Bank Marketing Analysis Approach Using Logistic Regression of Machine Learning

Kalp Pawar, Nitesh Yadav, Nandita Mendhe, Sivani Koduri,

Nidhi Lal

Dept. of Computer Science and Engineering
IIIT Nagpur, India

kalp.pawar@cse.iiitn.ac.in, nitesh.yadav@cse.iiitn.ac.in
nandita.mendhe@cse.iiitn.ac.in, sivanikoduri17@gmail.com
nidhi.2592@gmail.com

Abstract— The banking institutions have marketing departments with more staff than most banks and credit unions have total employees. And yet institutions of all sizes tend to face marketing challenges due to improper insights. This proposed technique facilitates useful data explications for the banking sector to avoid customer attenuation. Customer relations is one of the most important factors that are to be analyzed in today's competitive business environment. This paper analyzes the machine learning techniques like Logistic Regression, Random Forest classifiers and its applications in the marketing sector.

Keywords— **Logistic Regression, Random Forest, Bank Marketing, Grid Search, Classification.**

I. Introduction

Recently, machine learning has attracted considerable attention in the real world because it can provide cutting edge analysis for marketing [1] which is very beneficial for the business to customer marketing. Machine learning works by creating models that are trained on data through various machine learning techniques such as logistic regression, etc. These trained models then used to serve the purpose of predicting the result. In many studies, the impact of marketing activities on relationship quality has played a vital role [2,4]. Data mining techniques have a tremendous amount of applications in the banking sector [5]. These could be very helpful though data mining has certain drawbacks and issues because data mining is not an easy task it has a problem of performance and requires parallel and distributed algorithms and handling of relational and complex types of data. We use machine learning for bank marketing and analysis. We trained a machine learning model on the dataset [6] to predict the result. We use logistic regression technique of machine learning to predict the bank marketing to give good marketing insights to marketers.

II. Related Work

Several studies have been delved into by many researchers in the phenomenon of Bank Direct Marketing with various marketing techniques existing nowadays. Decision Support Systems [9-10]. The Scikit-learn [8] is a Python framework comprising of various machine learning algorithms for both supervised and unsupervised learning. This Scikit-learn package is focussing to bring machine learning to beginners of various disciplines by using a general-purpose high-level language i.e python.

There have been many techniques used for statistical modeling such as Logistic regression[7] which is used to model the probability of a certain class or event existing such as pass or fail, win or lose, alive or dead. This can be further utilized to model several classes for experiments such as predicting whether an image containing a particular object. Every object which is being detected in the image would then be assigned a probability on a scale from 0 to 1. Logistic regression is used in multidimensional fields in the healthcare domain it is used for drug testing, cancer detection, etc. In stock prediction for a particular stock to buy or sell, etc. In malware detection and many more applications ahead.

III. Technologies Used

Machine learning techniques aim to automatically learn and recognize patterns from large datasets. There is a great variety of machine learning techniques available within the literature which makes the classification more and more difficult. This paper divides the literature into an artificial neural network (ANN) based and optimization-based techniques.

Table 1 shows that variations of ANNs and hybrid systems are very popular in the literature. There is a clear trend to use established ANN models and enhance them with new training algorithms and combine ANNs with emerging technologies into complex hybrid systems[3].

Technology	Number	Publications
Artificial Neural Network based	21	[10], [11], [12], [13], [14], [15], [16],[17],[18],[19], [20],[21],[22],[23], [24],[25],[26],[27], [28],[29],[30]
Optimisation & Evolutionary techniques	4	[31],[32],[33],[34]
Hybrid or Multiple	15	[35],[36],[37],[38], [39],[40],[41],[42], [43],[44],[45],[46], [47],[48],[49]

Table 1: Reviewed papers classified by machine learning technique

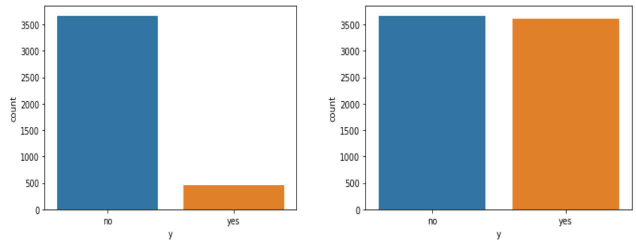
IV. Proposed Work

A. Data Preparation

The given raw data[6] is related to the direct marketing campaigns of a Portuguese banking institution. The classification goal is to predict if the client will subscribe to a term deposit.

