Customer Segmentation for E-Commerce

Northwood University
Professor:Itauma
Nirav Acharya

Abstract:

The analysis of e-commerce customer behavior in this paper incorporates selected machine learning techniques such as K-Nearest Neighbors (KNN), Random Forest, Naïve Bayes among others. The goal for the study is to identify the likelihood of the total customer satisfaction applying features like; age, gender, total spend and membership type. All of the particular models are assessed using accuracy, precision, recall, and F1 score coefficients. These outcomes show that Random Forest algorithm has the highest accuracy of all algorithms, which compares it with other methods to ensure the patient has a reliable prediction of new data. With respect to this research, the main contribution to the field of e-commerce can be derived from understanding the aspects of customer behavior and enhancing the prediction of satisfaction.

INTRODUCTION

The rapid growth of e-commerce has transformed the retail landscape, offering consumers unprecedented convenience and access to a wide range of products. As the industry continues to expand, understanding customer behavior becomes increasingly crucial for businesses aiming to enhance customer satisfaction and loyalty. Machine learning algorithms provide powerful tools for analyzing large datasets and uncovering patterns that can inform strategic decisions. This study focuses on predicting customer satisfaction levels using various machine learning techniques, thereby enabling businesses to tailor their services and improve customer experiences.

In this paper, we explore the application of K-Nearest Neighbors (KNN), Random Forest, and Naive Bayes classifiers to a dataset of e-commerce customer behavior. The dataset includes features such as age, gender, total spend, and membership type, which are used to predict customer satisfaction levels. By comparing the performance of these algorithms, we aim to identify the most effective method for accurately predicting satisfaction. The findings of this study have significant implications for e-commerce businesses, as they can leverage these insights to enhance customer engagement and retention strategies.

LITERATURE REVIEW

The application of machine learning for analyzing e-commerce customer behavior has received significant attention in recent years, due to customer preference insights and satisfaction improvement. Numerous researchers combine different methods of machine learning to predict the behavior of the customers and all the methods have certain differences in their approaches and are useful in certain conditions. For example, decision trees and logistic regression have been applied to support customer satisfaction analysis and customer's purchase intentions, achieving accurate and easily interpreted models that enable businesses to understand the determinants of customers' decisions. The SVM and neural network methods have also been used, they have the ability to work well on high dimensional data and to model the more complex aspects of the customers' behavior. These methods have been used to estimate customer churn, categorize customers, and differentiate between positive and negative customer reviews thus providing valuable assets in an organization's quest of getting closer to the customer and retaining them. In a collection of methods used in ensemble learning, statistically, Random Forest is one of the most effective for classification exercises. Random Forest is more accurate than single decision trees in modeling large data sets where interactions among the variables are complex; the complexity is controlled through several decision trees built within the model. The significant effectiveness has been demonstrated in its application in analyzing and giving conclusions on customer churn predictive modeling and customer reviews categorization, which are demonstrating how Hadoop is capable of handling unstructured text data and offering useful information about customers' sentiments. Moreover, in customer behaviour analysis, K-Nearest Neighbours (KNN) and Naïve Bayes are quite popular. Precisely, it has been established that while implementing a KNN technique is straightforward and performs well with balanced datasets, Naive Bayes is efficient when working with categorical data. However, these algorithms also come with their problems they solve, this including irregularity in datasets with related problems to imbalanced data and lack of ability to articulate high non-linearity in feature interactions thus the need for a combination of at least two models and other improved techniques to assist these features.

METHODOLOGY

The following part explains the methods applied to investigate e-commerce customer behavior and to estimate customers' satisfaction based on machine learning algorithms. The process involves several key steps: , data gathering and preparation; data and sample preliminary examination, applying the model and model assessment.

