Bank Marketing with the help of Machine Learning

—

By Shubh Mehtani

INTRODUCTION

Marketing is a technique of exposing the target clients to a product via suitable systems and channels. It ultimately facilitates the way to buy the product or service. The overall aim is to increase the sales of products and services for the enterprise, businesses and financial institutes. Telemarketing is a form of direct marketing in which a salesperson approaches the customer face-to-face or via phone call to persuade him to buy the product. In the Banking sector, marketing is the backbone to sell its product or service. Banking advertising and marketing are mostly based on an intensive knowledge of objective information about the market and the actual client needs for the bank's profitable manner. Making right decisions in organizational operations is sometimes proved a great challenge where the quality of the decision really matters. Decision Support Systems (DSS) are classified as a particular class of computerized facts and figures that helps the organization or administration in their decision-making actions.

DSS uses statistical data to overcome the deficiencies and helps the decision makers to take the right decision. Data mining (DM) plays vital role to support the Decision support systems which are based on the data obtained from the data mining models: rules, patterns and relationship. Data mining is the process of selecting, discovering, and modeling high volume of data to find and clarify unknown patterns. The objective of data mining in decision support systems is to suggest a tool which is easily accessible for the business users to analyze the data mining models. A specific technology used within the DSS is Machine learning (ML) that combines data and computer applications to accurately predict the results. The fundamental principle of machine learning is to construct algorithms that can obtain input data and then predict the results or outputs by using statistical analysis within a satisfactory interval. ML allows the DSS to obtain new knowledge which helps it to make right decisions. Machine learning can be mainly classified into 2 categories i.e. supervised and unsupervised learning. In supervised learning, the output of the algorithm is already known and we use the input data to predict the output. In contrast, in unsupervised learning, we have only input data whereas no corresponding output variables are selected.

Following are some of the articles which have solved the same problems . Section 3 presents the description of the data set of the Pourtuguese baking institution. Section 4 is a representation of the architectural diagram of the framework and further is the algorithms explained which are used in the project. And at last the result section.

Background

This section explains the previous research work which have been already done in classification using ML techniques.

The data which is used in this study work is the data of customers of a Portuguese banking institution. The similar data set has been used in Moro et al. (2011, 2014). In Moro et al. (2011), the aim of this study was to find the model that can increase the success rate of telemarketing for the bank. The statistical techniques of data mining which have been used in their research are Support Vector Machine (SVM), Decision Tree (DT) and Naive Bayes. The performance of these models was checked through the Receiver Operator Characteristics (ROC) curve (detail of ROC curve is given in section 5). Among all these statistical techniques, SVM comes up with the most efficient results. Regarding attributes, Call duration was the most relevant feature which states that longer calls tend increase the success rate. After that month of contact, number of contacts, days since last contact, last contact result and first contact duration attributes respectively.

In Moro et al. (2014), objective of the study was to predict the success of bank telemarketing. Data set which they used in their research was consists of 150 attributes and is complete data 3 set of the period

2008 to 2013. They compare the 4 data mining models i.e. Logistic Regression (LR), Decision Tree, Support Vector Machine and Neural Network (NN). The best result was obtained by the neural network while decision trees discloses that probability of success in inbound calls are greater.

Statistical learning algorithms have successfully been used in many research problems for classification. For example, Qi et al. (2018) conducted a research to find out the fault diagnosis system for reciprocating compressors. Reciprocating compressors are extensively used in petroleum industry. Data was taken from oil corporation (5 years operational data) and uses the Support Vector Machine to analyze it. They come up with the results that SVM accurately predicts the 80% right classification to find the potential faults in compressor.

Similarly, Gil & Johnsson (2010) did a research in medical field for diagnosing the urological dysfunctions using SVM. In this research data was collected from the 381 patients who are suffering from a number of urological dysfunctions to check the overall worth of decision support system. The fivefold cross validation has been utilized for the robustness. The outputs of this study describe that for the purpose of classification SVM attained the accuracy of 84.25%

Nogami et al. (1996) utilized the machine learning in decision support system. In their research they introduce the air traffic management for the future which can manage the flight schedule and flow of air traffic professionally. Their system involves many decision makers and utilized it with the neural network. They require such system which can make the optimal decision in the critical situation. Their simulation studies prove that system which is based on neural network is performed more efficiently than the current air traffic system.

Another research by Cramer et al. (2017) the machine learning methods are used in time series for rainfall prediction. Data was derived from the 42 cities including climatic features.They tried Support vector regression, NN, and k nearest neighbors. After performing these methods they come up with the results that machine learning methods have predictive accuracy.

Wang & Summers (2012) used the machine learning in field of radiology. They used it for the neurological disease diagnosis images, medical image segmentation and MRI images. They come-up with the results that machine learning identifies the complex patterns. It also helps the radiologists to make right decisions. Furthermore, they suggest that development of technology in machine learning is an asset for long term in the field of radiology.

Machine learning algorithms are also used in the field of applied mathematics. For instance, Barboza et al. (2017) did a research to predict the models for developing of bankruptcy by using the SVM and random forest methods. The data was taken from the Salomon Center database & Compustat about North American firms from period 1985 to 2013 with observations of more than 10,000. After applying SVM

and RF techniques they compare the results with the ordinary used methods such as discriminant analysis and logistic regression. They concluded that ML techniques are come up with 10% averagely more accurate results than usual methods.

To find the risk factors about failure of banks Le & Viviani (2017) conducted a research. In their study, a sample of 3000 US banks was analyzed by using 2 traditional statistical methods i.e. discriminant analysis and logistics regression. Then they compare these methods with the machine learning methods i.e. SVM, ANN and k-nearest neighbors. The results of this study illustrate that ANN and k-neighbors method gives the accurate predictions as compared to the 4 traditional methods.

Dataset Description

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit.The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

There are four datasets:bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014] , bank-additional.csv with 10% of the examples (4119), randomly selected from 1), and 20 inputs, bank-full.csv with all examples and 17 inputs, ordered by date (older version of this dataset with less inputs), and bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).

The smallest datasets are provided to test more computationally demanding machine learning algorithms (e.g., SVM).

No preprocessing was required in the project as the dataset was already clean.

We have performed outlier analysis, missing value detection, data manipulation and used correlation matrix for feature selection.

Feature	**Explanation**	**Units**	**Data Type**	**Range**
Age	Age of the person	Years	int	[18, …, 95]
Job	Type of Job the person has	categorical	string	for eg. technician, admin, etc.
Marital	Marital status of the person	categorical	string	for eg. single, married, divorced,

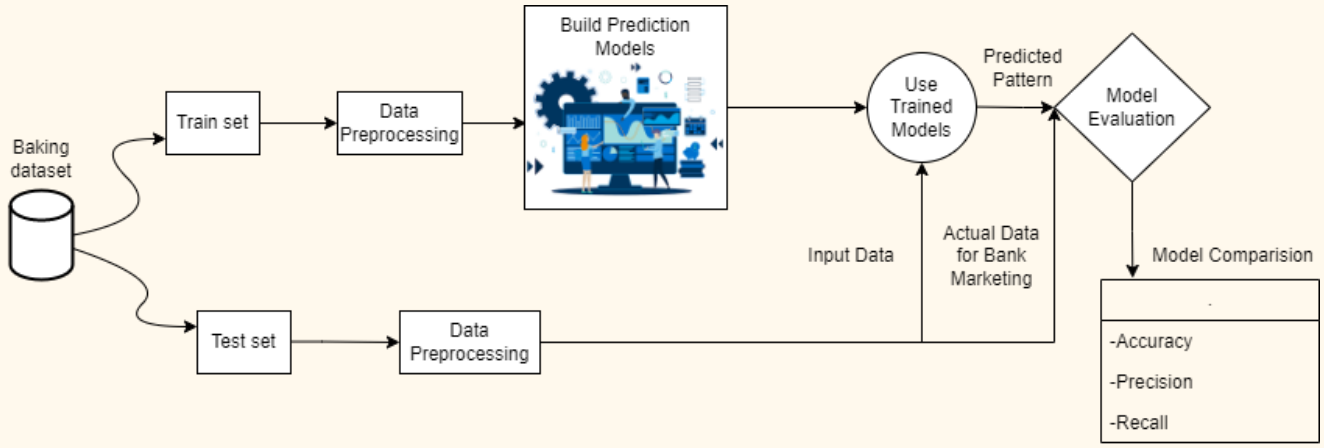
				etc.
Education	Education level of the person	categorical	string	For eg. primary, secondary, tertiary
Default	Does the person have credit in default?	Boolean	binary	0,1
Balance	Bank balance of the person	Euros	int	[-8019, …, 102127]
Housing	Does the person have a housing loan	Boolean	binary	0, 1
Loan	Does the person have a personal loan	Boolean	binary	0, 1
Contact	Contact communication type of the person	Boolean	binary	0, 1
Day	Last contact day of the person	discrete	int	[1, …, 31]
Month	Last contact month of the person	categorical	string	For eg. jan, feb, mar, etc.
Duration	Last contact duration of the person	Seconds	int	[0, …, 4918]
Campaign	number of contacts performed during this campaign and for this client	Contacts	int	[1, …, 63]
Pdays	number of days that passed by after the client was last contacted from a previous campaign	Days	int	[-1, …, 871]
Previous	number of contacts performed before this campaign and for this client	Contacts	int	[0, … ,275]
Poutcome	outcome of the previous marketing campaign	boolean	binary	0, 1

Y	Has the client subscribed to a term deposit?	Boolean	binary	0, 1

Correlation Matrix

Looking at it, we observe that there is a moderate correlation between the days and the previous ones. (r=0.51)

Architecture

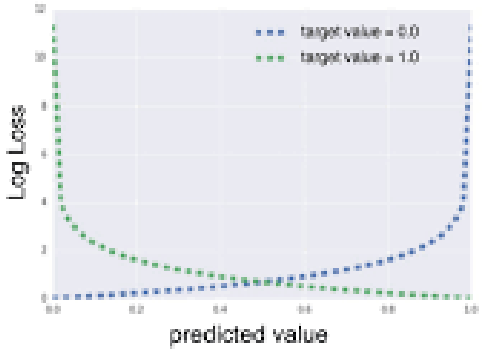


In ths project, a bank marketing dataset was collected from a portuguese bank institute. This dataset is divided into 2 main parts: one is the training set and another is the test dataset. The training dataset is used in training the prediction models with different algorithms to predict the desired output. The remaining part that is the test dataset is used as the insput data which is the actual data for the bank marketing project. This output generated is then used for the model evaluation comparing it with the hep of model parameters such as accuracy, precision recall,sensitivity and other factors to determine its correctness about the prediction.