1. First step: Making the Data Symmetric

- The final output can either be ‘yes’ or ‘no’. In the given dataset 87% of tuples belonging to the class ‘no’. Therefore the data is highly Skewed.
- To make the data symmetric, we took the tuples belonging to minority class(i.e. Class ‘yes’) and duplicated them until both the classes have an equal number of tuples in them.



Raw data Processed data

Fig 1: Graph of Balancing data

2. Second step: Splitting ‘pdays’ feature into 2 features

- The entries of ‘pdays’ feature are as follows: ○ More than 90% of the entries are 999. (i.e. Client was never contacted)
  ○ The remaining entries are in the range 0 - 20.
- Because of this, when we standardize the data, the entries corresponding to 999 will become 1 and remaining entries will be very close to zero(they all will be almost the same).
- To avoid this we split the feature into 2 features where one feature contains information about whether the client was contacted before or not(binary feature), and another feature will contain information about how long ago the client was contacted(if the client was never contacted we put 30 in this field instead of 999).

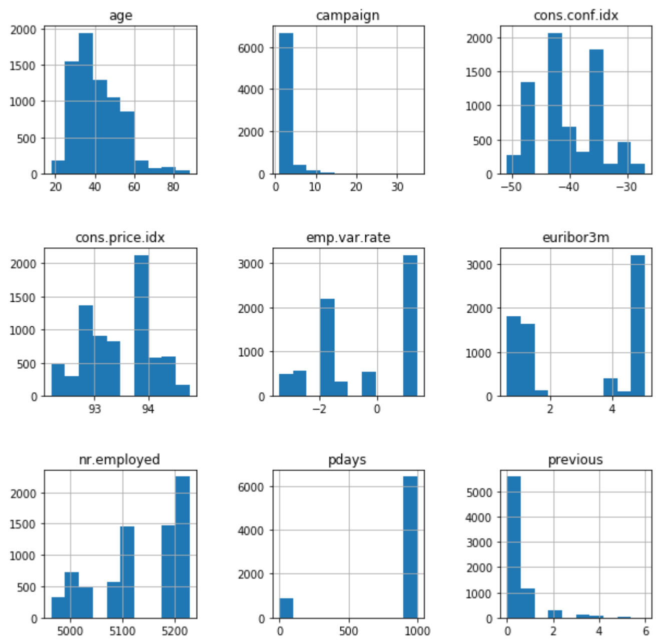


Fig 2: Plots of the finalized dataset
Count vs Input variables

3. Third step: Standardizing the data

- This step was done :
  - For faster convergence.
  - To ensure that variables measured at different scales do not contribute differently for analysis.

B. Predicting the outcome

After finalizing the data the model is trained using the Logistic Regression Model.

1. Logistic Regression Models:
   The central mathematical concept that underlies logistic regression is the logit—the natural logarithm of an odds ratio.

2. Input Variables:
   Selecting the right input variables is very important for machine learning techniques. Even the best machine learning technique can only learn from the input if there is actually some kind of correlation between input and output variables.

input variables used for prediction -

[

'age', 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'day_of_week', 'campaign', 'pdays', 'previous','p outcome', 'emp.var.rate', 'cons.price.idx', 'cons.conf.idx', 'euribor3m', 'nr.employed'

]

C. Classification :

We tried out the following classifiers:

- Logistic Regression(with Linear and Polynomial Features)

- Random Forest Classifier Here is a table comparing the accuracy, precision, recall, and f1_score of all the models.

V. Result and Discussion

The performance of the proposed technique named logistic regression is evaluated in terms of accuracy, precision, F1 score and recall. The results are as shown in Tables 2 and 3.And comparison shown in fig 4.

The Random Forest classifier provides the best accuracy among all the classifiers used with an accuracy of 0.963353.

