Applied Machine Learning [Final Project Proposal (Group 19)]

Gursifath Bhasin (gb2760) Jonathan Benghiat (jb4653) Lukas Wang (bw2712)
Qi Meng (qm2162) Siddhant Pravin Mahurkar (sm5129)

Introduction / Background

We intend to compare and contrast a few different models applied to the same dataset, leveraging several machine learning paradigms to determine which best applies to our problem & associated data set. We will begin with exploratory data analysis and then apply ML techniques to the problem of predicting sales volume based on a fairly high dimensional dataset that spans multiple years. Our hope is to garner a deeper understanding of why different models and paradigms are better suited to this kind of problem.

Description of Dataset

We will employ the *Brazilian E-Commerce Public Dataset by Olist* in Kaggle. This dataset is an ecommerce dataset of order with 100 thousand orders from 2016 to 2018 in Brazil. The dataset describes multiple features of the orders at marketplaces, including order status, customer locations, price, etc. The schema of the dataset is as shown in Figure 1.

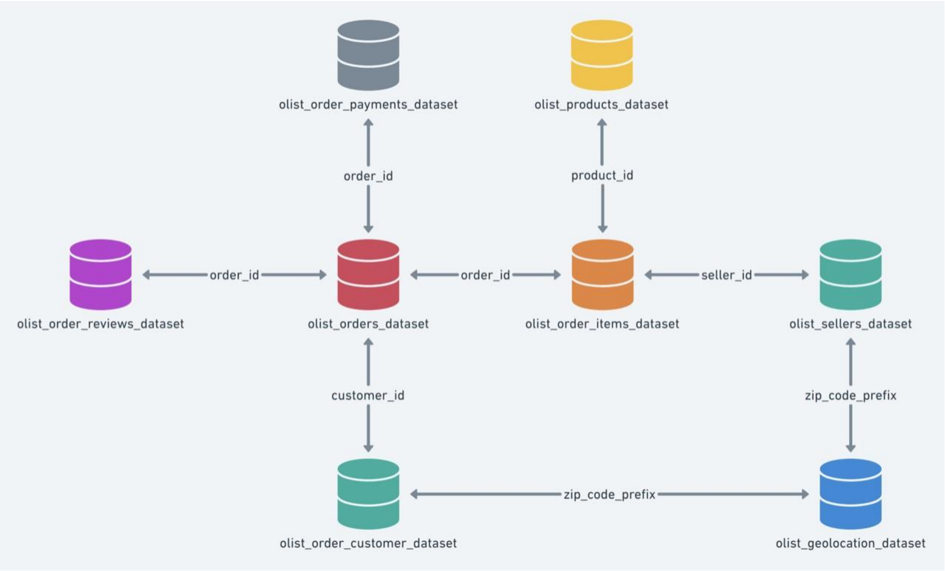

Figure 1: Data Schema

There are a total 8 dataset, containing 50 columns. Since the data is obtained from real orders, we find that the data is imbalanced from day to day. In order to fulfill our goal of a robust time series prediction, we need some further processing on the dataset, which will be described in the following section.

Tentative Approach

We plan to solve the problem in two class of models and compare their results to select the best one:

Class 1: Traditional machine learning model

Considering the dataset contains not only time series data but other regular data, we can treat them all as regular data and use models such as logistic regression, SVM, random forest, XGBoost, LightGBM to do predictions and compare results.

Class 2: Time series model + Deep Neural Network (DNN)

In order to use the information for time series data, we plan to use RNN / LSTM model to handle time series data and use DNN to tackle other data. Then, we concatenate the results and use another DNN to generate the final prediction.