Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/74/20240473

Research on sales strategies for target customers of electric vehicles based on gradient boosting ensemble trees

Kaiwen Yang

Chinese University of Hong Kong Shenzhen, School of Science and Engineering, Guodong province, 518172, China

121090701@link.cuhk.edu.cn

Abstract. This paper focuses on the automotive industry, especially the new energy vehicle sector, which is vital for the economy and needs careful market sales decisions. It analyzes data from 1,964 potential customers, including their satisfaction and personal information. The study uses machine learning to create a math model and give sales advice. The data is first cleaned without changing its original structure, and then normalized using the Z-Score method. For any outlier values in the data, the mean of similar items is used to fill in gaps. Specifically, for 500 missing entries about the number of children (B7), the paper groups customers by age, calculates the frequency of children in each group, and uses these frequencies to fill in missing values.Using the cleaned data, the XGBoost algorithm ranks the factors influencing customer decisions. For Brand 1, the top factors are economy (a3), work experience (B10), and safety performance (a4). For Brand 2, comfort (a2), disposable income (B15), and mortgage expenses (B16) are key. For Brand 3, car loan expenses (B17), battery technology (a1), and again mortgage expenses (B16) are most important.The paper also addresses imbalanced data by using an XGBoost binary classification model. This model predicts the likelihood of buying an electric vehicle using the Sigmoid function on tree node weights. It evaluates the model using Precision, Recall, and Area Under the Receiver Operating Characteristic Curve (AUC), which works well even with imbalanced data. The Precision and Recall rates for all three brands are above 90%, with AUC over 85% for Brands 1 and 2, and over 75% for Brand 3. This shows that the model is effective and can be used in real-world situations.

Keywords: Electric Vehicles, Sales Strategy, XGBoost, Dynamic Programming, Feature Selection

1. Introduction

As China's rapid economic and social development, higher national demands are being placed on green, environmentally friendly initiatives, including energy conservation and emission reduction. Currently, both domestic and international research in the transportation sector has achieved certain results in terms of energy conservation and emission reduction [1-3]. The new energy vehicle industry, a strategic and emerging sector within the automotive industry, not only addresses energy and environmental issues but also boasts a broad market prospect. Simultaneously, a key marketing challenge is how the electric vehicle industry, compared to traditional automobiles, can better reach consumers and penetrate the market, necessitating exploration, research, and scientific decision-making in marketing strategies.

To study consumers' purchase intentions towards electric vehicles and develop corresponding sales strategies, sales department invited 1,964 target customers to experience and evaluate three brands of electric vehicles. These included a joint venture brand, an independent brand, and a new power brand. The customers were asked to score the performance indicators of these vehicles, as shown in Table 1, with scores ranging from 0 to 100.

Here's a summary of the table in the format requested:

Table 1. Target Customer Experience Performance Indicator

Code	Indicator
a1	Battery Technology Performance (Battery durability and charging convenience)
a2	Comfort (Environmental friendliness and seating space)
a3	Economy (Energy consumption and value retention)
a4	Safety Performance (Brakes and driving visibility)
a5	Power Performance (Hill climbing and acceleration)
a6	Driving and Handling Performance (Turning and stability at high speeds)
a7	Overall Appearance and Interior Design
a8	Overall Configuration and Quality

Firstly, performing data cleaning, identifying outliers and missing values and their treatment methods. Descriptive statistical analysis is conducted, including a comparative analysis of target customer satisfaction with different car brands. In this target customer experience event, some customers purchased the electric cars they tried (purchase represented by 1, non-purchase by 0). Using previous research results, customer mining models are established for different brands of electric vehicles and evaluate the excellence of the models. Then the models are used to predict the likelihood of 15 target customers purchasing electric cars. Finally, based on previous research conclusions, sales strategy recommendations for the sales department can be raised.

2. Problem description

2.1. Data cleaning

This paper starts by examining the data on electric vehicle performance (indicators a1-a8) and personal characteristics of potential customers (indicators B1-B17) to find any outliers or missing values. For performance indicators a1-a8, any score not between 0 and 100 is seen as an outlier. For personal characteristic indicators B1-B17, values that are not realistic are also considered outliers. These outliers make up about 0.2% of the total data, which has 1,964 samples, so their effect is small. The outliers are removed, and the average value for each feature in the sample data is calculated. To keep the data set complete and accurate for training models later, the outliers are replaced with these mean values. The paper also looks at missing values and finds 500 missing in feature B7 (number of children born). By checking the National Bureau of Statistics website, there's a clear link between the number of children and different age groups. So, the paper divides the data into age groups, calculates how often children are born in each group, and uses these frequencies as the chance of having children to fill in the missing B7 values.