Metric	Logistic Regression with Linear Features	Logistic Regression with Features of Degree=2	Logistic Regression with Features of Degree=3	Logistic Regression with Features of Degree=3 after applying Grid Search
Accuracy	0.731104	0.81585	0.931287	0.93586
Precision	0.617375	0.829945	0.99353	1.00000
Recall	0.794293	0.804659	0.882594	0.88543
F1 Score	0.694748	0.817106	0.934783	0.93923

Table 2: Comparison between logistic regression degree 1,2 and 3

Metric	Logistic Regression with Features of Degree=3 after applying Grid Search	Random Forest
Accuracy	0.93586	0.963353
Precision	1.00000	1.00000
Recall	0.88543	0.931153
F1 Score	0.93923	0.964349

Table 3: Comparison between logistic regression degree 3 and random forest classifier

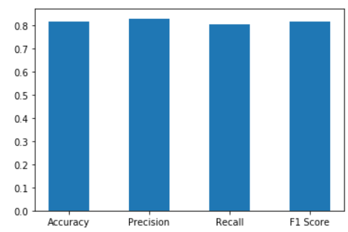


Fig 4 :(i) Logistic Regression degree=1

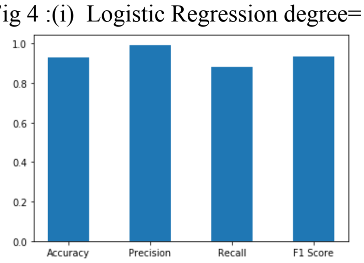


Fig 4 :(ii) Logistic Regression degree=2

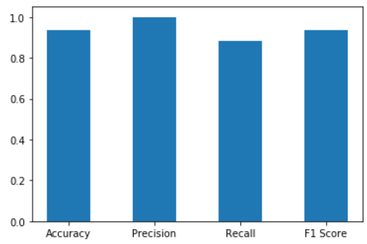


Fig 4 :(iii) Logistic Regression degree=3

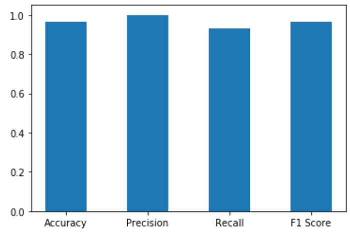


Fig 4 :(iv) Logistic Regression using grid search

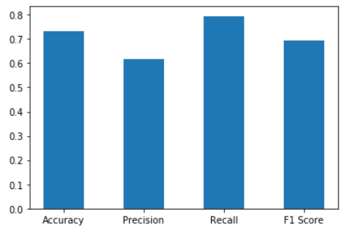


Fig 4 :(v) Random forest classifier

VI. Conclusions

Machine Learning has attracted increasing attention from the viewpoint of providing useful insights, especially for the banking sector. The use of machine learning techniques for analyzing bank marketing has limitations. To solve these limitations, we used the dataset[6] by using a content-centric concept. Random forest classifier based machine learning technique is suited for bank marketing. As the accuracy of the machine learning model is important. It has an advantage that manages computing resources because the learning function of the random forest provides faster and better results. The proposed technique meets the requirements of the machine learning model with the support of banking data. In this term, it achieves high accuracy, scalability, reliability and efficient to deal with any kind of customers.

VII. References

[1] Sundsøy, Pål, Johannes Bjelland, Asif M. Iqbal, and Yves-Alexandre de Montjoye. "Big data-driven marketing: how machine learning outperforms marketers' gut-feeling." In International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, pp. 367-374. Springer, Cham, 2014.

[2] Al-Alak, Basheer A. "Impact of marketing activities on relationship quality in the Malaysian banking sector." Journal of Retailing and Consumer Services 21, no. 3 (2014): 347-356.

[3] BjoernKrollner,BruceVanstone,GavinFinney. ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

[4] Huang, Zan, Hsinchun Chen, Chia-Jung Hsu, Wun-Hwa Chen, and Soushan Wu. "Credit rating analysis with support vector machines and neural networks: a market comparative study." Decision support systems 37, no. 4 (2004): 543-558.

[5] Chitra, K., and B. Subashini. "Data mining techniques and its applications in the banking sector." International Journal of Emerging Technology and Advanced Engineering 3, no. 8 (2013): 219-226..

[6] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[7] Shashidhara, Bhuvan M., Siddharth Jain, Vinay D. Rao, Nagamma Patil, and G. S. Raghavendra. "Evaluation of machine learning frameworks on bank marketing and Higgs datasets." In 2015 Second International Conference on Advances in Computing and Communication Engineering, pp. 551-555. IEEE, 2015.

[8] Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel et al. "Scikit-learn: Machine learning in Python." Journal of machine learning research 12, no. Oct (2011): 2825-2830.

