Oracle SmartBridge Project Kasturi Harkanchi
Machine Learning through java 4/5/2021

INTRODUCTON:

Project Description:

Marketing to potential clients has always been a crucial challenge in attaining success for banking institutions. It’s not a surprise that banks usually deploy mediums such as social media, customer service, digital media and strategic partnerships to reach out to customers. But how can banks market to a specific location, demographic, and society with increased accuracy? With the inception of machine learning - reaching out to specific groups of people have been revolutionized by using data and analytics to provide detailed strategies to inform banks which customers are more likely to subscribe to a financial product.

So the goal is to build the best classification model to predict whether a client will subscribe to a bank term deposit or not

Predictive analytics can help banks by providing deep insights into customer needs; launch innovative products and services; deliver personalized and stellar experiences; lead to new business models; and transform using new processes and technology. Related benefits of predictive analytics include:

- **Deep insights into customer needs**—Better customer insights enable lenders to more effectively target their customers with relevant and thoughtful services at the appropriate moment.
- **Launch innovative products and services**—More effective redesign of products and services based on customer research, segmentation and analysis, enhanced portfolio strategies and pricing.

Literature Survey:

Problems before predictive analytics :

- credit scoring
- finance mangement
- managing budget
- fraud activities
- loan procedures are very long

After predictive analytics come into view:

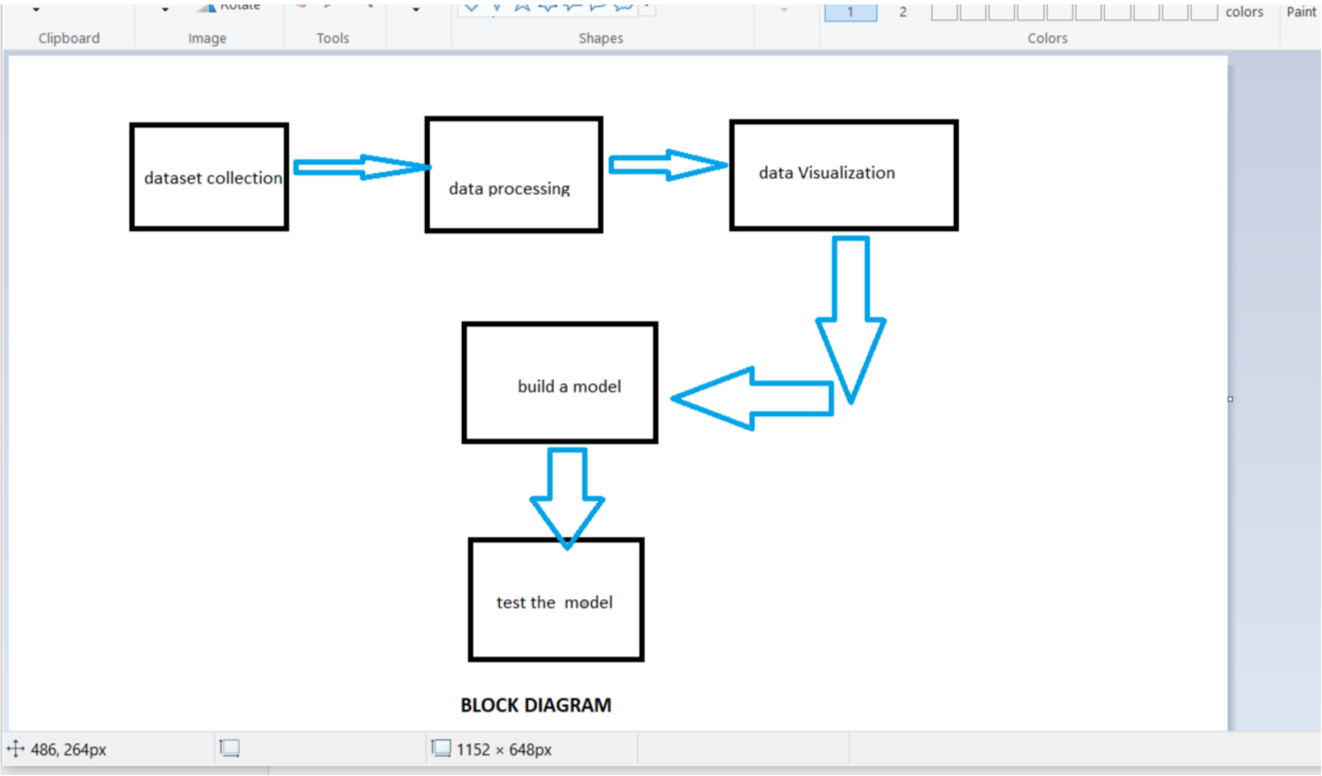
As the computers getting smarter every day, so does our tasks are easing out. Most importantly the crucial sectors such as healthcare, financial institutions, and more demand high attention as these sectors involve the critical consumer databases and historical transactions with the aim of predicting the future. Thus, predictive analytics could aid in cutting down the costs as well as enhance the overall banking experience.

