Comparison of machine learning algorithm for Santander dataset

Yudhistira Arie Wijaya, Nana Suarna, Iin, Ryan Hamonangan, Ninin R

STMIK IKMI Cirebon, Cirebon, Indonesia

*yudhistira010471@gmail.com

Abstract. The dataset for Santander banks is released on kaggle.com to decide whether the customer makes a transaction or not. The classes in this dataset are 2 with 200,000 entries in records. Earlier experiments using the regression algorithm led to a precision rate of 89%. In this analysis, the best accuracy value from the algorithm was obtained by using 6 different algorithms, namely Support for the Vector Machine (SVM), Neural Network (NN), Naive Bayes (NB), Decision Tree (DT). Before performing the data mining with the algorithm, preprocessing is carried out using a normalizing technique using the range transformation method with values 0 and 1. From the study, the best results were obtained in a Decision Tree 96.03% accurate algorithm, 95.82%, and 95.71%, 95.38%, 90.42%, 90.42%, and Naive Bayes 14.69%. The algorithms of the Decision Tree are 95.03%, 95.71% and 92%. Except for the Naïve Bayes algorithm, the precise value is better than previous study.

1. Introduction

Mechanical learning is currently not only used for automation, but also for the future, customer loyalty, standard recovery rate details. [1], maintenance of customers, credit support planned, extortion discovery and financial leaders revealed and harmed [2]. Banks store a lot of customer records. This information can be used to build and maintain specific customer relationships and partnerships solely for certain banking products or deals. Information mining in financial processes has become popular for applying diagrams and forecasts [3].

Financial notification is a key reason for the decisions of a large number of speculators, officials and others needing accounting details, such as the good articulation of business conduct, financial position and social obligations of the listed organization and others [4]. After the financial crisis of Asia in 1997 several cases of falsified cash have been circulating in the US and Taiwan. The Enron case in 2001, the Abit computer case, the pro computer case, the info disk, and the Technology Summit case in Taiwan in 2004, and the Global Com case of 2003. Given this problem the option to differentiate between fraudulent behavior, before it happens, turns out to be significant. Information mining is a crucial method in the management and classification of complex information requests. Most of the material covered in an organized model should be defined and considered as well as information on valuable ways to include references to leadership. Informative mining has various capacities, such as classification, membership, grouping and estimation[5].

2. Dataset and Methods
2.1. Dataset

Dataset are taken from the Kaggle site competition for Santander transaction prediction [1][6]. The dataset has two different files of 200,000 entries each. The distinction between the two files is that the one file already has a target, used for the model, while the other file lacks a target variable to predict. You can upload the results on the competition website to see if forecasts are correct. Both datasets are fitted with an ID column and 200 variables to forecast while a further column consists of goals. The datasets have a total of 800,000 entries, which is relatively high, according to information. [7][4].

2.2. Pre-Processing

Pre-processing data is a process which manipulates raw data which can be further implemented in the dataset. Incompleteness, non-permanence and data errors need to be dealt. The data are typically incomplete, complex, double and noisy in the real world. It may not contain enough attribute values or just aggregate data. There can be errors and outliers in the data, and variations in the attribute names can also have issues, which may impact the overall distribution of the data. As these data issues can lead to inaccurate forecasts, we use a pre-processing approach to address the question before it happens. As this approach is excellent. It means that the data set is completely clean during pre-processing, with the train and test data set not having a single missing value. There have been no incoherent, redundancies or interruptions. Moreover, all attributes have a floating (numeric) value and are not categorical [5].

2.3. Support Vector Machine (SVM)

The strength of the SVM is its capacity to analyze patterns of data classification in an equal manner and accuracy. With their vast flexibility and multiple data science scenarios, including research into brain disorders, SVM has become a commonly used method for classification. [8].