[9] Moro, Sergio, Raul Laureano, and Paulo Cortez. "Using data mining for bank direct marketing: An application of the crisp-dm methodology." In *Proceedings of European Simulation and Modelling Conference-ESM'2011*, pp. 117-121. EUROSIS-ETI, 2011.

[10] Moro, Sérgio, Paulo Cortez, and Paulo Rita. "A data-driven approach to predict the success of bank telemarketing." *Decision Support Systems* 62 (2014): 22-31.

[11] Abraham, A., Nath, B. & Mahanti, P. K. (2001), Hybrid intelligent systems for stock market analysis, in 'Proceedings of the International Conference on Computational Science-Part II', Springer-Verlag, London, UK, pp. 337–345.

[12] Bekiros, S. D. & Georgoutsos, D. A. (2008), 'Direction-of-change forecasting using a volatility based recurrent neural network', Journal of Forecasting 27(5), 407–417.

[13] Chen, A.-S., Leung, M. T. & Daouk, H. (2003), 'Application of neural networks to an emerging financial market: forecasting and trading the Taiwan stock index', Comput. Oper. Res. 30(6), 901– 923.

[14] Chen, Y., Dong, X. & Zhao, Y. (2005), 'Stock index modeling using eda based local linear wavelet neural network', International Conference on Neural Networks and Brain 3, 1646–1650.

[15] de Faria, E., Albuquerque, M. P., Gonzalez, J., Cavalcante, J. & Albuquerque, M. P. (2009), 'Predicting the brazilian stock market through neural networks and adaptive exponential smoothing methods', Expert Systems with Applications

[16] Hamid, S. A. & Iqbal, Z. (2004), 'Using neural networks for forecasting volatility of s&p 500 index futures prices', Journal of Business Research 57(10), 1116–1125.

[17] Hanias, M., Curtis, P. & Thalassinos, J. (2007), 'Prediction with neural networks: The Athens stock exchange price indicator', European Journal of Economics, Finance and Administrative Sciences 9, 21–27.

[18] Jaruszewicz, M. & Mandziuk, J. (2004), 'One day prediction of nikkei index considering information from other stock markets', International Conference on Artificial Intelligence and Soft Computing 3070, 1130–1135.

[19] Lee, T.-S. & Chen, N.-J. (2002), 'Investigating the information content of non-cash-trading index futures using neural networks', Expert Systems with Applications 22(3), 225–234.

[20] [19] Leigh, W., Hightower, R. & Modani, N. (2005), 'Forecasting the new york stock exchange composite index with past price and the interest rate on condition of volume spike', Expert Systems with Applications 28(1), 1–8.

[20] Liao, Z. & Wang, J. (2009), 'Forecasting model of global stock index by stochastic time effective neural network', Expert Systems with Applications

[21] Ning, B., Wu, J., Peng, H. & Zhao, J. (2009), 'Using the chaotic neural network to forecast stock index', Advances in Neural Networks 5551, 870–876.

[22] Pan, H., Tilakaratne, C. & Yearwood, J. (2005), 'Predicting the Australian stock market index using neural networks exploiting dynamical swings and Intermarket influences', Journal of research and practice in information technology 37(1), 43–55.

[23] Roh, T. H. (2007), 'Forecasting the volatility of stock price index', Expert Systems with Applications 33(4), 916–922.

[24] Shen, J., Fan, H. & Chang, S. (2007), 'Stock index prediction based on adaptive training and pruning algorithm', Advances in Neural Networks 4492, 457–464.

[25] Slim, C. (2004), 'Forecasting the volatility of stock index returns: A stochastic neural network approach', Computational Science and Its Applications 3045, 935–944.

[26] Stansell, S. R. & Eakins, S. G. (2004), 'Forecasting the direction of change in sector stock indexes: An application of neural networks.', Journal of Asset Management 5(1), 37–48.

[27] Thawornwong, S. & Enke, D. (2004), 'The adaptive selection of financial and economic variables for use with artificial neural networks', Neurocomputing 56, 205–232.

[28] Witkowsk , D. & Marcinkiewicz, E. (2005), 'Construction and evaluation of trading systems: Warsaw index futures', International Advances in Economic Research 11(1), 83–92.

