TEKNOKOM : Jurnal Teknologi dan Rekayasa Sistem Komputer
Vol. 5, No. 2, September 2022, pp. 119 - 124
ISSN: 2686-3219, DOI: 10.31943/teknokom.v5i2.77

APPLICATION OF CLASSIFICATION ALGORITHM FOR SALES PREDICTION

Sendi Permana[1*], Rosadi[1], Nikki[1]
[1]Computer Science Study Program, Universitas Nusa Mandiri, Indonesia

ARTICLE INFO

History of the article:
Received July 5, 2022
Revised July 26, 2022
Accepted August 1, 2022
Published September 3, 2022

Keywords:
Classification
Machine Learning
Prediction
Sale

ABSTRACT

Increasing sales results is a desired target for all companies both at home and abroad. The company has a wide variety of products to offer. This paper (to fulfill a Business Intelligence course assignment) is the result of an experiment from data (keaggle) about consumer demand for products during the 2013-2015 period, then based on this data we try to predict to classify product sales, in order to make it easier for companies to classification for sales predictions. To find out the sales of the best-selling products, data mining classification techniques are used, namely XGBoost, Decision Tree, Random Forest, Linear Regression, and Nave Bayes. Based on the test results of the five classification techniques, the XGBoost model is the best with the data training value producing an RMSE value of 0.68% and data testing of 0.79%. This method is also better than the results of previous studies.

Correspondece:
Sendi Permana,
Computer Science Study Program,
Universitas Nusa Mandiri,
Email : 14207003@nusamandiri.ac.id

INTRODUCTION

Indonesia is a country with the largest mini market in Southeast Asia. Retail sales in Indonesia increased by 130% between 2010 and 2015, with 43,826 stores opening in mid-2019. A number of factors are driving this growth, including the growth of a young middle class and an increasingly congested life, especially in big cities, where easy access and Shopping convenience is very important for consumers[1].

With increasing internet penetration and greater exposure to brands, products and services, consumers in the region are becoming smarter and more informed when making purchasing decisions. Offers tailored to local tastes and buying behavior are critical to maintaining business continuity. In this field, intelligent systems and data science can help solve market and consumer segmentation problems [2].

In this study, the classification method is used to solve the problem of predicting sales in the next period. The purpose of this study is to make it easier for the company to classify for sales predictions so that it can reduce the error rate in the difference in data calculations that often occur [3].

LITERATURE REVIEW

XGBoost

Extreme Gradient Boosting or XGBoost is a decision tree or regression tree based enhancement algorithm[4]. Figure 1 shows an overview of the regression tree-based boosting algorithm. The first estimation results are obtained from the learning process of the first tree from the training data. The second tree carries out the learning process from the training data, where the value of |Y-Y1| is the difference between the true label and the predicted label from the previous step. The third tree performs the learning process from the data and produces an estimate of Y3. This can effectively reduce the error value.

Related Research

There is a lot of research in the field of forecasting, and the methods vary widely. The model developed so far mainly focuses on two aspects, namely time series methods and machine learning. The XGBoost algorithm is one of the more popular techniques in the amplifier group due to its good convergence properties. There is a lot of research on the use of XGBoost for forecasting and market segmentation.

XGBoost can be used to predict customer loyalty or churn with high accuracy. XGBoost can give better results than other methods. XGBoost

produces a good level of generalization in stock price predictions, so it can predict the opening price correctly. Applying XGBoost to malware detection achieves high accuracy. Many machine learning methods perform poorly when handling high-dimensional data.

The second study conducted by Ma et al.,(2018) entitled "Estimating Warehouse Rental Price using Machine Learning Techniques" aims to predict warehouse rental prices on the market. To get the best model, this research uses four methods, namely Linear Regression, Random Forest Regressor, Regression Tree, and Gradient Boosting Regression Trees. The results of this study indicate that the Random Forest Regressor method has the highest accuracy and shows the results that the distance variable from the city center has a major influence on the accuracy of warehouse rental price predictions [5].

