Case Study of Model Selection on Customer Information Task Based on Machine Learning Algorithms

Qingyu Cai *

Department of accounting, Wenzhou-Kean University, Zhejiang, China

*Corresponding author: caiqin@kean.edu

Abstract. In today's era of high-speed development, more and more companies are finding customers with different needs in the market. Due to the large market size, each company cannot tailor its market for each customer, it's difficult for them to predict the customer's need, so market segmentation has emerged. This paper will use a case study about an automotive company to choose a best prediction modelusing the information of existing products and customers. The company divides the customer into 4 segmentations A, B, C and D. The quantitative method of study will be used to find the relationship between products and customers. Logistic Regression, KNN, SVM, Random Forest, and Decision Tree are used to compute the accurate rate. Decision Tree model was found to be the most accurate and the accuracy is 53%. In this paper, business objectives were defined, features and distribution of data were explored, data were processed, relevant features were selected, data were modeled, and accurate values between five different models were calculated. These steps can help the company find the nearest algorithmic model that allows it to use the best marketing strategy for its customers.

Keywords: Market segment; Customer needs prediction; Logistic regression; KNN; SVM; Random Forest; Decision Tree.

1. Introduction

In the era of big data, the speed of information exchange is getting faster and faster, which influences managers to consider consumer preferences and needs when making marketing decisions. Nowadays, most companies develop their products based on a mainly customer-oriented direction, and customer behavior can help managers evaluate whether they can earn the benefit with their business strategies and constantly modify them to achieve profitable revenues [1]. In modern research, many studies are based on machine learning to solve problems such as automotive cost [2] and patent technology prediction [3]. However, dealing with a large amount of consumer preference data and correctly classifying consumers has been a challenge for many company managers in running their businesses.

This paper uses information about the known customers of automotive companies from the Kaggle website and choose an appropriate model. During the process of finding model, the independent variables are gender, marital status, age, education information, profession, work experience, spending score, number of family members and anonymized category for the customer. The dependent variable is target customer segment of the customer. In the following, five models, including Logistic Regression, SVM, KNN, Random Forest, and Decision Tree, are used to compare three metrics: accuracy, precision, and recall, to select the best model for customer segmentation prediction.

2. Related work

Markets are changing from time to time because of the internet age, and in the past, data often relied on manual, which made it difficult for a large amount of data to be used by managers. This section reviews the literature on customer segmentation and machine learning models in business as a predictive customer classification model analysis.

Modern companies often classify customers based on customer values, needs, preferences, and other factors that can provide customers with targeted products and services to improve customer

satisfaction. Companies integrate a variety of marketing and quantitative analysis to identify customer lifetime value, current value, and customer loyalty as segmentation criteria, which allows companies to allocate resources better and develop corporate strategies [4]. Companies can use psychological analysis of consumer information as well as analysis of their consumer behavior itself. As active customers will have loyalty to the brand, managers focus on the marketing management, future planning, and development strategies of the company to increase customers' personal experience with the product through the affection between the different categories of customers and the brand [5]. There are many reasons for the diversity of supply, including production facilities, similar uses in product design, different manufacturer status of products; different product research and development by the competitor; factors such as consumer price sensitivity, product materials, and package size. Due to these factors, companies must develop different products and marketing strategies related to other customers to meet the needs of consumers [6]. For different psychology and behavior, it allows companies to provide various promotional tools and product quality content in the production and sales process according to different criteria such as customer's buying style and size.

With the development of machine learning, it has been widely used in data science and automation. Still, it is not particularly widely used in economic business. The application of machine learning in business will be presented through the literature below. Machine learning and economics are relevant to the differences in objectives, methods, and settings. The machine learning literature employs unsupervised learning methods, supervised learning methods of regression and classification, and matrix completion methods applied to specific class problems [7]. Supervised learning is one of the main approaches to machine learning, which provides researchers with more precise criteria for model optimization. According to Nasteski, 10 supervised learning methods being empirically demonstrated on a large scale, including SVM, logistic regression, neural nets, random forests and decision trees [8].

3. Methodology

3.1 Data

The data used in this study was obtained from the Kaggle website by Kash, who uploaded this database on August 17, 2022[9]. 8068 objects were grouped as train data and 2627 as test data. Customer information included ID, gender, marital status of the customer, age, graduate status, profession, work experience, spending score, family size, anonymized category, and target customer segment. In the process of cleaning the data, there are many missing values in the database, as deleting the missing values directly will make a lot of data lost. Thus, the data of numeric type is filled using the average value and the data of categorical type which is missing is deleted directly. After cleaning, 7699 objects in the train data and 2488 objects in the test data.

