Classification Models for Bank Marketing Campaign: Towards Smart Bank Marketing

Ahmad Freij

American University in the Emirates, UAE.

Email: Afreij790@gmail.com

Abstract

In this paper, we have proposed two models of marketing classification which are Support Vector Machine (SVM) and Linear regression, these two models are the most popular and useful models of classification. In this paper, we represent how these two models are used for a case study of a bank marketing campaign, the dataset is related to a bank marketing campaign, and for Applying the machine learning models of classification, the RapidMiner software was used.

Keywords: Bank Marketing, Machine Learning, Artificial Intelligence, Smart E-Banking, Business Intelligence, Classification, E-Marketing.

1. Introduction

The extensive use of technology has led the world to capture some very insightful data. This data includes very critical information such as consumer behavior. If this data is processed correctly, we can predict the future. However, before we reach there, it is important to classify data into smaller chunks based upon the data characteristics and then use it for further processing. Classification allows for the categorization of a group of data and observes the patterns or trends associated with it. Without classification, processing the data or extracting useful information can become difficult. It's a way of better organizing the data. It is different in a way that we can identify both the similarities and differences existing within the data set and understand diversity in a better way.

DOI:
https://doi.org/10.54216/AJBOR.050102
Received: April 22, 2021 Accepted: September 14, 2021

Machine learning is a learning mechanism for the computers in which we offer our devices the required access to the information and permit them to acquire knowledge for themselves using that data. The three main ways of machine learning are Supervised Learning, Unsupervised Learning, and Reinforcement Learning.

For supervised learning, algorithms are trained using labeled instances, such as an input where the output results are known. E.g., a piece of equipment may have data sets labeled "F" (failed) or "P" (pass). The learning algorithm collects a collection of inputs and outputs that are valid. To find mistakes, the algorithm compares the current output with the correct outputs and then respectively modifies itself.

For unsupervised learning, it can be used for data that has no specific labeling. The consumer does not provide the "right answer" to the device. The algorithm should be able to decide what is being interpreted. The aim is to search the database to detect a sequence. On transactional results, unsupervised learning works well. It is widely used in advertisement campaigns because it allows advertisers to segment consumers based on common characteristics.

Reinforcement learning is guided learning focuses on getting "right" responses to a situation, reinforcement learning focuses on optimizing incentives rather than obtaining "correct" answers. This technique aids in the integration of machine learning into gaming, robots, and mapping. Algorithms use trial and error to figure out what to do next and which acts would produce the best results.

2. Related work

Classification is a method for categorizing any data into a defined and distinctive variety of groups, each with its label. Moreover, classification in machine learning is a supervised learning approach in which the ML algorithms learn from the input data supplied to them and then use what they've learned to identify new information for effective analysis and estimation.
For a more in-depth analysis, ML classification means that a set of data can be categorized into different classes. Voice recognition, image classification, bio-metric authentication text categorization, and other implementations of classification tasks are examples [10].
Machine learning algorithms for classification are known as "supervised" or "unsupervised" learning in particular. The expression "supervise" here refers to the presence of a testing dataset with the "right" properties for the algorithm to operate from, rather than human interaction. These approaches may be further categorized based on their objectives and the framework to which they are applied.
The styles of classification techniques used in machine learning are determined by the model viability and the skills of the ML developer. Classification's primary goal is to define classification, mention its algorithms, and clarify regression analysis and sigmoid likelihood. It also explains support vector machines, polynomial kernels, and the kernel trick, as well ask-nearest neighbors and KNN classification. With an example, Classification analyses Kernel support vector machines and implements the nave Bayes classifier [8].
It explains how to use a decision tree classifier and illustrates how to use a random forest classifier to make machine learning algorithms more realistic and efficient in the real world. People often associate classification with clustering, but the two are very distinct. In clustering, the aim is not to determine the desired class as in classification, but rather to affiliate program kinds of objects by assuming the happiest situation: all items in the same category should be identical, and no two items from different classes should be dissimilar.
Classification is the process of identifying various objects based upon their common characteristics and them grouping them

DOI:
https://doi.org/10.54216/AJBOR.050102
Received: April 22, 2021 Accepted: September 14, 2021

by associating them with terminology. We see classification in one form or another in everything that is around us. This includes the plants, animals, insects if we talk about nature, and on other hand, if we talk about the digital world, we deal with classification in terms of spam emails and those which are not spam.