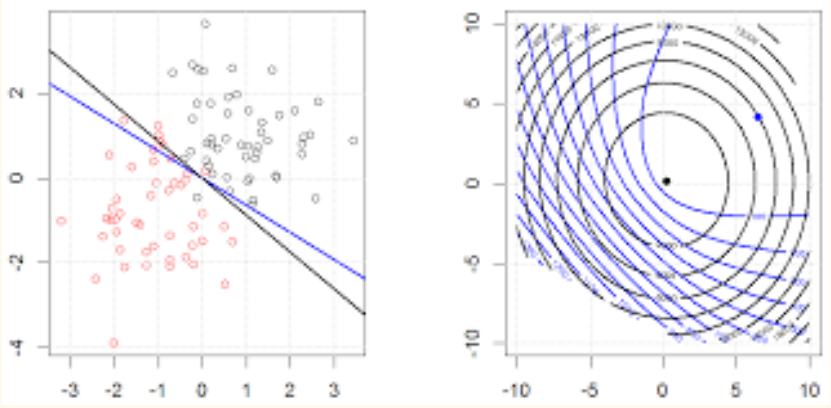
Algorithms used

Logistic regression without regualrization: Logistic regression is a statistical method that is used to model a binary response variable based on predictor variables. Although initially devised for two-class or binary response problems, this method can be generalized to multiclass problems.

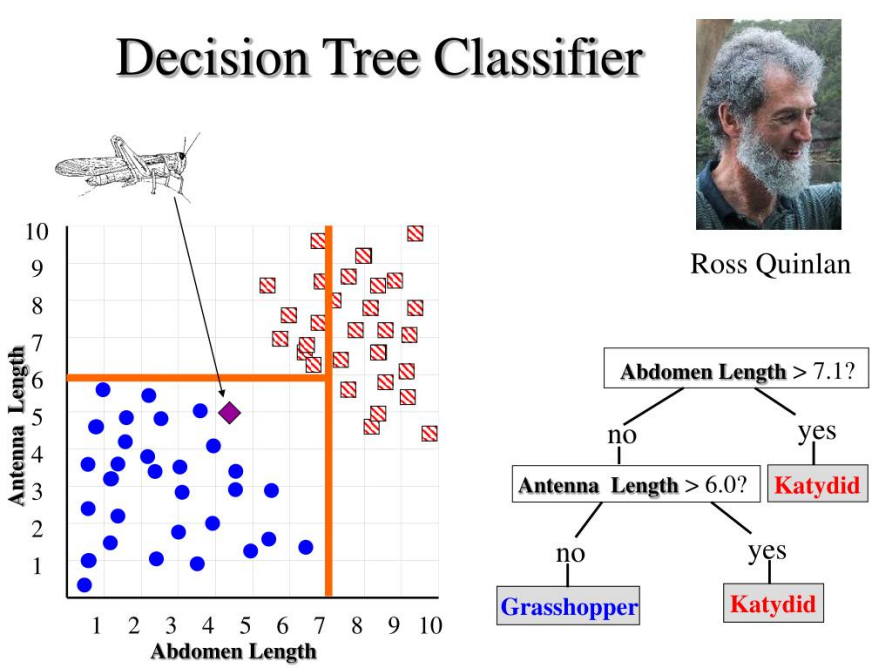
$$LogLoss = \sum_{(x,y)\in D} -y\,log(y') - (1-y)\,log(1-y')$$



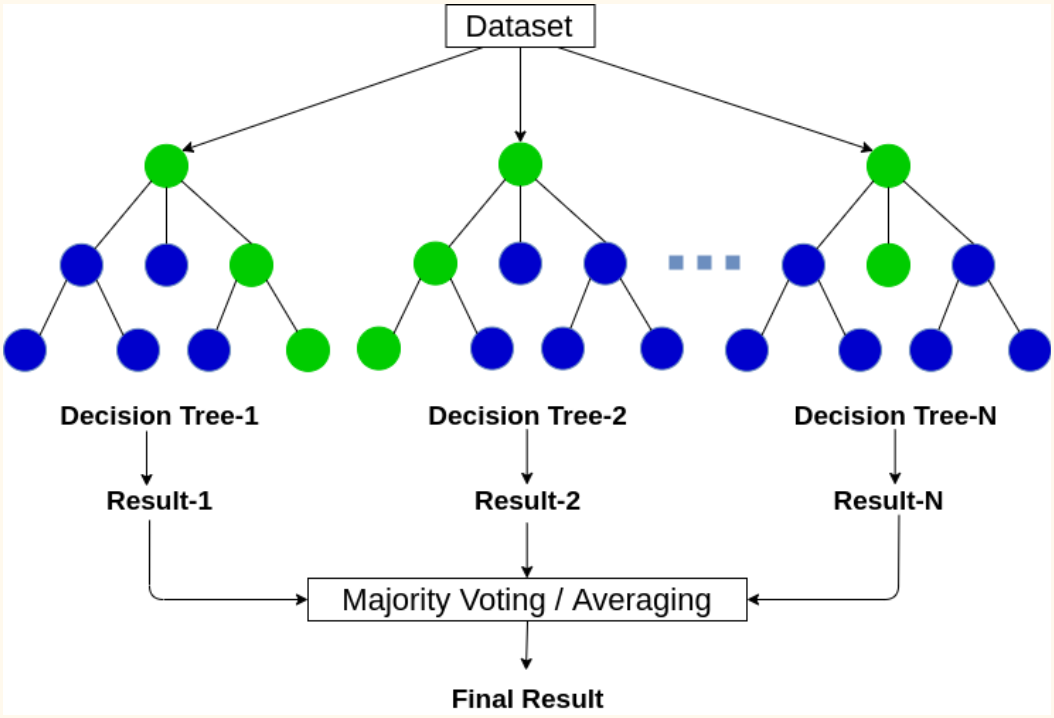
Logistic regression with regularization: Ridge and Lasso regularizations are also known as ‘shrinkage’ methods, because they reduce or shrink the coefficients in the resulting regression. This reduces the variance in the model: as input variables are changed, the model’s prediction changes less than it would have without the regularization. Why would you want to reduce the variance of a model? To avoid overfit.



Decision Trees (DTs): Decision Trees play s significant role in the field of data mining as they are really fast to construct as compared to the other data mining methods. They can easily handle the data even if it comprises of mixture of numeric and categorical predictor variables (Friedman et al. (2001)). The algorithm of DTs is to split the data-set to accomplish a homogenous classification for the dependent variable. At every part, objective of algorithm goes for diminishing the entropy of the dependent variable in the subsequent datasets by selecting the ideal part from various explanatory variables.



Random Forest for Classification: Random forest can also be used for the purpose of classification. It is one of the most broadly used machine learning algorithm for classification. Either the response variable is continuous or categorical, it works in both cases. According to Friedman et al. (2001) random forests starts to become stable at around 200 trees, whereas at 1000 trees the boosting of this still keeps on improving. If trees are much smaller or there is a presence of shrinkage then process of boosting starts to reduce. The important function of the RF is the utilization of out of bag (OOB) samples. For each value zi = (xi , yi), in which term zi not appears that makes the RF predictor by averaging only those trees which are consistent to the bootstrap samples. The OOB error estimate is then nearly identical of that which is getting by N fold cross validation. In contrast to the other non linear estimators, it is possible to fit the RF in one sequence with the cross validation being completed. The training can be finished if the OOB error stabilizes itself.



K Neighbors Classifier: The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.

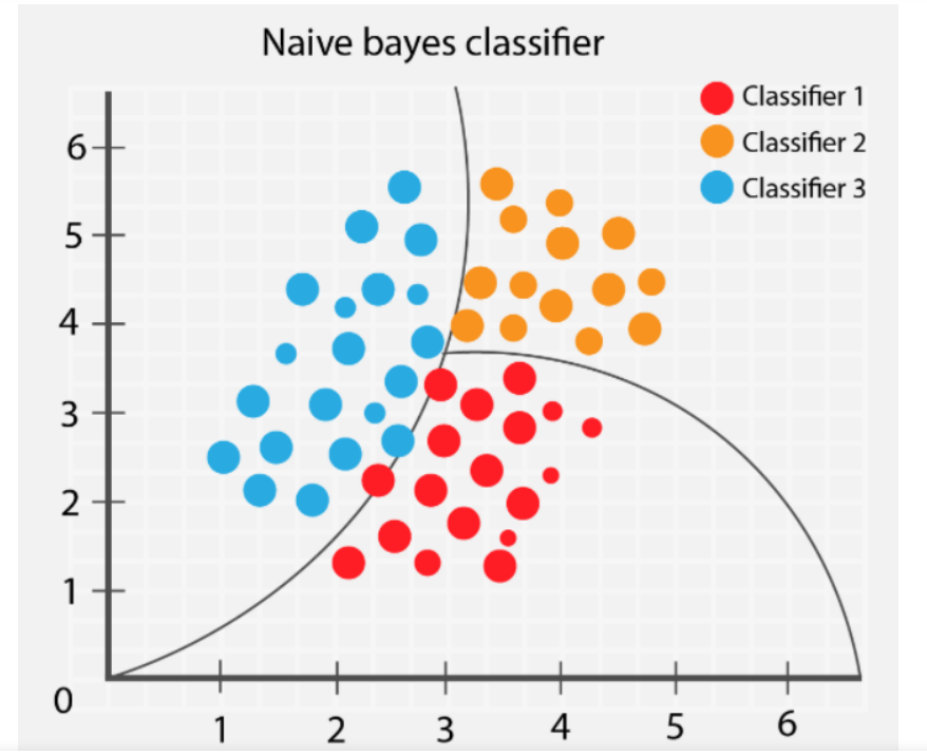
Linear SVM: Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.