3.2 Logistic Regression Model

Estimating the likelihood of something happening and solving binary classification problems can usually be achieved by logistic regression methods, a machine learning method. Logistic regression models are models with a fixed number of parameters that output categorical predictions and make no assumptions for a linear relationship between the dependent and independent variables [10]. In the logistic regression model, the observations are independent of each other. The equation for multidimensional linear regression is

$$y = a_1 * x_1 + a_2 * x_2 + \cdots + a_n * x_n = \sum a_i x_i \quad (1)$$

x is the variable, a is the model parameter, which is the intercept of the regression model on the vertical y-axis and also a constant value, and y is the model output. The coefficient of a is the slope of the regression line. Based on increases in the value of the independent variable by one unit more or less, the result increases, meaning the more significant the coefficient of a, the greater the corresponding independent variable on the results [11].

3.3 KNN Model

The KNN algorithm is a supervised machine learning algorithm that can be used for classification and regression prediction problems, requiring only the number of nearest neighbors to be defined and no other parameters[12]. First of all, choose the appropriate formula for calculating the distance, the common distance definitions in modern academia are Minkowski distance, Euclidean distance, Manhattan distance, Chebyshev distance, and Hamming distance. The K value represents the number of classifications. A small K value makes the model complex and easy to overfit in the choice of K value, while a large K value makes the model simple and underfit. After defining the distance and K values in advance, for any new sample that needs to define a category, a new sample is classified as the one with the most categories among the K samples with the closest distance to that sample. The equation is

$$y = \arg\max_{cj} \sum (x_i, y_i) \in N_k(x) f_{Cj}(y_i) \tag{2}$$

The performance of KNN methods may be affected by the choice of K values and distance measurements. The flowchart of KNN model is

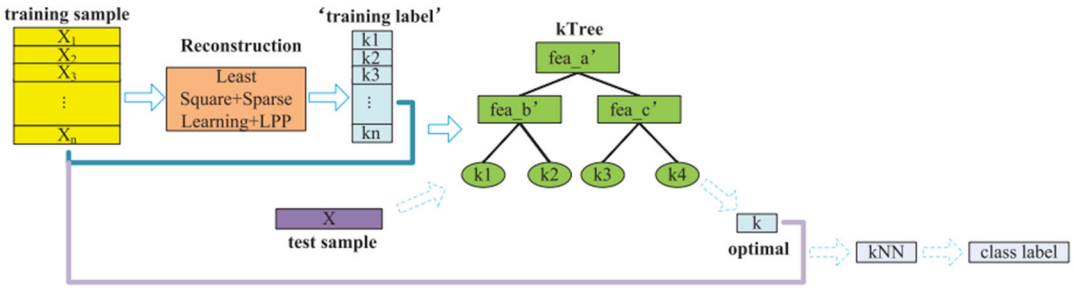

Fig.1 The flowchart of KNN Model [13]

3.4 SVM Model

SVM is Support Victory Machines, built on the principles of statistical VC and structural risk minimization that can be used for classification, regression, and outlier detection. For nonlinear problems, SVM nonlinearly maps the training data to higher dimensional feature space [14]. Support Vector Machine is a binary classification model and basic model is defined as the separating hyperplane that divides the training data set and has the largest geometric interval. w*x + b=0is the separation hyperplane, and there are infinitely many such hyperplanes for linearly separable data sets. SVM models are classified into linearly divisible support vector machines, linear support vector machines, and nonlinear support vector machines. Linearly separable support vector machine applies to training data linearly separable, and linear support vector machine applies to training data approximately linearly separable, with the samples linearly separable after removing the outliers. Nonlinear support vector machine applies to training data linearly non-separable [15].