Support vector machines (SVMs, also known as support vector networks) are guided classification algorithms that process data for regression and classification analytics and come with related learning algorithms. An SVM training algorithm creates a model that attributes new examples to one of 2 groups, rendering it a non-probabilistic linear programming classifier, given a collection of training examples, each labeled for methods are divided into two groups. It's a binary linear classifier that is non-probabilistic.

An SVM model is a description of the examples as points in space, mapped such that the instances of the different groups are separated by a large distance. New examples are then traced into that same domain and classified according to which aspect of the distance they fall on.

For example, SVM algorithms have been widely used for protein remote homology detection in recent years. These algorithms have long been used to differentiate between biological sequences. For example, gene classification, patients classified by their genes, and a variety of other biological issues [9].

In Machine Learning, linear regression is classified as a supervised learning algorithm. It is broadly used for predictive analysis, and it does so use the regression technique. What is the logistic regression, exactly? Well, regression is a technique for expressing the relationship among variables to better understand trends over time [5].

It is used in Analyzing engine efficiency in motor vehicles using test data. In biological systems, least squares regression is used to model causal connections among parameters. Scientists working with climate data analysis may benefit from Regression analysis. Consumer empirical studies and consumer survey outcome studies can also benefit from linear regression. In astronomical observation, linear regression is widely used. Astronomical data analysis uses a range of mathematical approaches and practices, and there are large libraries of languages such as python devoted to physics and astronomy data processing.

3. Proposed Models

For implementing the model RapidMiner software will use to analyze the chosen dataset, which is related to bank marketing, where the direct marketing campaign is being carried by the Portuguese banking organization, that we will extract meaningful analysis from it by applying various data processing models which are classification models such as support vector machine and linear regression models.

The following figures represent the mechanism of each model:

DOI:
https://doi.org/10.54216/AJBOR.050102
Received: April 22, 2021 Accepted: September 14, 2021

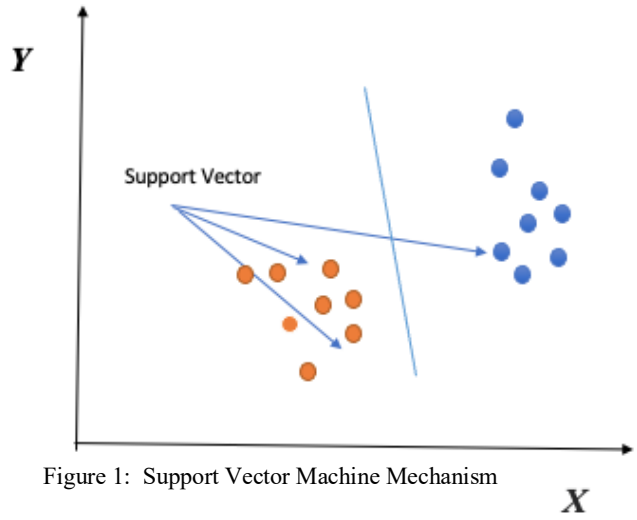

Figure 1: Support Vector Machine Mechanism

SVM (Support Vector Machine) is a supervised machine learning algorithm that can be used to solve classification and regression problems. It is, however, mostly used to solve classification problems. The value of each function is the value of a particular coordinate in this algorithm, which plots each data object as a point in n-dimensional space (where n is the number of features you have). Then we conduct classification by locating the hyper-plane that better separates the two classes [4].