RBF SVM: RBF kernels are the most generalized form of kernelization and is one of the most widely used kernels due to its similarity to the Gaussian distribution. The RBF kernel function for two points X_1 and X_2 computes the similarity or how close they are to each other. This kernel can be mathematically represented as follows:

$$K(X_1, X_2) = exp(-\frac{||X_1 - X_2||^2}{2\sigma^2})$$

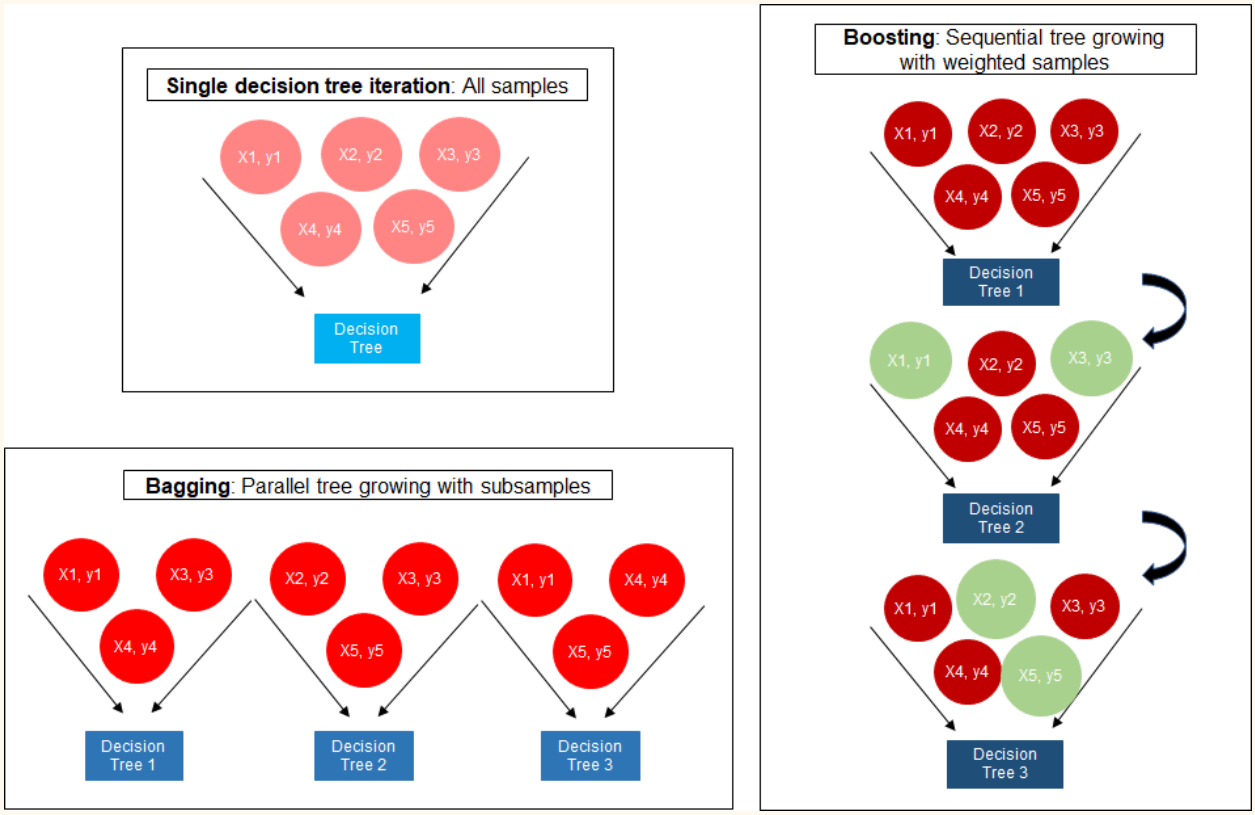
where, ‘σ’ is the variance and our hyperparameter $||X_1 - X_2||$ is the Euclidean (L_2-norm) Distance between two points X_1 and X_2

Naive Bayes:Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.This algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in *text classification* that includes a high-dimensional training dataset. It is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

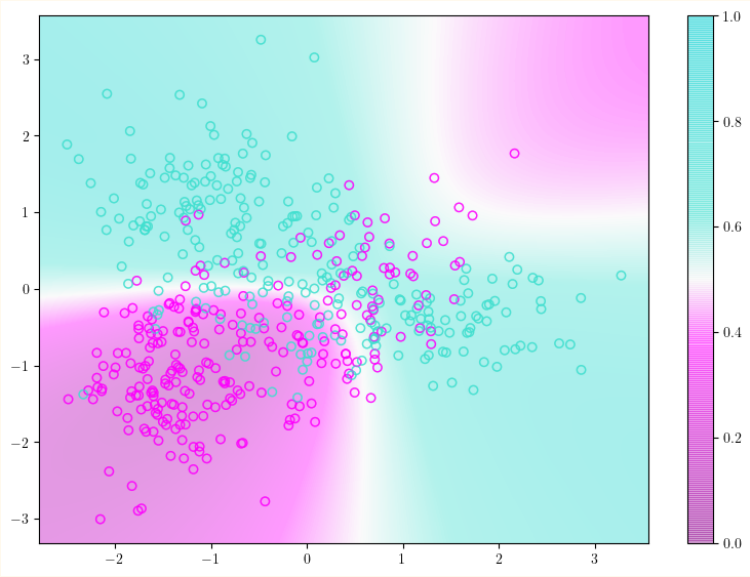


AdaBoost: AdaBoost also called Adaptive Boosting is a technique in Machine Learning used as an Ensemble Method. The most common algorithm used with AdaBoost is decision trees with one level that means with Decision trees with only 1 split. These trees are also called Decision Stumps.What this algorithm does is that it builds a model and gives equal weights to all the data points. It then assigns

higher weights to points that are wrongly classified. Now all the points which have higher weights are given more importance in the next model. It will keep training models until and unless a lowe error is received.

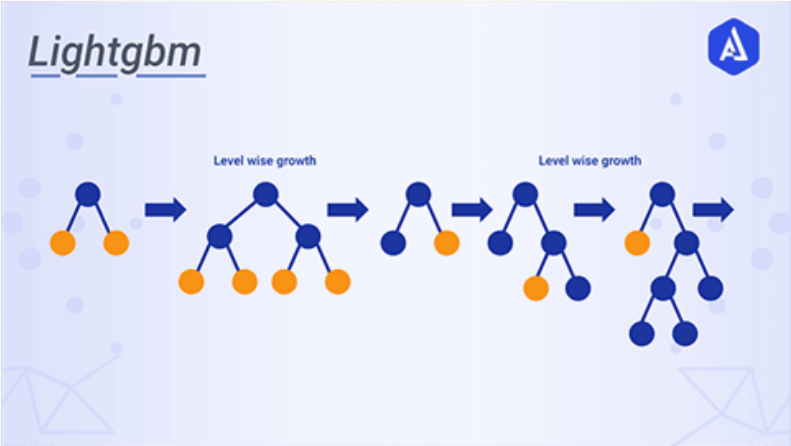


Quadratic Discriminant Analysis (QDA): Quadratic Discriminant Analysis is a generative model.QDA assumes that each class follow a Gaussian distribution. The class-specific prior is simply the proportion of data points that belong to the class. The class-specific mean vector is the average of the input variables that belong to the class. The class-specific covariance matrix is just the covariance of the vectors that belong to the class.

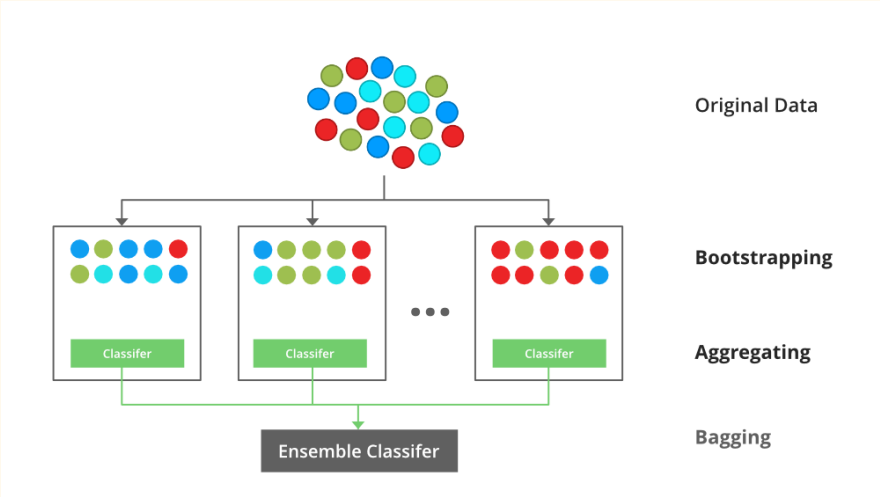


Gradient boosting classifiers: Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting. Gradient boosting models are becoming popular because of their effectiveness at classifying complex datasets. The Gradient Boosting Classifier depends on a loss function. A custom loss function can be used, and many standardized loss functions are supported by gradient boosting classifiers, but the loss function has to be differentiable.

Light Gradiemt Boosting Machine: Light GBM is a fast, distributed, high-performance gradient boosting framework based on decision tree algorithm, used for ranking, classification and many other machine learning tasks.it is based on decision tree algorithms, it splits the tree leaf wise with the best fit whereas other boosting algorithms split the tree depth wise or level wise rather than leaf-wise. So when growing on the same leaf in Light GBM, the leaf-wise algorithm can reduce more loss than the level-wise algorithm and hence results in much better accuracy which can rarely be achieved by any of the existing boosting algorithms.

