know the best parameters to use while working with a particular model.

It allows us to use the best available parameters which suits the model with the current data. This allows us to split data manually and carry out validation and test predictions with also preventing from problems like overfitting. It is usually called K-fold cross validation in which k means the number of splits of data.

Evaluation of Models

The models are evaluated by various ways to decide between the models that which model is the best for a particular type of data. For evaluation, several method:

- **Accuracy**

Accuracy is calculated based on confusion matrix. Confusion matrix compare the predicted results with the true results and find the number of observation which are:

True Positive (TP) – The observation which is predicted true and are true.

True Negative (TN) - The observation which is predicted false and are false.

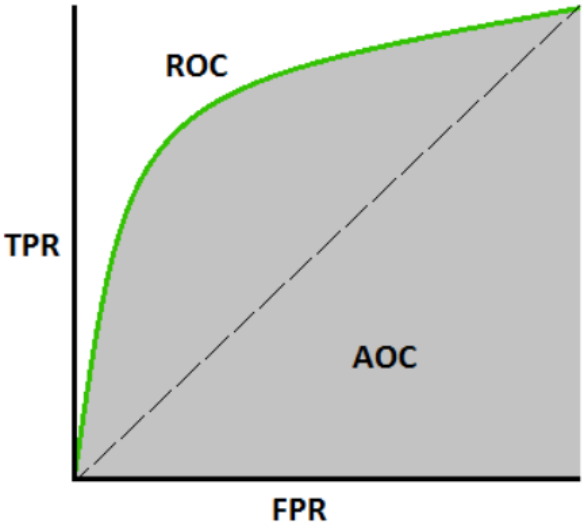
False Positive (FP) – The observation which are predicted true but are false.

False Negative (FN) - The observation which are predicted false but are true.

Accuracy = (TP + TN) / (TP + TN + FP + FN)

- **AUC**

It is the AUC ROC curve we use to evaluate model. The curve is between True Positive Rate (TPR) and False Positive Rate (FPR). The more the curve is towards the TPR, the better the model is.



TPR = TP/(TP + FN)

FPR = FP/(TN + FP)

- **Mean Square Error (MSE)**

Mean square error as the name suggests, it is mean of squares of the difference between the predicted value and the respective actual value. The less the MSE, the good the model.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

- **Residual Sum of Squares (RSS)**

RSS is the sum of squares of difference between predicted and real value. It is calculated by below formula:

$$RSS = \sum_{i=1}^{n}(y_i - f(x_i))^2$$

Experimental Setup

We were given a dataset of bank clients and we need to train and evaluate the models by using the given data so as to get the best model.

1. Imported Libraries
2. Data Exploration
3. Checked NAs present in the data and resolved them with the appropriate values such as median or most occurred value or mean
4. Converted binary variables into 0 and 1
5. Converted categorical variables into dummy variables and dropped the original columns
6. Basetable was created
7. Data Exploration
8. Exploring feature importance and selected the most important ones by different methods like Linear Regression, Boruta, stepwise forward backward selection, xg boost, rpart method.
9. Took intersection of these features from all methods and selected them for further model training

10. Model: Logistic Regression
    - Created partitions of data
    - Trained data and predicted on test data with different threshold value such as 0.5, 0.25
    - Calculated confusion matrix, accuracies, precision, sensitivity, specificity, AUC
    - Ran a cross-validation and got the optimum model results with the most AUC 0.808

11. Model: Decision Tree
    - Ran a model on a full data and then reduced data
    - Plotted a tree

- Explored results and calculated AUC
- Pruned the tree to avoid overfitting and plotted analysis about tree
- The tree was already better than pruned so no need to prune
- Ran a cross-validation and got the optimum model results.

12. Model: SVM
    - Ran SVM model with different cost values
    - Tuned the cost parameter using cross validation
    - Got the best cost parameter and used it to run SVM model to get the AUC

13. Model: KNN
    - Ran KNN model with different k values to check the results
    - Ran a cross validation to get the best score AUC
    - Ran a cross validation with multiple k values to get the best k values
    - Extracted the best model and its best AUC

14. Model: Random Forest
    - Created partitions of data
    - Ran model and calculated MSE
    - Calculated AUC
    - Ran a cross validation and evaluated model

Result and Comparison of Models

```
[1] "AUC of Logistic Regression = 0.7842233"
[1] "AUC of Decision Tree = 0.744511811721982"
[1] "AUC of Random Forest = 0.793158903066153"
[1] "AUC of KNN = 0.9011"
[1] "AUC of SVM = 0.686186820980773"
```

We have calculated multiple parameters for the evaluation of models like MSE, RSS, Confusion Matrix, Accuracy but we have used AUC as a standard evaluation technique to compare the models with each other.

We split the dataset into train and test data subset and predicted the values and compared them with the actual values given in the dataset. The AUC values shows that KNN model is the best model for the given data. The result is calculated by optimizing the model by using cross-validation to get the best parameters. The models in the order of their AUC are:

1. KNN
2. Random Forest
3. Logistic Regression
4. Decision Tree
5. SVM

Conclusion

Different models can be used for the prediction but the selection of which model to use depends on the data and the output we want. Models can perform better than each other with different data. Also, models can be evaluated and compared to decide which model to be used for better predictions. Further, more optimized results can be found by running cross validation and evaluation strategies so to get best parameters for the best models.

References

- Class Codes and slides
- Towardsdatascience.com
- Geeksforgeeks.org
- Youtube