Individual observation coordinates make up Support Vectors. The Classifier is a boundary that separates the two groups (hyper-plane/line) the most.

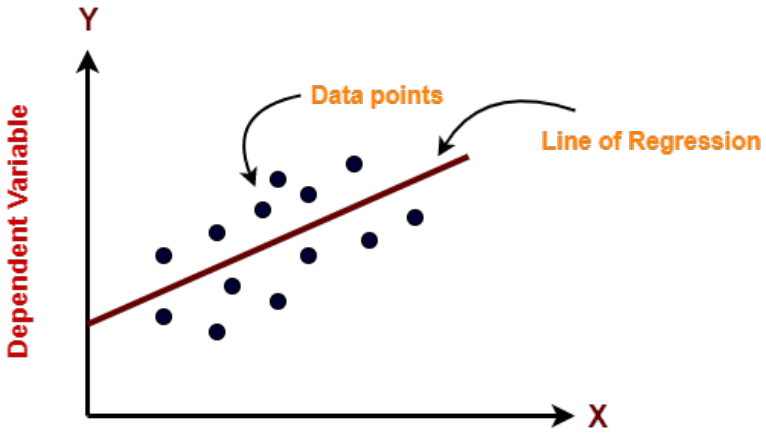

Figure 2: Linear Regression Mechanism

DOI:
https://doi.org/10.54216/AJBOR.050102
Received: April 22, 2021 Accepted: September 14, 2021

As stated earlier, linear regression is a predictive modeling technique for determining the association between the dependent and independent variables. The equation for determining the correlation between the variables is as follows:
Y = b0 + b1* x (Y is the dependent variable and x is the independent variable)
Here is a diagram that shows the linear relation between the certain dependent and the independent variable. The red line denotes the best fit line for the data points [2].
When we add decision-making computation to a database, linear regression represents machine learning. It is divided into two sections. To match the data, we need to approximate a component and use that approximation to guide automated decision-making. We use linear regression to estimate a database engine with a linear function that broadly summarizes a given data set. We've also discovered a particular sequence as a result of this. We have developed learning if we use the pattern to guide our choices[7].
There are two types of linear regression: simple linear regression and multiple linear regression. A variable quantity characterizes simple linear regression. And, as the name implies, multiple linear regression is characterized by a large number of independent variables (more than 1). You'll be able to change a polynomial or regression toward the mean when you're looking for the easiest match rows.

4. Implementation Details

4.1 About RapidMiner

RapidMiner is software that is used to analyze a dataset and extract meaningful analysis from it by applying various data processing models. It takes in data as input and provides numerous operators that can be applied to that dataset to get useful insights as output. It is a widely used tool for making predictions. Many notable companies are known to use this tool to create a prototype of their model and validating it [5].

4.2 Dataset

The dataset being used for this model has been downloaded from the Machine Learning Repository by the Center for Machine Learning and Intelligent Systems. The dataset falls under the category of Bank Marketing. The access link to the dataset is as follows: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing#

The bank-full.csv file has been used for this process. Diving a bit deep into the dataset, the data gives us information about the direct marketing campaign being carried by the Portuguese banking organization. The campaign was run by making phone calls to customers. There are certain variables taken into account such as age, balance, loan, etc to determine the effect on the customer decision and whether or not they subscribed to bank term deposit in the end. The final decision is also our dependent variable y against which the effect of various variables will be studied using the linear regression model and support vector machine method [5].
One thing to note here is, the selected data type is multivariate with a certain level of classification. The type of data inside is real with about 17 attributes within the datasheet. Out of these 17 attributes, we will use 8 for our models.

4.3 Step by Step Process

1. Visit https://archive.ics.uci.edu/ml/datasets.php

DOI:
https://doi.org/10.54216/AJBOR.050102
Received: April 22, 2021 Accepted: September 14, 2021

2. Scroll down to **Bank Marketing** and click on it.

3. Click on the **Data Folder** option.

4. Click on **bank.zip** file to download it. Extract the files.

5. Open RapidMiner.

6. Import **bank-full.csv** file from the extracted files. You will be able to see the datasheet upon successfully importing the file.

