Proceedings of the 6th International Conference on Statistics: Theory and Applications (ICSTA'24)
Barcelona, Spain – August 19 - 21, 2024
Paper No. 142
DOI: 10.11159/icsta24.142

Cluster-Based Customer Satisfaction Analysis

Elena Barzizza[1], Riccardo Ceccato[1], Luigi Salmaso[1]
[1]Departement of Management Engineering, University of Padova,
Stradella San Nicola, 3, 36100, Vicenza, Italy
elena.barzizza@phd.unipd.it; riccardo.ceccato.1@unipd.it; luigi.salmaso@unipd.it

Abstract – This study introduces a new and innovative approach to evaluate the customer satisfaction. Understanding customer satisfaction is crucial for companies that want to success in nowadays highly competitive markets. The use of tools such as Big Data Analytics and Machine Learning can help companies to identify key drivers of satisfaction and predict the impact of potential improvements. Traditional methods of gathering customer satisfaction data, such as questionnaires and online reviews, offer valuable insights, but analyzing this data comprehensively poses challenges such as understanding the impact of several different variables related to the same product or service aspect. Therefore, innovative approaches like clustering analysis are proposed. Our study proposes a novel machine learning-based tool for studying customer satisfaction, enabling the identification of impactful drivers, facilitating product comparisons, and providing insights into areas for improvement. In summary, understanding and enhancing customer satisfaction are essential for business success. By leveraging advanced analytical methodologies, companies can gain deeper insights into customer preferences and make informed decisions.

***Keywords*:** Machine learning, Big Data, Clustering, Customer Satisfaction,

1. Introduction

To gain competitive advantages, companies should consistently introduce innovation, enhance customer segmentation, and offer additional services. Certainly, customer relationships serve as a strategic asset for the company, requiring diligent monitoring similar to that of the typical physical assets. By enhancing services, customers become essential contributors to enhancing company performance (McColl-Kennedy, J., & Schneider, U.): there is a clear link between customer satisfaction and company performances (Yeung & Ennew). Both managerial and marketing theories, along with practitioners, concur that business success is influenced by customer relationships, which necessitate monitoring and management.

A commonly emphasized concept for companies aiming to succeed in the market is customer satisfaction. The customer satisfaction is defined in the literature as "consumer's fulfillment response. It is a judgment that a product or service feature, or the product or service itself, provided (or is providing) a pleasurable level of consumption-related fulfillment, including levels of under- or over-fulfillment" (Oliver). From a more mathematical perspective, we can define this concept as the ratio between customer expectations and perceptions. (Kracklauer, Quinn & Seifert). In other words, it measures the ability of a product or service to fulfill consumer expectations (Ngai, Xiu & Chau) To enhance customer satisfaction, a company should strive to eliminate customer dissatisfaction, aiming ultimately for zero customer complaints (Kondo). Certainly, a customer complains when they are dissatisfied with a particular product or service they have purchased or received.

In today's highly competitive markets, the ability to manage and improve customer satisfaction is a critical task for maintaining or enhancing a company's competitiveness (Patterson). We are currently in the era of digital transformation, where the availability of vast amounts of data is exceedingly high. Utilizing tools like Big Data Analytics can assist companies in comprehending data more effectively and leveraging this information to strengthen their decision-making processes, thereby enhancing the customer experience and concurrently supporting company efficiency. (D. Polding & Eizaguirre Dieguez). Certainly, the utilization of such tools can aid companies in identifying the key drivers that have the greatest impact on customer satisfaction. This valuable insight guides crucial business decisions (Tama). It's evident that there is a need to introduce new and powerful analysis tools to handle this task. Machine learning represents a promising solution that can predict the impact of various drivers on overall customer satisfaction.

