Predicting a Customer Churn Risk Group with Customer Attributes

Kiwoon Hong
School of Information
kiwoon2799@naver.com

Abstract

This report addresses the heightened competition in the restaurant industry, emphasizing the significance of customer retention amid economic challenges and the proliferation of mobile platforms. Using data from the UCI Machine Learning Repository, the project employs logistic regression, support vector machines, random forest, and naive Bayes classifiers to predict customers at risk of giving low ratings.Logistic regression outperforms other techniques, showcasing superior accuracy and AUC values. Additionally, the analysis reveals a correlation between high customer drink levels and lower restaurant ratings. In essence, the project provides actionable insights for restaurants to proactively address potential customer churn and optimize marketing strategies based on rating patterns and drinking habits.

1 Introduction

The National Restaurant Association predicts a lasting change for the restaurant industry, stating that 2022 will mark a "new normal." Challenges in rebounding and intense competition for workers are expected to persist, as outlined in the association's recent 2022 State of the Restaurant Industry report.[1] In this competitive environment, customers can quickly find information about different restaurants through mobile apps and websites. As a result, the likelihood of a restaurant customer churning has increased, and restaurants need to anticipate it. Among the paramount considerations for restaurants in addressing customer churn, ratings emerge as a critical focal point. The influence of restaurant star ratings significantly outweighed other factors in customer decision-making, revealing that website visitors frequently influenced by the sheer volume of written reviews.[2]

Another aspect that restaurants should pay attention to is customer characteristics. Advances in machine learning and artificial intelligence have made it possible to analyze customer groups individually. Therefore, it is crucial to understand the rating distribution based on customer characteristics. Anticipating the risk of churn and providing personalized marketing can prevent customer defection, leading to customer retention. Retaining existing customers is the least expensive way to maximize revenue on a recurring basis.[3] The prediction of ratings based on customer profiles, including budget, family information, interests, age group, etc., can serve as a pivotal tool in this regard.

In this study, to address this issue, customers were categorized into churn-risk and non-churn-risk groups based on their ratings. Machine learning classifiers are effective at categorizing these risk groups, and can also identify customer characteristics that have a significant impact on churn. In this study, the following machine learning classifiers were employed: (1) logistic regression, (2) linear support vector machine, (3) random forest, and (4) naive Bayes. To select the optimal model, the performance of each model was evaluated. Following the employment of the models, overfitting was mitigated using Leave-One-Out Cross-Validation (LOOCV). This study will serve as an exploratory investigation employing various models to attempt classification based on customer profiles.

2 Related Work

There are a lot of previous works on customer churn prediction using machine learning methods. On the methodological side, Xia[4] and one other researcher applied Support Vector Machines (SVM) for predicting customer churn. They conducted a comparative analysis with cases of customer churn prediction in domestic and international airlines, comparing it with artificial neural networks, decision trees, logistic regression, naive Bayes classifiers, and others. The results revealed that this approach exhibited superior performance in terms of accuracy, precision, recall, and F1 score, making it an effective method for predicting customer churn. Also, Malikireddy[5] and another researcher addressed the application of ensemble classifiers, specifically bagging and boosting in Random Forest, for predicting customer churn, highlighting the potential for achieving higher performance. However, they also discussed the issue of overfitting associated with these methods.

There has been a consistent association between low satisfaction and negative evaluations. and Raza[6] found that dissatisfaction with healthcare services, staff behavior, and profit-driven attitudes are important factors in low ratings of hospitals. This could be a link between ratings and customer churn.

Predicting restaurant ratings is also an active area of research. Kaviya[7] and other researchers designed an emotion analysis system for automatic restaurant evaluation, aiming to assist people in choosing restaurants they would enjoy. The sentiment analysis system for restaurant reviews evaluates restaurants based on user-provided feedback. Kulkarni[8] and other researchers predicted restaurant ratings by considering various factors such as reviews, location, average cost, cuisine type, and restaurant category. In this paper, various prediction models, including Support Vector Machine (SVM), Random Forest, Linear Regression, XGBoost, and Decision Tree, were employed.

Numerous other previous studies focus on text mining of customer reviews. Therefore, this study takes on an exploratory nature regarding rating prediction based on customer profiles, signifying the potential for proactive marketing actions in response.