[29] Zapranis, A. (2006), 'Testing the random walk hypothesis with neural networks ', Artificial Neural Networks 4132,664–671.

[30] Zhu, X.,Wang, H., Xu, L. & Li, H. (2008), 'Predicting stock index increments by neural networks: The role of trading volume under different horizons', Expert Syst. Appl. 34(4), 3043–3054.

[31] Kim, M.-J., Min, S.-H. & Han, I. (2006), 'An evolutionary approach to the combination of multiple classifiers to predict a stock price index', Expert Systems with Applications 31(2), 241–247.

[32] Majhi, R., Panda, G., Majhi, B. & Sahoo, G. (2009), 'Efficient prediction of stock market indices using adaptive bacterial foraging optimization (abfo) and bfo based techniques', Expert Systems with Applications 36(6), 10097–10104.

[33] Majhi, R., Panda, G., Majhi, B. & Sahoo, G. (2009), 'Efficient prediction of stock market indices using adaptive bacterial foraging optimization (abfo) and bfo based techniques', Expert Systems with Applications 36(6), 10097–10104.

[34] Zhang, X., Chen, Y. & Yang, J. Y. (2007), Stock index forecasting using pso based selective neural network ensemble, in 'International Conference on Artificial Intelligence', pp. 260–264.

[35] Abraham, A., Philip, N. S. & Saratchandran, P. (2003), 'Modeling chaotic behavior of stock indices using intelligent paradigms', Neural, Parallel Sci. Comput. 11(1 & 2), 143–160.

[36] Armano, G., Marchesi, M. & Murru, A. (2005), 'A hybrid genetic-neural architecture for stock indexes forecasting', Information Sciences 170(1), 3–33.

[37] Chen, Q.-A. & Li, C.-D. (2006), 'Comparison of forecasting performance of ar, star and ann models on the chinese stock market index', Advances in Neural Networks 3973, 464–470.

[38] Chen, Y., Abraham, A., Yang, J. & Yang, B. (2005), Hybrid methods for stock index modeling, in 'International Conference on Fuzzy Systems and Knowledge Discovery', Springer Verlag, pp. 1067–1070.

[39] Chun, S.-H. & Kim, S. H. (2004), 'Automated generation of new knowledge to support managerial decision-making: a case study in forecasting stock market', Expert Systems 21(4), 192–207.

[40] Fu, J., Lum, K. S., Nguyen, M. N. & Shi, J. (2007), 'Stock prediction using fcmac-byy', Advances in Neural Networks 4492, 346–351

[41] Huang, S.-C. & Wu, T.-K. (2008), 'Integrating ga based time-scale feature extractions with svms for stock index forecasting', Expert Systems with Applications 35(4), 2080–2088.

[42] Huang, W., Nakamori, Y. & Wang, S.-Y. (2005), ' Forecasting stock market movement direction with support vector machine', Computers & Operations Research 32(10), 2513–2522.

[43] Jia, G., Chen, Y. & Wu, P. (2008), 'Menn method applications for stock market forecasting', Advances in Neural Networks 5263, 30–39.

[44] Kim, K.-J. (2004), 'Artificial neural networks with feature transformation based on domain knowledge for the prediction of stock index futures', Intelligent Systems in Accounting, Finance & Management 12(3), 167–176.

[45] Leung, M. T., Daouk, H. & Chen, A.-S. (2000), 'Forecasting stock indices: a comparison of classification and level estimation models', International Journal of Forecasting 16(2), 173–190.

[46] Niu, F., Nie, S. & Wang, W. (2008), 'The forecasts performance of gray theory, bp network, svm for stock index', International Symposium on Knowledge Acquisition and Modeling pp. 708–712.

[47] Perez-Rodriguez, J. V., Torra, S. & Andrada-Felix, J. (2005), 'Star and ann models: forecasting performance on the Spanish ibex-35 stock index', Journal of Empirical Finance 12(3), 490–509.

[48] Wang, W. & Nie, S. (2008), 'The performance of several combining forecasts for stock index', International Seminar on Future Information Technology and Management Engineering 0, 450– 455.

[49] Wu , Q., Chen, Y. & Liu, Z. (2008), Ensemble model of intelligent paradigms for stock market forecasting, in 'Proceedings of the First International Workshop on Knowledge Discovery and Data Mining', IEEE Computer Society, Washington, DC, USA, pp. 205–208.