2.2. Data cleaning process

In the survey of target customer satisfaction and personal information, errors may occur in data recording or entry. In such cases, the existing information of the data should be fully considered to fill in missing values. After screening, it was found that feature B7 (number of children born) has 500 missing values. According to the National Bureau of Statistics, the number of children born is strongly associated with age groups. Based on the aforementioned data query and analysis, the age groups were first divided into 60s, 70s, 80s, 90s, 00s, and then, using the number of births in B7 as a variable, a statistical analysis was performed by age group, with the results shown in Table 2.

Table 2. Different Age Groups' Remaining Number of Children Born Statistics

Age	Brand 1	Brand 2	Brand 3	Null	Sum of all people
Born after 1960	4	1	0	1	6
Born after 1970	70	25	1	7	103
Born after 1980	873	170	5	177	1225
Born after 1990	283	32	0	312	627
Born after 2000	0	0	0	3	3
Sum	1230	228	6	500	1964

2.3. Descriptive Statistical Analysis of the Target Customer Dataset

A descriptive statistical analysis model is built based on the Mpai platform, using statistical indicators such as mean, extreme values, standard deviation, quartiles, and the coefficient of variation (CV) to analyze the target customer data. Line graphs comparing the mean, standard deviation, and coefficient of variation are shown in Figure 1, while Figure 2 displays a line graph of the coefficient of variation. Features a1-a8 and B8 have smaller coefficients of variation, whereas the others are larger, confirming that after data cleaning, the target customers' satisfaction scores for different performance indicators are consistent with subjective experiences, and the personal feature information aligns with objective reality.

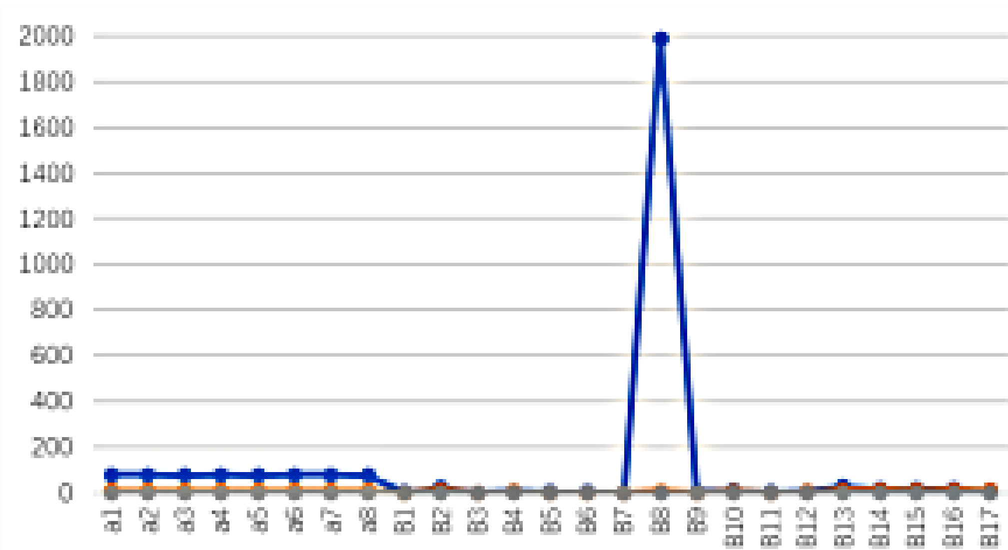

Figure 1. Mean, Standard Deviation, Coefficient of Variation Line Graph (Mean for blue, Standard for orange, CV for grey)

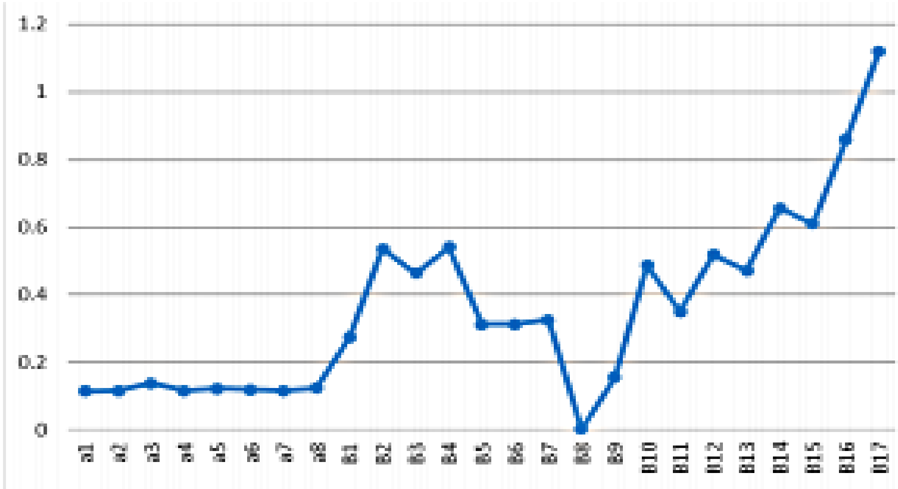

Figure 2. Coefficient of Variation Line Graph

2.4. The use of XGBoost

XGBoost is an advanced version of the Gradient Boosting Trees algorithm, designed for improved performance. Its main approach is to use the gradient boosting algorithm, which means it gets better as more decision trees are added. This process aims to minimize the difference between predicted and actual results by continually improving the objective's effectiveness.

The model uses the ensemble tree algorithm to create a boosting tree with the gradient boosting method. This helps measure how important different features are in the model. For example, in a customer dataset with features a1-a8 and B1-B17, XGBoost calculates the importance of each feature. In each decision tree, a feature's importance is determined by where it splits the tree. Features are considered more important if they are used near the tree's root or chosen in more boosting trees.

The importance of each feature in a tree is shown by how it's weighted and counted at the leaf nodes. The overall importance of a feature is the average across all the trees in the model. This is shown in Figure 3. XGBoost also looks at how much each feature improves the predictions, using the Gini coefficient to measure the average information gain from all splits that use that feature.

$$\mathrm{Gini} = 1 - \sum_{k=1}^{|y|} (p_k)^2 \tag{1}$$

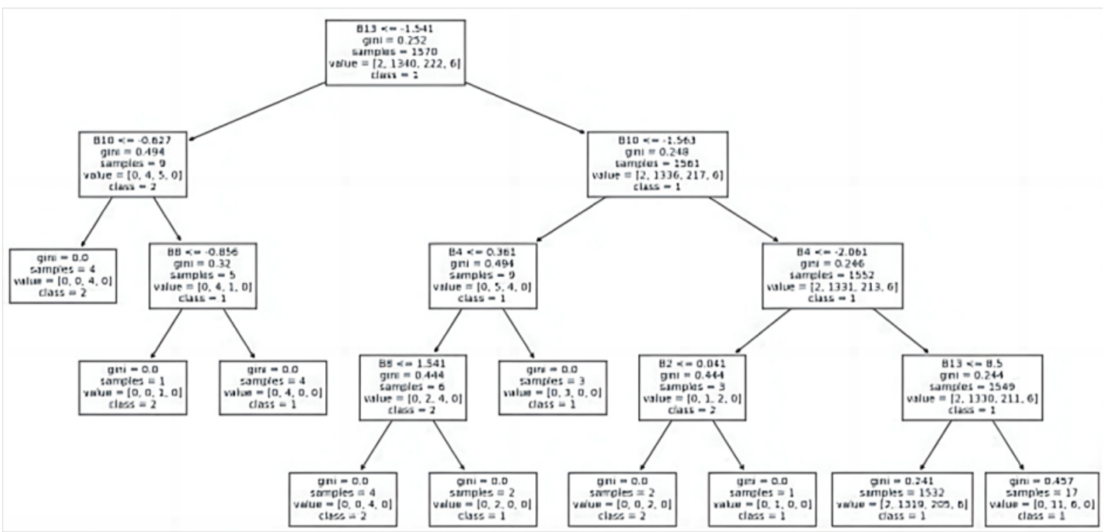

Figure 3. Schematic Diagram of One of the Decision Trees in the Established XGBoost

DOI: 10.54254/2755-2721/74/20240473

2.5. Establishment of the XGBoost Customer Mining Model

XGBoost is an ensemble decision tree model [8] and as a supervised learning algorithm, it comprises several key components: the model, parameters, objective function, and optimization algorithm. Suppose there are n samples and m features, where xi and yi respectively represent the i-th sample and its category label. The variable q represents the structure of each tree, mapping samples to their corresponding leaf nodes. T denotes the number of leaf nodes in a tree, and f(x) corresponds to the tree structure q and the leaf node weights w. The final prediction value is obtained by summing up the results of k trained trees, which can be expressed as:

$$\mathrm{yi} = \sum_{k=1} \mathrm{fk(xi)}, \mathrm{fk} \in \mathcal{F} \tag{2}$$

$$\mathrm{D} = \{(\mathrm{xi}, \mathrm{yi})\}(|\mathrm{D}| = \mathrm{n}, \mathrm{xi} \in \mathrm{R}^{\mathrm{m}}, \mathrm{yi} \in \mathrm{R}) \tag{3}$$

$$\mathscr{F} = \{\mathrm{f(x)} = \mathrm{wq(x)}\}(\mathrm{q}: \mathrm{Rm} \rightarrow \mathrm{T}, \mathrm{w} \in \mathrm{RT}) \tag{4}$$

In equation (5), the first term represents the logistic loss error, and the second term represents the L2 regularization term. Expanding and simplifying equation (5) using a second-order Taylor expansion yields:

$$\mathrm{L(t)(q)} = -0.5 * \sum_{j=1}^{T} \frac{(\sum_{i \in Ij} \mathrm{gi})^2}{\sum_{i \in Ij} \mathrm{hi} + \lambda} + \mathrm{YT} \tag{5}$$

The equation above is used to measure the structure of the t-th tree, where a smaller function value indicates a better tree structure. XGBoost predicts the leaf node weights, and these weights are transformed through a Sigmoid function to output the corresponding probability for binary classification, i.e.,

$$\mathrm{Sigmoid} = \mathrm{P} = \frac{1}{1+e^{y}} \tag{6}$$

2.6. Model performance evaluation and prediction results

This article is about deciding if a potential customer will buy something, which is a simple yes-or-no (binary) problem. It looks at how well predictions match actual results. A tool called a confusion matrix helps understand the model's accuracy. It has four parts: True Positives (TP) are correct guesses of a "yes", False Positives (FP) are wrong "yes" guesses, True Negatives (TN) are correct 'no' guesses, and False Negatives (FN) are missed "yes" guesses.

Three measures evaluate the model: Recall (or sensitivity), Precision, and AUC (Area Under the Receiver Operating Characteristic Curve). Recall shows how many correct guesses there are compared to all true cases. Precision is the ratio of correct guesses to all guesses made. AUC is a score between 0.1 and 1 that shows how good the model is at making predictions, where closer to 1 is better. AUC is useful because it's not thrown off by imbalanced data; it relies on the confusion matrix.

The article uses the confusion matrix to analyze different brands, with results shown in Figure 4. It also uses ROC Curve graphs for each brand category, plotting False Positives (FP) against True Positives (TP).

DOI: 10.54254/2755-2721/74/20240473

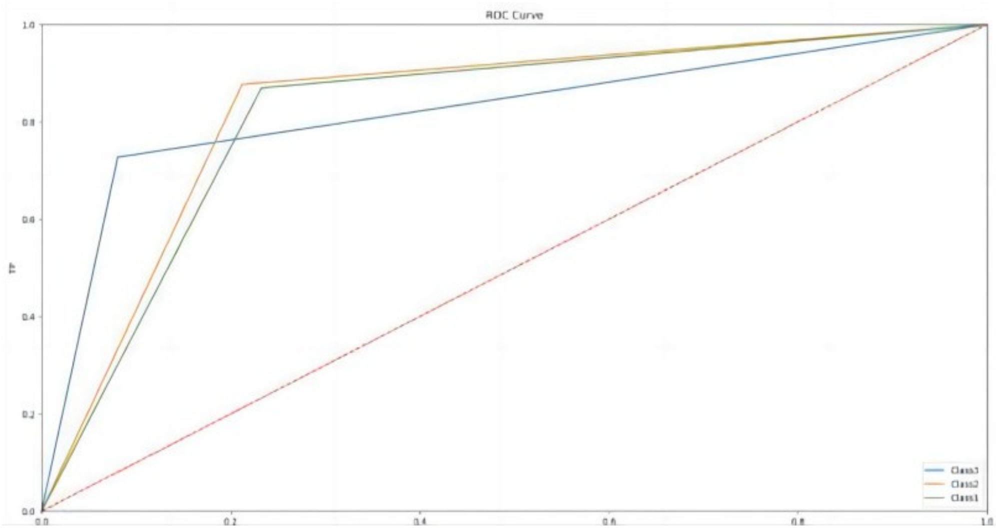

Figure 4. ROC curves for the three different brands.

Using the XGBoost customer mining model, predictions are made on preprocessed data to obtain the probability of purchase. A threshold of 0.5 is used as the standard: if the probability is greater than or equal to 0.5, it is considered a purchase; otherwise, it is not. This determines the final purchase decision. Table 3 shows the results of the likelihood of purchase for the pending customers.