Row No.	age	job	marital	education	default	balance	housing	loan
1	58	management	married	tertiary	no	2143	yes	no
2	44	technician	single	secondary	no	29	yes	no
3	33	entrepreneur	married	secondary	no	2	yes	yes
4	47	blue-collar	married	unknown	no	1506	yes	no
5	33	unknown	single	unknown	no	1	no	no
6	35	management	married	tertiary	no	231	yes	no
7	28	management	single	tertiary	no	447	yes	yes
8	42	entrepreneur	divorced	tertiary	yes	2	yes	no
9	58	retired	married	primary	no	121	yes	no
10	43	technician	single	secondary	no	593	yes	no
11	41	admin.	divorced	secondary	no	270	yes	no
12	29	admin.	single	secondary	no	390	yes	no
13	53	technician	married	secondary	no	6	yes	no
14	58	technician	married	unknown	no	71	yes	no
15	57	services	married	secondary	no	162	yes	no
16	51	retired	married	primary	no	229	yes	no
17	45	admin.	single	unknown	no	13	yes	no
18	57	blue-collar	married	primary	no	52	yes	no
19	60	retired	married	primary	no	60	yes	no
20	33	services	married	secondary	no	0	yes	no
21	28	blue-collar	married	secondary	no	723	yes	yes

Figure 3: Importing the Dataset

7. Navigate to the **Design** tab.

8. Drag and drop the **bank-full** database into the design canvas.

9. Click on the search bar in the operator's section and type **Set Attributes**. Drag and drop this operator in to the canvas.

10. From **Select Attributes** section to the right, set the following values:

    a. **Attribute filter type**: subset

    b. **Attributes**: Select the following:

DOI:
https://doi.org/10.54216/AJBOR.050102
Received: April 22, 2021 Accepted: September 14, 2021

i. Age
ii. Balance
iii. Day
iv. Duration
v. Campaign
vi. Pday
vii. Previous
viii. Y

11. Since the y is a dependent variable, we need to define that. Click on the search bar in the operator's section and type **Set Role**. Drag and drop this operator in to the canvas.

12. From **Set Role** section to the right, set the following values:

    a. **Attribute name**: y

    b. **Target Role**: Label

13. Since the value of y is not numeric and we need to convert it to numeric, click on the search bar in the operator's section and type **Map**. Drag and drop this operator in to the canvas.

14. Define the following mapping:

    a. Replace yes with 1

    b. Replace no with 0

15. Now, we need to apply the regression model on our dataset. Click on the search bar in the operator's section and type **Linear Regression**. Drag and drop this operator in to the canvas.

16. To apply the model, click on the search bar in the operator's section and type **Apply Model**. Drag and drop this operator in to the canvas.

17. Connect all the operators with each other as shown in the figure below:

DOI:
https://doi.org/10.54216/AJBOR.050102
Received: April 22, 2021 Accepted: September 14, 2021

Figure 4: Applying Classification Models

18. For **Support Vector Machine**, SVM, Repeat the same steps from 1-14 as above. Click on the search bar in the operator's section and type **Support Vector Machine**. Drag and drop this operator in to the canvas.

19. Click on the play button in blue located at the top left of the screen.

20. RapidMiner will process the data the way the operators are defined and give out results if everything is defined right. The results for the Linear Regression Model are shared below:

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
age	0.001	0.000	0.025	1.000	5.794	0.000	****
balance	0.000	0.000	0.040	0.999	∞	0	****
duration	0.000	0.000	0.391	0.998	91.081	0	****
campaign	-0.003	0.000	-0.031	0.989	-7.191	0.000	****
pdays	0.000	0.000	0.076	0.989	15.825	0	****
previous	0.008	0.001	0.056	0.986	11.722	0	****
(Intercept)	-0.051	0.006	?	?	-8.742	0	****

Figure 5: Linear Regression Model Results

DOI:
https://doi.org/10.54216/AJBOR.050102
Received: April 22, 2021 Accepted: September 14, 2021

Similarly, the result for SVM is shared below:

5. Analysis of Results

From the results of the linear regression model, we observe the coefficients and conclude that age, campaign, and previous variables affect the dependent variable y. Balance, duration, and pdays have no impact as the coefficient is zero. Age and previous has a positive relationship with y and campaign has a negative relationship since the coefficient has a negative sign. Their relationship can be further tested using t-stat method. The overall graph has a decreasing slope since the y-intercept is negative, which is the dependent variable that indicates whether the client has subscribed to the term deposit. From the results of support vector machine analysis, the number of support vectors is 45211. The SVM model calculated for us the weights of each independent variable including age, balance, duration, day, campaign, pdays, and previous. Among all these, age, balance, pdays and previous have negative weights, and day, duration, and campaign have positive highest weight.

Figure 6: Support Vector Machine Model Results

6. Conclusion

To conclude, machine learning is the future since it helps in using the real data in a productive manner which in turn helps in automating the tasks with much precision and serves back humanity. To make the machine learn, the first step is to classify the data so that machine can categorize and match what kind of data is it dealing with by checking from the reference data. There are many classification methods. The method that one should go with depends more upon the use case that you need to deal with. This could be supervised or unsupervised learning and the use of the reinforcement method. Where linear regression helps

DOI:
https://doi.org/10.54216/AJBOR.050102
Received: April 22, 2021 Accepted: September 14, 2021

in understanding the relationship between the independent and dependent variables, SVMs are used with the purpose of classification and detecting the outliers within the dataset with much precision. Where linear regression cannot be considered for a nonlinear relationship, SVMs cannot be used with a huge number of data set. Each method has its strengths and caveats that act as a tradeoff. However, most of these methods are widely used for studying the datasets in a better way and to extract useful information from them.

References

[1] Sammut, C., & Webb, G. I. (Eds.). (2011). Encyclopedia of machine learning. Springer Science & Business Media.

[2] Bonaccorso, G. (2017). Machine learning algorithms. Packt Publishing Ltd, 24(2), 1065-1069.

[3] Noble, W. S. (2006). What is a support vector machine?. Nature Biotechnology, 24(12), 1565-1567.

[4] Mitchell, T. M. (1999). Machine learning and data mining. Communications of the ACM, 42(11), 30-36.

[5] Srinivasan, K., & Fisher, D. (1995). Machine learning approaches to estimating software development effort. IEEE Transactions on Software Engineering, 21(2), 126-137.

[6] M. S. Acharya, A. Armaan, and A. S. Antony, "A comparison of regression models for prediction of graduate admissions," in 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), 2019, pp. 1-5.

[7] Z. Zhang, Y. Li, L. Li, Z. Li, and S. Liu, "Multiple linear regression for high-efficiency video intra coding," in ICASSP 2019-2019 IEEE

[8] A. K. Prasad, M. Ahadi, B. S. Thakur, and S. Roy, "Accurate polynomial chaos expansion for variability analysis using optimal design of experiments," in 2015 IEEE MTT-S International Conference on Numerical Electromagnetic and Multiphysics Modeling and Optimization (NEMO), 2015, pp. 1-4

[9] J. Wolberg, Data analysis using the method of least squares: extracting the most information from experiments: Springer Science & Business Media, 2006.

[10] T. Bakibayev and A. Kulzhanova, "Common Movement Prediction using Polynomial Regression," in 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), 2018, pp. 1-4

DOI:
https://doi.org/10.54216/AJBOR.050102
Received: April 22, 2021 Accepted: September 14, 2021