The third study was conducted by (Čeh et al., 2018) entitled “Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments”[6]. This study aims to analyze the prediction performance of the machine learning random forest method with multiple regression methods in predicting apartment prices. The dataset used is apartment transaction data for real estate sales in the city of Ljubljana, Slovenia in 2008-201. From this study, the results of the Rsquare value, sales ratio, average percentage error (MAPE), coefficient of dispersion (COD)) revealed that the random forest method had much better results. The fifth study conducted by (Borde et al., 2017) entitled "Real Estate Investment Advising Using Machine Learning" aims to compare several methods, namely the gradient descent method, K-nearest neighbor regression and random forest regression in predicting real estate prices. shows from the calculation of MAPE, RMSE and MAE errors the random forest method has the smallest error value[7].

RESEARCH METHOD

Proposed Model

The proposed model is a comparison between those generated by the XGBoost Algorithm, Decision Tree, Random Forest, Linear Regression, Naïve Bayes.

XGBoost

XGBoost was one of our early preferred algorithms. Statistically, it is the most commonly used model for Kaggle competitions. It provides system optimization through parallelization and hardware optimization. XGBoost has an advantage over general gradient enhancement as it provides regularization via a combination of ridge regression and lasso regression. It also handles different types of Sparsity Patterns in data is valid [8].

XGBoost is a regression tree with the same decision rules as the classic decision tree. In a regression tree, each internal node represents a value for the attribute test, and a leaf node with a score represents a decision. The output is the sum of all the scores predicted by the K-tree, as shown below.

$$\gamma = \sum_{k=1}^{K} fk(xi), fk \in F \quad (1)$$

where x t is the i-th sales training sample, fk is the k-th tree score, and F is the function space containing all the regression trees. XGBoost uses the same gradient enhancements as the Gradient Boosting Machine (GBM), but with slight improvements to the regularization objective that reduce model complexity.[9]

Decision Tree

Often called a decision tree. It is like a tree structure where there are internal nodes that describe attributes, the results are represented by each branch. The attribute test of each leaf describes a class. The C4.5 algorithm is the ID3 algorithm. Thanks to this development, the C4.5 algorithm has the same basic working principle as the ID3 algorithm[10].

Random Forest

Random Forest (RF) is a method that can improve the accuracy of the results because the generation of children for each node is done randomly[11]. This method is used to extract attributes and data randomly to build a decision tree consisting of root nodes, internal nodes, and leaf nodes, in accordance with applicable regulations. The root node is the node at the top, or commonly referred to as the root of the decision tree. An internal node is a branch node that has a minimum of two outputs and only one input. While the leaf node or terminal node is the last node that has only one input and no output [12]. The decision tree first calculates the entropy value as a determinant of the level of attribute impurities and the value of information acquisition. The entropy value is calculated using Equation 1, while the information gain value is calculated using Equation 2 [13].

Linear Regression

Regression method is a statistical method that uses the development of a mathematical relationship between variables to make predictions, the dependent variable (Y) and the

independent variable (X) [14]. The dependent variable is the variable that affects or affects and the independent variable is the causal variable or affects. If the independent variable is known, then the value of the dependent variable can be predicted. Usually, sales or demand for a product is represented as a large dependent variable or its value is influenced by an independent variable, and linear regression is one of the methods used in production to predict or predict quality and quantity characteristics [15].

Naive Bayes

The Naive Bayes classifier algorithm is a tool for writers to solve existing problems. The advantage of using the naive bayes classifier is that this method only requires a little training data to determine the estimated parameters needed in the classification process [11]. In the naive bayes classifier approach, which separates constant string data from continuous numeric data, this difference can be seen when determining the probability value for each criterion, both criteria for string data values and for standard numeric data values [16].

