Influencing Factor Analysis for Automobile Customers Based on One-Way ANOVA and Decision Tree

Yifan Chen[1,a,*]

[1]Pittsburgh Institute, Sichuan University, Chengdu, China
a. chenyifan2526@stu.scu.edu.cn
**corresponding author*

Abstract: Customer-related companies are always doing researches on the customer behaviors in order to promote the profitability, and one of the most representative corporations is the automobile companies. For long time they are studying on the customer segments and using different methods to treat them. Therefore, this paper studies the influencing factors of the customer spending level and the classification method to determine different customer segments and give predictions for future customers. Firstly, method of One-Way ANOVA is used to determine the significance of influence on each factor, and the result shows that attributes including Age, Gender, Graduated, Married, Work Experience and Family Size have a determining impact on the customer spending level. Then, based on all these variables included, a decision tree model is constructed using CART method with Gini Index and Pruning for cost complexity method. The depth of the tree is 5 and the overall accuracy reaches 81.6%. Compared with other decision tree model, CART tree gives a high accuracy level with relatively simple tree structure. These results could assist automobile companies to make appropriate customer segments and develop corresponding strategies, which could be further expanded to other customer-targeted fields.

Keywords: decision tree, One-Way ANOVA, customer segments, influencing factor

1. Introduction

1.1. Background

The research of customer behaviors is always the focus of companies which are connected with customers [1]. Automobile companies are one of the more representative companies. For them, targeted advertising and publicity strategy based on customer behavior prediction could largely reduce the time and financial cost in the process of sales [2]. Thus, it is of importance to give predictions based on existing customer information. Considering that, many automobile companies generate and group by their customer data from historical sales and orders to do researches on the influencing factors and customer classification models for their consuming potentials. It is a never-end problem to precisely market for a specific customer segment.

1.2. Related Research

Many analysists and researchers have dug into the customer classification based on the existing data from various kinds of companies.

In electricity business field, it is of vital importance to identify the consumption patterns of customers regarding the existing similar patterns. Unsupervised machine learning methods has been used in the consumption behavior recognition, including hierarchical clustering, K-means and fuzzy K-means. The researches done by Chicco et al. show that classification methods and some clustering techniques are very useful in assisting the classification of electricity customers [3]. It suggests the usability of customer classification in future consumption pattern recognition and prediction.

For customer segmentation, the life time analysis and customer value are of high importance. Also, customer satisfaction, loyalty and profitability have an increasing value in the business environment today. Through the research of customer value, potential value and loyalty, the customer segmentation could be made in a more balanced way, which assist a lot in customer classification. Then, by studying the characteristics of each segment, the corresponding strategies could be tailored and developed [4]. It also confirms the importance of the classification of customers based on customer data.

To give a classification model for personal credit, Zhao et al. designs a decision tree model with boosting algorithm. They give metrics including accuracy, precision, recall, and so on to evaluate the effect of the decision tree model constructed. The result shows that decision tree with boosting algorithm could lift the performance but increase the time spent. Both conventional and boosting decision tree could give a relatively satisfying results [5]. Their researches show that it is feasible and useful to use decision tree model with C5.0 algorithm or CART method, which could reach a high accuracy level and gives satisfying outcome.

Current researches have shown the importance of customer classification in market competition, the practice of customer segments based on existing customer data and the feasibility of the introduction of decision tree into the customer classification process.

1.3. Objective

Based on the researches above, this paper intends to study on the influencing factors which affect the spending level of automobile customers significantly and then give a practicable decision tree model for the prediction of new customer spending level based on the existing historical data and customer information. It could be helpful for the customer behavior analysis and future customer spending level prediction and thus, assist the automobile companies develop appropriate strategies for different customer segments.

2. Methodology

2.1. Source of Data

A dataset from Kaggle [6] is used to do the automobile customer research and classification based on variables such as age, family size, profession and so on. The dataset contains 8, 069 records of data with 13 features, including Customer ID, which is not relevant to the study in this paper. Each record gives information of one specific customer. This paper uses it to construct the classification model.

