CS 539 Machine Learning Project Proposal
Delivery Detector

Ronit Kapoor, Yiqun Duan, Noushin Khosravi Largani, Jinqin Xiong

Function:

Our project will predict if the delivery will reach on time. We need a tool so that customers can know if their delivery will be delayed.

Who would benefit from this:

This has the potential to increase customer satisfaction, because the customer is less likely to be frustrated with a late package if they get an alert saying that the package will be late versus not knowing at all if the package will be delayed. Companies will also benefit from this because customer satisfaction will increase, leading to more usage of the company's delivery services. Furthermore, our tool can provide insights on feature importance which will benefit data scientists with further studies on this topic.

Does tool already exist:

On codeproject.com, there exists a tool where a camera can see if a package has arrived at a customer's house. Then, the system can email the user alerting the user that the package has arrived, and the customer can go retrieve it. Our delivery detector is different. We don't use computer vision to determine if a package will have arrived or not, rather we use different metrics to make a prediction if a delivery will be delayed. These metrics include, warehouse block, more of shipment, customer rating, and more.

How we will build this:

1. EDA and feature engineering
   Our analysis will start from Exploratory Data Analysis (EDA) to get comprehensive statistical insights into each feature and the distribution of data. In addition, prior to model construction, it is important to pre-process the data (feature engineering). Categorical data, represented as strings, will be converted into digital inputs by One-hot Encoding. And numerical data will be normalized because of the disparate scales.
2. Machine learning model
   To predict on-time delivery, we opt for employing a decision tree classifier and a logistic regression model due to their efficacy in binary classification tasks. Additionally, we plan to explore the potential enhancement in performance and accuracy by implementing bagging trees.
   To predict customer ratings, we plan to train both regression and classification models to determine the most effective approach. Regarding regression, we intend to explore linear regression and lasso regression. For classification, we will evaluate KNN with $k \geq 5$, given the five levels of rating scores.

What existing resources we can use:

We can utilize the scikit-learn machine learning library in Python, which provides a wide range of tools for data analysis. Additionally, for data preprocessing tasks such as normalization, one-hot encoding, and train-test splitting, we can import pandas and numpy.

How we will demonstrate:

In our website, users will enter statistics about the product based on these answers. We will be able to make a prediction if the delivery will make it on time or not. Furthermore, we will utilize the True Positive Rate and False Positive Rate to make our website more accurate as time goes on.

Data Introduction

The dataset includes 12 features (columns). Key information includes:

ID: A distinct identifier (categorical variable).

Warehouse Block: The company divides warehouse into blocks A to E (categorical variable).

Mode of Shipment: Methods for transportation, including flight, road, etc. (categorical variable)

Customer Care Calls: The frequency of customer inquiries (ordinal variable).

Customer Rating: One of the target variables, indicating customer satisfaction assessment in range 1 to 5 (ordinal variable).

Cost of the Product: To help with pricing and profitability evaluation (numerical variable).

Prior Purchases: Customers' purchase history track (numerical variable).

Product Importance: Classify to low, medium and high (ordinal variable).

Gender: To examine preferences and patterns of shopping (categorical variable).

Discount Offered: Examine the impact of discounts on sales (numerical variable)

Weight in Grams: The logistical aspect of shipping (numerical variable).

Reached on Time: One of the target variables, indicating whether the product is delivered within the expected time (categorical variable)

Reference

1. https://www.kaggle.com/datasets/willianoliveiragibin/on-time-delivery
2. https://www.jaggaer.com/press-release/first-ai-ontime-delivery-predictor/