RESULTS AND DISCUSSION

Sales data is public data obtained from Kaggle, from 2013 – 2015 and the data obtained are 2935849 Sales data, 22170 item data, 84 category data and 60 store data. Prior to predictive modeling, several methods were used for research, these methods can be seen in Figure 1 below:

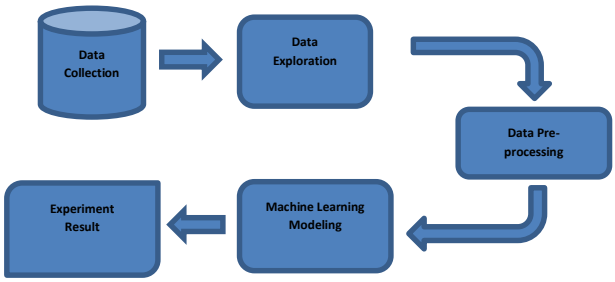

Figure 1. Research Flow

a. Data collection
   At this stage the author takes data from kaggle
b. Data Exploration
   After the data is obtained, the next step is the data exploration stage, where at this stage the data is explored and identified missing values or empty data that will cause inaccurate predictions.
c. Data Pre-processing
   At this stage, before we use the data for prediction purposes, the data must be preprocessed beforehand because not all the attributes in the data will be used for further processing so that the data used will suit our needs.
d. Machine Learning Modeling
   To make a prediction with Machine Learning, of course, it is necessary to choose the right Machine Learning model to process the data we have so that the results obtained are in line with expectations.
e. Experiment Results
   After carrying out several stages of data processing and then the data is processed, at this stage an evaluation of the experimental results is carried out using a classification algorithm to predict prices with the performance metric used is based on the Root Mean Squared Error (RMSE).

Translate Data

The data obtained is data in Russian, and to facilitate the research, the first step in this research is to translate the data into Indonesian. The features that are translated are category names and item names.

Figure 2. Wordcloud before Translation

It can be seen in Figure 2, which is a Wordcloud of store name data and category names before being translated into Indonesian using the googletrans translator function and the translation results can be seen in Wordcloud Figure 3. as follows.

Figure 3. Wordcloud after Translation

Data Analysis Exploration

Data Exploration Analysis was carried out in order to know the characteristics of the data before testing the model. The results of data exploration to see the relationship between stores, items and item prices can be seen in Figure 4 below.

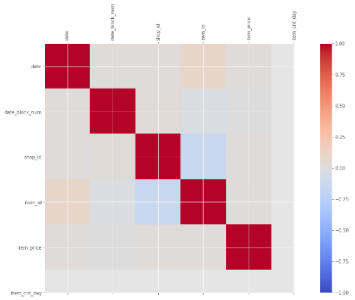

Figure 4. Relationship Between Shop, Item and Item Price

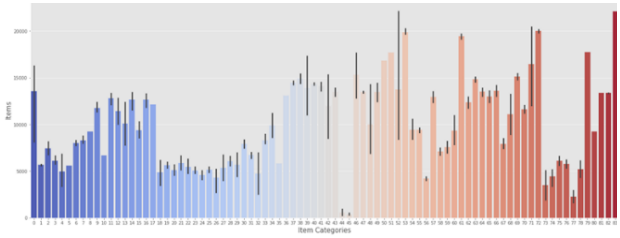

Figure 5. Number of Items Sold

Based on Figure 5 the most sales of items in each category can be seen in the item category id 83 which shows a fairly significant bar chart. Next is the analysis of data items sold every month from February 2013 – October 2015 as follows.

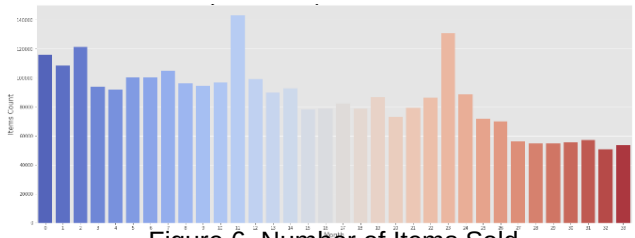

Figure 6. Number of Items Sold

Figure 6 shows the results of item sales data analysis with the highest number of item sales shown in the 12th month and the second highest rank is in the 24th month.