3 Procedure

3.1 Representation

I utilized the "Restaurant & customer Data" [9] provided by the UCI Machine Learning Repository. This dataset was obtained from a recommendation system prototype and was employed for the task of generating the top n restaurants based on customer preferences. The dataset consists of nine parts: three restaurant-related datasets, one geographic dataset, one overall rating dataset, and three user-related datasets. Since this study is interested in the characteristics of customers, I utilized the datasets 'userprofile' for their profile, 'userpayment' for their payment method, and 'rating' for their output. I arbitrarily selected features from the userprofile dataset that could directly affect user ratings, and combined them with the payment methods of customers from the userpayment dataset to complete the input features.

Table 1: Processed input features

Features	**Type**	**Example**
Upayment	Categorical	cash
smoker	Categorical	true
drink_level	Categorical	social drinker
dress_preference	Categorical	formal
ambience	Categorical	family
transport	Categorical	on foot
marital_status	Categorical	single
hijos	Categorical	independent
age	Numeric	23
interest	Categorical	technology
personality	Categorical	hard-worker
budget	Categorical	low

Table 1 above represents the features of the input data. Each categorical feature underwent label encoding or ordinal encoding as described in the paper. The original feature name for "age" is "birth_year," and it has been transformed to represent the age adjusted to the year of data creation, which is 2012. Due to the predominance of categorical features in the input features, there are limitations. When a category has more than three values, one-hot encoding is preferable. However, to prevent an excessive increase in the number of features, as five or more features have three or more values each, label encoding was applied. This approach, while mitigating the issue of an overwhelming number of features, has its limitation in potentially introducing biases in setting weights.

Risk of churn, the output of the model, was designed based on the ratings. From the 'rating_final' dataset, ratings were extracted for each user. In cases where a user provided multiple reviews, the average value of those reviews was used. Figure 1 below represents a distribution of average ratings among users divided into intervals. Subsequently, the group with average ratings between 0 (inclusive) and 1 (exclusive) was classified as the 'churn risk group' for binary classification. The final setup resulted in two binary output groups: 'churn' and 'normal'. Figure 2 illustrates the distribution of the 'normal' group and the 'churn' group.

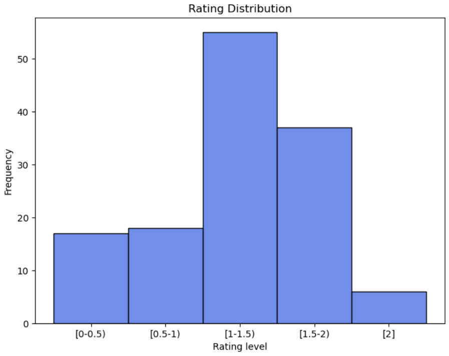

Figure 1: Distribution of customer ratings

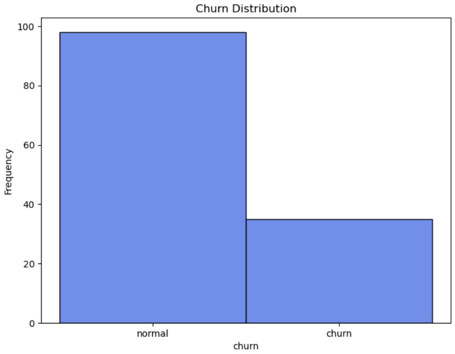

Figure 2: Distribution of a churn risk group

After selecting features from the data, I obtained 132 customer records. Subsequently, by removing NaN (Not a Number) values, I ended up with a final dataset of 119 entries. Due to the limited size of the dataset, concerns about overfitting arose. To address this, I evaluated the performance of the first model using three different methods: splitting the dataset into training, validation, and test sets with a ratio of 70%, 15%, 15%; employing K-fold cross-validation; and utilizing Leave-One-Out Cross-Validation (LOOCV). After assessing the potential for overfitting, I opted to adopt only LOOCV for the final evaluation.

3.2 Algorithm/equations

In this work, four classification methods were employed based on related work. Starting with a fundamental model, the goal was to explore various models, progressively applying them to the task.

