Capstone 2 Project Proposal

Problem: Predicting Customer Satisfaction

Many companies today measure customer satisfaction. Some of the quick ways of collecting the feedback from customers is through their reviews or by surveying them. These are very important customer service metrics, but are not typically being used to improve operations or help reduce customer churn.

Instead of waiting until the customer interaction is over for feedback/ review, companies can predict how likely an operation is to receive a good or bad rating while they are still in contact with the customer.

My main hypothesis is that the product and how the order was fulfilled influences customer review rating.

In this project, machine learning techniques will be applied to the dataset to predict customer review ratings.

Client : This project can be used by any online retail company

Questions To Explore:

- Which features will produce the best model for predicting customer ratings
- Which products categories are more prone to customer dissatisfaction
- Which City has more customers
- How long it took orders to be delivered

DATASET:

Brazilian E-Commerce Public Dataset by Olist

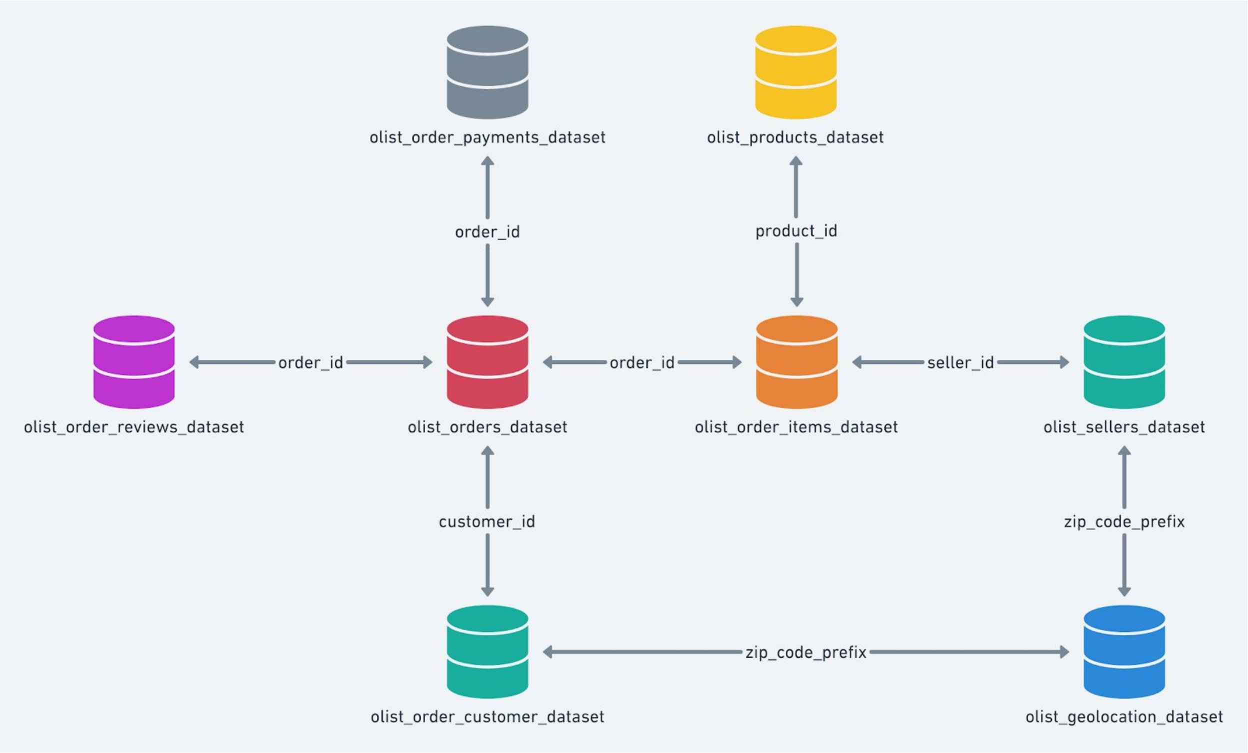
https://www.kaggle.com/olistbr/brazilian-ecommerce

This is a Brazilian ecommerce public dataset of orders made at Olist Store. The dataset has information of 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. It’s features allows viewing an order from multiple dimensions: from order status, price, payment and freight performance to customer location, product attributes and finally reviews written by customers.

This is real commercial data, it has been anonymised, and references to the companies and partners in the review text have been replaced with the names of Game of Thrones great houses.

Data Schema

The data is divided in multiple datasets for better understanding and organization.



Datasets and data descriptions:

1. Customers Dataset

Customer_id : key to the orders dataset. Each order has a unique customer_id.
Customer_unique_id: unique identifier of a customer.
Customer_zip_code_prefix: first five digits of customer zip code
Customer_city: customer city name
Customer_state: customer state

2. Geolocation Dataset

Geolocation_zip_code_prefix: first 5 digits of zip code

Geolocation_lat: latitude

Geolocation_lng: longitude

Geolocation_city: city name

Geolocation_state: state

3. Order Items Dataset

Order_id: order unique identifier

Order_item_id: sequential number identifying number of items included in the same order.

Product_id: product unique identifier

Seller_id: seller unique identifier

Shipping_limit_date: Shows the seller shipping limit date for handling the order over to the logistic partner.

Price: item price

Freight_value: item freight value item (if an order has more than one item the freight value is splitted between items)

4. Payments Dataset

Payment_sequential: a customer may pay an order with more than one payment method. If he does so, a sequence will be created to accommodate all payments.

Payment_type : method of payment chosen by the customer.

Payment_installments : number of installments chosen by the customer.

Payment_value: transaction value.

5. Order Reviews Dataset

review_comment_title:Comment title from the review left by the customer, in Portuguese.

Review_comment_message: Comment message from the review left by the customer, in Portuguese.

Review_creation_date : Shows the date in which the satisfaction survey was sent to the customer.

Review_answer_timestamp: Shows satisfaction survey answer timestamp.

6. Order Dataset :

Order_approved_at :Shows the payment approval timestamp.

Order_delivered_carrier_date: Shows the order posting timestamp. When it was handled to the logistic partner.

Order_delivered_customer_date : Shows the actual order delivery date to the customer.

Order_estimated_delivery_date: Shows the estimated delivery date that was informed to customer at the purchase moment.

7. Products Dataset:

product_photos_qty:number of product published photos

Product_weight_g: product weight measured in grams.

Product_length_cm: product length measured in centimeters.

Product_height_cm:product height measured in centimeters.

Product_width_cm: product width measured in centimeters.

8. Sellers Dataset

Seller_id: seller unique identifier

Seller_zip_code_prefix: first 5 digits of seller zip code

Seller_city: seller city name

Seller_state :seller state

9. Category Name Translation

Product_category_name: category name in Portuguese

Product_category_name_english : category name in English

Outline: I will be using various data wrangling techniques to clean the data and analyze for any outliers in the data. Next, I will run an analysis to see which features will work best in the model, and then finally create a predictive model that will using features selected. The accuracy of the model will determine the best model.

Deliverables: The deliverables will be code, a machine learning model, a final report, and a slide deck.