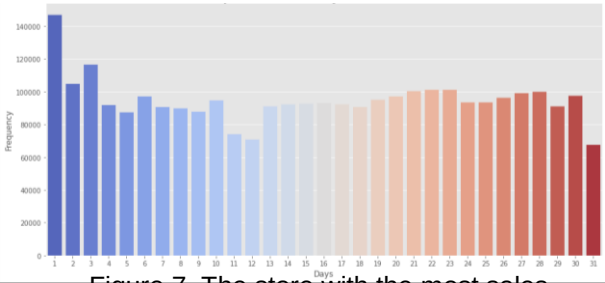

Figure 7. The store with the most sales (in days)

Figure 7 Shows The highest average daily sales are shown significantly on the 1st of more than 140000 sales. But the average number of daily sales is not stable so that the sales per day are very diverse.

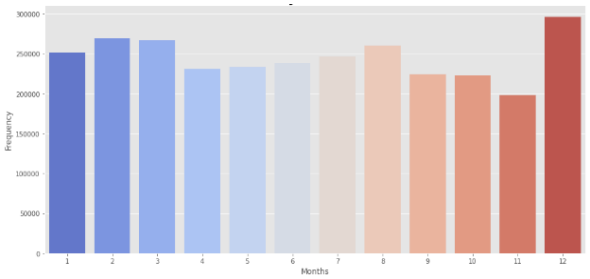

Figure 8. The store with the most sales (in months)

Based on Figure 8, the store with the most sales can be seen in the December sales frequency of almost 30000 sales. And the year with the highest sales frequency is shown in 2013, it can be seen in Figure 9 following.

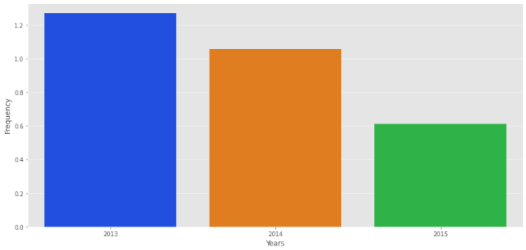

Figure 9. The store with the most sales (in Years)

Feature Engineering

Based on the exploration data analysis process, the resulting features for prediction are made by making master data by combining sales data, items and category items into one training data. Next, make a name for each feature that will be used. After the renaming, the next step is to create a sales feature based on 'date_block_num', 'shop_id', 'item_id', create item_cnt_month based on 'shop_id', 'item_id', create item_price_unit based on item_price/ item_cnt results, increase items based on item_price-hist_min_item results items based on the results of hist_max_item_price - item_price and make shifts to items 'shop_id', 'item_category_id', 'item_id' and fill 0 values for each shift with Missing Value.

Preprocessing data

Before the data mining process can be carried out, it is necessary to do a preprocessing stage, at this stage data cleaning is carried out to produce a clean dataset so that it can be used in the next stage, namely mining. Then before doing algorithm modeling, first the data used is filled with features that are shifted and emptied by 0

then the test data is filled with the mean value for the missing value.

Model Testing

In order to get the expected results, in this study several experimental models were carried out so that the expected results would be more optimal. The tests carried out included the following:

a. Testing with XGBoost algorithm

The architecture used in the XGBoost algorithm as an optimization of prediction results in order to produce a lower error value can be seen in table 1 below.

Table 1. Parameter XGBoost

No	Parameter	Optimal Value
1	max_depth	10
2	n_estimator	250
3	eta	0.5
4	min_child_weight	500
5	subsample	0.7
6	colsample_bytree	0.7

In addition to the architecture used, in table 1 there is feature engineering used in the XGBoost algorithm modeling, feature engineering is made so that the prediction results are more optimal by getting fewer error values.