2.1.1. Dependent Variable

In the dataset used, a categorical variable is used to indicate the spending level of customers, with values of “High”, “Average” and “Low”. These values are originally assigned in the dataset by the automobile company.

2.1.2. Independent Variable

There are many other factors which influence the spending level, and thus may be included in our classification model. For example, the family size of the customers may affect the decision made by them due to different daily needs of automobile, and the difference in age may lead to various concepts in consuming, which also impacts the spending on automobiles. All the variables that may impact the spending level in automobile consumption are listed as the independent variables that may be used in the classification model, including Gender, Married, Age, Graduated, Profession, Work Experience and Family Size.

2.2. Data Processing

After first scanning of the dataset, these records with missing values in columns like Work Experience, Family Size are directly excluded out of our research. Thus, a total 6969 records of data are used in the model construction.

In convenience of calculation and regression, the values in the column “Spending Score” are labeled with numerical values using equal interval method 0-1 scaling. Thus, the original value “Low” is 0, “Average” is 0.5 and “High” is 1.

In the classification model construction process, 70% of the datasets are randomly selected as the training sets and the other 30% is used as the testing sets [7].

2.3. Decision Tree Model

To construct an accurate classification model for automobile customers, the Classification and Regression Tree (CART) model is used to build the Decision Tree [8]. The process contains two main steps. Firstly, the splitting attributes are selected based on Gini Index, and then some nodes are pruned using the method of pruning for cost complexity to avoid overfitting. The detailed description is in the following part.

2.3.1. Selection of Splitting Attribute.

The splitting attribute chosen is based on Gini Index, which gives the impurity level of the data [9]. Smaller the index is, the more information it could give. The formula is shown below:

$$Gini = \sum_{i=1}^{n} p(x_i)\big(1 - p(x_i)\big) = \sum_{i=1}^{n} p(x_i)^2 \qquad (1)$$

where $p(x_i)$ is the percentage of the number of corresponding records in the total dataset.

In every splitting node, the classification which gives the smallest Gini Index is chosen, and the division of dataset follows the rules in Table 1 [10].

Table 1: Division rules for selecting splitting attribute.

Type of Attribute	Division	Selection
Categorical	Every combination into two groups.	The combination giving the smallest Gini Index.
Numerical	Every average value of two values.	The demarcation giving the smallest Gini Index.

According to the division rules listed, for every splitting, the binary tree is constructed based on the variable and the division of the smallest Gini Index.

2.3.2. Pruning for the Model

To avoid overfitting to the existing dataset, the Pruning for Cost Complexity method is used [11].

$$\alpha = \frac{C(t) - C(T_t)}{|T_t| - 1} \tag{2}$$

$$C(t) = r(t) \times p(t) \tag{3}$$

where α measures the surface error rate of a certain node, r(t) shows the error rate of nodes and p(t) is the percentage of data sets following that node. The node giving the smallest value of α is then pruned.

Based on the methods mentioned above, the Binary Decision Tree model could be constructed.

2.4. Evaluation Metric

The Decision Tree model is evaluated by the accuracy of classification [12], which is defined as follows:

$$\text{Accuracy} = \frac{\#\text{Correct Classification}}{\#\text{Records}} \tag{4}$$

Accuracy is the ratio of correct classification over the number of total records. It measures the ability of the model of giving the correct classification. The range is accuracy is [0, 1]. The larger the ratio, the more accurate the classification model [13].

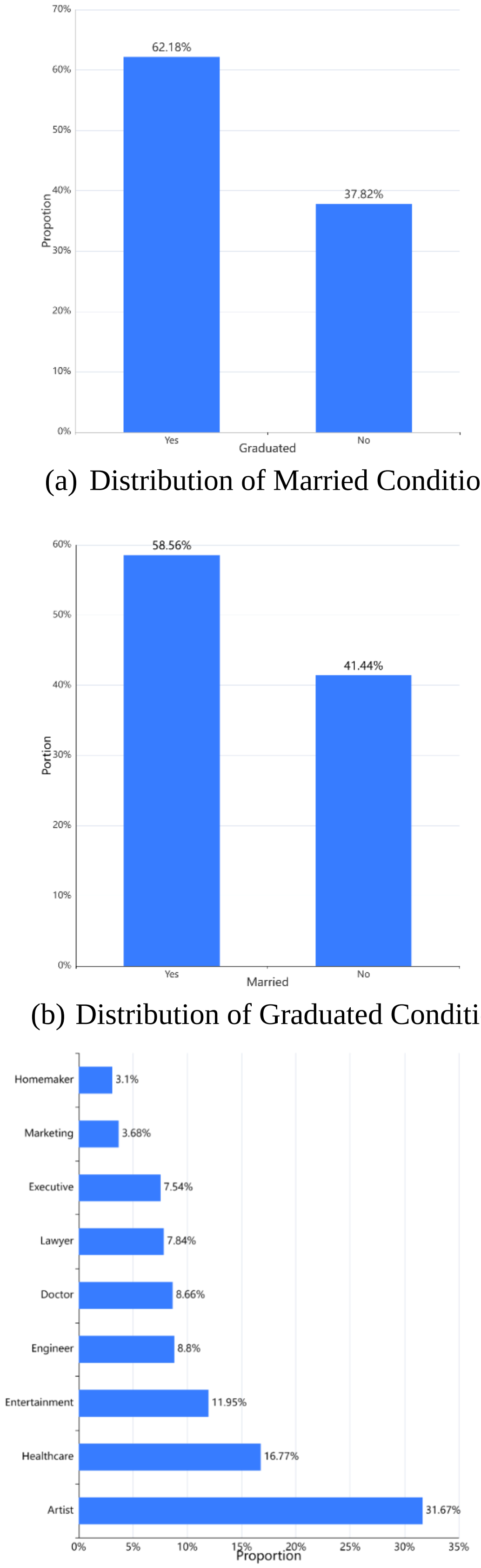
3. Results and Discussion

3.1. Data Visualization

Table 2: Statistics of important numerical variables.

Variable	Maximum	Minimum	Mean	SD
Age	89	18	43.469	16.531
Work Exp	14	0	2.633	3.403
Family Size	9	1	2.844	1.528
Spending Score	1	0	0.275	0.0137

The Statistics of some important numerical variables are listed in Table 2. The age of customers ranges from 18 to 89, with a mean value of 43.469 and standard deviation of 16.531; the work experience is from 0 to 14 in years; the family size is from 1 to 9; and the spending score, which is assigned artificially with value of 0, 0.5, 1, has a mean value of 0.275 and standard deviation of 0.0137.



(a) Distribution of Married Condition

(b) Distribution of Graduated Condition

(c) Distribution of Profession

Figure 1: Distributions of categorical variables (Photo credit: Original).

Fig. 1 shows the distributions of some important categorical variables, including married condition, graduated condition and profession. It is shown that about 58.56% of the customers are married, and 62.18% of the customers are graduated. For profession condition, over 30% of the customers are artists, while the least profession is homemaker.

3.2. Variable Selection

To identify the factors affecting the classification and promote the accuracy of the decision tree model, One-Way ANOVA method is used to select the variables entering the classification model.

Table 3: Results of One-Way ANOVA.

Variable	F-Statistics	P-Value
Gender	16.865	<0.001
Age	770.759	<0.001
Married	2903.503	<0.001
Work Experience	18.675	<0.001
Family Size	46.306	<0.001

Table 3 indicates that all the variables included in One-Way ANOVA test shows a large value of f-statistics, and thus the average of different groups by these variables is significantly different from the other. Therefore, all these variables are selected into the construction of decision tree model. Some other variables are excluded from the model, including Customer ID, Graduated, Customer Segment and Customer Category.

3.3. Model Training and Evaluation

The decision tree model is trained for the first splitting attribute after the calculation for the Gini Indexes with every different splitting attribute division, and then continue to form a full map of splitting attributes.

Next, the values of α are calculated for the pruning process. The node which gives too small the value of α is considered as overfitting for the existing dataset, and thus, is pruned.

