# Prediction of prospect customers response “YES” or “NO” to open a term deposit account of Bank Marketing Campaign.

# Using Classification and Regression Machine Learning tools

# In Python.

**By: Marina Golberg 501072689**

[**marina.golberg@ryerson.ca**](mailto:marina.golberg@ryerson.ca)

**Ryerson University: The Chang School of Continuing Education**

**CIND820: Big Data Analytics Project**

**Date: 10/18/2021**

**Instructor: Dr. Tamer Abdou**

1. **Abstract:**

Businesses these days are using every available platform to do marketing of their company, products, and services. They use telemarketing, email marketing, advertisements, and many other different tools to gain sales and expand their business. Telemarketing is still the most used method of increasing sales of small to big companies. It is a cost-effective and flexible marketing strategy that offers a high level of personal contact.

Predictive analytics uses data models, statistics, and machine learning to predict future events.

It's a discipline that helps you to analyze your marketing campaigns, assess their efficiency, and see possible improvements to lead an increase in sales in future.

In this capstone project, the theme is Classification and Regression. The goal of these classification models is to help the company more reliably predict future customer subscription before it occurs to secure deposits more effectively and increase customer satisfaction by reducing undesirable advertisements for certain customers.

This project will be able to answer the following questions:

* Will prospective customers respond "yes" or "no" to term deposit subscription?
* What type of customers are more likely to subscribe to term deposit and which feature has higher influence?
* What is the best time of year for a marketing campaign?

**Data Set Information:**

The dataset is publicly available for research from the UCI Machine Learning repository. It is about a direct marketing campaign of a Portuguese banking institution, based on more than one phone calls to access if the bank term deposit would be ('yes') or not ('no') subscribed. The details are described in (Moro et al., 2011).

Data can be found here: <https://archive.ics.uci.edu/ml/datasets/bank+marketing>

**Code and documentation for this project on GitHub repository as following:** <https://github.com/marinagolberg/CIND820-MarGolb.git>

**The tools that will be used for this project are:**

Jupyter notebook for Python environment, Pandas to perform data manipulation and analysis, NumPy will be used to perform a wide variety of mathematical operations on arrays, Seaborn, Plotly and Matplotlib will be used for data visualization statistical graphing and plotting, Sklearn will be used for classification and regression, these library contains a lot of efficient tools for machine learning and statistical modeling.

**Following is the summary of technics that will be used in this research to answer the questions:**

* **Subscription of customers to a term deposit.**

To answer first question, all data will be splited to training and testing set. Categorical data encoded to numerical labels and SMOTE will be used as an oversampling technique to balance the data. Classifiers to predict the results of “yes” or “no” are Logistic Regression, Decision Tree Classifier, Random Forest. Then I will compare the scores of the algorithms to choose the best one.

* **Customer profiling, feature influence on subscription, and campaign profiling.**

To answer the second and third questions. Feature selection techniques of Filter, Wrapper and Embedded Methods will be used to check the contribution of each attribute on explaining the client's subscription then the technics will be compared and tested to choose the best combination of attributes. The correlation will help to see the relationship between the attributes. Exploratory data analysis will be used to visually see the relationship between variables and the dependent variable "y" to build a campaign profile and customer profile.

|  |  |  |
| --- | --- | --- |
| Will prospective customers respond "yes" or "no" to term deposit subscription? | What type of customers is more likely to subscribe to term deposit and which feature has higher influence? | What is the best time in a year for marketing campaign? |
| Logistic Regression | Exploratory Data Analysis for customer profiling | Exploratory Data Analysis for  campaign profiling |
| Decision Tree Classifier | Feature selection(Filter, Wrapper) for feature influence on subscription |  |
| Random Forest | Correlation for Numerical var. |  |
|  |  |  |
|  |  |  |

1. **Literature Review:**

This literature review will discuss techniques and results that were found in previous researches and what will be the best approach for further analysis.