2.4. Naïve Bayes

Naïve Bayes is a probability and method classification technique [9]. In many complex real-world situations Naïve Bayes also performs much better [2]. Because of its simplicity, Naïve Bayes is a common machine learning model that allows all attributes to make an equal contribution to the final decision. This simplicity is equivalent to calculation efficiency, making Naïve Bayes attractive and suitable for a variety of fields. [2]. The three key elements of the Naïve Bayes Classifier are pre, post and class likelihood. [10]

2.5. Decision Tree

High classification accuracy and robustness are the advantages of a decision tree algorithm. The downside is that the decision tree is easily affected by the sample and in a decision tree the sub decision tree can be replicated several times. The issue of over-stitching can be resolved by cutting and k-fold cross validation technology. Excess branches can be cut off first by means of pruning. Excessive installation can also be stopped. The second issue concerns the pre-filter stage, which can be used to eliminate some of the irrelevant functions from the data pre-processing stage. This allows the decision tree to be smaller and avoids an incorrect decision tree problem. [11]. Perhaps due to its predictive capacity and readability, the decision tree is one of the most common classification models in computer literature [12].

2.6. K-Nearest Neighbor

There are several benefits, but also challenges, in the K-Nearest Neighbor (kNN) algorithm. Second, the rules for decision classification are too straightforward. The KNN algorithm uses a majority voting rules, meaning every closest neighbor has a single vote in the process of decision-making, ignoring variations in the classification of the various closest neighbors. Second, the selection method for the nearest neighbor is not accurate enough from the KNN algorithm. It only tests the similarity between the query and the example using the distance function and completely ignores its spatial distribution. And similarities only are seen from a question point of view, which is not enough to take advantage of the relation between the query and the training example. The third argument is that the KNN algorithm is k-sensitive. If k is too small, the accuracy of classification decreases and noise is amplified in noisy
situations. If k is too large, and the query in an unbalanced data set belongs to a category with less learning instances, excessive noise instances will be selected as the closest neighbor [13].

2.7. Linear Regression
The empirical regression algorithm focuses on physical methods because the reverse design of the process is unsuitable for recovery [14]. For regression, various methods such as linear regression, kriging, vector regression help or neural networks are available [15]. For a broad range of reasons, regression analysis uses funding. Including asset price models, asset price models and pricing theory, price prediction and scenario analysis, risk modelling, volatility forecasts and simulation of Monte Carlo, for example. Regression analysis [16].

3. Result and Discussion
3.1. Accuracy
Accuracy for the Support Vector Machine algorithm is 90.43% with the following details (Table 1). It was predicted 0 and it turned out to be 0 as many as 6858 records, it was predicted to be 0 and it turned out that 1 was 285 records. Predicted 1 and it turned out to be 0 as many as 443 records, it was predicted that 1 and turned out to be 1 as many as 16 records. With the class recall value for true 0 of 93.93% and 5.32% for true 1.

Table 1. Accuracy Algorithm Support Vector Machine

|          | true 0 | true 1 |
|----------|--------|--------|
| pred. 0  | 6858   | 20     |
| pred. 1  | 443    | 16     |
| class recall | 93.92% | 5.32% |

Accuracy to the Neural Network algorithm is 95.82% with the following details (Table 2). Predicted 0 and it turned out to be 0 as many as 7283 records, predicted 0 and turned out to be 1 as many as 300 records. Predicted 1 and turned out to be 0 as many as 18 records, it was predicted that 1 and it turned out that 1 was 1 record. With the class recall value for true 0 of 99.75% and 0.33% for true 1.

Table 2. Neural Net Accuracy Algorithm

|          | true 0 | true 1 |
|----------|--------|--------|
| pred. 0  | 7283   | 20     |
| pred. 1  | 10     | 1      |
| class recall | 99.75% | 0.33% |

Accuracy on the Naïve Bayes algorithm is 14.69% with the following details (Table 3). Predicted 0 and it turned out to be 0 as many as 836 records, it was predicted that 0 and it turned out that 1 was 20 records. It was predicted 1 and it turned out to be 0 as many as 6465 records, it was predicted 1 and it turned out that 1 was 281 records. With the class recall value for true 0 of 11.45% and 93.36% for true 1.