Finally, the constructed decision tree model is shown in Figure 2.

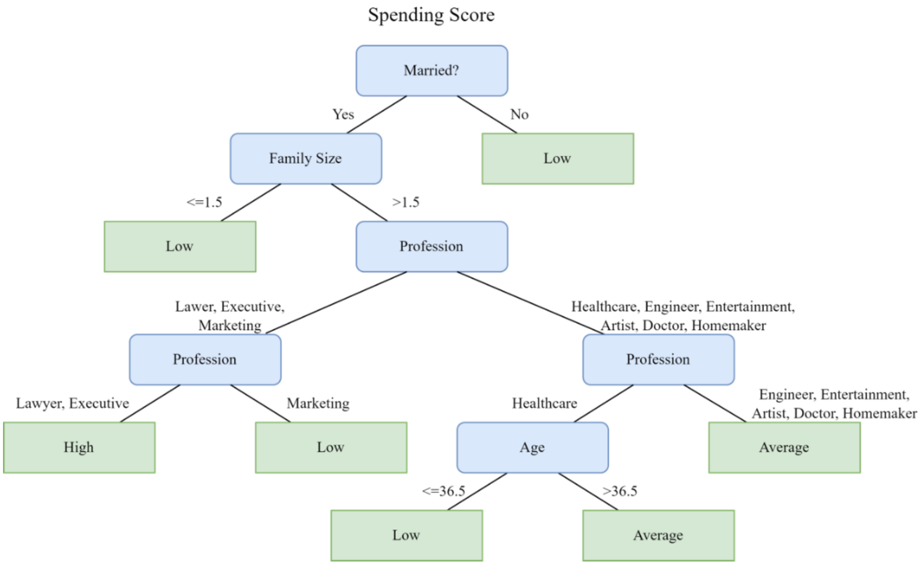


Figure 2: The decision tree of classification for spending score (Photo credit: Original).

In Fig. 2, the rounded rectangle in blue represents the condition judgement, and the corresponding condition is just below it, near the straight lines extending from the rectangle. The rectangle with right angles in green represents the classification results. The corresponding condition is just above it, near the straight line extending to it.

After learning the training sets for the maximum depth of 5, the final decision tree has a total of 13 nodes, with the maximum cases of 100 sets and minimum cases of 50 sets.

After the training process, the decision tree model is then examined by the testing sets to compute the accuracy in classification. The results are shown in Table 4.

Table 4: Accuracy of the decision tree model.

Observation\Predicted	Average	High	Low	Percent Correct
Average	1608	91	14	93.9%
High	340	648	34	63.4%
Low	664	117	3346	81.1%
Overall Percentage	38.1%	12.5%	49.5%	81.6%

Table 4 shows the accuracy of the decision tree constructed. The left column are the observations and the upper row is the predicted value. The percentage in the bottom row shows the partition of each spending level in the whole dataset. The right column indicates the percentage of the correct classification over the total sets of data.

From Table 4, it is shown that for classifying customers of average spending level, the model has the accuracy up to 93.9%; for classifying customer of high spending level, the model's accuracy is 63.4%; and for classifying customer of low spending level, the decision tree has the accuracy of 81.1%. Totally, the decision tree has the accuracy level of 81.6%.

3.4. Method Comparison

The Decision Tree based on CART fits the existing dataset more compared with binary tree like Quick, Unbiased, Efficient Statistical Tree (QUEST) method [14], which could only classify correctly 18.8% of the data sets of high spending level. The comparison is listed below.

Table 5: Comparison of Accuracy between CART method and QUEST method.

Spending Score	CART	QUEST
Average	93.9%	99.5%
High	63.4%	18.8%
Low	81.1%	70.7%
Overall Percentage	81.6%	70.1%

From Table 5, it is shown that although QUEST method has a higher accuracy in classification of average spending score than CART method, the accuracy in both low spending score and high spending score is much lower than CART method, especially for high spending score. Also, the overall accuracy differs a lot.

Therefore, CART method has a higher accuracy, and thus confirms the method selected.