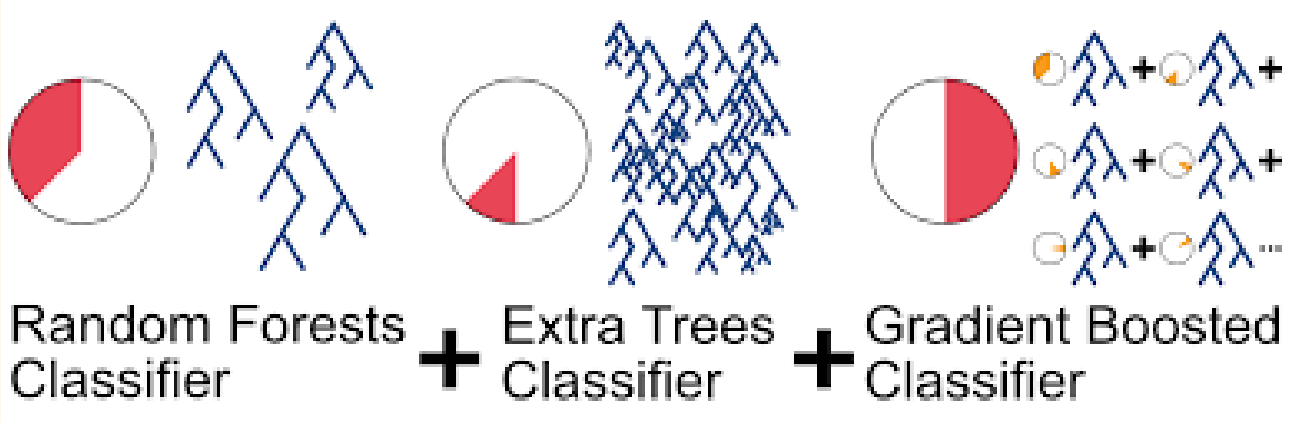


XGBoost: XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way.



Extra Trees classifier : Similar to a Random Forest classifier we have the Extra Trees classifier — also known as Extremely Randomized Trees. To introduce more variation into the ensemble, we will change how we build trees.Each decision stump will be built with the following criteria: all the data available in the training set is used to built each stump, to form the root node or any node, the best split is determined by searching in a subset of randomly selected features of size sqrt(number of

features). The split of each selected feature is chosen at random, and maximum depth of the decision stump is one.



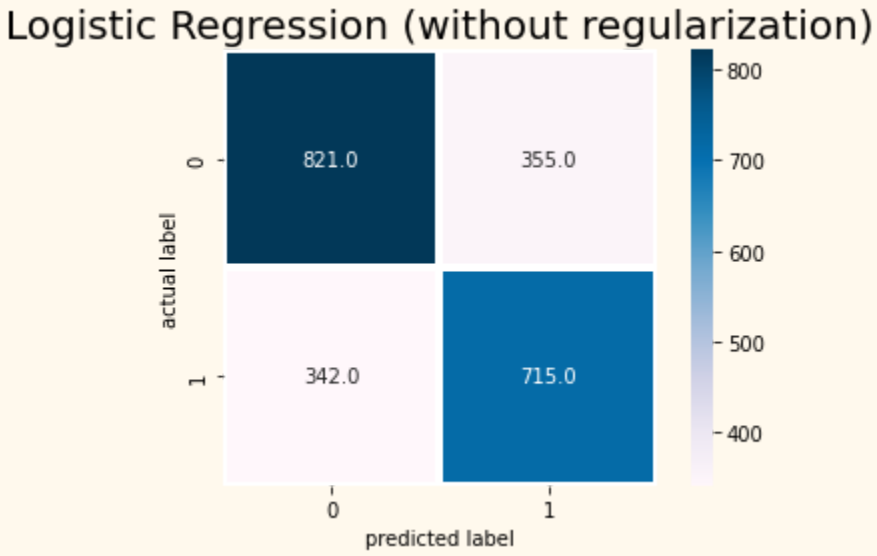
Algorithm	Accuracy	Precision	Recall	F1 Score	Sensitivity	Specificity	AUC (ROC)	Error Rate	TPR	FPR	Execution Time
Logistic Regression (without regularization)	68.78	66.82	67.64	67.23	67.64	69.81	68.72	31.21	67.64	30.18	2.34 ms
Logistic Regression (with regularization)	70.71	71.23	63.85	67.36	63.85	76.87	70.36	29.28	63.85	23.12	1.61 ms
KNeighbor Classifier	60.94	63.86	40.30	49.41	40.30	79.50	59.90	39.05	40.40	20.49	5.72 ms
Naive Bayes	68.38	71.80	54.68	62.08	54.68	80.69	67.69	31.61	54.68	19.30	1.13 ms
RBF SVM	56.47	56.13	36.80	44.45	36.80	74.14	55.47	43.52	36.80	25.85	1.08 ms
Linear SVM	49.48	47.25	57.71	51.95	57.71	42.09	49.90	50.51	57.71	57.90	2.11 ms
Decision Tree (informatio	69.50	73.15	56.19	63.56	56.19	81.46	68.82	30.49	56.19	18.53	1.05 ms

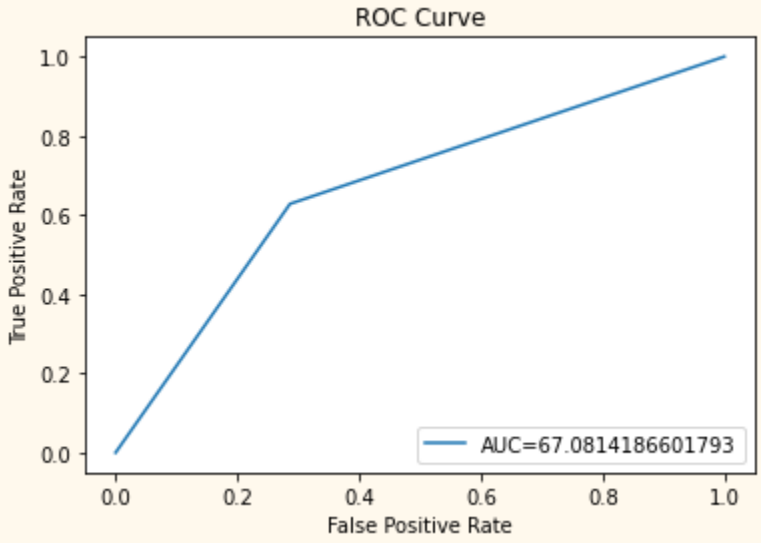
Algorithm	Accuracy	Precision	Recall	F1 Score	Sensitivity	Specificity	AUC (ROC)	Error Rate	TPR	FPR	Execution Time
n gain)											
Decision Tree (gini index)	70.53	78.37	52.12	62.61	52.12	87.07	69.60	29.46	52.12	12.92	1.04 ms
Random Forest	72.05	76.82	58.65	66.52	58.65	84.09	71.37	27.94	58.65	15.90	1.54 ms
Ada boost	71.69	75.14	60.07	66.71	60.07	82.14	71.10	28.30	60.07	17.85	1.94 ms
MLP Classifier	64.12	58.57	82.68	68.57	82.68	47.44	65.06	35.87	82.68	52.55	2.00 ms
Quadratic Discriminant Analysis	64.26	84.71	29.89	44.19	29.89	95.15	62.52	35.73	29.89	4.84	2.55 ms
Gradient boosting Classifier	70.98	73.86	59.88	66.14	59.88	80.95	70.41	29.01	59.88	19.04	1.89 ms
Light Gradient Boosting Machine	72.50	76.91	59.88	67.34	59.88	83.84	71.86	27.49	59.88	16.15	2.38 ms

Algorithm	Accuracy	Precision	Recall	F1 Score	Sensitivity	Specificity	AUC (ROC)	Error Rate	TPR	FPR	Execution Time
XG Boost	72.50	76.52	60.45	67.54	60.45	83.33	71.89	27.49	60.45	16.66	1.81ms
Extra Tree Classifier	67.30	66.33	62.81	64.52	62.81	71.34	67.08	32.69	62.81	28.65	1.61 ms

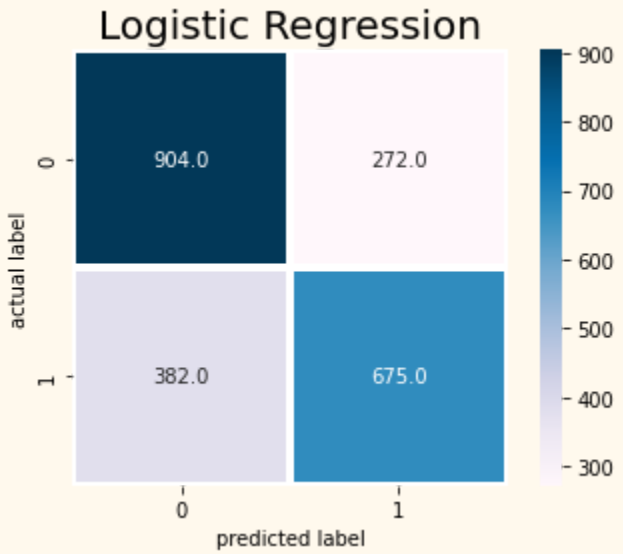
Confusion Matrices and ROC Curves

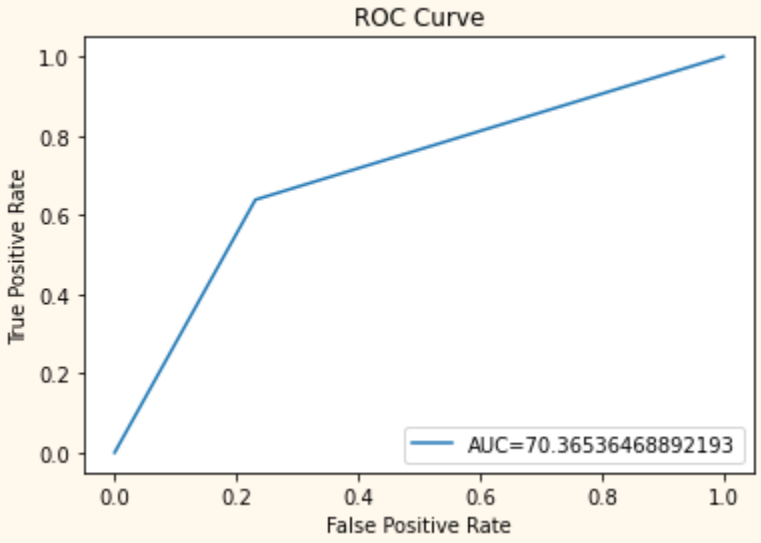
1. **Logistic Regression (without regularization)**