Predictive analytics is the process of using computer models to forecast future events using sophisticated programs and depend on certain efficient technologies such as artificial intelligence, data mining, and machine learning to process massive information. Using these technologies the model would analyze and predict the future happening based on the current conditions.

The predictive analytics could enhance the overall banking experience for the customers in several ways, it could find it unsettling that financial institutions that have no much information, and that rely on computers for decision making could affect one's life. On the other hand, computers are always available and would provide a similar service to every customer without being partial to any customers.

Theoretical Analysis:

Block diagram of Project:



Hardware and Software Requirements:

✠ Hardware requirements:

- If your tasks are small and can fit in a complex sequential processing, you don’t need a big system. You could even skip the GPUs altogether. A CPU such as i7–7500U can train an average of ~115 examples/second.
- 256 MB RAM
- 1 Gb hard free drive space

✠ Software Requirements:

- eclipse ide for java developers
- weka library

Experimental Analysis:

Data Processing:

Stages of Data Preprocessing are:

1. Data Cleaning
2. Data Integration
3. Data Reduction
4. Data Transformation

Data Cleaning:

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

In this dataset datacleaning is done in weka by selecting filters such as remove duplicates,missingvaluebyuserinput and applyting them to dataset.

- Remove duplicate is used for removing duplicate elements in a dataset.
- MissingvaluebyuserInput filter is used to replace missing values in dataset by user input.

Data Integration:

Data Integration refers to the process of unifying data from multiple data sources.

Data Reduction:

Data Reduction mechanism can be used to reduce the representation of the large dimensional data. By using a data reduction technique, you can reduce the dimensionality that will improve the manageability and visibility of data. Further, you can achieve similar accuracies.

Principal Component Analysis is also known as Karhunen-Loeve or K-L method, is used to reduce components to handlable attributes from a large number of the dimensions. In other words, Principal Component Analysis combines the important features of attributes and reduces the variables by introducing alternative variables. After the Principal Component Analysis is done, multiple dimensions can be represented into a manageable number of variables.

Data Transformation:

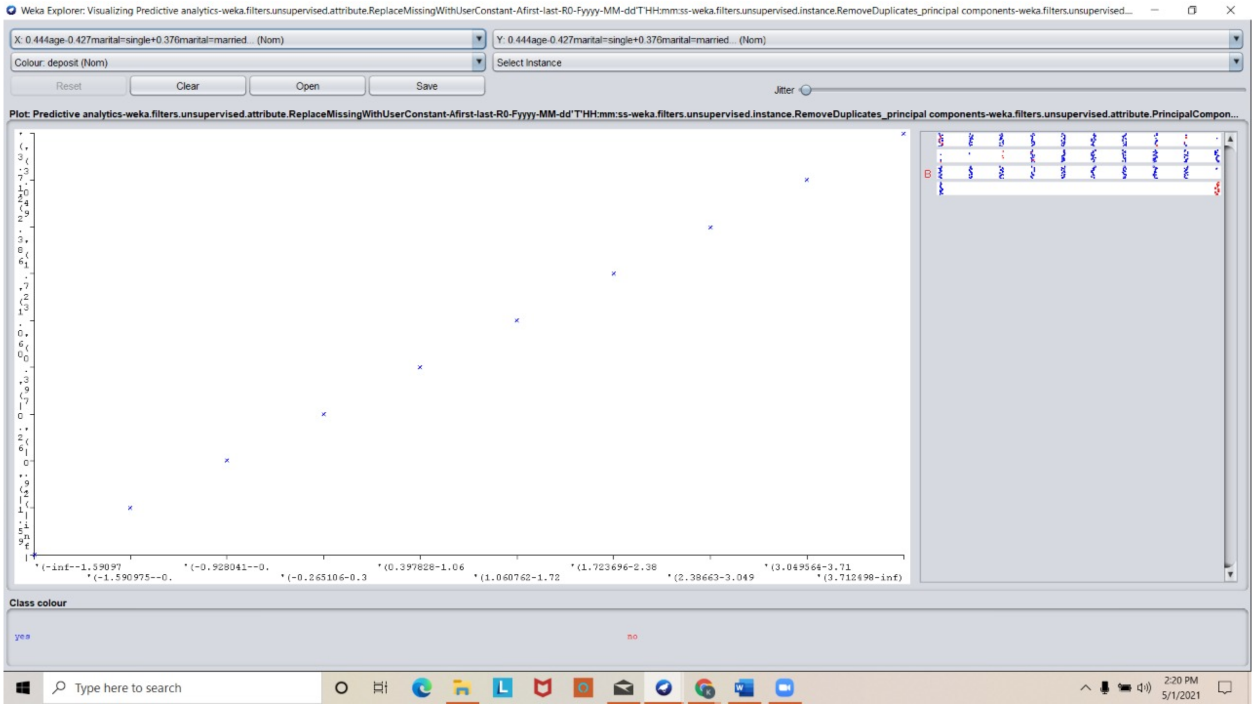
Data Transformation is the technique of converting data from one format to another.