Table 3. Accuracy of the Naïve Bayes Algorithm

|          | true 0 | true 1 |
|----------|--------|--------|
| pred. 0  | 836    | 20     |
| pred. 1  | 6465   | 281    |
| class recall | 11.45% | 93.36% |
The accuracy of the Decision Tree algorithm is 96.03% with the following details (Table 4). It was predicted 0 and it turned out to be 0 as many as 7300 records, it was predicted to be 0 and it turned out that 1 was 301 records. It was predicted 1 and it turned out that 0 was 1 record, it was predicted 1 and it turned out that 1 was 0 records. With the class recall value for true 0 of 99.99% and 0.00% for true 1.

Table 4. Accuracy Decision Tree Algorithm

|          | true 0 | true 1 | class precision |
|----------|--------|--------|-----------------|
| pred. 0 | 7300   | 301    | 99.99%          |
| pred. 1 | 1      | 0      | 0.00%           |

Figure 1 ROC Decision Tree Algorithm

Accuracy to the K-Nearest Neighbor algorithm is 95.71% with the following details (Table 5). It was predicted 0 and it turned out to be 0 as many as 7276 records, it was predicted that 0 and turned out to be 1 as many as 301 records. Predicted 1 and it turned out to be 0 as many as 25 records, it was predicted 1 and it turned out that 1 was 0 records. With the class recall value for true 0 of 99.66% and 0.00% for true 1.

Table 5. Accuracy of the KNN Algorithm

|          | true 0 | true 1 | class precision |
|----------|--------|--------|-----------------|
| pred. 0 | 7276   | 301    | 99.66%          |
| pred. 1 | 25     | 0      | 0.00%           |

Accuracy of the regression algorithm is 95.38% with the following details (Table 6). It was predicted 0 and it turned out to be 0 as many as 7246 records, it was predicted to be 0 and it turned out that 1 was 296 records. Predicted 1 and it turned out to be 0 as many as 55 records, it was predicted that 1 and it turned out that 1 was 5 records. With the class recall value for true 0 of 99.25% and 1.66% for true 1.

Table 6. Accuracy Regression Algorithm

|          | true 0 | true 1 | class precision |
|----------|--------|--------|-----------------|
| pred. 0 | 7246   | 206    | 99.25%          |
| pred. 1 | 55     | 5      | 1.66%           |

The accuracy of the six algorithms mentioned above is shown in Table 7. Of the six algorithms, it appears that the Decision Tree algorithm produces the best value of 96.03%. Next, the Neural Network algorithm is 95.82%, K-Nearest Neighbor is 95.71%, Regression is 95.38%, Support Vector Machine is 90.42% and Naïve Bayes is 14.69%.
Table 7 Algorithm accuracy results

| Algorithm                     | Accuracy |
|-------------------------------|----------|
| Decision Tree                 | 96.03%   |
| K-Nearest Neighbor            | 95.71%   |
| Naïve Bayes                   | 14.69%   |
| Neural Network                | 95.82%   |
| Regression                    | 95.38%   |
| Support Vector Machine        | 90.42%   |

4. Conclusion

From the results of this study, the comparison of the use of six Decision Tree algorithms, K-Nearest Neighbor, Naïve Bayes, Neural Network, Regression, and Support Vector Machine and adding a preprocessing process to Santander Customer data using normalize techniques with range transformation methods values 0 and 1 have an impact accuracy results. In previous studies using regression algorithms resulted in an accuracy value of 89%, while in this study yielded an accuracy value of 95.38%. For the best accuracy value of 96.03% the use of the Decision Tree algorithm.