**Predicting term deposit subscription from similar datasets**

The purpose of the first study was to find if the use of the Random Forest improves the performance of the Decision Tree for the bank customer marketing response prediction ([Olatunji](https://www-proquest-com.ezproxy.lib.ryerson.ca/indexinglinkhandler/sng/au/Apampa,+Olatunji/$N?accountid=13631),2016). Classification algorithms used for modelling were; Logistic Regression, Decision Tree, Naive Bayes and the Random Forest ensemble. These algorithms were applied to both the balanced and original bank data by ten-fold cross-validation method. Results derived from the experiment showed that the performance of the Random Forest improved when the data was balanced. The Decision Tree algorithm returned 76.6% area under Curve (AUC) and Classification Accuracy (CA) compared to Logistic Regression 75.7% and Naive Bayes 75.6%. The Random Forest had an AUC and CA value of 74.2%. There were no found improvements in Random Forest ([Olatunji](https://www-proquest-com.ezproxy.lib.ryerson.ca/indexinglinkhandler/sng/au/Apampa,+Olatunji/$N?accountid=13631" \o "Click to search for more items by this author), 2016). Therefore, second experiment was conducted and the results showed that the performance metrics of Random Forest increased with an increase of "n" to 200. ([Olatunji](https://www-proquest-com.ezproxy.lib.ryerson.ca/indexinglinkhandler/sng/au/Apampa,+Olatunji/$N?accountid=13631" \o "Click to search for more items by this author), 2016). The second study found that changing the number of trees has no significant effects on mean accuracy of the Random forest. Random Forest and k-Nearest Neighbor are proved to be the best classifiers for any type of dataset. (Singh et al., 2017).

A time-ordered split was performed, where the data was divided into four years of training (May 2008 to June 2012,) and one year of testing (July 2012 to June 2013). The authors' decision was to merge the two data sources that led to a large data set of 150 attributes, which could be potentially useful features (Moro et al., [2014](https://onlinelibrary-wiley-com.ezproxy.lib.ryerson.ca/doi/full/10.1111/exsy.12253#exsy12253-bib-0030)). Using a semi-automated feature selection procedure within the modeling stage, researchers selected a reduced set of 22 relevant features. The dataset was unbalanced. Four DM learning techniques were explored: logistic regression, decision trees, support vector machines (SVMs), and a neural network. The best result was achieved by the neural network AUC = 0.8 and ALIFT = 0.7(Moro et al., [2014](https://onlinelibrary-wiley-com.ezproxy.lib.ryerson.ca/doi/full/10.1111/exsy.12253#exsy12253-bib-0030)). Such a model was then improved by including customer lifetime‐value‐related features, increasing the performance to 83% of subscribers with the half better classified contacts (Moro et al., 2015b).

a bank telemarketing dataset was enriched with social and

economic context features, leading to a tuned model that enabled to reach 79% of the deposit subscribers by selecting the half better classified

clients

a bank telemarketing dataset was enriched with social and

economic context features, leading to a tuned model that enabled to reach 79% of the deposit subscribers by selecting the half better classified

clients Portuguese retail bank

The goal of Ghatasheh et al.’s research is to improve the performance of predicting the readiness of customers to subscribe for a term deposit in a highly imbalanced dataset. It proposes improved Artificial Neural Network models (i.e., cost-sensitive) to facilitate the dramatic influence of highly imbalanced data, without changing the original data samples. Authors created models that was compared to different machine-learning models evaluated and validated. Telemarketing dataset from a Portuguese bank is used. The model achieved the greatest prediction of 79% geometric mean and minimized misclassification errors to 0.192 Type I, and 0.229 of Type II. In conclusion, the Meta-Cost method improved the performance of the prediction model without imposing important processing overhead or changing original data samples (Ghatasheh et al., [2020](https://onlinelibrary-wiley-com.ezproxy.lib.ryerson.ca/doi/full/10.1111/exsy.12253#exsy12253-bib-0030)). Experiment proved that MetaCost reduces cost by large amounts compared to error-based classification according to Domingos (1999).