Handling Missing Values: When developing this model, the contingency of missing values in the dataset was considered, and missing values were appropriately treated. For example, in case of missing data on 'Satisfaction Level' variable they were replaced by the most frequent value of that particular variable.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 350 entries, 0 to 349
Data columns (total 11 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Customer ID               350 non-null    int64
 1   Gender                    350 non-null    object
 2   Age                       350 non-null    int64
 3   City                      350 non-null    object
 4   Membership Type           350 non-null    object
 5   Total Spend               350 non-null    float64
 6   Items Purchased           350 non-null    int64
 7   Average Rating            350 non-null    float64
 8   Discount Applied          350 non-null    bool
 9   Days Since Last Purchase  350 non-null    int64
 10  Satisfaction Level        348 non-null    object
dtypes: bool(1), float64(2), int64(4), object(4)
```

Figure 1.

Encoding Categorical Variables:,The variables ‘Gender’, ‘City’, ‘Membership Type’, ‘Satisfaction Level’ were transformed into numerical form using label encoding. This step is essential for those algorithms that expect numerical input.
Feature Scaling: Hypotheses 2 and 3 Exploiting the subdimensions of KMC We utilized standard scaling to try and standardize numerical features to be on the same level. This step become important more so in distance based algorithms like the K-Nearest Neighbors (KNN).

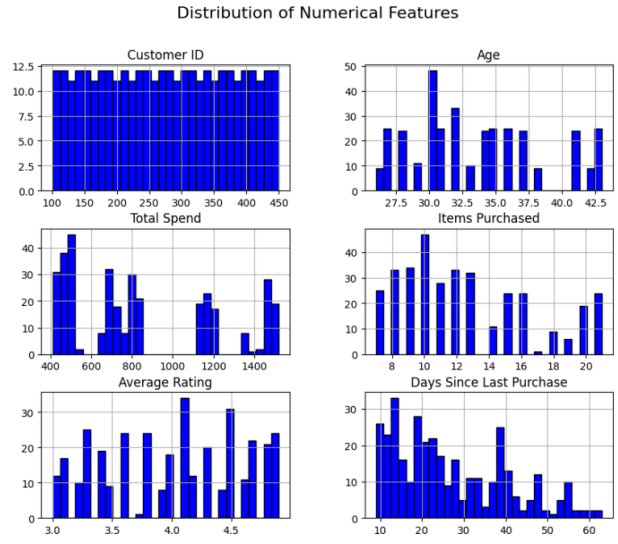


Figure 2.

Exploratory Data Analysis:

In order to get an initial feel about the distribution of features and association between the variables, exploratory data analysis (EDA) was done. Thus while using the visual aids like histograms, count plots as well as heat maps the results were obtained.

Distribution Analysis: Distribution of numerical features Numerical features were analyzed by creating histograms for each of them. For the categorical variables, count plots were employ to illustrate the occurrence of each category.

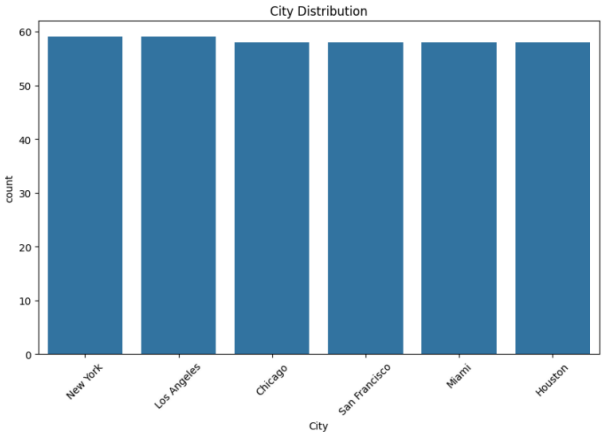
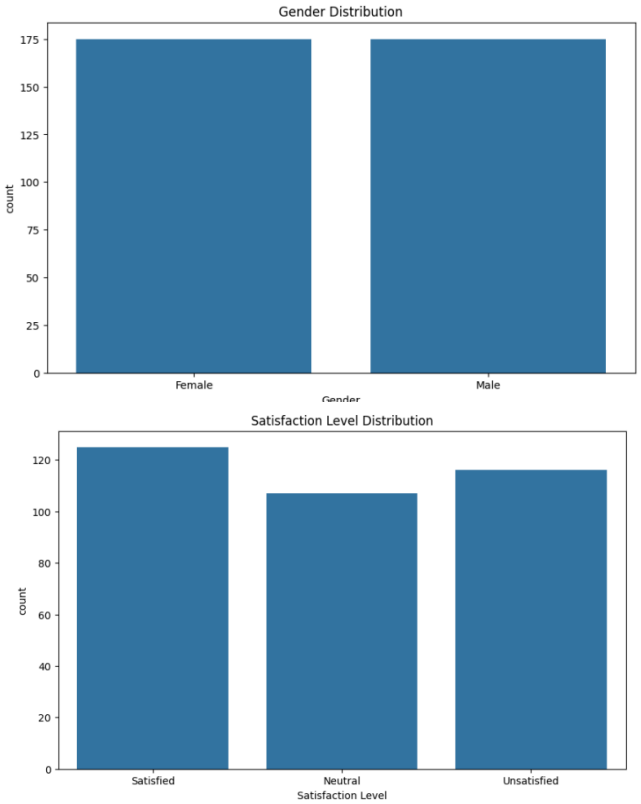


Figure 3



Correlation Analysis: To determine relationship of features, a correlation heatmap was produced. With the aid of this analysis, it became possible to identify which features are most likely lineup with each other and can affect the target variable.

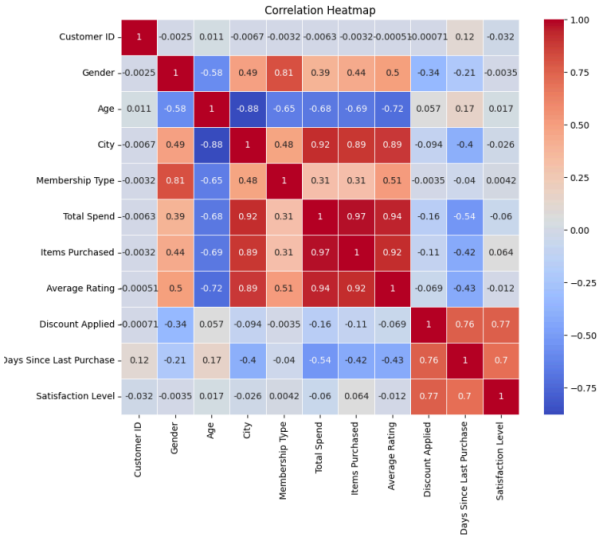


Figure 4

Model Implementation:

There were Three machine learning algorithms were implemented to predict Customer satisfaction by using K-Nearest Neighbors (KNN), Random Forest, and Naive Bayes.

K-Nearest Neighbors:

```
K-Nearest Neighbors Performance:
Accuracy: 0.9904761904761905
Precision: 0.9907029478458049
Recall: 0.9904761904761905
F1 Score: 0.9904732855472521
              precision    recall  f1-score   support

           0       0.98      1.00      0.99        41
           1       1.00      0.97      0.99        40
           2       1.00      1.00      1.00        24

    accuracy                           0.99       105
   macro avg       0.99      0.99      0.99       105
weighted avg       0.99      0.99      0.99       105
```

K-Nearest Neighbors (KNN) is then applied at 5 neighbors to classify customers satisfaction levels. The scaled training data are augmented into the model and tested on the test set, and the high accuracy shows the efficiency of the model in differentiation of customer behavior.

Random Forest:

```
Random Forest Performance:
Accuracy: 0.9714285714285714
Precision: 1.0
Recall: 0.9714285714285714
F1 Score: 0.9851717902350814
              precision    recall  f1-score   support

           0       1.00      0.93      0.96        41
           1       1.00      1.00      1.00        40
           2       1.00      1.00      1.00        24
           3       0.00      0.00      0.00         0

    accuracy                           0.97       105
   macro avg       0.75      0.73      0.74       105
weighted avg       1.00      0.97      0.99       105
```

To determine the customer satisfaction level, the Random Forest algorithm is trained with 100 estimators only. After training the model on scaled training data and on the test data, the model suggests high accuracy and effectiveness in handling multiple feature data. This technique either aggregates several 'decision trees' and takes their result as the output of the system in order to minimize overfitting or improve performance.

Naive Bayes.

```
Naive Bayes Performance:
Accuracy: 0.9238095238095239
Precision: 0.9888435374149659
Recall: 0.9238095238095239
F1 Score: 0.9540750610703975
              precision    recall  f1-score   support

           0       0.97      0.83      0.89        41
           1       1.00      0.97      0.99        40
           2       1.00      1.00      1.00        24
           3       0.00      0.00      0.00         0