- It is performed in weka through filters such as discretization,standardize,nominaltobinary
  Discretize will discretize the values according to a number of bins (n). Weka will simply cut the range of the values in n subsets, and give the value of the subset to the instances. This is if your attribute is really a continuous variable.
- Numeric To Nominal is to transform some Numeric values into a Nominal variable, if this attributes has few unique values. For example, if you have an ID attribute which clusters your dataset in few subsets, it may be wise to convert it into a Nominal attribute instead of treating it like a number. This applies for attributes which are not really continuous, but treated as numeric .

DATA VISUALIZATION:

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

The data visualization of this project is performed in weka



Weka Explorer: Visualizing Predictive analytics-weka.filters.unsupervised.attribute.ReplaceMissingWithUserConstant-Afirst-last-R0-Fyyyy-MM-dd'T'HH:mm:ss-weka.filters.unsupervised.instance.RemoveDuplicates_principal components-weka.filters.unsupervised...

X: 0.444age-0.427marital=single+0.376marital=married... (Nom)

Y: 0.444age-0.427marital=single+0.376marital=married... (Nom)

Colour: deposit (Nom)

Select Instance

Plot: Predictive analytics-weka.filters.unsupervised.attribute.ReplaceMissingWithUserConstant-Afirst-last-R0-Fyyyy-MM-dd'T'HH:mm:ss-weka.filters.unsupervised.instance.RemoveDuplicates_principal components-weka.filters.unsupervised.attribute.PrincipalCompon...

Model Building:

The algorithm used for building model is logistic regression.

LOGISTIC REGRESSION:

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

Mathematically, a logistic regression model predicts P(Y=1) as a function of X.

Metrics used:

Precision is a metric that quantifies the number of correct positive predictions made.

- Precision = TruePositives / (TruePositives + FalsePositives)

Recall is a metric that quantifies the number of correct positive predictions made out of all positive predictions that could have been made.

- Recall = TruePositives / (TruePositives + FalseNegatives)

F-Measure provides a way to combine both precision and recall into a single measure that captures both properties.

- F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Accuracy - Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.

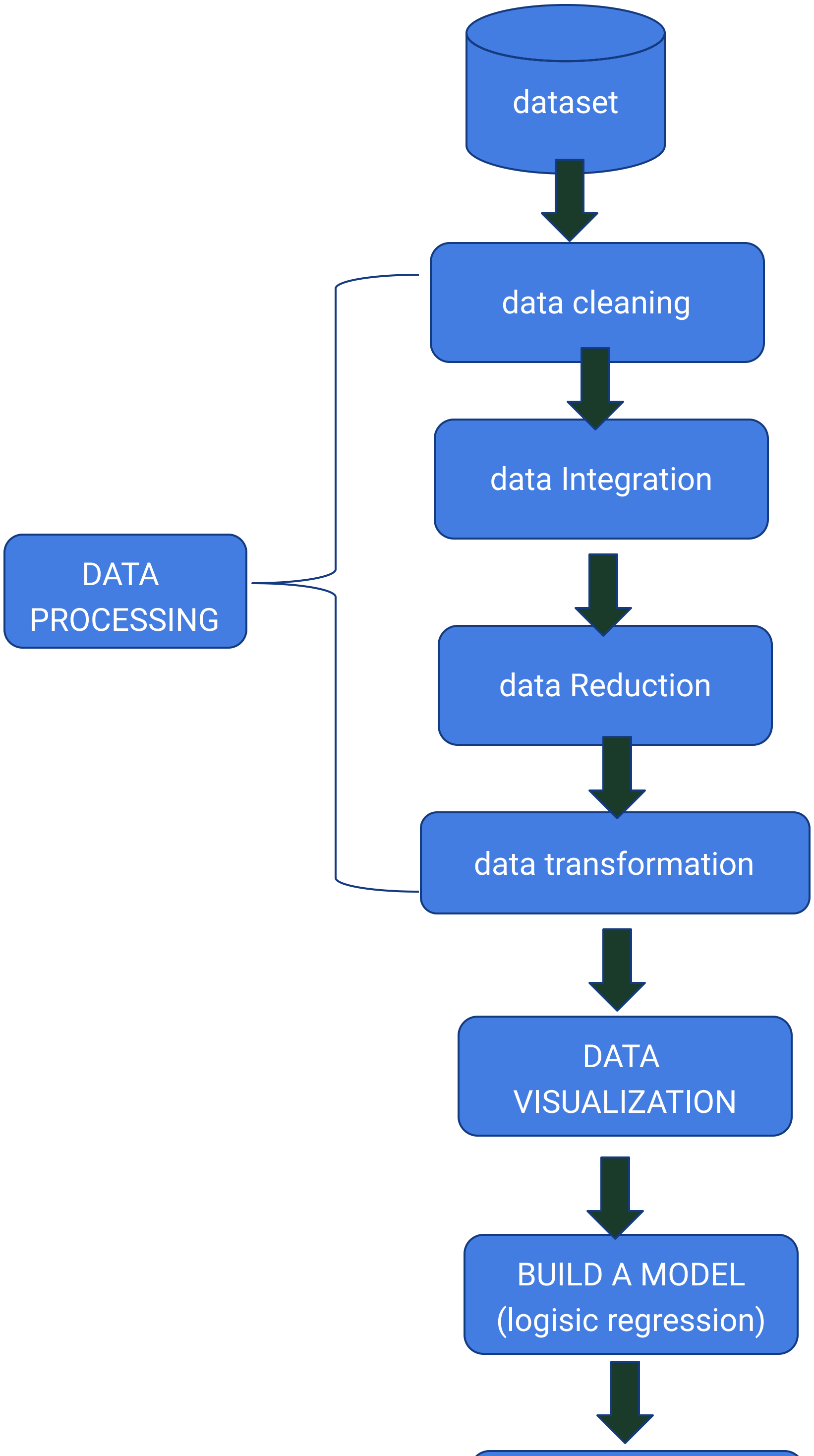
- Accuracy = TP+TN/TP+FP+FN+TN

Testing the model:

Now that we have trained our model lets see how it did on the data

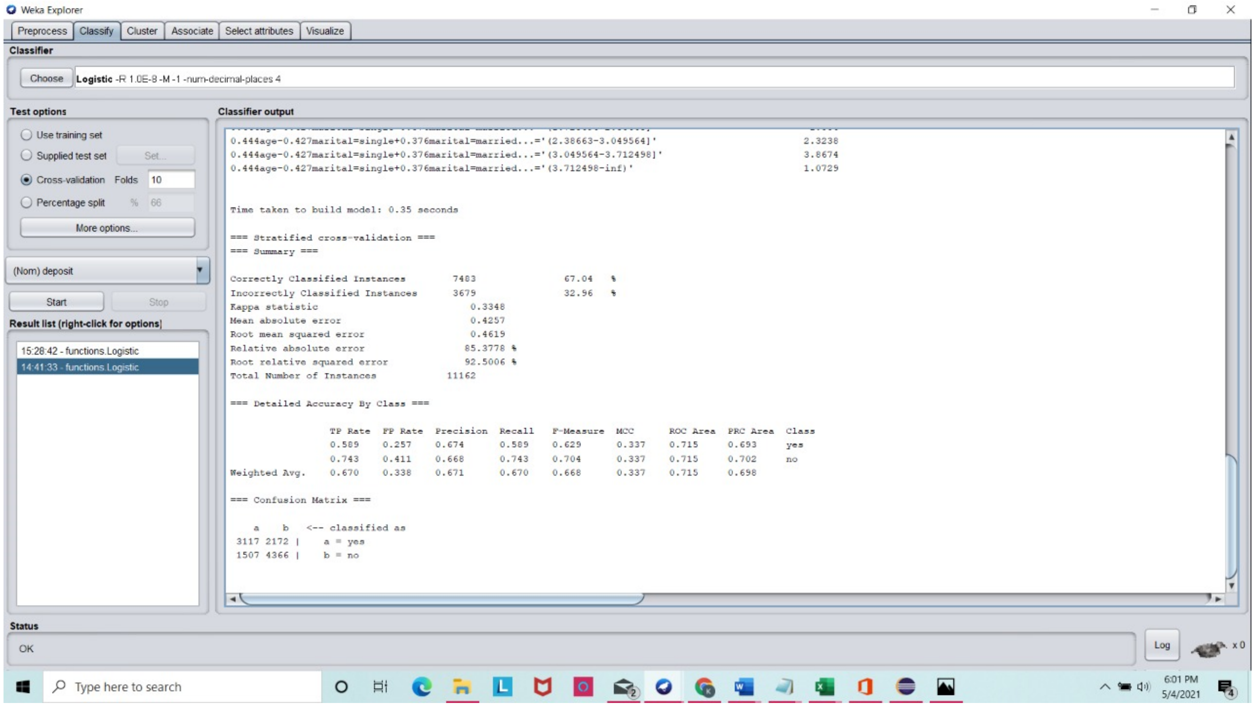
As you can see , the scores are very close which indicates that we avoided over-fitting. I should mention that this is a good indication that we have not over-fit the model, however it is not the end all be all. Next we'll discuss another method to prevent over-fitting of our data and hopefully improve our ability to generalize over new data.