3.2.1 Logistic Regression

As the baseline for this study, the logistic regression algorithm was employed. Logistic regression is a statistical method commonly used for binary classification tasks where the goal is to predict the probability of an instance belonging to a particular class.[10] Unlike linear regression, which predicts a continuous output, logistic regression uses a logistic function (sigmoid function) to map the predicted value to a range between 0 and 1. This output is then interpreted as the probability that an instance belongs to a positive class.The logistic function, also known as the sigmoid function, is expressed as follows:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

In this work, the initial approach involved employing logistic regression, a fundamental and well-suited model for datasets with limited size. The evaluation of the model's performance utilized three assessment techniques: the train-validation-test split, K-fold cross-validation, and Leave-One-Out Cross-Validation (LOOCV).

3.2.2 Linear Support Vector Machine

One of the objectives of this work is to explore a variety of models, and therefore, Support Vector Machine (SVM) was chosen as the next model. SVM is well-suited for datasets with numerous features, making it a fitting choice for the subsequent model. The primary goal of the SVM algorithm is to discover the most effective hyperplane within an N-dimensional space for separating data points into distinct classes in the feature space. This hyperplane is optimized to maximize the margin between the closest points of different classes. The dimension of the hyperplane is determined by the number of features in the dataset.[11] In this work, a linear kernel was employed. The linear kernel function for SVM is as follows:

$$K(w, b) = w^T x + b$$

3.2.3 Random Forest

The choice of the next model was determined to be Random Forest based on the findings in related work, which suggested its potential high performance in predicting customer churn.[5] The Random Forest algorithm is an extension of bagging, utilizing both bagging and feature randomness to create an ensemble of decision trees with uncorrelated relationships. Feature randomness, also known as 'random subspace method' or 'feature bagging,' ensures a low correlation between decision trees by generating random subsets of features.[12] Furthermore, this method is stable as it involves ensemble learning with multiple trees, and it allows for the visualization of Variable Importance for each feature. In this work, the number of decision trees in the random forest was set to 100.

3.2.4 Naive Bayes

The last model attempted was the Naive Bayes model. Since the features in the dataset were label-encoded, posing a risk, a model that disregards the relationships between variables was needed. The Naive Bayes assumption posits that the D-dimensional class-conditional distributions can be decomposed into a product of D univariate distributions.[13] For a given feature vector X, the conditional probability of class C is decomposed as follows:

$$P(X|C) = P(x_1|C) \cdot P(x_2|C) \cdot \ldots \cdot P(x_D|C)$$

3.3 Summary Figure

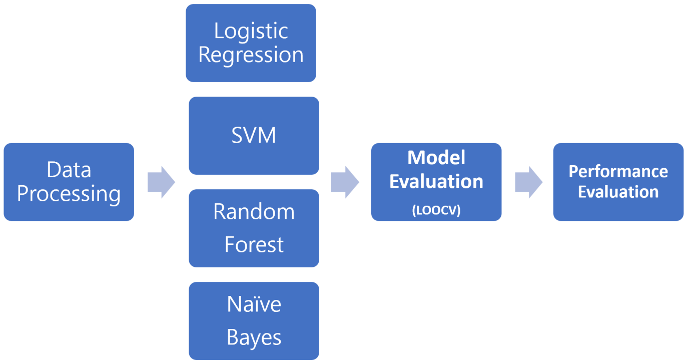

Figure 3: The summary figure of the project

In Figure 3, following model deployment, LOOCV was ultimately adopted as the method for generalization, and it was executed to assess the performance of each model. The performance evaluation included measurements of accuracy, ROC curve, and AUC value.

4 Evaluation

4.1 Metrics

I utilized LOOCV to train each model, measured the accuracy of each model, and evaluated the model performance by visualizing ROC curves and AUC values.

4.2 Individual Models

I initially measured the accuracy and AUC value of each model, and the results are presented in the table below.

Table 2: The accuracy and AUC value of indivudual model

Model	**Accuracy**	**AUC**
Logistic Regression	**0.7647**	**0.61**
Linear SVM	0.7479	0.59
Random Forest	0.7479	0.58
Naive Bayes	0.7143	0.59

Although each model does not exhibit a significant performance difference, both accuracy and AUC values slightly favor Logistic Regression. The following figures depict the ROC curves for each model.