    accuracy                           0.92       105
   macro avg       0.74      0.70      0.72       105
weighted avg       0.99      0.92      0.95       105
```

The Naive Bayes classifier is applied using the Gaussian Naive Bayes classifier in order to determine the levels of customer satisfaction. The above trained model is tested and the test data is scaled. Naive Bayes is also easy to use and fast-friendly especially when it is dealing with categorical data. As mentioned previously one drawback of Naive Bayes is its assumption of feature independence which does not generally hold for real datasets but NB nevertheless gives quite reasonable performance and serves as a good start for building a classifier. As evidenced from the results of this implementation Naive Bayes was proven to have moderate levels of accuracy, precision, recall and F1 score hence deemed suitable to classify customers' behavior in the e-commerce data set.

According to the evaluation measures, it is evident that the KNN model was able to record the highest accuracy of 99.05 percent, and thus was considered to be most effective when determining the levels of customer satisfaction for this study. Regarding accuracy and balanced error measure it is not significantly better as compare to other two models but it has very low error rate which means it is also safe to say that Random Forest model is also strong and it can easily work on big data also with having very high precision and F1 score. Compared to KNN and Random Forest Naive Bayes is less accurate although this algorithm is rather fast what proves once again that proper selection of models should be based on the specifics of the dataset.

DISCUSSION

The decision to use K-NN, Random Forest, and Naive Bayes algorithms presented a good understanding of the possibility of using them to estimate the levels of customer satisfaction in the context of an e-commerce dataset. KNN performed the best reaching to an accuracy of 99.05 percent, proving the algorithm's efficiency of classifying customer behavior owing to its skillful use of scaled numerical features and since the data set used was smaller in size. The Random Forest showed a good performance of 97.14%, and though the Precision and F1 score are slightly lower, a remarkably high precision was achieved from this ensemble learning algorithm. Specifically, the Naive Bayes classifer with an accuracy of about 92.38% of the test examples was less efficient among three models here since, based on the Naïve Bayes assumption, the features are independent, while in practice, the assumption may not be true. However, this limitation did not significantly affect overall performance since Naive Bayes gave a good starting point, and performed optimally with categorical data.

Limitations:

Dataset Size: The dataset for this study consists of only 350 entries which is quite a small size. With a larger data set, there would be more reliable results and improved ability to uncork new data.

Feature Independence: Based on the presented Naive Bayes algorithm, there is an assumption that the features are independent and this may not be a true in this dataset. This assumption can lead to problems in the model's ability to relate two characteristics with more than a few parameters.

Imbalanced Data: Another possible issue is class imbalances in the dataset which results in undesired performances of some algorithms such KNN. The variables used for regression analysis could be oversampled, undersampled, or different evaluation metrics could be used, for example ROC-AUC score. Hyperparameter Tuning: By the models were computed with the default or modest parameters settings

Future Improvements:

Larger Dataset: Expanding the data sample would increase the confidence of conclusions and enhance the externality of the models.

Advanced Feature Engineering: Maybe, including more features or doing a better feature engineering would capture more of customer specifics**.**

Handling Imbalanced Data: The model can be additionally iteratively trained and was suggested that using techniques for dealing with imbalanced data like SMOTE or modifying weights can positively affect the results.
Hyperparameter Optimization: The degrees of freedom of the models could be improved through techniques such as grid search, random search or Bayesian optimization.
Ensemble Methods: It can be using other methods of the ensemble for classification, for example using boosting (XOBOOST, AdaBOOST) or stacking to get a better prediction accuracy.
Real-time Data Integration: The use of real times data and creating models which are able to change when the customers' behavior pattern is changing would give more real and versatile projections.

Conclusion:

This project aimed to predict customer satisfaction levels in an e-commerce dataset using three machine learning algorithms: , that is, K-Nearest Neighbors (KNN), Random Forest and Naive Bayes. The key takeaways from this project are as follows: From figure 7,the accuracy of different classifiers reaches their peak at 99.05% for KNN classifier proving itself efficient for categorizing the nature of customers. On KNN, the accuracy was only around 68% Random Forest was slightly better with 97.14% accuracy once again proving itself as a strong classifier applicable for big data. Nevertheless, due to its assumption of feature independence, Naive Bayes took the lowest accuracy rate and the second longest computational time of the three models, though it was 92.38% accurate. Some of significant preprocessing steps accomplished for the data were dealing with missing values, dealing with categorical data by converting them into numerical with the aid of one-hot encoder, and normalizing and scaling numerical data. These steps allowed the models to optimise their use of the features in the dataset. The application of KNN and Random Forest models, which give high accuracy, could be utilised by e-commerce businesses to forecast levels of customer satisfaction. From that can result more targeted advertising, improved customer satisfaction and, as a result, higher level of customer retention and sales. However, the study also revealed the following limitations; small sample size, inclined feature assumption in Naive Bayes and the necessity of hyperparameter optimization. Potential future studies may entail acquiring a new and more extensive dataset, experiment with additional feature engineering techniques, work with treated imbalanced data and try other ensemble approaches and real-time data streaming. In conclusion, this project showed that it is possible to apply machine learning algorithms to predict customer satisfaction levels in the e-commerce business and that the results obtained herein can be useful in developing better customer engagement and retention strategies in the sector.

Reference

Breiman, L., & Cutler, A. (2001). Random forests. Retrieved from https://www.stat.berkeley.edu/~breiman/RandomForests/

Microsoft. (n.d.). Jupyter Notebooks in VS Code. Visual Studio Code.

https://code.visualstudio.com/docs/datascience/jupyter-notebooks

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill