Final Report
Predict Delivery Time For Online Order

Huy Hoang Vuong- Data Science Diploma- Jun 25, 2023

Business Problem

The market for online shopping is increasing every day, and the challenge of bringing the best experience for customers when shopping online is the focus of every business nowadays. One priority is helping customers answer “When will my order arrive?” as closely as possible. According to *ottomotors.com,* 17% of customers said they would not return to the business after only one late delivery, and more than half said they would put that company on their blocklist if they had the second or third time being late.

The project uses the record of the past to be able to calculate the estimated number of days needed to deliver the package to a customer and help answer the question:
“How might we use machine learning to let customers know the day they would receive their package when they shop online using historical business data?”

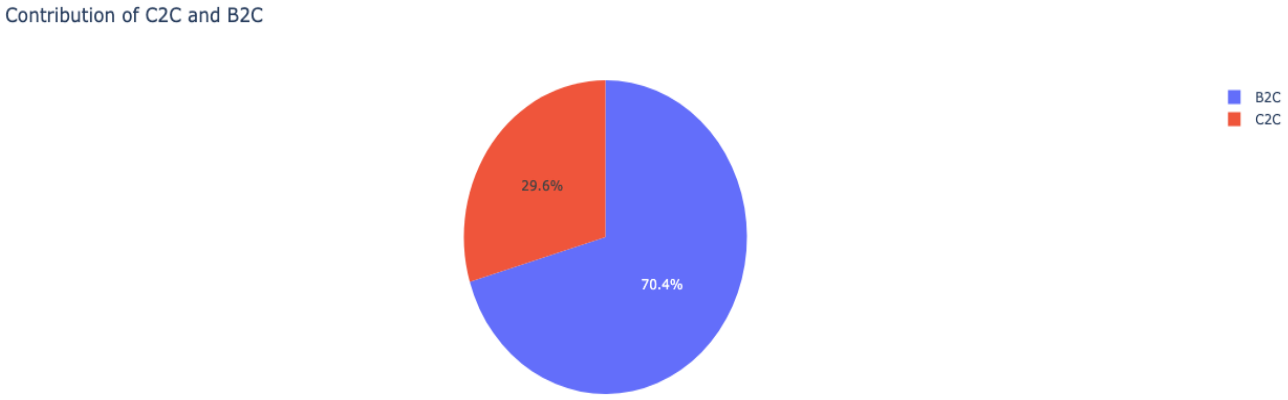
Data Source

The data to be used was given from eBay, one of the biggest platforms for shopping online, allowing business to sell their product on the platform (the consumer who would like to sell their product also can sell on this platform) in their machine learning challenge in 2021 to find out the solution help improve the delivery system and bring better customer experience when shopping on eBay, the data can be downloaded on Kaggle. The data has 15 million rows and records the customer shopping history on the eBay platform from 2017 to 2019.

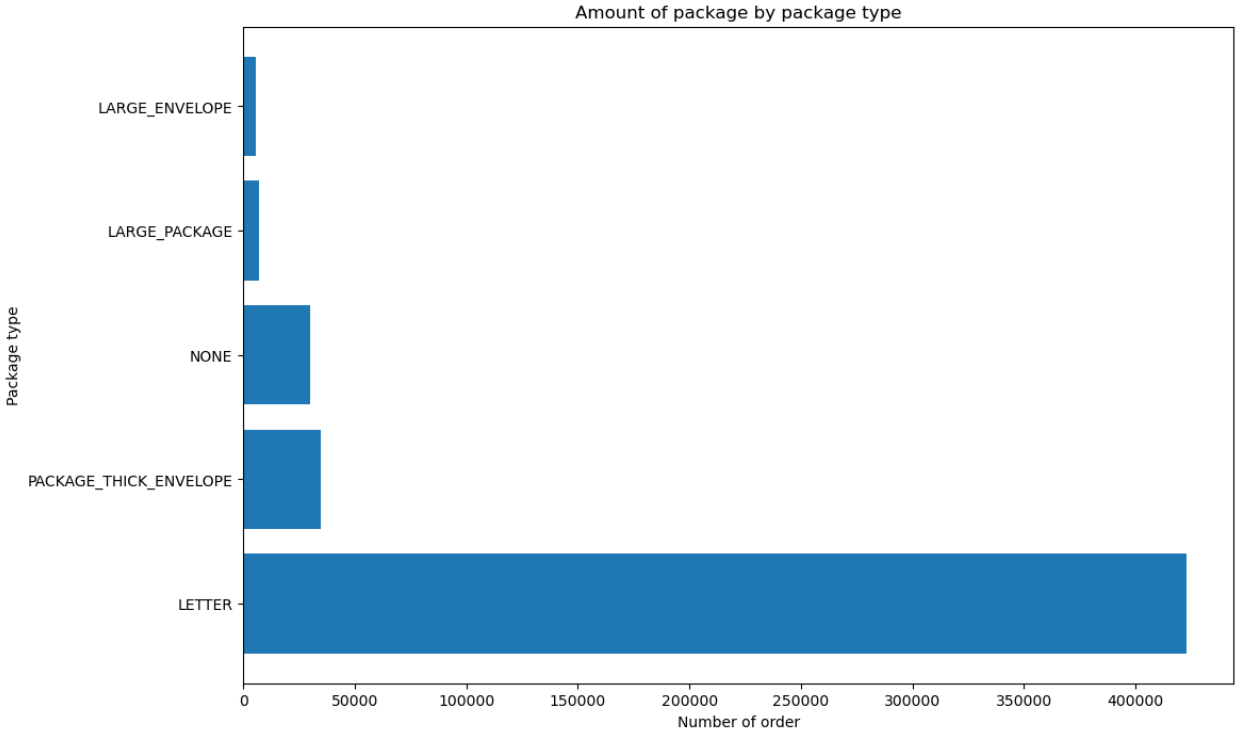
Exploratory Data Analyst(EDA)

