**E-commerce Customer Satisfaction Prediction**

**Table of Contents**

1. Introduction

2. Business Problem

3. Machine Learning Formulation

4. Business Constraints

5. Possible Metrics 6. Data Overview

7. Exploratory Data Analysis (EDA)

8. Feature Engineering

9. Machine Learning Models

10. Future Improvements

11. Deployment

12. References

---

**1. Introduction**

E-commerce has become an integral part of daily life, serving as a platform that connects numerous sellers with customers online. The key to success in e-commerce lies in providing exceptional service to customers. A critical measure of service quality is customer satisfaction, which is often gauged through feedback and ratings. Understanding and predicting customer satisfaction can significantly enhance business performance.

**2. Business Problem**

Currently, e-commerce platforms solicit feedback from customers via email after product delivery, allowing them to rate their experience on a scale of 1 to 5. However, many customers do not provide ratings or reviews. This raises the question: \*\*Can we predict the review score a customer might give before they actually submit it?\*\* Solving this problem can help e-commerce platforms improve their services and customer satisfaction.

For this case study, I utilized a dataset from Olist, a Brazilian e-commerce platform, which contains over 100,000 orders from 2016 to 2018. The goal is to predict customer satisfaction ratings based on historical data.

**3. Machine Learning Formulation**

To address the business problem, we can frame it as a supervised machine learning task. Since we aim to predict discrete ratings (1 to 5), this is a multi-class classification problem. Our objective is to predict customer ratings before they are submitted, which means we will exclude any data related to customer reviews.

**4. Business Constraints**

- There is no strict requirement for low latency, but predictions should be made quickly enough to be useful.

- Misclassifying low ratings (1, 2, 3) is particularly detrimental, as it can lead to customer loss.

**5. Possible Metrics**

- Multi-class confusion matrix

- Macro F1 score

- Multi-class log loss

- Balanced accuracy

**6. Data Overview**

The dataset is sourced from Kaggle and consists of nine CSV files, including customer data, order details, payment information, and product categories. The files can be loaded into Python using Pandas:

```python

import pandas as pd

customer = pd.read\_csv("olist\_customers\_dataset.csv")

geo\_location = pd.read\_csv("olist\_geolocation\_dataset.csv")

items = pd.read\_csv("olist\_order\_items\_dataset.csv")

payments = pd.read\_csv("olist\_order\_payments\_dataset.csv")

reviews = pd.read\_csv("olist\_order\_reviews\_dataset.csv")

orders = pd.read\_csv("olist\_orders\_dataset.csv")

products = pd.read\_csv("olist\_products\_dataset.csv")

sellers = pd.read\_csv("olist\_sellers\_dataset.csv")

translation = pd.read\_csv("product\_category\_name\_translation.csv")

```

After merging these files, we need to check for null values and irrelevant columns for our analysis.

**7. Exploratory Data Analysis (EDA)**

EDA is crucial for understanding the dataset and its features. The distribution of review scores reveals an imbalance, with a higher frequency of 5-star ratings compared to lower ratings. This J-shaped distribution is common in e-commerce ratings.

Distribution of Review Scores

The distribution of review scores can be visualized using a bar chart:

```python

import matplotlib.pyplot as plt

import seaborn as sns

sns.countplot(data=reviews, x='review\_score', palette='viridis')

plt.title('Distribution of Review Scores')

plt.xlabel('Review Score')

plt.ylabel('Count')

plt.show()

```

![Distribution of Review Scores](https://images.app.goo.gl/6gPjPaZkrNxc59Qc6)

Correlation Analysis

A correlation heatmap indicates that no features are highly correlated with the review score, suggesting that we need to create new features for better classification.

Payment Type Analysis

The analysis of payment types shows that credit cards are the most common payment method. A Chi-square test indicates a significant relationship between payment type and review score.

Pie Chart of Payment Types

A pie chart can be used to visualize the distribution of payment types:

```python

payment\_counts = reviews['payment\_type'].value\_counts()

plt.figure(figsize=(8, 8))

plt.pie(payment\_counts, labels=payment\_counts.index, autopct='%1.1f%%', startangle=140)

plt.title('Distribution of Payment Types')

plt.axis('equal') Equal aspect ratio ensures that pie is drawn as a circle.

plt.show()

```

![Distribution of Payment Types](https://images.app.goo.gl/6gPjPaZkrNxc59Qc6)

**8. Feature Engineering**

Feature engineering is vital for improving model performance. I created several time-based features, such as:

- \*\*Estimated Delivery Time\*\*: Difference between estimated delivery date and purchase timestamp.

- \*\*Actual Delivery Time\*\*: Difference between actual delivery date and purchase timestamp.

- \*\*Delivery Distance\*\*: Distance between seller and customer locations.