Typically, customer satisfaction studies involve collecting data through questionnaires administered to consumers. These questionnaires ask individuals who have tried a particular product to evaluate various aspects, both tangible and intangible, of their experience. Typically, these aspects are measured as numerical variables, often using Likert scales, or as

categorical variables. Another method to gather information about customers' experiences with a product or service is to extract data from online reviews. Typically, this vast amount of data is analyzed using descriptive statistics to gain an initial understanding of the insights it provides. Subsequently, statistical tools such as ANOVA or Z-tests may be employed for further analysis and inference (Bhat & Darzi). In other cases, researchers may employ multivariate regression analysis (Li, Liu, Tan & Hu; Zhao, Xu & Wang). Alternatively, machine learning tools can be utilized to handle both regression and classification tasks efficiently (Cavalcante Siebert, Bianchi Filho, Silva Júnior, Kazumi Yamakawa & Catapan; Goode & Moutinho; Zeinalizadeh, Shojaie & Shariatmadari; Tama).

The importance of the customer satisfaction analysis is multiple: a) by focusing on customer preferences, companies can enhance their marketing capabilities. (Bhat & Darzi; Ming), b) By comprehending customer expectations, companies can guide their research and development teams (Haslinah, Mutmainnah & Jalil) to develop products or services that are more closely aligned with customer needs and desires.

In this paper, we present a novel machine learning-based tool for studying customer satisfaction. This tool holds potential utility for both marketing and research and development purposes. Consumer data collected, for instance, through custom questionnaires, can be analyzed using this tool to achieve several objectives. Firstly, it provides a ranking of the drivers that have the greatest impact on customer satisfaction metrics such as overall liking, overall rating, or overall satisfaction, along with measures of their respective impacts. Secondly, it enables the comparison of two different prototypes or products/services, highlighting their strengths and weaknesses.

Typically, questionnaires assess various variables related to the same aspect of a product or service. However, this can lead to difficulties in understanding the impact and importance of each aspect individually. Therefore, it is necessary to find a method to consider groups of drivers collectively. In our proposed machine learning-based tool, we address this by constructing clusters of drivers prior to analysis. This approach allows for a more comprehensive understanding of the interrelationships between different aspects of customer satisfaction.

The article is structured as follows: in Section 2 the proposed methodology is illustrated in detail and final remarks are discussed in Section 3.

2. Methodology

2.1. Data transformation

To assess the perception of a specific product/service in the market, or prior to its launch, customized questionnaires can be distributed to a sample of customers. In these questionnaires, various aspects are evaluated using typically a 5-point or 10-point Likert scale, or even a slider scale. Generally, the lower value is associated with disagreement, while the highest indicates full agreement.

In most questionnaires, multiple questions are typically linked to the same aspect (i.e. the design). This can complicate the interpretation of results, as it can be difficult to determine whether a particular aspect is valued by the consumer. Quantifying the impact of this aspect on overall satisfaction can also prove difficult. One solution is to create clusters of drivers to group variables that relate to a common aspect of the product/service. We use the K-means clustering. The K-means clustering is a method for grouping similar data points into clusters or groups and it is available in ClustOfVar R package (Chavent, Kuentz, Liquet & Saracco). The functioning is quite straightforward. Firstly, we need to choose a number k of clusters we want to find (i.e. the number of aspects). We then initialize k points, called centroids, which represent the centers of the clusters. Then an iterative relocation algorithm performs a partitionning of a set of variables and give as an output our clusters, our set of synthetic variables. The data points within each cluster are similar to each other, and the data points between clusters are different.

Another challenge to address involves how to handle all this data. Typically, simply calculating the mean value of a rating can lead to misleading interpretations (Jones & Sasser). A viable solution to obtain more reliable results could be to use a top box measure (Morgan & Rego). This idea is supported by the understanding that consumers who consistently rate products or services with the highest scores, often referred to as "top box" consumers, are typically the most loyal and valuable. They demonstrate a higher probability of making repeated purchases (Jones & Sasser). Therefore, the outcome score obtained from the cluster analysis will be firstly transformed into binary variables: TB (Top Box score) – Other (Other score). Usually, the positive scores of the synthetic variables derived from a clustering algorithm are categorized as TB (Top Box), while the negative scores are classified as Other. With the outcome variable Y now binary, we encounter a binary classification problem.