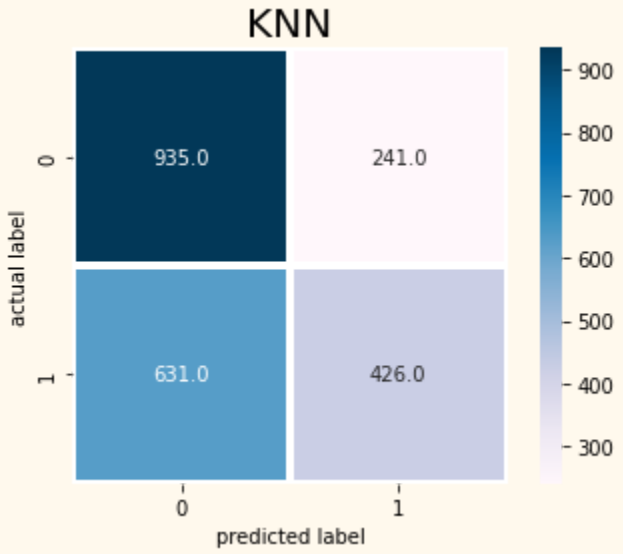


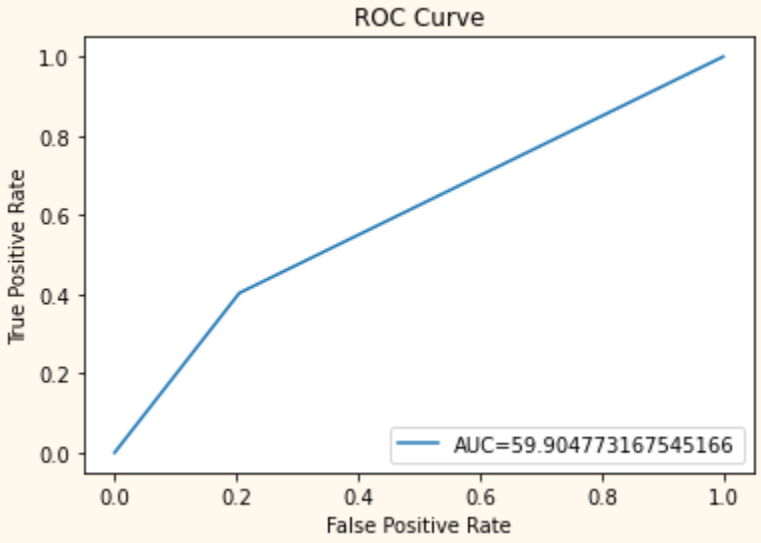
2. Logistic Regression (with regularization)



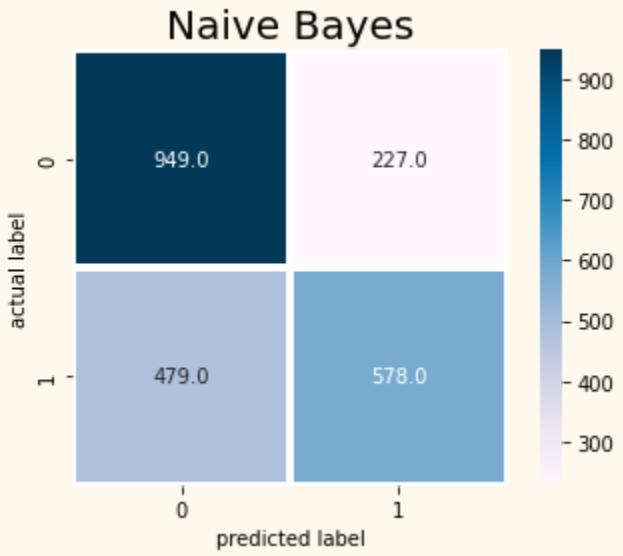


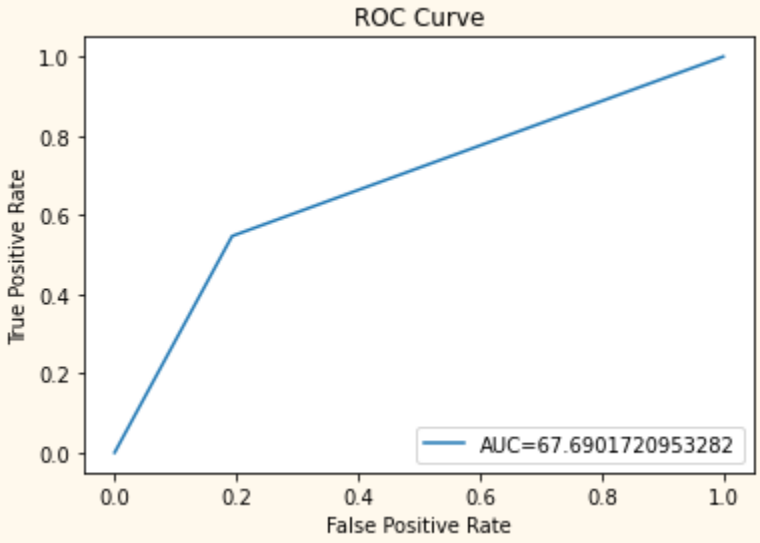
3. KNeighbor Classifier



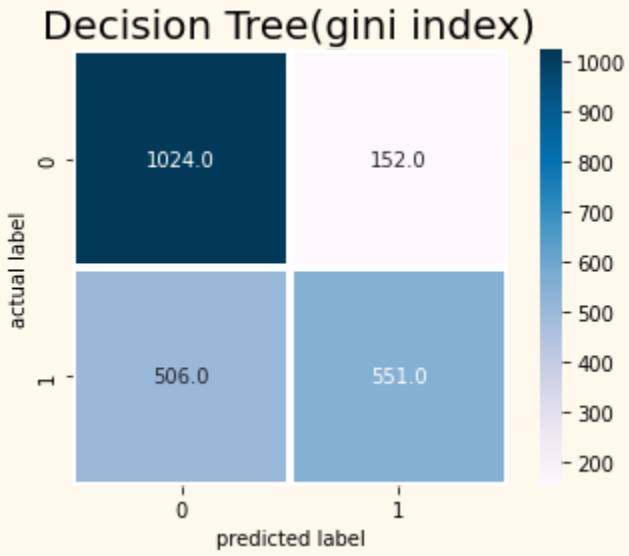


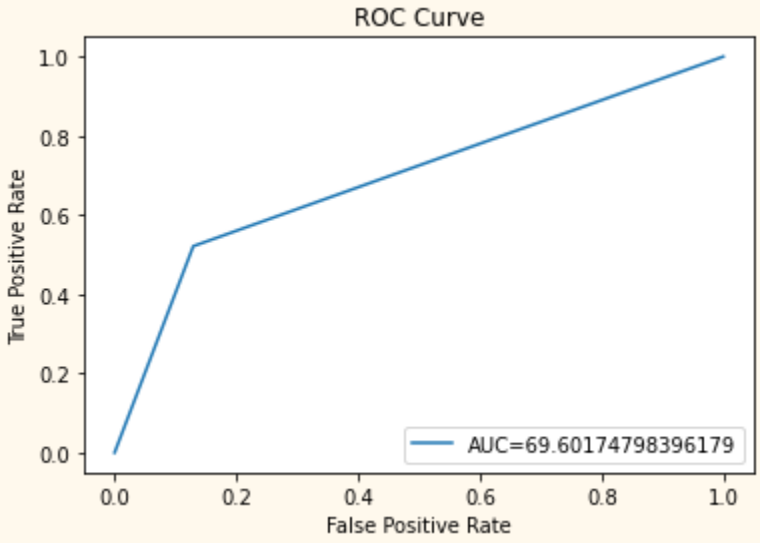
4. Naive Bayes



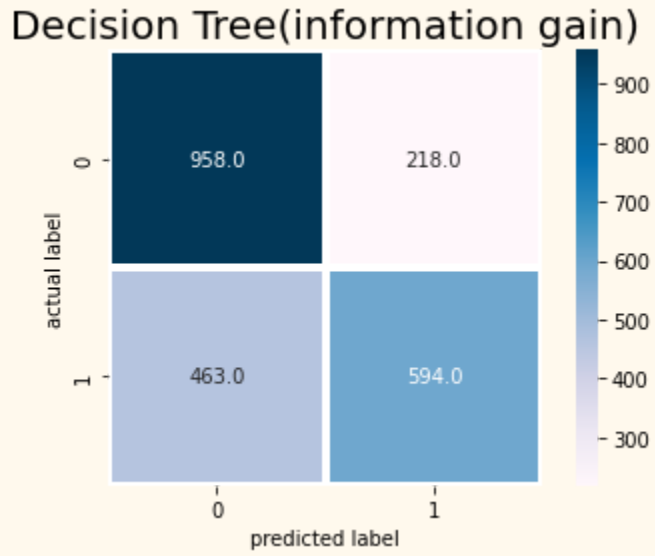


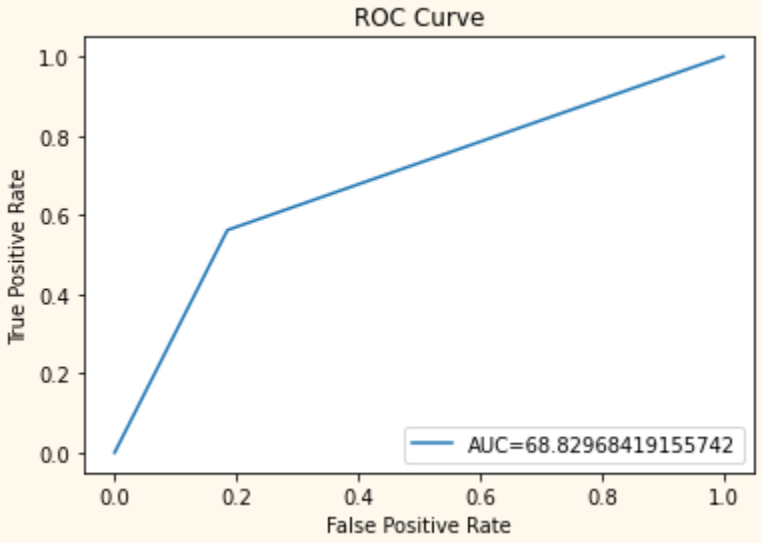
5. Decision tree (gini index)



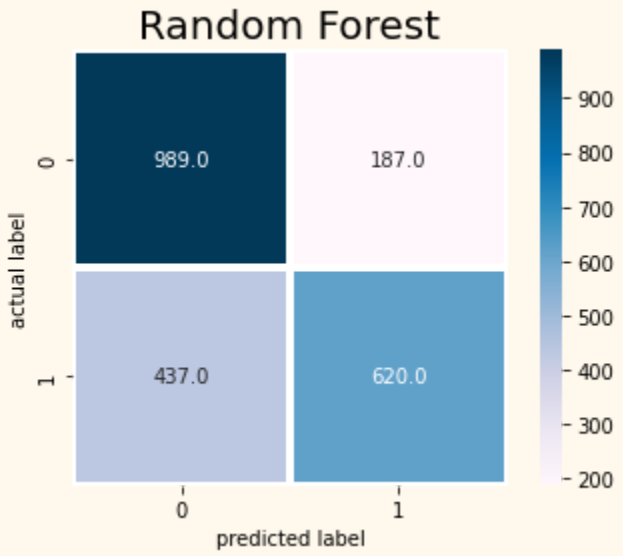


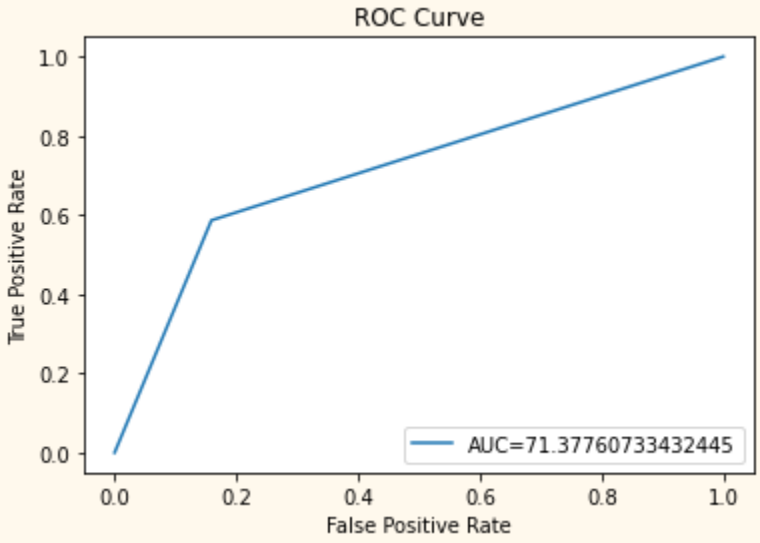
6. Decision tree (information gain)



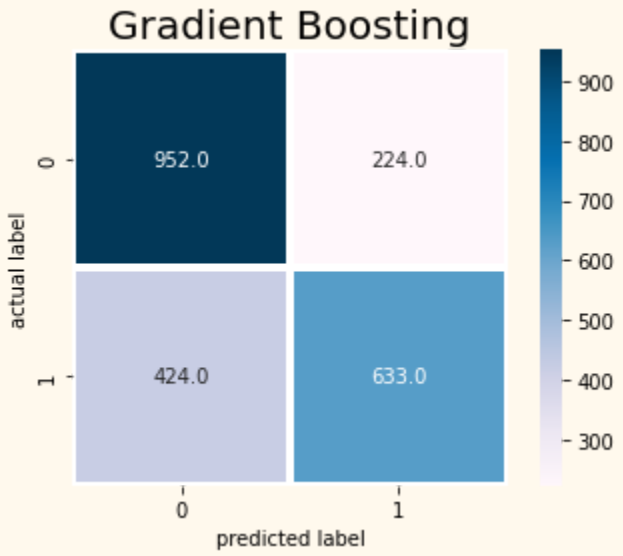


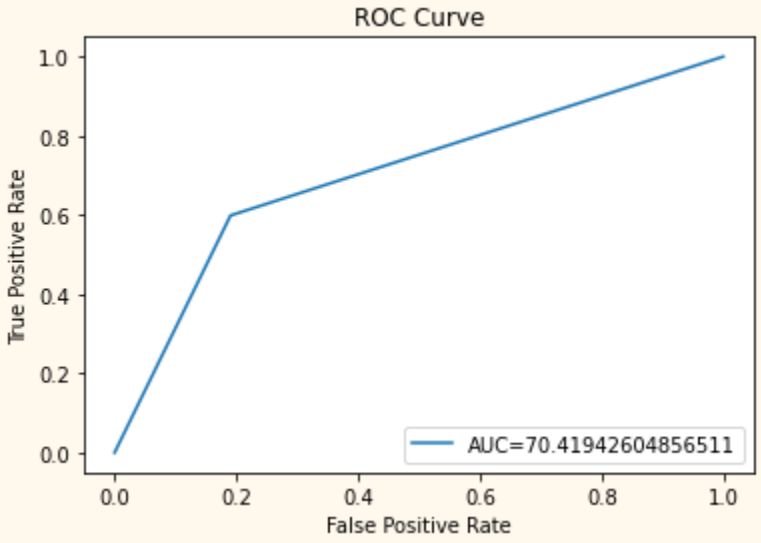
7. Random Forest



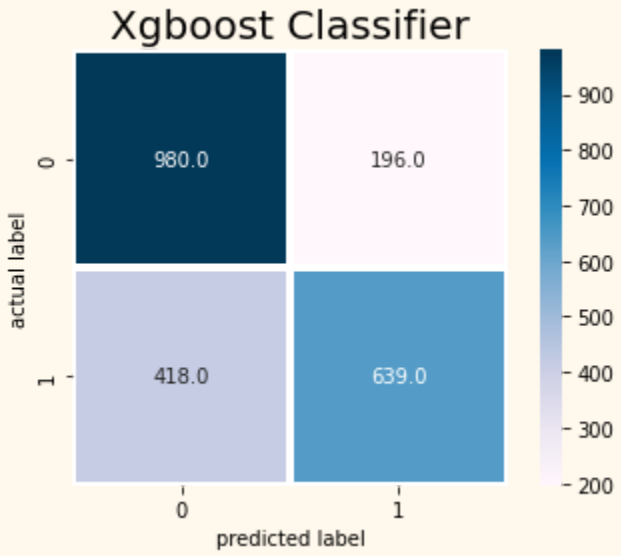


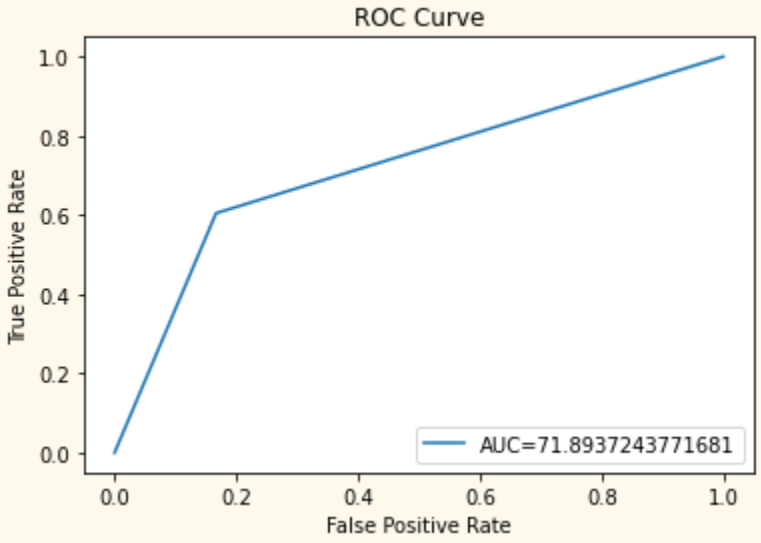
8. Gradient Boosting



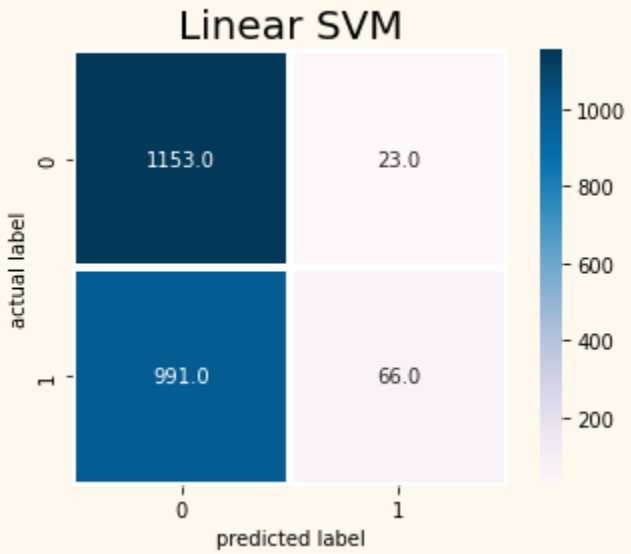

ROC Curve
True Positive Rate
False Positive Rate
AUC=70.41942604856511
0.0
0.2
0.4
0.6
0.8
1.0

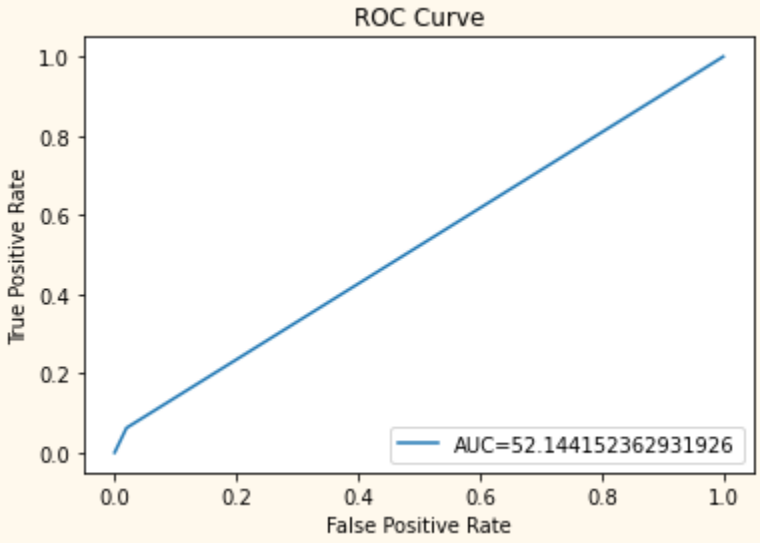
9. XGboost


Xgboost Classifier
980.0
196.0
418.0
639.0
actual label
predicted label
0
1
900
800
700
600
500
400
300
200

