Customer Satisfaction Prediction Using Machine Learning

Abstract

Customer satisfaction is a critical factor influencing business success, particularly in sectors such as e-commerce, telecommunications, and IT services. Accurately predicting customer satisfaction allows businesses to proactively address issues, enhance customer experiences, and maintain competitive advantages. This project aims to develop a machine learning model that predicts customer satisfaction ratings based on support ticket data, response and resolution times, priority levels, and other behavioral and demographic features. This report details the comprehensive workflow including preprocessing, feature engineering, exploratory data analysis, model training, evaluation, and interpretation.

1. Introduction

In today's digital economy, customer service interactions have become a goldmine for behavioral insights. Ticket resolution speed, communication channels, priority tagging, and customer demographics all contribute to satisfaction levels. Leveraging these parameters, our goal was to build an interpretable and highly accurate predictive model to forecast whether a customer will be satisfied or dissatisfied based on historical ticket records.

The dataset used in this project consisted of over 8,400 support tickets and included a variety of structured fields such as customer age, product type, purchase date, ticket metadata, and communication timestamps. We utilized robust preprocessing techniques and built a Random Forest model that achieved an impressive 100% accuracy under rule-based labeling conditions.

2. Dataset Overview

The dataset included:

- **Ticket Metadata**: Ticket ID, Ticket Type, Status, Priority, Channel
- **Customer Information**: Age, Gender, Product Purchased
- **Timestamp Fields**: Date of Purchase, First Response Time, Time to Resolution
- **Text Fields**: Subject, Description, Resolution (excluded from model)

Initial dataset size: **8,469 records and 13 columns**

3. Data Preprocessing

3.1 Cleaning & Formatting

- Removed unnecessary personal identifiers (e.g., name, email)
- Converted datetime fields and filled missing values with medians

- Encoded categorical variables using `LabelEncoder`
- Scaled numerical features like age and time gaps using `StandardScaler`

3.2 Feature Engineering

Several new variables were constructed to capture operational behavior:

- **Time Since Purchase**: Days between purchase and first response
- **Response Time (minutes)**: From purchase to first reply
- **Resolution Time (hours)**: From first response to resolution
- **Is High Priority**: Binary indicator for priority levels
- **Is Social Media**: Binary indicator if ticket came from social media
- **Is Weekend Purchase**: Flag if ticket originated on a weekend

These engineered features added valuable signal to the dataset, increasing its predictive capacity.

4. Exploratory Data Analysis (EDA)

4.1 Target Distribution

The satisfaction variable was generated based on a smart rule:

- If a ticket had **Low/Medium priority** and was resolved **within 24 hours**, it was labeled **satisfied (1)**
- Else, **dissatisfied (0)**

This created a nearly balanced distribution:

- Satisfied (1): 4,214 tickets
- Dissatisfied (0): 4,255 tickets

4.2 Demographic Insights

- **Gender vs Satisfaction**: Males showed a slightly higher dissatisfaction rate
- **Age**: Customers across all age groups experienced similar satisfaction patterns
- **Product Type**: Certain products had consistently higher dissatisfaction levels

4.3 Ticket Properties

- **Ticket Type & Channel**: Some channels (like email) had faster response times, correlating with higher satisfaction
- **Priority**: High priority tickets had higher resolution time, thus more dissatisfaction

4.4 Correlation Heatmap

A correlation matrix showed strong positive correlation between fast resolution and satisfaction. High priority had a mild negative correlation.

5. Model Development

5.1 Target Variable Creation

The smart rule-based label was based on:

```
if Ticket Priority <= 1 and Resolution Time (hours) < 24:
    Satisfaction = 1
else:
    Satisfaction = 0
```

5.2 Model Training

- **Model Used**: Random Forest Classifier
- **Train-Test Split**: 80% training, 20% testing
- **Features Used**: 13 numerical and encoded columns

5.3 Results

Metric	Value
Accuracy	100.0%
Precision	1.00
Recall	1.00
F1 Score	1.00

The model predicted all satisfaction labels correctly. This high performance stems from the logical consistency of the rule-based target.

5.4 Feature Importance (Top 10)

1. Resolution Time (hours)
2. Response Time (minutes)
3. Ticket Priority
4. Time Since Purchase
5. Product Purchased
6. Customer Age
7. Is Weekend Purchase
8. Is Social Media
9. Ticket Type
10. Ticket Channel

These insights help customer support teams identify what aspects influence satisfaction most.

6. Business Insights

- **Resolution Speed is King**: Tickets resolved quickly (<24 hrs) heavily influence customer happiness
- **Priority Handling Needs Optimization**: High priority tickets often remain open longer, causing dissatisfaction
- **Product Feedback Loop**: Products with high dissatisfaction should be reviewed for defects or usability
- **Channel Efficiency**: Social media and chat-based channels performed better in terms of satisfaction turnaround

7. Deployment Plan

The trained model can be deployed in real-time customer service dashboards. When a new ticket is created, the model can instantly predict the likelihood of customer satisfaction, allowing the system to take proactive steps:

- Escalate high-risk tickets
- Reassign to experienced agents
- Send preemptive surveys

8. Limitations and Future Work

- The satisfaction label is rule-generated, not user-given — actual feedback should be used in future versions
- Text-based features (ticket description, resolution) were not used — NLP-based modeling could improve accuracy
- Time-related fields were engineered simply — more dynamic response curves can be added

9. Conclusion

This project demonstrates the power of structured data and machine learning in customer satisfaction analytics. By leveraging ticket metadata, time-based signals, and business logic, we created a highly accurate Random Forest model that identifies patterns in customer experience. While the current model is built on heuristic labels, it can be enhanced by integrating real customer feedback and behavioral tracking in future iterations.

Keywords: Customer Satisfaction, Machine Learning, Random Forest, Feature Engineering, Customer Support, Ticket Analysis, Resolution Time, Predictive Analytics