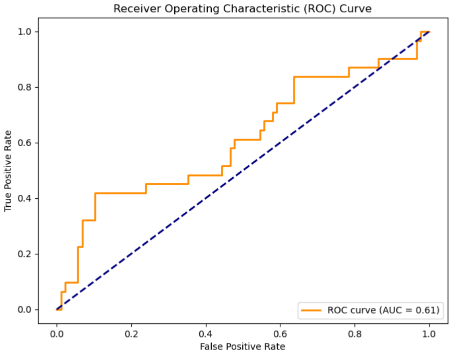

Figure 4: ROC curve of logistic regression

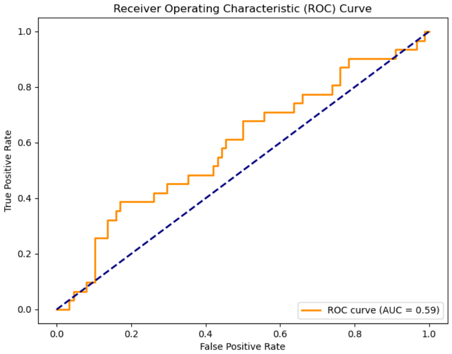

Figure 5: ROC curve of SVM

5 Results

5.1 Best Model

As visualized in Section 4, my baseline model, Logistic regression, exhibited the best performance with an accuracy of 0.7647. SVM and Random Forest showed similar performances, while the Naive Bayes model demonstrated the lowest performance, suggesting dependencies among input features. Various interpretations can be drawn from the superior performance of the Logistic Regression model. The simplicity of the data structure, dominated by categorical features, may have facilitated effective linear separation. [14] Additionally, the project's limitation of not undergoing model tuning could have favored simpler models, contributing to their impressive performance. Furthermore, due to the limited amount of data and its straightforward structure, I identified a limitation where the accuracy of the SVM and Random forest models coincided.

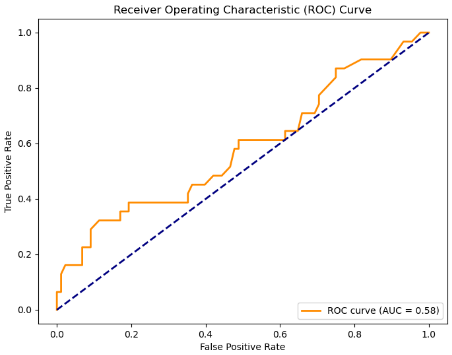

Figure 6: ROC curve of random forest

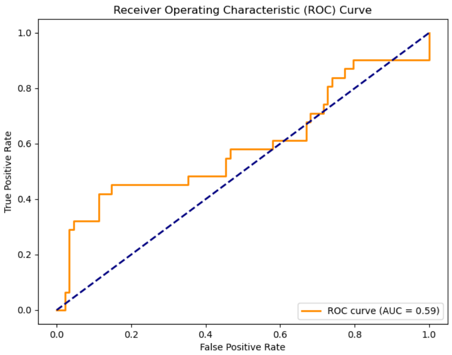

Figure 7: ROC curve of Naive bayes

5.2 Features

By using the coefficients of Logistic Regression and SVM, as well as the feature importance of Random Forest, I can explore which features contribute to low ratings and, consequently, form the risk group of customer churn.The following figures visualize these findings.

