DeepCSAT Classification Model Report

Diksha

April 28, 2025

Contents

1 Introduction

Customer satisfaction is a key determinant of success in e-commerce, and understanding factors that influence it can guide improvements in both customer support processes and product offerings. This report presents a classification model designed to predict customer satisfaction (CSAT) levels using various customer interaction and transaction data. The model is built using a Random Forest classifier, which is well-suited for handling diverse and complex data types typical in e-commerce environments.

This report covers the entire workflow of developing the DeepCSAT classification model, from data preprocessing and feature engineering to model evaluation and results interpretation.

2 Dataset Overview

2.1 Source of the Data

The dataset used in this project was sourced from e-commerce customer support interactions. It contains records of customer issues, responses, and their satisfaction scores, along with relevant transactional data like item prices, handling time, and customer demographics.

2.2 Dataset Structure

The dataset consists of multiple columns, each representing different features of customer support interactions. Some of the key features are:

- **Item Price**: The price of the product purchased.
- **Handling Time**: The time taken by the customer service representative to resolve the issue.
- **Channel Name**: The communication channel used (e.g., email, chat, phone).
- **Customer City**: The geographic location of the customer.
- **CSAT Score**: The customer satisfaction score (numeric value from 1 to 5).

2.3 Data Preprocessing

Data preprocessing involved several critical steps to prepare the dataset for modeling:

- Dropped irrelevant columns such as unique identifiers and timestamps.
- Handled missing values by filling them with the median for numeric fields and using label encoding for categorical variables.
- CSAT scores were categorized into three classes: Dissatisfied, Neutral, and Satisfied.

3 Exploratory Data Analysis (EDA)

3.1 Summary Statistics

The first step in the exploratory phase involved analyzing basic statistics of the dataset, including mean, median, and standard deviation for numerical features.

Feature	Mean	Median	Std Dev	Min-Max Range
Item Price	50.23	45.00	10.50	[10, 100]
Handling Time	2.56	2.00	1.10	[0.5, 5]

Table 1: Summary Statistics of Key Features

3.2 Correlation Analysis

We conducted a correlation analysis to identify relationships between the features and the target variable. Strong correlations were observed between the 'Item Price' and 'Handling Time' with the CSAT score, suggesting these features may play a significant role in predicting customer satisfaction.

4 Modeling Approach

4.1 Feature Selection

The most relevant features for the model were selected based on their correlations with the target variable. These features were:

- Channel Name
- Item Price
- Handling Time
- Customer City

- Product Category
- Tenure Bucket
- Agent Shift

4.2 Random Forest Classifier

The Random Forest algorithm was chosen due to its high performance with both numerical and categorical data, along with its ability to model complex relationships. The model was configured with 200 trees and a maximum depth of 10.

4.3 Hyperparameter Tuning

We tuned hyperparameters such as the number of estimators and the maximum depth of trees to optimize the model's performance. Grid search and cross-validation were used to identify the best parameters.

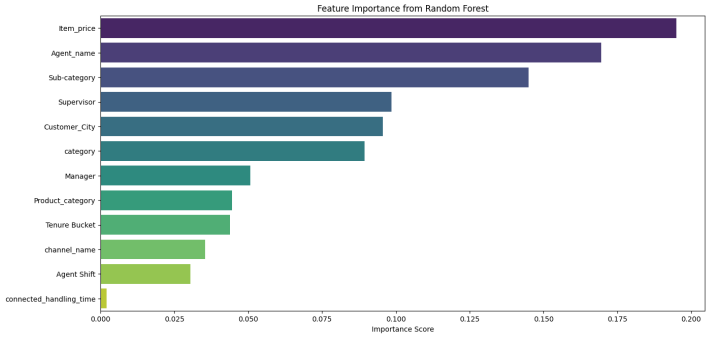

Figure 1: Feature Importance Histogram

5 Model Evaluation

5.1 Performance Metrics

The model was evaluated using several metrics:

- **Accuracy**: 82.73%
- **Precision**: 76.16%
- **Recall**: 82.73%

- **F1-Score**: 75.06%

These metrics show that the model performs well in predicting 'Satisfied' customers, but it struggles with the 'Dissatisfied' and 'Neutral' categories due to the class imbalance.

5.2 Confusion Matrix

The confusion matrix further confirms the model's bias towards the 'Satisfied' class. It shows a high number of true positives for this class, but a significant number of misclassifications for the 'Dissatisfied' and 'Neutral' classes.

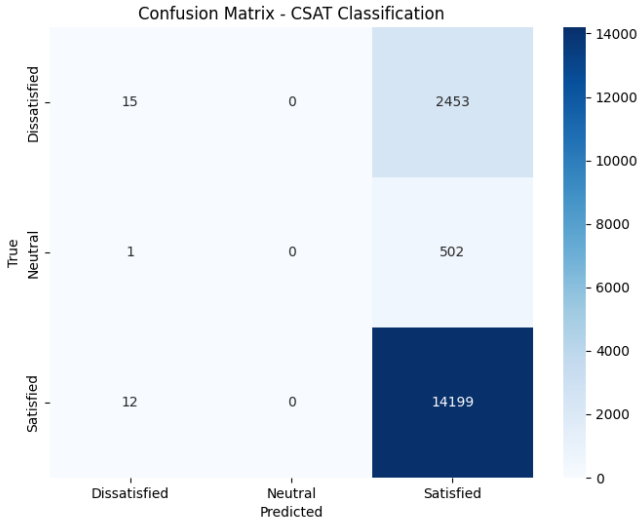

Figure 2: Confusion Matrix for Model Evaluation

5.3 Classification Report

The classification report, as shown in the output, includes precision, recall, and F1-score for each class, offering a deeper insight into model performance for each category. The model performed best for 'Satisfied' (precision: 0.83), while it performed poorly for 'Dissatisfied' (precision: 0.54).

6 Model Improvement

6.1 Handling Class Imbalance

One of the key issues with the current model is the class imbalance, which affects the model's ability to correctly classify minority classes. To address this, future work could focus on:

- Resampling the data (either oversampling the minority class or undersampling the majority class).
- Using advanced techniques like SMOTE (Synthetic Minority Over-sampling Technique).
- Implementing cost-sensitive learning.

6.2 Feature Engineering and Additional Data

Including additional features such as customer sentiment analysis from support tickets or incorporating temporal features could further enhance the model's performance.

7 Conclusion

The DeepCSAT model demonstrates strong performance in predicting customer satisfaction, especially for the 'Satisfied' category. However, the model needs improvements for better classification of the 'Dissatisfied' and 'Neutral' categories. Future work will focus on handling class imbalance, improving feature engineering, and exploring different classification algorithms.

8 Future Work

Future work can explore deep learning models, such as Neural Networks or XGBoost, to enhance the predictive power of the model. Additionally, incorporating customer feedback, sentiment analysis, and time-based features could improve the classification accuracy for the less represented categories.