2.2. Variable impact

The primary objective of our methodology is to assess the influence of a driver, X_j on the binary response variable Y. We analyze all observations across two distinct datasets, D_{1j} and a D_{2j}, employing a suitable machine learning model in each to predict the probability of Y being in the TB category. D_{1j} is a hypothetical dataset derived from the original training set, where all actual levels of X_j are replaced with TB. Similarly, in D_{2j} is constructed by replacing the actual levels of X_j with Other. The predicted TB probability for observation i in dataset D_{kj} is denoted as $\hat{P}_{kji}$ and $\bar{P}_{kj}$ is the retrieved average probability. The difference $\bar{P}_{1j} - \bar{P}_{2j}$ denote the impact of X_j on Y.

To accurately predict Y, it's essential to employ a well-suited machine learning model that effectively captures all existing relationships within the data. Various machine learning models can be utilized and compared to determine the most effective one based on the insights provided by their performance metrics. When handling classification tasks, the AUC error metric enables us to assess the overall discriminatory power of the model (Bradley). To mitigate the risk of overfitting, it's crucial to employ appropriate resampling techniques, such as cross-validation or bootstrapping. In our framework, we opt for a 10-fold cross-validation approach for both hyperparameter tuning and model selection. We calculate the cross-validated AUC for each model, selecting the one with the highest value.

2.3. Area for advancement

Understanding how overall customer satisfaction would change if specific drivers, such as design, were improved can be of interest for a company. In simpler terms, this refers to measuring the enhancement in Y resulting from improving a specific input X_j. We refer to it as area for advancement.

Similarly to the previous section we consider two different datasets: the original training set T and D_{1j}. In each we predict the TB probability of Y. Be $\hat{P}_i$ the predicted TB probability for observation i in dataset T and $\hat{P}_{1ji}$ the predicted TB probability for observation i in dataset D_{1j}. Respectively, the average probabilities are denoted as $\bar{P}$ and $\bar{P}_{1j}$. We can retrieve the area for advancement in Y due to X_j as $\bar{P}_{1j} - \bar{P}$.

As before, there is the need to find out the best performing model.

We can illustrate the area for advancement for each impactful clustered-driver on overall customer satisfaction on the y-axis of a graph. On the x-axis, we can illustrate the gap for each of these impactful drivers, defined as the difference in terms of TB percentage for a specific driver between two versions of a product. This approach enables the comparison of two different products by highlighting their strengths and weaknesses, as well as the potential for improvement. Each driver in the graph can be represented as a bubble, with the radius of the bubble proportional to the size of the effect of the specific driver on overall customer satisfaction.

3. Conclusion

In this research, we introduced a machine learning-driven approach to conduct customer satisfaction analysis. Specifically, it offers valuable insights in several aspects. Firstly, it effectively identifies clusters of variables influencing customer satisfaction and quantifies their impact. Secondly, it facilitates product comparison, identifying strengths and weaknesses. Lastly, this methodology predicts the potential impact on overall customer satisfaction if specific product aspects were enhanced, thereby offering crucial guidance for future improvement efforts. A case study is currently underway.

References

[1] Bhat, Suhail Ahmad, & Mushtaq Ahmad Darzi. «Exploring the influence of consumer demographics on online purchase benefits». FIIB Business Review, vol. 8, fasc. 4, 2019, pp. 303–16.

[2] Bradley, Andrew P. «The use of the area under the ROC curve in the evaluation of machine learning algorithms». Pattern recognition, vol. 30, fasc. 7, 1997, pp. 1145–59.