The authors collected two variants of the bank marketing data set to predict whether a client subscribes to a term deposit in Portuguese financial institution (Krishna et al., 2019). The two experimental results show that the Deep Neural Network classifier has outperformed four existing classifiers. On the first data set after preprocessing, researchers applied feature selection using attribute subset selection (that include Forward Selection, Backward Elimination, Decision Tree), a method based on information gain parameter. The top 10 features have been chosen. Results are as follows: Decision Tree 88.99%, Naïve Bayes 86.64%, Support Vector Machines 89.82%, K-NN 88.65%, Deep Neural Network 91.15%. On the second data set after preprocessing, researchers applied feature extraction using Principal component analysis method based on a cumulative variance parameter. The top 3 principal have been chosen. Results are as follows: Decision Tree 83.15%, Naïve Bayes 84.16%, Support Vector Machines 87.95%, K-NN 86.76% (Krishna et al., 2019).

**Feature contribution on the subscription success, campaign and customer profiling.**

[Olatunji](https://www-proquest-com.ezproxy.lib.ryerson.ca/indexinglinkhandler/sng/au/Apampa,+Olatunji/$N?accountid=13631), divided attributes to two groups, numerical and categorical, then he tested their contribution to success of deposit subscription. By running numerical features on Decision Tree (DT), [Olatunji](https://www-proquest-com.ezproxy.lib.ryerson.ca/indexinglinkhandler/sng/au/Apampa,+Olatunji/$N?accountid=13631) found that feature duration is the root node that is the most important variable. The next feature selected by DT were poutcome, then month and contact. Similar results were obtained for the Logistic Regression. The author also demonstrated categorical features contribution by conducting Correspondence analysis. Results showed that customers in the management and technician cadre responded positively to a term deposit. Divorced and single clients responded more positively to the campaign than married clients. Clients with post-secondary education have a better subscription percentage than customers with elementary education. The months of September, November, March, and April were found to have higher subscription rates than other months. Finally, clients without bank loans were more likely to correspond with the "yes” response than those with bank loans ([Olatunji](https://www-proquest-com.ezproxy.lib.ryerson.ca/indexinglinkhandler/sng/au/Apampa,+Olatunji/$N?accountid=13631" \o "Click to search for more items by this author), 2016).

Junfeng et al.’s constructed a Decision Tree algorithm to analyze factors that affect customer subscription to fixed deposits of an Australian bank. The data set was similar to Portuguese banks. The attributes were selected by the highest information gain Entropy. The author found three factors that significantly affect customers' subscriptions are the number of employees, duration, and month, which greatly reduce the range of clients that banks push to subscribe for long-term deposits, and are beneficial to improving the efficiency of banks (Junfeng et al., 2019).

Parlar et al. used two feature selection methods, information gain and Chi-square, to select the important features. The results were compared with Naive Bayes's supervised machine learning algorithm. This study found that a reduced set of features improves classification performance. The ten highest-ranked features are duration, poutcome, month, pdays, contact, previous, age, job, housing, and balance (Parlar et al., 2017).

Most machine learning algorithms designed for classification assume that there is an equal number of examples for each observed class. This is not always the case in practice, and datasets that have a skewed class distribution are referred to as imbalanced classification problems. Most algorithms are overloaded by the majority class at a time they are learning from highly unbalanced data, so the false-negative (FN) measure is always high. In the past, most researchers have introduced many methods to deal with unbalanced data, most of them focus on resampling techniques, and another one was cost-sensitive learning (CSL). [Thai-Nghe](https://ieeexplore.ieee.org/author/38277551900) et al. demonstrated in their study that one of the technics improve the classifier performance, and another one reduce the misclassification costs ([Thai-Nghe](https://ieeexplore.ieee.org/author/38277551900) et al., 2010). The studies mentioned above showed better performance results when the different classification algorithms were optimized. The most common performance evaluation metric is the AUC, but some authors indicated classification error rates. Random Forest ensemble, k-Nearest Neighbor (kNN), and Neural Network is known to improve the classification accuracy and considered as a good classifiers for any type of dataset. Authors found that a reduced set of features improves classification performance and defines importance of attributes that explains the customer behavior on subscriptions for long-term deposits.

**Tools and Technics**

Data mining refers to extracting or mining knowledge from large amounts of data, which is stored in various repositories. One of the popular tasks of data mining is Classification, which assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data (Krishna et al., 2019). Introduction to data mining was reviewed to obtain fundamental concepts and background for understanding each data mining technique, followed by more advanced concepts and algorithms (Steinbach et al., 2005).