3.5 Random Forest Model

A random forest has multiple decision trees. When a prediction is needed for a sample, each tree in the forest is counted to predict that sample. Random forests are a scheme proposed by Leo Breiman in the 2000s, in which each tree in the set is grown according to random parameters, and the final prediction is obtained from the set[16].The random forest construct is built from multiple decision trees. Suppose a sample needs to make a prediction; each tree in the forest calculates the prediction for that sample and then, through a voting method, chooses the final predictive result. Random Forests is to take features and samples randomly so that each tree in the forest has both similarities and differences. Random forest is an effective algorithm and is mostly used in classification and regression problems. Two factors determine the random forest error rate: the correlation rate of any two trees and the strength of any tree itself. The higher the correlation, the higher the error rate [17].

3.6 DecisionTree Model

The decision tree method is a popular data mining method that classifies populations into branch-like segments and can efficiently handle large, complex data sets without imposing complex parameter structures [18]. A decision tree is a tree-like structure where the internal nodes represent judgments on attributes showing a mapping relationship between object attributes and object values, the branches represent the output of judgment results, and the leaf nodes represent classification results. Making a Decision Tree starts with variable selection. Many variables in the experiment are not well correlated, so it is essential to assess the importance of the variables and select the most relevant input variables to inform subsequent studies. Here are two ways decision tree model todeal with missing data: the first way is to put missing values as a single category and then analyzed together with other categories, or it can build target variables with a large number of missing values, make predictions, and replace these missing values with predicted values [19].

4. Experimental results and analysis

4.1 Data analysis

Table 1. Statistic Description of customer information

Statistics	Age	Work Experience	Family Size
count	7669.00	7669.00	7669.00
mean	43.51	2.64	2.85
std	16.69	3.23	1.5
min	18.00	0.00	1.00
25%	31.00	0.00	2.00
50%	40.00	1.00	2.85
75%	53.00	4.00	4.00
max	89.00	14.00	9.00

According to Table 1, most respondents are between 25 and 45 years old, and the average age lies at 43.5 years old. The average working age is 2.6 years, and the household size is mostly distributed among 2-3 people. Among each type of customer, the number of male customers is larger than female customers, and the spending score of most of them is low.