References

[1] E. Echaniz, C. Q. Ho, A. Rodriguez, and L. dell’Olio, “Comparing best-worst and ordered logit approaches for user satisfaction in transit services,” *Transp. Res. Part A Policy Pract.*, vol. 130, no. January, pp. 752–769, 2019, doi: 10.1016/j.tra.2019.10.012.
[2] U. N. Dulhare, “Prediction system for heart disease using Naïve Bayes and particle swarm optimization,” *Biomed. Res.*, vol. 29, no. 12, pp. 2646–2649, 2018, doi: 10.4066/biomedicalresearch.29-18-620.
[3] K. R. Tiwari, “Customer Transaction Prediction 15.06.2019 —,” 2019.
[4] R. Mohammed, J. Rawashdeh, and M. Abdullah, “Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results,” *2020 11th Int. Conf. Inf. Commun. Syst. ICICS 2020*, pp. 243–248, 2020, doi: 10.1109/ICICS49469.2020.239556.
[5] P. Mishra, A. Biancolillo, J. M. Roger, F. Marini, and D. N. Rutledge, “New data preprocessing trends based on ensemble of multiple preprocessing techniques,” *TrAC - Trends Anal. Chem.*, vol. 132, p. 116045, 2020, doi: 10.1016/j.trac.2020.116045.
[6] S. Ilyas, S. Zia, U. M. Butt, S. Letchmunan, and Z. un Nisa, “Predicting the future transaction from large and imbalanced banking dataset,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 1, pp. 273–286, 2020, doi: 10.14569/ijacsa.2020.0110134.
[7] S. Guo, K. Hayakawa, and C. Link, “Customer Transaction Prediction,” *Diabetes*, vol. 60, no. 8, pp. 1789–1796, 2016, doi: 10.7282/t3-0j57-mt16.Terms.
[8] D. A. Pisner and D. M. Schnyer, *Support vector machine*. Elsevier Inc., 2019.
[9] H. Muhamad, C. A. Prasojo, N. A. Sugianto, L. Surtiningsih, and I. Cholissodin, “Optimasi Naïve Bayes Classifier Dengan Menggunakan Particle Swarm Optimization Pada Data Iris,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 4, no. 3, p. 180, 2017, doi: 10.25126/jtiik.201743251.
[10] A. Jamain and D. J. Hand, “The Naive Bayes Mystery: A classification detective story,” *Pattern Recognit. Lett.*, vol. 26, no. 11, pp. 1752–1760, 2005, doi: 10.1016/j.patrec.2005.02.001.
[11] H. F. Zhou, J. W. Zhang, Y. Q. Zhou, X. J. Guo, and Y. M. Ma, “A feature selection algorithm of decision tree based on feature weight,” *Expert Syst. Appl.*, vol. 164, no. August 2020, p. 113842, 2021, doi: 10.1016/j.eswa.2020.113842.
[12] S. Maliah and G. Shani, “Using POMDPs for learning cost sensitive decision trees,” *Artif. Intell.*, vol. 292, p. 103400, 2021, doi: 10.1016/j.artint.2020.103400.
[13] Z. Pan, Y. Wang, and Y. Pan, “A new locally adaptive k-nearest neighbor algorithm based on discrimination class,” *Knowledge-Based Syst.*, vol. 204, p. 106185, 2020, doi: 10.1016/j.knosys.2020.106185.

[14] N. Panigrahi and B. S. Das, “Evaluation of regression algorithms for estimating leaf area index and canopy water content from water stressed rice canopy reflectance,” *Inf. Process. Agric.*, no. xxxx, 2020, doi: 10.1016/j.inpa.2020.06.002.

[15] A. Kämper, L. Leenders, B. Bahl, and A. Bardow, “AutoMoG: Automated data-driven Model Generation of multi-energy systems using piecewise-linear regression,” *Comput. Chem. Eng.*, vol. 145, 2020, doi: 10.1016/j.compchemeng.2020.107162.

[16] R. L. Kissell, “Linear Regression Models,” in *Algorithmic Trading Methods*, Elsevier, 2021, pp. 151–173.