Silipo et al. explores some of the most commonly used techniques for dimensionality reduction, for example removing data columns with too many missing values, removing low variance columns, reducing highly correlated columns, applying Principal Component Analysis (PCA), investigating Random Forests, Backward Feature Elimination, and Forward Feature Construction.

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering, and dimensionality reduction via a consistence interface in Python. The Python programming language is one of the most popular languages for scientific computing. Documentation of Scikit-learn provides a ∼300 page user guide including narrative documentation, class references, a tutorial, installation instructions, as well as more than 60 examples, some featuring real-world applications. It has a wide assortment of well-established algorithms, with integrated graphics. It's relatively easy to install, learn, and use (<https://scikit-learn.org/stable/>).

[Pandas](https://pandas.pydata.org/docs/#module-pandas) is an open source, library providing high-performance, easy-to-use data structures and data analysis tools for the [Python](https://www.python.org/) programming language. It offers data structures and operations for manipulating numerical tables and time series (<https://pandas.pydata.org/docs/>).

The NumPy library contains multidimensional array and matrix data structures. NumPy can be used to perform a wide variety of mathematical operations on arrays. It adds powerful data structures to Python that guarantee efficient calculations with arrays and matrices and it supplies an enormous library of high-level mathematical functions that operate on these arrays and matrices (<https://numpy.org/doc/>).

Data mining is a useful tool, an approach that combines exploration and discovery with confirmatory analysis. I will use Data mining tools for the exploration and analysis of medium data set in order do to answer research questions. Python is free and available software that will be used for implementing each step of modeling. In order to implement each step in Python, different packages were downloaded and installed as below.

**Methodology**

To answer the given research question, methods on Figure 2.0 would be most appropriate to find an answer. Literature on related topics suggests that those methods are most appropriate.

Figure 2.0: Overall Research Procedure

Bank Marketing Data Set

USI Data Repository

Data Exploration and Visualization

Categorical Variables encoding

Splitting data

To Testing 30%

Splitting data

To Training 70%

Data Cleaning

Class Balance (Oversampling SMOTE)

Feature Selection, Extraction

Feature Training (Filter (Mutual Information), Wrapper (Forward Selection), Embedded methods (LASSO Regularization)

Feature Testing

Algorithms Evaluation ROC\_AUC, Precision, Recall, F1, PR\_AUC, Brier Score

Classification Algorithms LR, DT, RF (12 times will run)

Classification Results

I will be running for each Feature Selection technic RF, DT and LR and separately running all Features without Feature selection (12 times RF, DT and LR will run) than I will built 2X2 confusion matrix and calculate the rates as follows ROC\_AUC, Precision, Recall, F1, PR\_AUC, Brier Score for comparition of the models results and the best combination will be chosen..

# Data Description

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. The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y). Data can be found here: <https://archive.ics.uci.edu/ml/datasets/bank+marketing>

Data has 41,188 rows and 21 attributes with highly unbalanced class label.

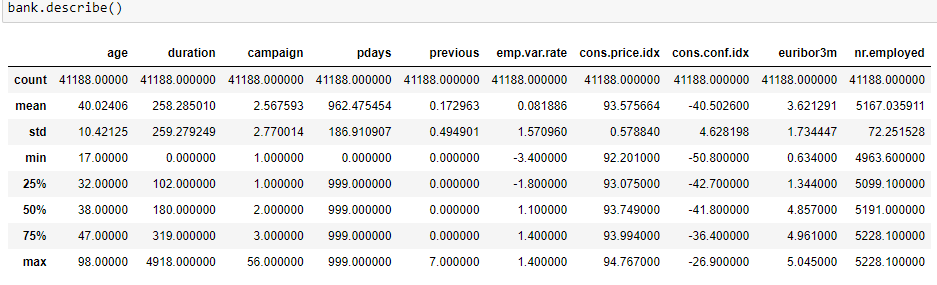
The positive class “Yes” of target variable “Y” is 4,640 observations that is 11.3%

The negative class “No” of target variable “Y” is 36,548 observations that is 89.7%

Input variables (independent variables):  
# bank client data:  
1 - age (numeric)  
2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')  
3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)  
4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')  
5 - default: has credit in default? (categorical: 'no','yes','unknown')  
6 - housing: has housing loan? (categorical: 'no','yes','unknown')  
7 - loan: has personal loan? (categorical: 'no','yes','unknown')  
# related with the last contact of the current campaign:  
8 - contact: contact communication type (categorical: 'cellular','telephone')  
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')  
10 - day\_of\_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')  
11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.  
# 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')  
# 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)  
Output variable (dependent variable):  
21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

Below you can see descriptive statistics for numerical variables.