Table 2. Best features used in XGBoost

No	Parameter	Optimal Value
1	Shop_mean	3275
2	Shop_item_mean	1113
3	Item_count_mean	818
4	Item_count_shifted3	677
5	Item_count	626
6	Item_count_shifted1	609
7	Item_trend	603
8	Item_count_shifted2	483
9	Item_count_std	288
10	Mean_Item_count	269

b. Testing With Decision Tree

Decision tree model testing is done after preprocessing the data and splitting the data with a comparison of 80% of training data and 20% of testing data. The test is carried out with the help of sklearn with the cross-validation function n_splits = 5.

c. Testing With Random Forest

The comparison of the data used in the Random Forest model testing is also the same as the previous model, namely 80% training data and 20% testing data. In Random Forest, the test also uses the help of a sklearn setup with many trees or n_estimator as much as 20 and random_state = 0 which means that seeds with random numbers are not used.

d. Testing With Linear Regression

Testing with linear regression also uses the help of sklearn and the LinearRegression function by determining the value of the Root Mean Squared Error (RMSE) on 80% of the training data and 20% of the testing data.

e. Testing With Naïve Bayes

Predictions based on the performance of the Root Mean Squared Error (RMSE) value in nave Bayes are also used with the help of sklearn and the GaussianNB function.

Results

Based on the results of experiments that have been carried out, the best model produced is using XGBoost when compared to the performance of other models using data training and data validation which are not much different. The RMSE validation results from the value experiment using the XGBoost Training RMSE model reached 0.689928548 for the Testing RMSE value reaching 0.790317290, for the Decision Tree model the RMSE Training value reached 1.177446199 and the Testing RMSE value reached 1.471040079. Then to see the detailed results of the RMSE Training and RMSE Testing scores on the Random Forest, Linear Regression and Naive Bayes models, see table 3 below.

Table 3. Parameters Model

No	Model	Training RMSE	Testing RMSE
1	XGBoost	0.689928548	0.790317290
2	Decision Tree	1.177446199	1.471040079
3	Random Forest	1.178820744	1.378878961
4	Linear Regresi	1.629092056 3	1.350822234
5	Naïve Bayes	2775.944893	2763.692180

CONCLUSION

Based on the results of experiments carried out in the previous discussion using the XGBoost , Decision Tree, Random Forest, Linear Regression and Naïve Bayes models, it can be concluded that the best performance is by experimenting with XGBoost. This can be seen based on the lowest Root Mean Squared Error (RMSE) value in the experiment using the XGBoost model on the training data resulting in an RMSE value of 0.68% and in the testing data, which is 0.79%. Suggestions for the next research

is optimization at the selection stage so that the attributes can be reduced. Thus, it is expected that the value of accuracy and precision will increase.

SUGGESTIONS

Based on the discussion that has been described in the previous chapter, the suggestions can be given for prediction development sales are as following:

1. Adding attributes for data Product sale.
2. Using another method to make sales predictions such as k-means and Support Vector Machines.

REFERENCES

[1] A. K. Aslam, J. Teknik, P. Wilayah, D. A. N. Kota, F. Sains, and D. A. N. Teknologi, "Pengaruh pertumbuhan minimarket terhadap minat dan kebiasaan belanja masyarakat di kelurahan tamamaung kota makassar," 2017.

[2] R. Siringoringo, R. Perangin-angin, and M. J. Purba, "Segmentasi Dan Peramalan Pasar Retail Menggunakan Xgboost Dan Principal Component Analysis," METHOMIKA J. Manaj. Inform. dan Komputerisasi Akunt., vol. 5, no. 1, pp. 42–47, 2021, doi: 10.46880/jmika.vol5no1.pp42-47.