[3] Cavalcante Siebert, L., Bianchi Filho, J. F., Silva Júnior, E. J. D., Kazumi Yamakawa, E., & Catapan, A., «Predicting customer satisfaction for distribution companies using machine learning». International Journal of Energy Sector Management, vol. 15, fasc. 4, 2021, pp. 743–64.

[4] Chavent, M., Kuentz, V., Liquet, B., & Saracco, L., ClustOfVar: An R package for the clustering of variables. arXiv preprint arXiv:1112.0295, 2011.

[5] D. Polding, Robert, & Maria Eizaguirre Dieguez. «An Investigation into the Effectiveness of Big Data in Organizations, the Use of Customer Data, and the Ethical Implications of the Data Economy». 2021 International Symposium on Electrical, Electronics and Information Engineering, 2021, pp. 599–607.
[6] Goode, Mark, & Luiz Moutinho. «The effects of consumers age on overall satisfaction: An application to financial services». Journal of Professional Services Marketing, vol. 13, fasc. 2, 1996, pp. 93–112.
[7] Haslinah, A., Mutmainnah, A., & Jalil, I. Z., «Product development analysis of dragon fruit praline chocolate by using Quality Function Deployment (QFD) method». AIP Conference Proceedings, vol. 2595, fasc. 1, AIP Publishing, 2023.
[8] Jones, Thomas O., & W. Earl Sasser. «Why satisfied customers defect». Harvard business review, vol. 73, fasc. 6, 1995, pp. 88-.
[9] KONDO, Yoshio. Customer satisfaction: how can I measure it?. Total Quality Management, 2001, 12.7-8: 867-872.
[10] Kracklauer, Alexander H., D. Quinn Mills, and Dirk Seifert. "Customer management as the origin of collaborative customer relationship management." Collaborative customer relationship management: taking CRM to the next level. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004. 3-6.
[11] Li, H., Liu, Y., Tan, C. W., & Hu, F., «Comprehending customer satisfaction with hotels: Data analysis of consumer-generated reviews». International Journal of Contemporary Hospitality Management, vol. 32, fasc. 5, 2020, pp. 1713–35.
[12] Mccoll-Kennedy, Janet; Schneider, Ursula. Measuring customer satisfaction: why, what and how. *Total quality management*, 2000, 11.7: 883-896.
[13] Ming, Gu. «Application research of customer big data analysis for online shop based on smart cloud platform tools». 2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA), IEEE, 2022, pp. 1142–45.
[14] Morgan, Neil A., & Lopo Leotte Rego. «The value of different customer satisfaction and loyalty metrics in predicting business performance». Marketing science, vol. 25, fasc. 5, 2006, pp. 426–39.
[15] Ngai, E. W., Xiu, L., & Chau, D. C., «Application of data mining techniques in customer relationship management: A literature review and classification». Expert systems with applications, vol. 36, fasc. 2, 2009, pp. 2592–602
[16] Oliver, Richard L. A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of marketing research*, 1980, 17.4: 460-469.
[17] Patterson, Paul G. «Expectations and product performance as determinants of satisfaction for a high-involvement purchase». Psychology & Marketing, vol. 10, fasc. 5, 1993, pp. 449–65.
[18] Tama, Bayu Adhi. «DATA MINING FOR PREDICTING CUSTOMER SATISFACTION IN FAST-FOOD RESTAURANT.» Journal of Theoretical & Applied Information Technology, vol. 75, fasc. 1, 2015.
[19] Yeung, Matthew CH, & Christine T. Ennew. «From customer satisfaction to profitability». Journal of strategic marketing, vol. 8, fasc. 4, 2000, pp. 313–26.
[20] Zeinalizadeh, N., Shojaie, A. A., & Shariatmadari, M., «Modeling and analysis of bank customer satisfaction using neural networks approach». International Journal of Bank Marketing, vol. 33, fasc. 6, 2015, pp. 717–32.
[21] Zhao, Y., Xu, X., & Wang, M., «Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews». International Journal of Hospitality Management, vol. 76, 2019, pp. 111–21.