Figure 1.0: Descriptive statistics



SD bigger then mean (duration, campaign, previous, emp.var.rate, cons.conf.idx) - high variation between values, and abnormal distribution for data. A smaller standard deviation indicates that more of the data is clustered about the mean while a larger once indicates the data are more spread out. I will be changing the shape of the distribution to Normal with Power Transformer.

## **What Is a Term Deposit?**

A term deposit is a fixed-term investment that includes the [deposit of money](https://www.investopedia.com/terms/d/deposit.asp) into an account at a financial institution. Term deposit investments usually carry short-term maturities ranging from one month to a few years and will have varying levels of required minimum deposits.

The investor must understand when buying a term deposit that they can withdraw their funds only after the term ends. In some cases, the account holder may allow the investor early termination—or withdrawal—if they give several days notification. Also, there will be a penalty assessed for early termination (<https://www.investopedia.com/terms/t/termdeposit.asp>).

# Reference for Literature review

# [Apampa, Olatunji](https://www-proquest-com.ezproxy.lib.ryerson.ca/indexinglinkhandler/sng/au/Apampa,+Olatunji/$N?accountid=13631), (2016). Evaluation of Classification and Ensemble Algorithms for Bank Customer Marketing Response Prediction. [Journal of International Technology and Information Management](https://www-proquest-com.ezproxy.lib.ryerson.ca/pubidlinkhandler/sng/pubtitle/Journal+of+International+Technology+and+Information+Management/$N/51771/DocView/1926957604/fulltext/B0B6C76C55724573PQ/1?accountid=13631); San Bernadino[Vol. 25, Iss. 4,](https://www-proquest-com.ezproxy.lib.ryerson.ca/indexingvolumeissuelinkhandler/51771/Journal+of+International+Technology+and+Information+Management/02016Y10Y01$232016$3b++Vol.+25+$284$29/25/4?accountid=13631" \o "Click to search for more items from this issue) (2016): 85-100. <https://www.proquest.com/docview/1926957604/fulltext/784C53BAAFA14040PQ/1?accountid=13631>

# [Sérgio Moro,](https://www-sciencedirect-com.ezproxy.lib.ryerson.ca/science/article/pii/S016792361400061X" \l "!) [Paulo Cortez](https://www-sciencedirect-com.ezproxy.lib.ryerson.ca/science/article/pii/S016792361400061X#!) , [Paulo Rita](https://www-sciencedirect-com.ezproxy.lib.ryerson.ca/science/article/pii/S016792361400061X#!) (2014). A data-driven approach to predict the success of bank telemarketing.

https://www-sciencedirect- com.ezproxy.lib.ryerson.ca/science/article/pii/S016792361400061X

1. Moro, S., Cortez, P., & Rita, P. (2015b). Using customer lifetime value and neural networks to improve the prediction of bank deposit subscription in telemarketing campaigns. Neural Computing and Applications, 26(1), 131–139.
2. Nazeeh Ghatasheh, Hossam Faris, Ismail AlTaharwa , Yousra Harb and Ayman Harb . Business Analytics in Telemarketing: Cost-Sensitive Analysis of Bank Campaigns Using Artificial Neural Networks. Appl. Sci. 2020, 10, 2581; doi:10.3390/app10072581

[Business Analytics in Telemarketing Cost-Sensitive Analysis of Bank Campaigns Using Artificial Neural Networks.pdf](file:///C:\RYERSON\CIND820Project\Literature\Articles\Articles%20that%20was%20summurzed%20for%20project\Business%20Analytics%20in%20Telemarketing%20Cost-Sensitive%20Analysis%20of%20Bank%20Campaigns%20Using%20Artificial%20Neural%20Networks.pdf)

1. Pedro Domingos (1999). Meta cost a general method for making classifiers cost –sensitive. Artificial Intelligence Group Instituto Superior Tecnico Lisbon 1049-001, Portugal.

<https://homes.cs.washington.edu/~pedrod/papers/kdd99.pdf>

# [Nguyen Thai-Nghe](https://ieeexplore-ieee-org.ezproxy.lib.ryerson.ca/author/38277551900), [Zeno Gantner](https://ieeexplore-ieee-org.ezproxy.lib.ryerson.ca/author/38273889600), [Lars Schmidt-Thieme](https://ieeexplore-ieee-org.ezproxy.lib.ryerson.ca/author/38276650400) (2010). Cost-sensitive learning methods for imbalanced data Member, IEEE <https://ieeexplore-ieee-org.ezproxy.lib.ryerson.ca/stamp/stamp.jsp?tp=&arnumber=5596486>

1. Asmita Singh, Malka N. Halgamuge, Rajasekaran Lakshmiganthan (2017). Impact of Different Data Types on Classifier Performance of Random Forest, Naïve Bayes, and K-Nearest Neighbors Algorithms. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 12, 2017 <https://minerva-access.unimelb.edu.au/bitstream/handle/11343/216910/2017_Asmita_Different_Data.pdf>
2. Tuba Parlar , Songul Kakilli Acaravc (2017). Using Data Mining Techniques for Detecting the Important Features of the Bank Direct Marketing Data**.** [**International Journal of Economics and Financial Issues**](https://www.proquest.com/pubidlinkhandler/sng/pubtitle/International+Journal+of+Economics+and+Financial+Issues/$N/816338/PagePdf/2270076746/fulltextPDF/B53E2FA044422APQ/1?accountid=13631)**; Mersin**[Vol. 7, Iss. 2,](https://www.proquest.com/indexingvolumeissuelinkhandler/816338/International+Journal+of+Economics+and+Financial+Issues/02017Y01Y01$232017$3b++Vol.+7+$282$29/7/2?accountid=13631) (2017): 692-696.

<https://www.proquest.com/docview/2270076746?pq-origsite=summon&accountid=13631>

1. Guo Junfeng, Hou Handan (2019). Statistical Decision Research of long-term deposit subscription in banks based on Decision Tree. 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS). [IEEE Xplore Full-Text PDF: (ryerson.ca)](https://ieeexplore-ieee-org.ezproxy.lib.ryerson.ca/stamp/stamp.jsp?tp=&arnumber=8669650)
2. Chittem Leela Krishna, Poli Venkata Subba Reddy (2019). Deep Neural Networks for the Classification of Bank Marketing Data using Data Reduction Techniques. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-3, September 2019. <https://www.researchgate.net/profile/Leela-Krishna-Chittem/publication/336287356_Deep_Neural_Networks_for_the_Classification_of_Bank_Marketing_Data_using_Data_Reduction_Techniques/links/5d9a1d5da6fdccfd0e7eeeb1/Deep-Neural-Networks-for-the-Classification-of-Bank-Marketing-Data-using-Data-Reduction-Techniques.pdf>
3. Michael Steinbach, Pang-Ning Tan, and Vipin Kumar (2005). Introduction to Data Mining. 1-769.
4. Rosaria Silipo, Iris Adae, Aaron Hart, Michael Berthold (2014). Seven Techniques for Dimensionality Reduction, KNIME.com [knime\_seventechniquesdatadimreduction.pdf](https://www.knime.com/sites/default/files/inline-images/knime_seventechniquesdatadimreduction.pdf)
5. Alexandre Gramfort, Vincent Michel….. Matthieu Perrot Edouard Duchesnay(2011). Scikit-learn: Machine Learning in Python, by Fabian Pedregosa , Gael Varoquaux . Journal of Machine Learning Research 12 (2011) 2825-2830

<https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf>

Source code, binaries, and documentation can be downloaded from: https://scikit- learn.org/stable/

1. Pandas documentation https://pandas.pydata.org/docs/

### NumPy Documentation <https://numpy.org/doc/>