The EDA process provides an overview of the data and finds some basic information that could help bring insight into what the data is trying to say. There are also some steps, include: filled up missing values, correcting the uncommon value, changing the type of date variables,
Some basic charts below will let us know how many business types, package types, and trending of business by month.

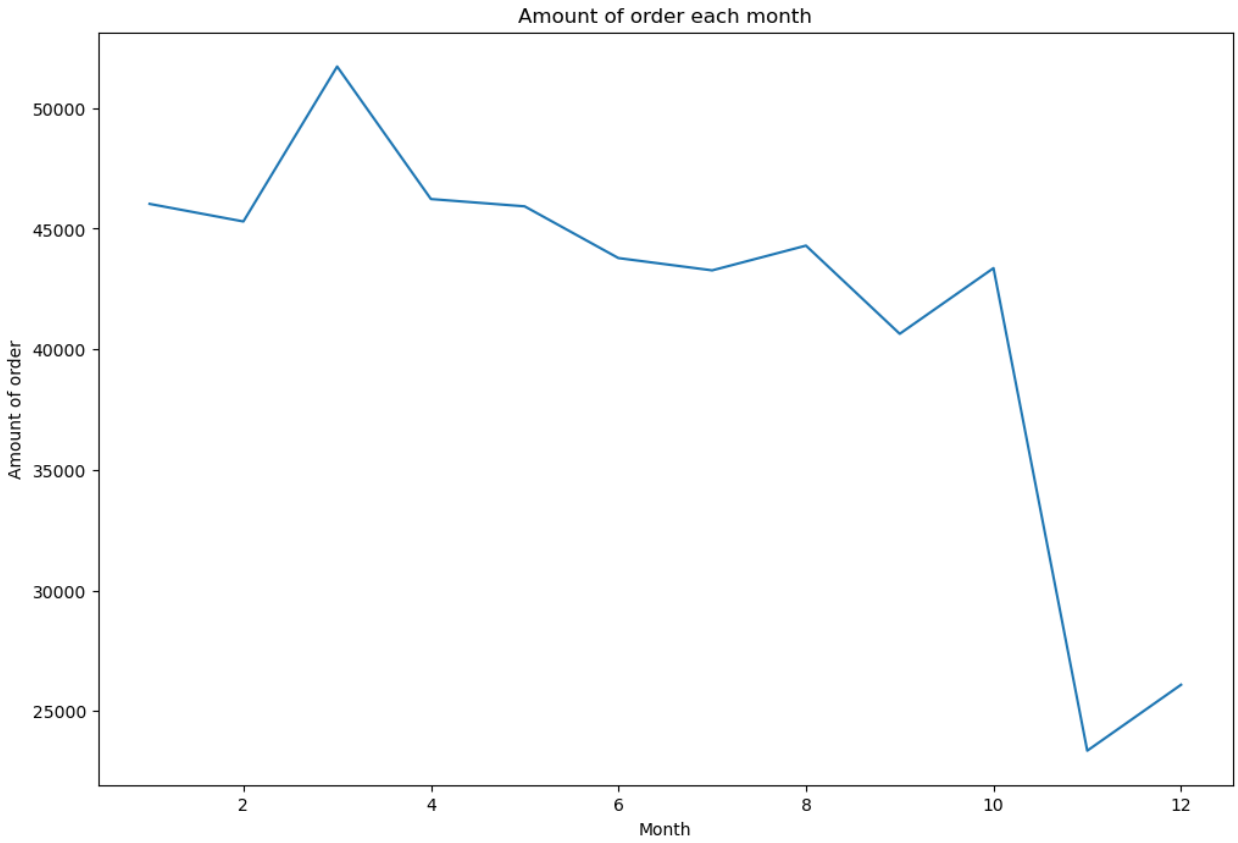
There are 70.4 percent of the transaction is from Business to Consumer, and 29.6 percent is from Consumer to Consumer.