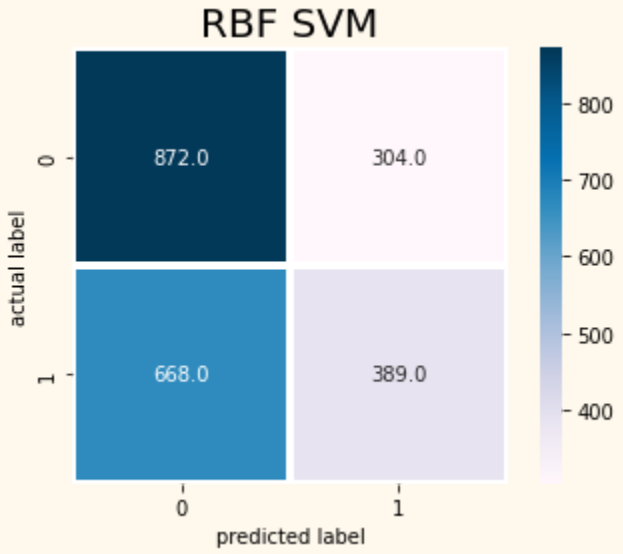


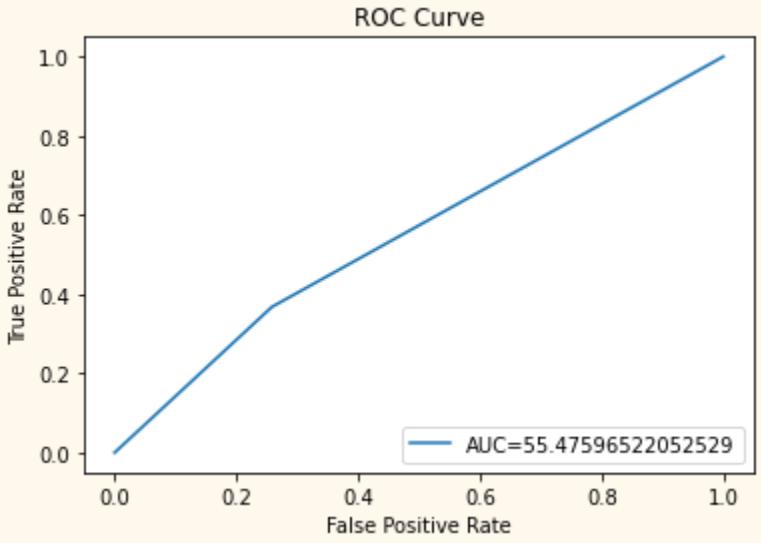
10. Linear SVM



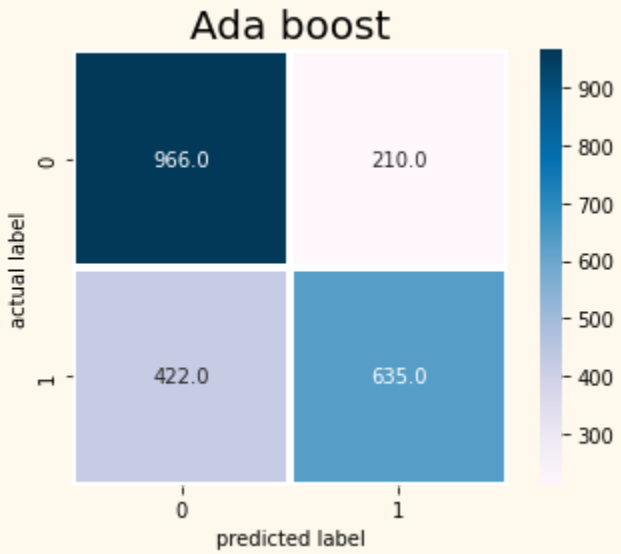

ROC Curve
True Positive Rate
False Positive Rate
AUC=52.144152362931926
0.0
0.2
0.4
0.6
0.8
1.0

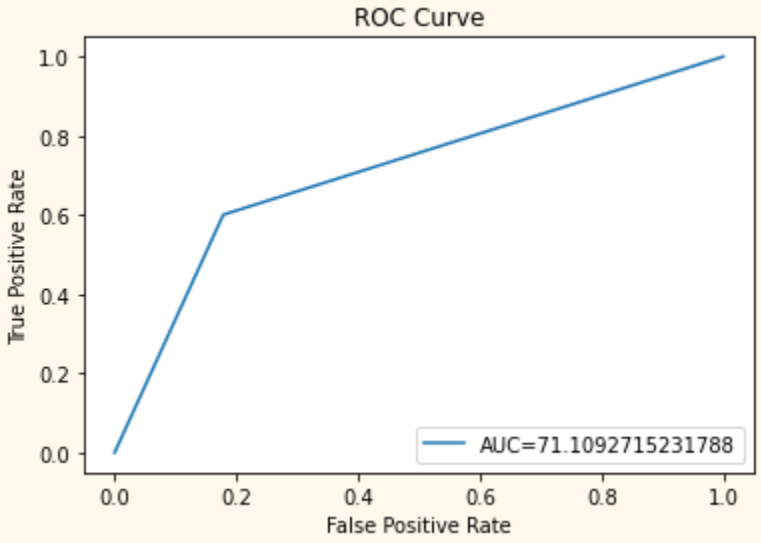
11. RBF SVM


RBF SVM
872.0
304.0
668.0
389.0
actual label
predicted label
0
1
800
700
600
500
400

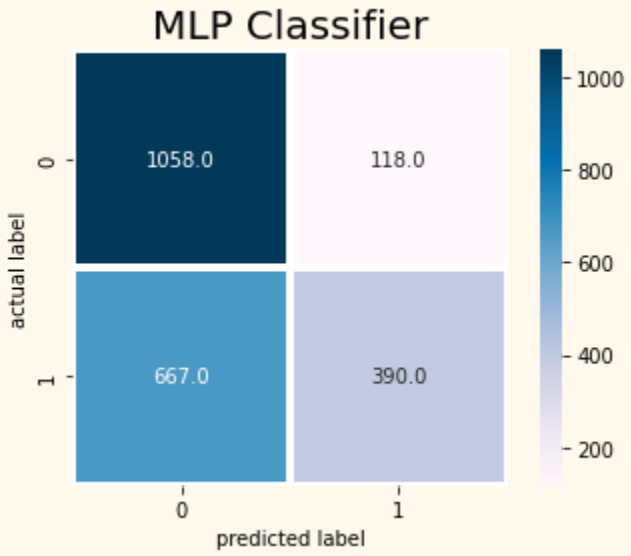

ROC Curve
True Positive Rate
False Positive Rate
AUC=55.47596522052529
0.0
0.2
0.4
0.6
0.8
1.0

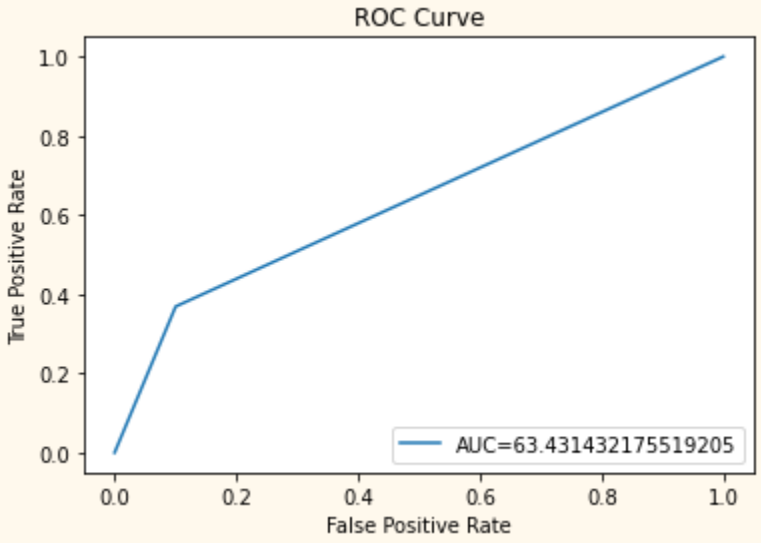
12. Ada boost


Ada boost
966.0
210.0
422.0
635.0
actual label
predicted label
0
1
900
800
700
600
500
400
300

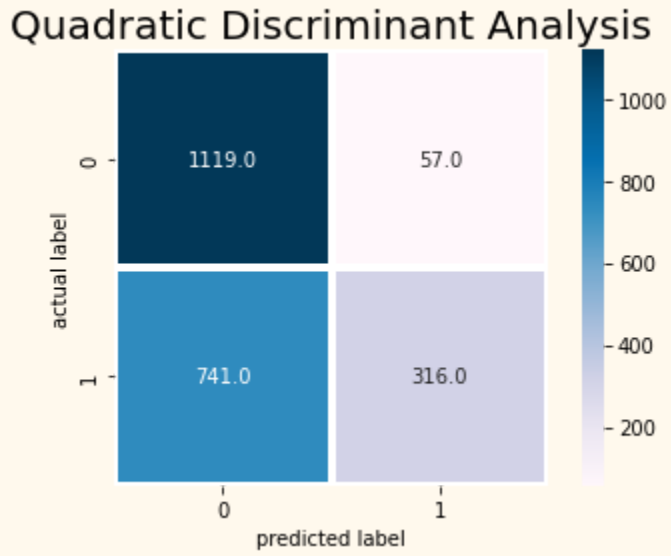


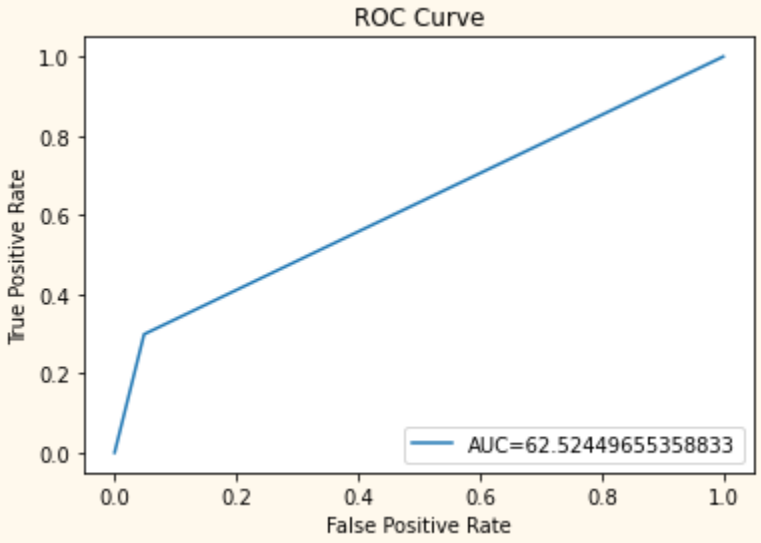
13. MLP Classifier



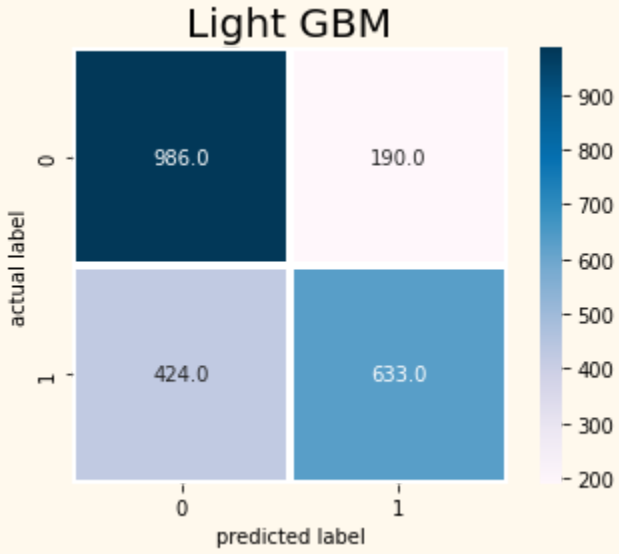


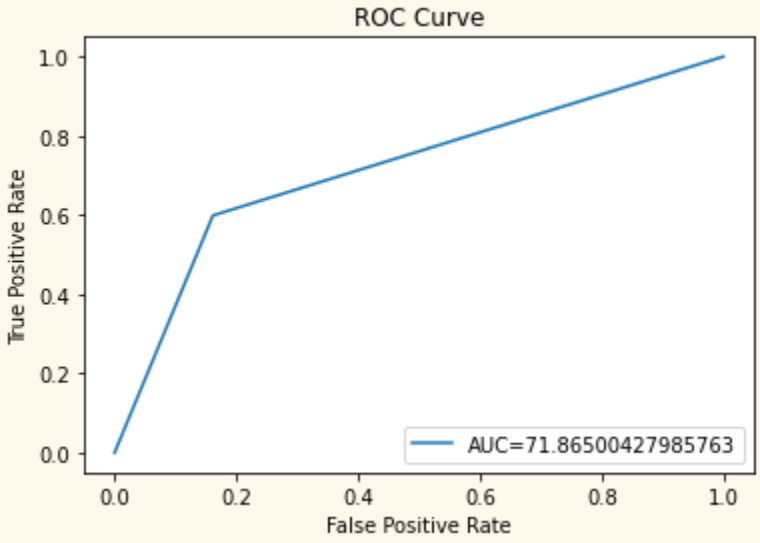
14. Quadratic Discriminant Analysis



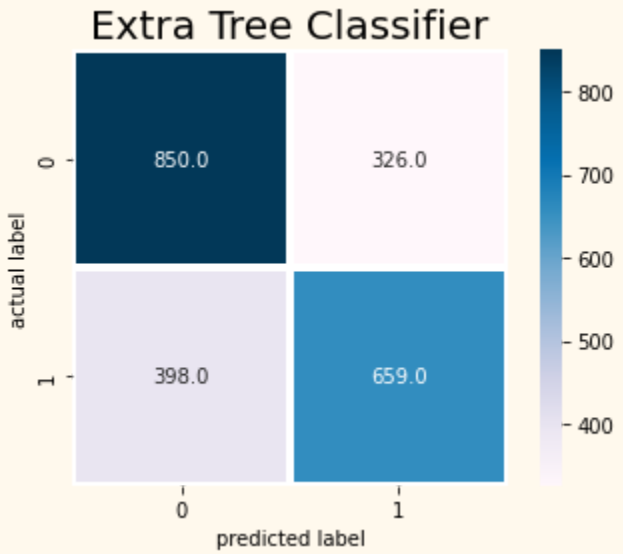


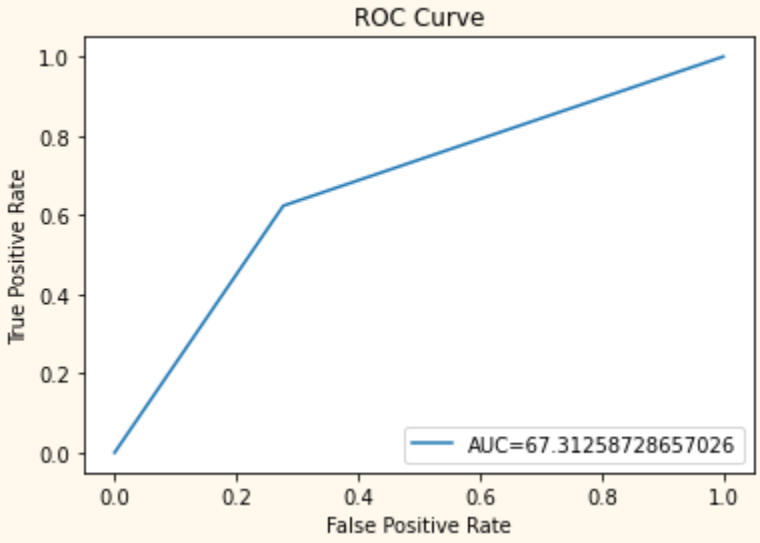
15. Light Gradient Boosting Machine





16. Extra tree classifier





Conclusion

This research demonstrate the different data mining methods which is a great tool in the decision making. For this study work, real data was taken from the data base of University of California, Irvine website which is open source. In general there are 2 steps involved in this thesis work. In first step the process of outlier analysis and feature selection have been done using correlation matrix. In the second step, the best subset of variables which are selected by these methods are go through for the classification. For classification purpose there are 4 computational algorithms for classification have been used that are Support vector Machines, Decision Trees, and Random Forest for classification, Artificial neural network. The aim was to check whether these classification methods give same number of accuracy and performance by using the feature selection approach. The results indicated that we do not need to go for the full model as reduced subset of 10 variables can provide almost the same accuracy. Regarding feature subset, the best subset of variables is chosen by the xgboost classification, also xgboost comes up with the most accurate results with 72.5% accuracy on the subset selected by XGboost. Area under curve of ROC shows the performance of 0.7189 for this selected subset. Moreover, RF reveals that the most impacting attribute is duration, afterwards there are balance and age respectively.

References

Barboza, F., Kimura, H. & Altman, E. (2017), 'Machine learning models and bankruptcy prediction', Expert Systems with Applications 83, 405–417.

Bardsley, W. E., Vetrova, V. & Liu, S. (2015), ‘Toward creating simpler hydrological models: A lasso subset selection approach’, Environmental Modelling & Software 72, 33–43.

Beucher, A., Møller, A. & Greve, M. (2017), ‘Artificial neural networks and decision tree classification for predicting soil drainage classes in denmark’, Geoderma .

Cramer, S., Kampouridis, M., Freitas, A. A. & Alexandridis, A. K. (2017), ‘An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives’, Expert Systems with Applications 85, 169–181.

Friedman, J., Hastie, T. & Tibshirani, R. (2001), The elements of statistical learning, Vol. 1, Springer series in statistics New York. Genuer, R., Poggi, J.-M. & Tuleau-Malot, C. (2010), ‘Variable selection using random forests’, Pattern Recognition Letters 31(14), 2225–2236.

Genuer, R., Poggi, J.-M. & Tuleau-Malot, C. (2015), ‘Vsurf: An r package for variable selection using random forests.’, R Journal 7(2). Gil, D. & Johnsson, M. (2010), ‘Using support vector machines in diagnoses of urological dysfunctions’, Expert Systems with Applications 37(6), 4713–4718.

Izetta, J., Verdes, P. F. & Granitto, P. M. (2017), ‘Improved multiclass feature selection via list combination’, Expert Systems with Applications 88, 205–216.

Jaiswal, J. K. & Samikannu, R. (2017), Application of random forest algorithm on feature subset selection and classification and regression, in ‘Computing and Communication Technologies (WCCCT), 2017 World Congress on’, IEEE, pp. 65–68.

Karouni, A., Daya, B. & Bahlak, S. (2011), ‘Offline signature recognition using neural networks approach’, Procedia Computer Science 3, 155–161.

Kohavi, R. & John, G. H. (1997), ‘Wrappers for feature subset selection’, Artificial intelligence 97(1-2), 273–324. Lawless, J. & Singhal, K. (1978), ‘Efficient screening of nonnormal regression models’, Biometrics pp. 318–327.

Le, H. H. & Viviani, J.-L. (2017), ‘Predicting bank failure: An improvement by implementing machine learning approach on classical financial ratios’, Research in International Business and Finance . Moro, S., Cortez, P. & Rita, P. (2014), ‘A data-driven approach to predict the success of bank telemarketing’, Decision Support Systems 62, 22–31.