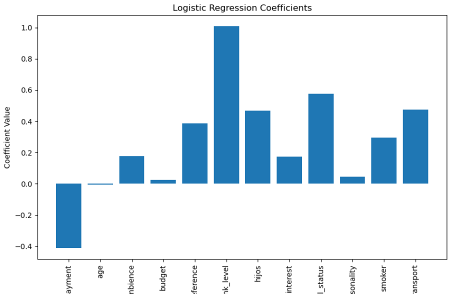

Figure 8: Coefficients of Logistic regression

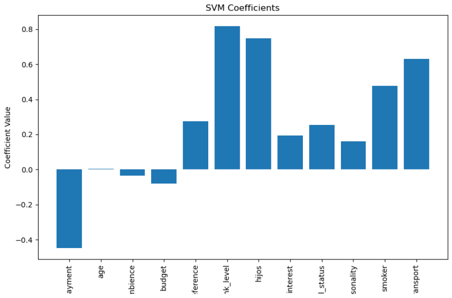

Figure 9: Coefficients of linear SVM

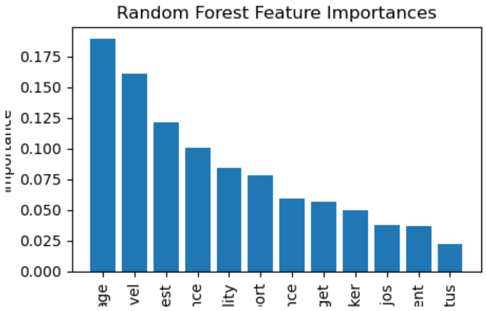

Figure 10: Feature importance of random forest

Among the three figures, the most notable factor is the drink level. It suggests that higher drink levels are likely to be associated with lower ratings. Additionally, the likelihood of giving lower ratings decreases with multiple payments, which could be natural as it implies repeated visits to the same restaurant.

6 Conclusion

A logistic regression model performed best at predicting customer churn. Moreover, each model indicates that specific features may play a crucial role in the prediction. Customers give different ratings based on their drinking preferences suggests that understanding customer drinking levels could be a valuable approach for formulating proactive marketing strategies.

However, this work has limitations, including the use of a small generative dataset, a limited number of numeric variables, and the absence of one-hot encoding. Additionally, the lack of hyperparameter tuning for the models is another limitation. Nevertheless, the exploratory nature of predicting ratings based on customer characteristics is valuable, and experimenting with various models provided insights.

In future work, I will explore customer churn prediction using more general customer characteristics and extend industries from restaurant to others.

References

[1] Vanessa Yurkevich. The restaurant business will probably never recover from Covid | CNN Business — cnn.com. `https://www.cnn.com/2022/02/02/business/restaurants-pandemic-effects/index.html`. [Accessed 08-12-2023].

[2] Michael Luca. Reviews, reputation, and revenue: The case of yelp.com. *Harvard Business School NOM Unit Working Paper*, 12(016), 2016.

[3] Praveen Lalwani, Manoj Kumar Mishra, Jyoti S Chadha, and Pankaj Sethi. Customer churn prediction system: a machine learning approach. *Computing*, 103(10):3057–3081, 2021.

[4] Guo-en Xia and Wei-dong Jin. Model of customer churn prediction on support vector machine. *Systems Engineering - Theory & Practice*, 28(1):71–77, 2008.

[5] Venkata Pullareddy Malikireddy and Madhavi Kasa. Customer churns prediction model based on machine learning techniques: A systematic review. 2021.

[6] Arif Raza and Ranjit Dehury. Dissatisfaction factors that influence customers to give low online rating to hospitals. *Asia Pacific Journal of Health Management*, 16(3), 2021.

[7] K Kaviya, C Roshini, V Vaidhehi, and J Dhalia Sweetlin. Sentiment analysis for restaurant rating. In *2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM)*, pages 89–93. IEEE, 2017.

[8] Atharva Kulkarni, Divya Bhandari, and Sachin Bhoite. Restaurants rating prediction using machine learning algorithms. *International Journal of Computer Applications*, 175(16):7–11, 2020.

[9] Rafael Medelln and Juan Serna. Restaurant consumer data. UCI Machine Learning Repository, 2012. DOI: https://doi.org/10.24432/C5DP41.

[10] Wikipedia. Logistic regression — Wikipedia, the free encyclopedia, 2023. [Online; accessed 7-December-2023].

[11] Support Vector Machine (SVM) Algorithm - GeeksforGeeks — geeksforgeeks.org. `https://www.geeksforgeeks.org/support-vector-machine-algorithm/`. [Accessed 08-12-2023].

[12] What is Random Forest? | IBM — ibm.com. `https://www.ibm.com/topics/random-forest`. [Accessed 08-12-2023].

[13] Simon Rogers and Mark Girolami. *A First Course in Machine Learning, Second Edition*. CRC Press, 2016.

[14] Advantages and Disadvantages of Logistic Regression - GeeksforGeeks — geeksforgeeks.org. `https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/`. [Accessed 08-12-2023].