Most of the order is in small sizes like Letter size; some packages are in None type, which will be considered in unknown size.



Most of the transactions happened in March, and the unusual thing is that there were only a few transactions were made in November and December, which should be busy for most businesses.



Feature Engineering

In the Feature Engineering step, we convert categorical columns into numbers, such as binary columns b2c_c2c and package_type, to numbers by ordinal encoding.

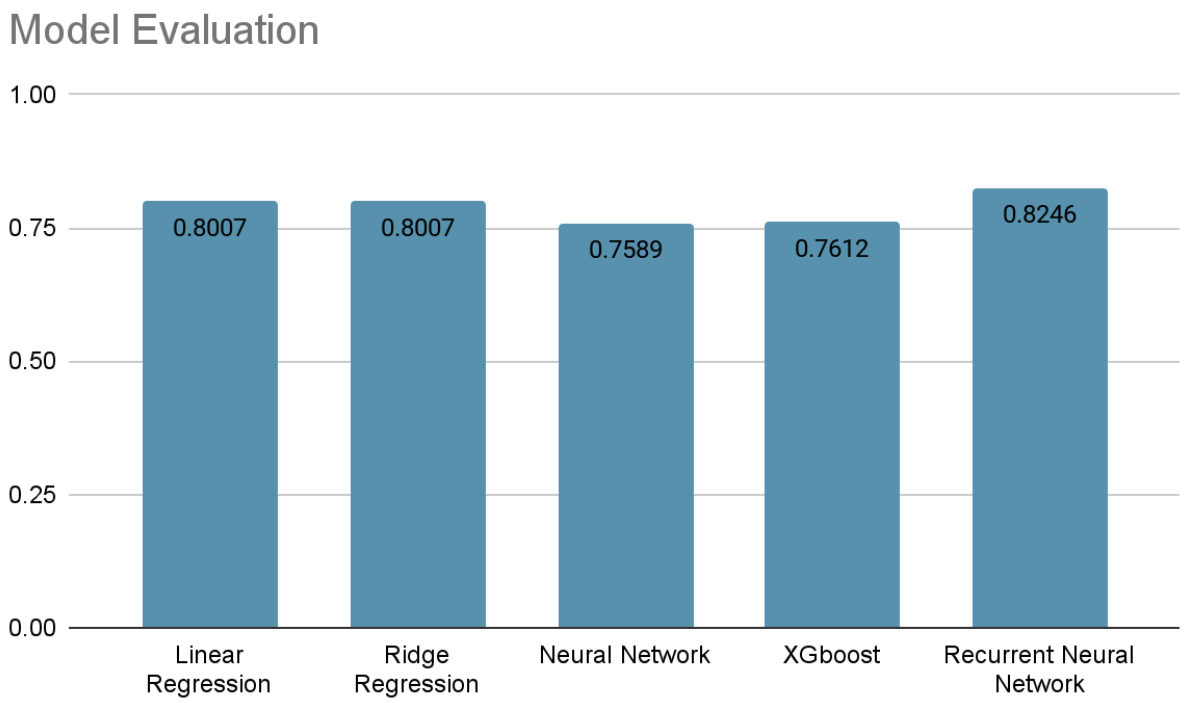
Create a new column using existing columns: calculate the distance between buyer and seller from their zip code, and subtract the order date to year, month, and day separately.
Feature handling_days, shipping_day, and total_day were added to bring more information and help build the model.

Model and insights

Our model focuses on calculating the estimated days needed to deliver the package for a customer as closely as possible. Data were split in train and test to process with the basic model; we used **GridsearchCV** to find the best optimal model and prevent bias with a split train dataset to train and validation part. Interm of models, we used: Linear

Regression, Ridge Regression, XGboost Neural network and Recurrent Neural Network.

Finding:



Neural Network has the lowest loss, 0.7589, compared to the other four models. This model also has quite a good accuracy when 72.0035 percent of the predicted outcome makes the order deliver on-time or early.

Next Steps:

The final model did a good job predicting the day order will be delivered, but there are still some extra data steps that can be added, which could help to improve.

- Predict the handling day, which can help businesses have early warning when the order is likely to be late and inform customer if needed.
- Consider the holiday during the delivery period, which could impact the delivery process.