Moreover, for other decision trees which are not binary, the accuracy is very close to the CART one, and thus we prefer the simpler one.

3.5. Limitation

Despite the relatively high accuracy of the decision tree model, limitations also exist.

This paper only analyzes one dataset from an automobile company. Therefore, the data volume could be further expanded, including the number of customers, the number of brands and different regions.

When doing data cleaning, now the records with missing values are directly removed, which may be further improved and interpolated with other classification models or interpolation methods.

This paper uses One-Way ANOVA based on linear regression. However, the values of R^2 of the linear regressions are mostly lower than 0.5, indicating the fitting effect to be not very high.

4. Conclusion

Historical customer data has a significant impact on the study of customer behavior and customer segments and decision tree model could be very useful in the classification of customer spending level. Based on One-Way ANOVA, influencing factors of customer spending are recognized including Gender, Age, Married, Graduated, Profession, Work Experience and Family Size. All these significance levels are far smaller than 0.05, indicating a very high determining effect on customer spending level. After including these factors into the CART decision tree using Gini index and method of Pruning for cost complexity, the decision tree constructed reaches depth of 5 and consists of attributes including Married, Family Size and Profession. The overall accuracy is over 80%, which outperforms that of QUEST method and is similar to the accuracy of other non-binary method. Thus, the simple one is preferred. This paper could assist customer-related companies to find factors affecting the spending level and develop appropriate strategies for different customer segments. Then, they could predict future behavior of certain customers and thus promote the profitability of companies. The research could be further expanded to other fields, where customer is the target of the market, and thus appropriate classification could help reach their goals.

Reference

[1] *Lim, Weng Marc., et al.: Past, present, and future of customer engagement. Journal of Business Research (2021).*

[2] *Yan, Qingyou., et al.: Research on real purchasing behavior analysis of electric cars in Beijing based on structural equation modeling and multinomial logit model. Sustainability 11(20), 5870 (2019).*

[3] *Chicco, Gianfranco., Roberto Napoli., and Federico Piglione.: Comparisons among clustering techniques for electricity customer classification. IEEE Transactions on power systems 21(2), 933-940 (2006).*

[4] *Kim, Su-Yeon., et al.: Customer segmentation and strategy development based on customer lifetime value: A case study. Expert systems with applications 31(1), 101-107 (2006).*

[5] *Zhao, Long., Sanghyuk Lee., and Seon-Phil Jeong.: Decision tree application to classification problems with boosting algorithm. Electronics 10(16), 1903 (2021).*

[6] *Kulia, Akashdeep.: Automobile Customer. 9(17), (2021). Accessed on 2.27 (2023).*

[7] *Waldner, François., Damien C. Jacques., and Fabian Löw.: The impact of training class proportions on binary cropland classification. Remote Sensing Letters 8(12), 1122-1131 (2017).*

[8] *Rutkowski, Leszek, et al.: The CART decision tree for mining data streams. Information Sciences 266, 1-15 (2014).*

[9] *Daniya, T., M. Geetha., and K. Suresh Kumar.: Classification and regression trees with Gini index. Advances in Mathematics: Scientific Journal 9(10), 8237-8247(2020).*

[10] *Chandra, B., P. Paul Varghese.: Fuzzifying Gini Index based decision trees. Expert Systems with Applications 36(4), 8549-8559 (2009).*

[11] *Zhou, Xinlei., Dasen Yan.: Model tree pruning." International Journal of Machine Learning and Cybernetics 10, 3431-3444 (2019).*

[12] *Zhang, Xudong., et al.: Prediction accuracy analysis with logistic regression and CART decision tree. Fourth international workshop on pattern recognition. Vol. 11198. SPIE, 2019.*

[13] *Ba'abbad, Ibrahim., et al.: A Short Review of Classification Algorithms Accuracy for Data Prediction in Data Mining Applications. Journal of Data Analysis and Information Processing 9(3), 162-174 (2021).*

[14] Djordjevic, Dejan., et al: Predicting entrepreneurial intentions among the youth in serbia with a classification decision tree model with the QUEST algorithm. Mathematics 9(13),1487 (2021).