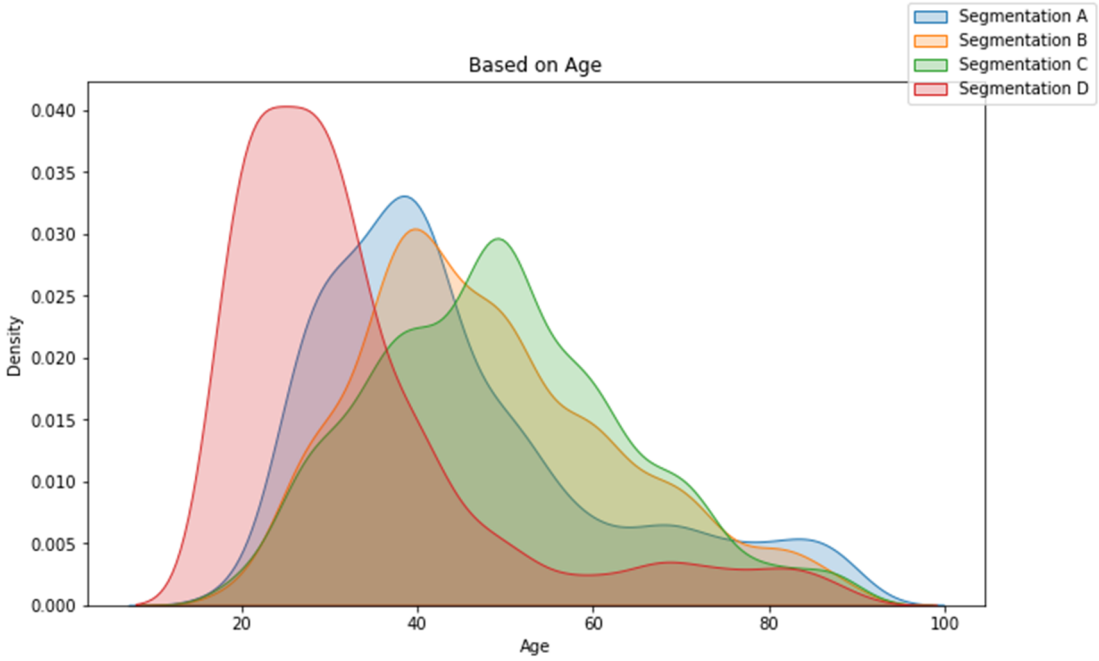

Fig.2 Distribution of different segmentation at different ages

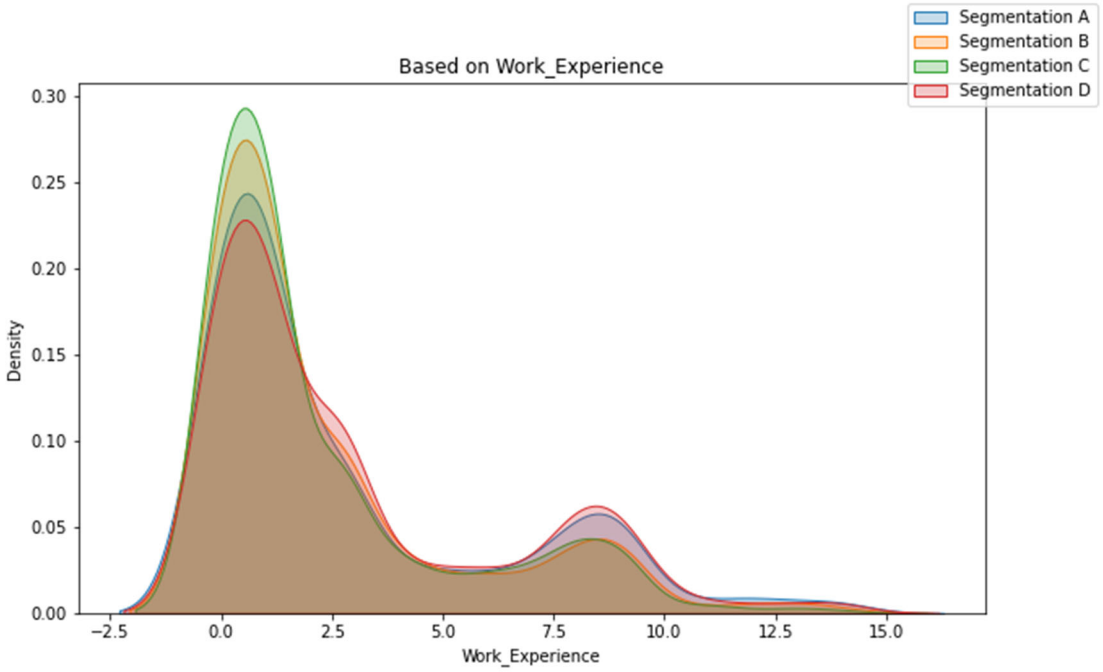

Fig.3 Distribution of different segmentation at different work experience year

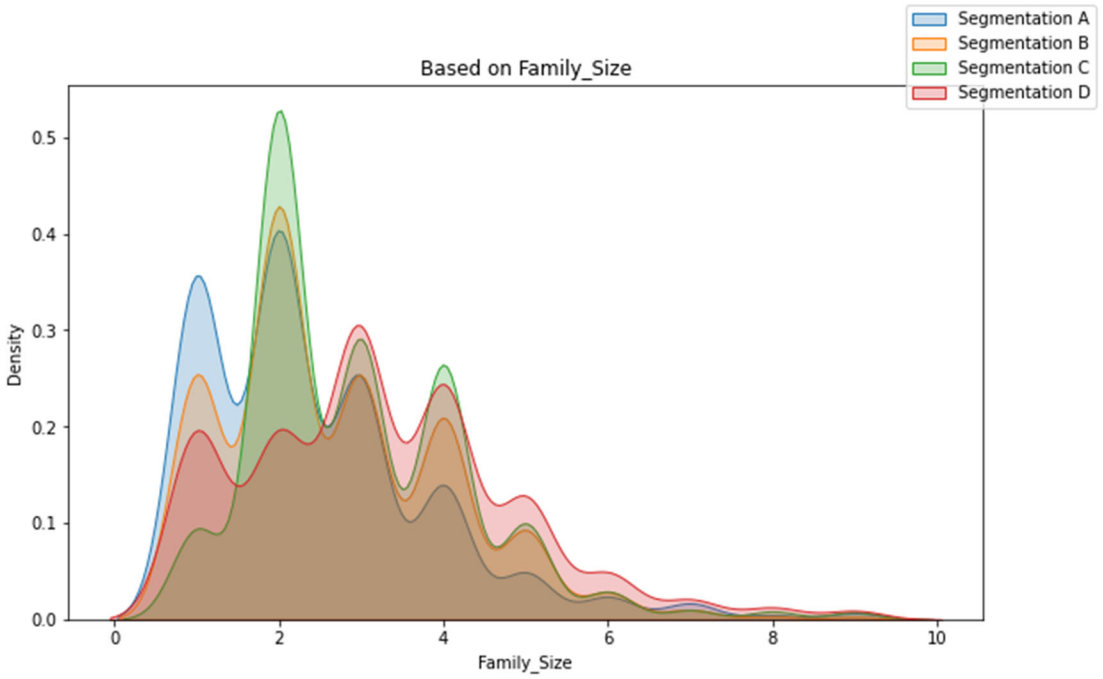

Fig.4 Distribution of different segmentation at different family size

Based on Figure 2, Figure 3 and Figure 4, information can be gotten that the segmentation A is concentrated in the age of 30-40, B is concentrated in the age of 40-50, C is concentrated in the age of 50, and D is concentrated in the age of 20.The years of work were all concentrated between 0-1 year and 8-9 years, and the household size of 1 person was more distributed in A, and two persons in B, C, and D.

4.2 Correlation between variable

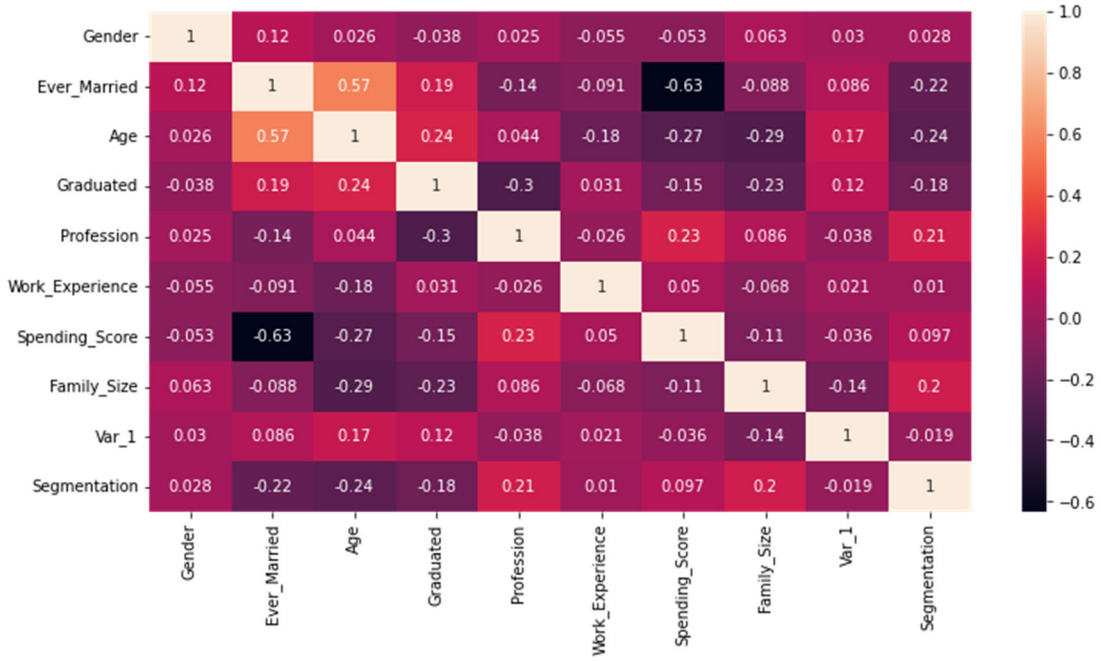

Fig.5 The correlation between different variables

The different shades of color in the heat map represent the shades of association between the different correlation coefficients. The scale on the right side and the different color representations show the coefficients, and the variable names are on the left and bottom sides. The lighter the color, the greater the relationship between the two variables is represented, and a higher correlation means strong multicollinearity [20]. One of the two variables may be considered for exclusion when performing feature engineering to avoid overfitting due to multicollinearity. From Figure 5, information can be gotten that there is no association between each variable and no need to eliminate it.

4.3 Logistic Regression Model

Table 2. Logistic Regression Model's classification report

Parameters	precision	recall	f1-score	support
A	0.43	0.46	0.44	379
B	0.34	0.16	0.22	334
C	0.48	0.56	0.51	388
D	0.62	0.74	0.67	433
Accuracy			0.50	1534
Macro avg	0.46	0.48	0.46	1534
Weighted avg	0.47	0.50	0.48	1534
LR accuracy	49.67%			

According to Table 2, this paper used the logistic regression model in Python and calculated that the model accuracy is 49.67%.

4.2 KNN Model

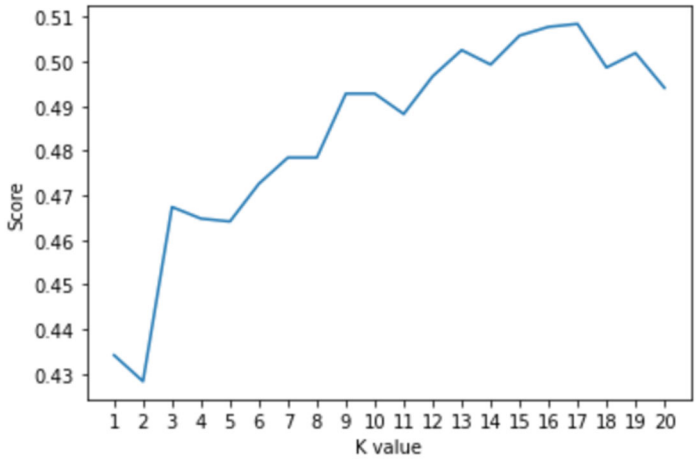

Fig.6 KNN Model's accuracy

From the Figure 6 is KNN model's accuracy and the best accuracy is 50.85%.

4.4 SVM Model

Table 3. SVM Model's classification report

Parameters	Precision	recall	f1-score	support
A	0.35	0.34	0.34	379
B	0.30	0.13	0.18	334
C	0.48	0.53	0.50	388
D	0.54	0.76	0.63	433
Accuracy			0.46	1534
Macro avg	0.42	0.44	0.41	1534
Weighted avg	0.43	0.46	0.43	1534
SVM accuracy	45.76%			

According to Table 3, the SVM model accuracy is 45.76%.

4.5 Random Forest Model

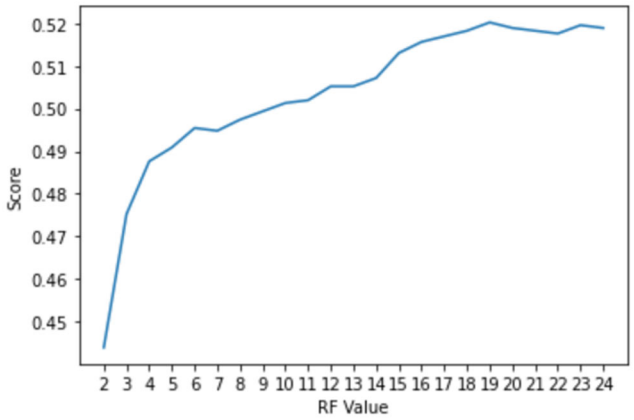

Fig.7 Random Forest Model's accuracy

Figure 7 is Random Forest model's accuracy and the best accuracy is 52.02%.

4.6 Decision Tree Model

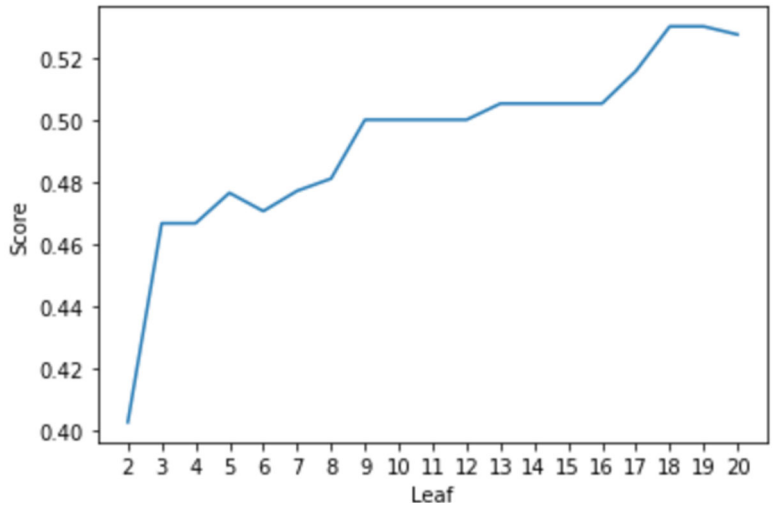

Fig.8 Decision Tree Model's accuracy

Figure 8 is Decision Tree Model's accuracy and the best accuracy is 53%.

4.7 Model comparison

Table 4. Accuracy of different models

Model	Accuracy
Decision Tree	53.00%
Random Forest	52.02%
KNN	50.85%
Logistic Regression	49.67%
SVM	45.76%

By ranking the accuracy of the above models, it is found that the Decision tree model has the highest accuracy and its accuracy is 53%.

5. Conclusions

This article have helped the manager of the automotive company to select the most accurate machine learning model and also review what customer segmentation and machine learning models in the economy are. Five models, including Logistic Regression, KNN, SVM, Random Forest, and Decision Tree, are used in this paper. There are 8067 customer objects in the training set and 2627 objects in the testing set. Because there were missing data, the data cleaning method of filling the quantitative data with the average and dropping the qualitative data directly was used. Transform quantitative data into qualitative data by normalizing the data. This study drawsthe heatmap to find the relationship between different variables. Five models are then briefly described and calculated for the available data.Bycalculation, the Decision Tree model has the highest accuracy and accuracy is 53%. In this paper, the algorithmic model combines marketing and machine learning algorithms. It helped the company to find a suitable algorithmic model, which laid the foundation for the subsequent company to find the right marketing strategy for its customers.

References

[1] Hosseini, M., &Shabani, M. (2015). New approach to customer segmentation based on changes in customer value. Journal of Marketing Analytics, 3(3), 110-121.

[2] Bodendorf, F., Merbele, S., & Franke, J. (2019). Predictive Cost Analytics of Vehicle Assemblies Based on Machine Learning in the Automotive Industry.

[3] Lee, C. W., Tao, F., Ma, Y. Y., & Lin, H. L. (2022). Development of Patent Technology Prediction Model Based on Machine Learning. Axioms, 11(6), 253.

[4] Sari, J. N., Nugroho, L. E., Ferdiana, R., &Santosa, P. I. (2016). Review on customer segmentation technique on ecommerce. Advanced Science Letters, 22(10), 3018-3022.

[5] Hultén, B. (2007). Customer segmentation: The concepts of trust, commitment and relationships. Journal of Targeting, Measurement and Analysis for Marketing, 15(4), 256-269.

[6] Smith, W. R. (1956). Product differentiation and market segmentation as alternative marketing strategies. Journal of marketing, 21(1), 3-8.

[7] Metsalu, T., &Vilo, J. (2015). ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. Nucleic acids research, 43(W1), W566-W570.

[8] Nasteski, V. (2017). An overview of the supervised machine learning methods. Horizons. b, 4, 51-62.

[9] Kash. 2022. Customer Segmentation Classification. Retrieved from https://www.kaggle.com/datasets/kaushiksuresh147/customer-segmentation

[10] Boateng, E. Y., &Abaye, D. A. (2019). A review of the logistic regression model with emphasis on medical research. Journal of data analysis and information processing, 7(4), 190-207.

[11] Stoltzfus, J. C. (2011). Logistic regression: a brief primer. Academic emergency medicine, 18(10), 1099-1104.

[12] Xiong, L., & Yao, Y. (2021). Study on an adaptive thermal comfort model with K-nearest-neighbors (KNN) algorithm. Building and Environment, 202, 108026.

[13] Zhang, S., Li, X., Zong, M., Zhu, X., & Wang, R. (2017). Efficient kNN classification with different numbers of nearest neighbors. IEEE transactions on neural networks and learning systems, 29(5), 1774-1785.

[14] Adankon, M. M., &Cheriet, M. (2009). Model selection for the LS-SVM. Application to handwriting recognition. Pattern Recognition, 42(12), 3264-3270.

[15] Wang, L. (Ed.). (2005). Support vector machines: theory and applications (Vol. 177). Springer Science & Business Media.

[16] Biau, G. (2012). Analysis of a random forests model. The Journal of Machine Learning Research, 13(1), 1063-1095.

[17] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[18] Adankon, M. M., &Cheriet, M. (2009). Model selection for the LS-SVM. Application to handwriting recognition. Pattern Recognition, 42(12), 3264-3270.

[19] Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130.

[20] Metsalu, T., &Vilo, J. (2015). ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. Nucleic acids research, 43(W1), W566-W570.