Table 3. Result of three evaluations

Evaluation	Brand1	Brand2	Brand3
Recall	0.95	0.98	0.96
Accuracy	0.94	0.98	0.96
ROC	0.88	0.89	0.76

3. Sales advice

(1) **Switch to Renewable Energy:** Electric vehicle companies should use the change in energy sources, environmental concerns, and government rules to grow. The industry is now influenced by both government policies and market trends.

Improve Vehicle Performance: Electric vehicle brands should work on making their vehicles perform better to stay competitive and successful in the market. They should also try to appeal to a wider range of customers.

(2) **Focus on Different Market Needs**: Companies should make products for different types of customers. For instance, a brand might make a more affordable electric vehicle that appeals to wealthier customers. They should keep improving their products to attract customers with less money.

(3) **Work Together and Learn**: Sales teams need to understand what affects sales, find out what technology is missing, and work with other companies, both local and international, to learn better production and marketing methods. They should also find more places to sell their products.

(4) **Targeted Marketing**: Use a specific marketing strategy that fits the product, brand, market, and customers. This includes knowing who to sell to, focusing on groups of customers, making it easy for them to be happy with the product, using real-life experiences to show off the product's features, and making the target customers more interested in buying.

(5) **Use Government Support**: Because of issues like battery life, safety, high costs for maintenance and replacement, and the limited appeal of expensive models, brands should take advantage of government subsidies and incentives to encourage people to buy their products.

(6) **Promote and Build Brand**: Sales departments should increase their promotional efforts and work on making their brand known. They should hold events and promotions to guide customers,

DOI: 10.54254/2755-2721/74/20240473

build a strong brand image, and increase trust. They should also support the growth of famous electric vehicle brands.

4. Conclusion

This study created models to predict how likely customers are to buy electric vehicles based on how happy they are with different brands and their personal details. They used the XGBoost and gradient boosting algorithms for this. These models can be used in other industries to find potential customers and create tailored sales plans. For instance, knowing how satisfied people are with different smartphone brands can help guess their buying intentions. This lets manufacturers plan how much to produce to increase profits and grow their brands.Companies can also use customer satisfaction to see how well their products are doing and focus on making them better. This helps keep the market running smoothly and stay competitive.However, the current model is quite simple. It might be better to use more detailed evaluation methods. XGBoost, the method used here, sorts features before starting, which can be slow with big datasets. It builds decision trees in a way that reduces the chance of making the model too specific to the data (overfitting), but this can be inefficient if many parts of the tree don't really help the predictions.

References

[1] Zhang, H., Ji, H., & Zhang, H. (2017). Application of XGBoost algorithm in e-commerce product recommendation. Internet of Things Technology, 7(02), 102-104.

[2] Zhang J. H., Zhou J., Hu G. P., et al. Scenario analysis of the energy demand and CO2 emission reduction potential of the urban transport system of Beijing through 2030 [J]. Frontiers of Energy and Power Engineering in China, 2010, 4(04): 459-468.

[3] Pasaoglu G., Honselaar M., Thiel C. Potential vehicle fleet CO2 reductions and cost implications for various vehicle technology deployment scenarios in Europe [J]. Energy Policy, 2012, 40: 404-421.

[4] Lian, K. Q. (2018). Research and Analysis of Ensemble Tree Algorithms Based on Boosting, Doctoral dissertation, China University of Geosciences (Beijing).

[5] Huo H, Wang M, Zhang X, et al. Projection of energy use and greenhouse gas emissions by motor vehicles in China: Policy options and impacts [J]. Energy Policy, 2012, 43: 37-48.

[6] Bai, P. F., An, Q., de Rooij, N. F., Li, N., & Zhou, G. F. (2017). Internet credit personal credit evaluation method based on multi-model fusion. Journal of South China Normal University (Natural Science Edition), 49(06), 119-123.

[7] Li, S. (2019). Research on Credit Risk Assessment Algorithm Based on Unbalanced Data Processing and Feature Selection, Doctoral dissertation, Northwest University.

[8] Ye, Q., Rao, H., & Ji, M. (2017). Commercial sales forecasting based on XGBoost. Journal of Nanchang University (Science Edition), 41(03), 275-281.

[9] Zhou, Z. H. (2016). Machine Learning. Tsinghua University Press.

[10] Chen, Y. T. (2021). Comparative analysis of ensemble learning algorithms: Random Forest and Gradient Boosting Decision Tree. Computer Knowledge and Technology, 17(15), 32-34.