FLOWCHART:



RESULT:
Build a model: Logistic regression

```
Time taken to build model: 0.12 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       7483            67.04 %
Incorrectly Classified Instances      3679            32.96 %
Kappa statistic                    0.3348
Mean absolute error                  0.4257
Root mean squared error                0.4619
Relative absolute error              85.3778%
Root relative squared error            92.5006%
Total Number of Instances            11162

=== Detailed Accuracy By Class ===


                  TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC
Area  Class
                   0.589    0.257    0.674      0.589    0.629      0.337    0.715     0.693     yes
                   0.743    0.411    0.668      0.743   0.704      0.337    0.715     0.702    no
Weighted Avg  0.670   0.338    0.671      0.670    0.668      0.337    0.715     0.698

=== Confusion Matrix ===

    a    b   <-- classified as
 3117 2172 |    a = yes
 1507 4366 |    b = no
```



Logistic -R 1.0E-8 -M -1 -num-decimal-places 4

Cross-validation Folds 10

Percentage split % 66

(Nom) deposit

15:28:42 - functions.Logistic
14:41:33 - functions.Logistic

```
0.444age-0.427marital=single+0.376marital=married...='(2.38663-3.049564]'                2.3238
0.444age-0.427marital=single+0.376marital=married...='(3.049564-3.712498]'               3.8674
0.444age-0.427marital=single+0.376marital=married...='(3.712498-inf)'                    1.0729


Time taken to build model: 0.35 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        7483               67.04   %
Incorrectly Classified Instances      3679               32.96   %
Kappa statistic                          0.3348
Mean absolute error                      0.4257
Root mean squared error                  0.4619
Relative absolute error                 85.3778 %
Root relative squared error             92.5006 %
Total Number of Instances            11162

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.589    0.257    0.674      0.589    0.629      0.337    0.715     0.693     yes
                 0.743    0.411    0.668      0.743    0.704      0.337    0.715     0.702     no
Weighted Avg.    0.670    0.338    0.671      0.670    0.668      0.337    0.715     0.698

=== Confusion Matrix ===

    a    b   <-- classified as
 3117 2172 |    a = yes
 1507 4366 |    b = no
```

Test the model:

Correctly Classified Instances 1864 83.4751 %
Incorrectly Classified Instances 369 16.5249 %
Kappa statistic 0.6682
Mean absolute error 0.2548
Root mean squared error 0.3498
Relative absolute error 51.0922 %
Root relative squared error 70.0658 %
Total Number of Instances 2233

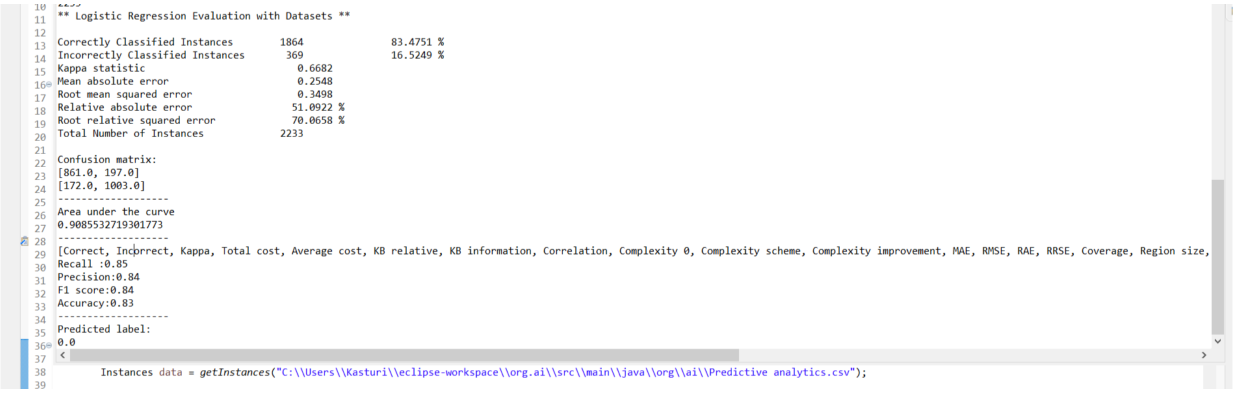
Confusion matrix:
[861.0, 197.0]
[172.0, 1003.0]

Area under the curve
0.9085532719301773

[Correct, Incorrect, Kappa, Total cost, Average cost, KB relative, KB information, Correlation, Complexity 0, Complexity scheme, Complexity improvement, MAE, RMSE, RAE, RRSE, Coverage, Region size, TP rate, FP rate, Precision, Recall, F-measure, MCC,

ROC area, PRC area]
Recall :0.85
Precision:0.84
F1 score:0.84
Accuracy:0.83

Predicted label:
0.0



```
** Logistic Regression Evaluation with Datasets **

Correctly Classified Instances        1864               83.4751 %
Incorrectly Classified Instances       369               16.5249 %
Kappa statistic                          0.6682
Mean absolute error                      0.2548
Root mean squared error                  0.3498
Relative absolute error                 51.0922 %
Root relative squared error             70.0658 %
Total Number of Instances             2233

Confusion matrix:
[861.0, 197.0]
[172.0, 1003.0]
-------------------
Area under the curve
0.9085532719301773
-------------------
[Correct, Incorrect, Kappa, Total cost, Average cost, KB relative, KB information, Correlation, Complexity 0, Complexity scheme, Complexity improvement, MAE, RMSE, RAE, RRSE, Coverage, Region size,
Recall :0.85
Precision:0.84
F1 score:0.84
Accuracy:0.83
-------------------
Predicted label:
0.0

        Instances data = getInstances("C:\\Users\\Kasturi\\eclipse-workspace\\org.ai\\src\\main\\java\\org\\ai\\Predictive analytics.csv");
```

Advantages:

*Stay one step ahead in performance

After all, it's a *forecasting* technology.

It will take all of your campaign data (and previous) to determine what's most likely to fail or succeed in the future

*It saves time and energy

And, where does that usually go?

Testing, research, etc. You know the drill.

Moreover, HubSpot found that collecting/organizing data, handling emails, and managing land pages to be the most time-consuming tasks in a bankers's w

Bank processes Automation:

Advanced analytics provides critical insights to bussinesses.such insights allow banks to recognize the banking that will benefit significantly from automation.

Decoding Customers Sentiment:

Customer analytics through predictive tools provide customers choices and behaviour to the bank.

Disadvantages:

1. It can be intimidating to adopt

2. You have to spend time using it for its full effect

Applications:

1. Customer Segmentation

2. Fraud management & prevention

3. Risk modeling

4. Identifying the main channels of transactions (ATM withdrawal, credit/debit card payments

5. Customer Lifetime Value (LTV)

6. Feedback management

Conclusion:

predictive analytics not only offer a range of applications for the banking sector but represent an integral part of the financial industry as a whole. With a growing knowledge of technology and what it has made possible, customer expectations are now higher than ever.

Going forward, it is highly unlikely any serious contender in the financial world will survive without a well-designed strategy for the implementation of predictive analytics.

So, client will not subscribe to a bank term deposit

Future Scope:

Predictive analytics have already proved their value in predicting customer interactions and questions, but there is still more to come.As the market develops further, prescriptive analytics will come to the fore,which will further provide employees with intelligence on what they should do next or tips when engaging with customers to improve interactions.

References:

https://machinelearningmastery.com/java-machine-learning/

appendix:

https://drive.google.com/file/d/1dSvyFRdfPtN4kgRdKs2zKqYir2-RMCAd/view?usp=sharing