[3] R. N. Hay's, Anharudin, and R. Adrean, "Sistem Informasi Inventory Berdasarkan Prediksi Data Penjualan Barang Menggunakan Metode Single Moving Average Pada Cv.Agung Youanda," Protekinfo, vol. 4, no. 5, pp. 29–33, 2017.

[4] R. Siringoringo, R. Perangin-angin, and J. Jamaluddin, "Model Hibrid Genetic-Xgboost Dan Principal Component Analysis Pada Segmentasi Dan Peramalan Pasar," METHOMIKA J. Manaj. Inform. dan Komputerisasi Akunt., vol. 5, no. 2, pp. 97–103, 2021, doi: 10.46880/jmika.vol5no2.pp97-103.

[5] A. N. Rachmi, "Xgboost Pada Klasifikasi Customer Churn," 2020.

[6] M. Čeh, M. Kilibarda, A. Lisec, and B. Bajat, "Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments," ISPRS Int. J. Geo-Information, vol. 7, no. 5, 2018, doi: 10.3390/ijgi7050168.

[7] S. Borde, A. Rane, G. Shende, and S. Shetty, "Real Estate Investment Advising Using Machine Learning," Int. Res. J. Eng. Technol., vol. 4, no. 3, pp. 1821–1825, 2017, [Online]. Available: https://irjet.net/archives/V4/i3/IRJET-V4I3499.pdf.

[8] J. Statistika, F. Matematika, D. A. N. Ilmu, P. Alam, and U. I. Indonesia, "Implementasi Metode Extreme Gradient Boosting (Xgboost) untuk Klasifikasi pada Data Bioinformatika (Studi Kasus : Penyakit Ebola , GSE 122692)," 2020.

[9] R. Rismala, L. Novamizanti, K. Nur Ramadhani, Y. Siti Rohmah, S. Parjuangan, and D. Mahayana, "JEPIN (Jurnal Edukasi dan Penelitian Informatika) Kajian Ilmiah dan Deteksi Adiksi Internet dan Media Sosial di Indonesia Menggunakan XGBoost," vol. 7, no. 1, pp. 1–11, 2021.

[10] A. Puspita, "Prediksi Kelahiran Bayi Secara Prematur dengan Menggunakan Algoritma C . 45," vol. II, no. 1, pp. 11–16, 2016.

[11] B. Dan, R. Forest, and P. Klasifikasi, "Analisis perbandingan kinerja cart konvensional, bagging dan random forest pada klasifikasi objek: hasil dari dua simulasi," vol. 12, no. 2, pp. 1–12, 2019, doi: 10.14710/medstat.12.1.1-12.

[12] V. O. L. N. O. Desember, N. I. Prabawati, and M. F. Duskarnaen, "Kinerja Algoritma Classification a nd Regression Tree (Cart) da lam Mengklasifikasikan Lama Masa Studi Mahasiswa y ang Mengikuti Organisasi d i Universitas Negeri Jakarta Avalaiable at : Avalaiable at :," vol. 3, no. 2, pp. 139–145.

[13] S. Saadah and H. Salsabila, "Prediksi Harga Bitcoin Menggunakan Metode Random Forest," J. Komput. Terap., vol. 7, no. 1, pp. 24–32, 2021.

[14] P. Kasus and D. Pemasaran, "Development of Multipolynomial Regression Model."

[15] G. N. Ayuni and D. Fitrianah, "Penerapan metode Regresi Linear untuk prediksi penjualan properti pada PT XYZ," J. Telemat., vol. 14, no. 2, pp. 79–86, 2019, [Online]. Available: https://journal.ithb.ac.id/telematika/article/view/321.

[16] M. Guntur, J. Santony, and Y. Yuhandri, "Prediksi Harga Emas dengan Menggunakan Metode Naïve Bayes dalam Investasi untuk Meminimalisasi Resiko," J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 2, no. 1, pp. 354–360, 2018, doi: 10.29207/resti.v2i1.276.