Prediction Analysis

Disclaimer

I do not have a background in Data Science. When speaking with Maggie, I expressed my interest in the Data Engineer position but let her know my lack of experience. I have extensive experience in the actual data engineering portion of the position. That is ETL, data storage, data retrieval, etc. I am open to learning and assisting with data science work.

Solution Approach

My goal for this task was to use the proper Python tools to prepare and perform a simple Linear Regression. Given the data for the merchant code, the potential variables were user_id and some kind of manipulation of datetime. This could include denoting the day of week, but I nixed this option because there were not consistent sales given a weekday. The difficult part of this regression was creating a prediction given data where the purchase amounts have large variability. I assumed that the lack of a data entry for a specific date would denote that the purchase amount is zero. It is possible that the store in question was not open that day, thus zero-purchase days are impossible. After consuming the raw data, I added these zero-purchase dates.

The data was further prepared by aggregating on the calendar date given the datetime field. The user_id column was omitted, though I could see how an advanced analysis may consider it. Using scikit-learn, I performed a simple LinearRegression model. The X variable was simply an indexed value of the data, while the Y was the purchase amount. I understand that this is not an ideal X, but it seemed to be the best option. After fitting the data, it returned constant values for predicted purchases. I assume I may have made a mistake due to my lack of experience.

Room for Improvement

It's possible that the correct prediction would consider the possibility of user_id being a predictor. As mentioned, I added zero-purchase days which significantly drove down the average per day. Given the provided dataset, there is a span of 79 days but only 28 records. So almost two thirds of the days had zero purchases. A separate analysis could be to create a scenario where the store would randomly be closed (using a Poisson distribution). On open days, those zero-purchase days would be ignored. Furthermore, the number of consecutive zero-purchase days could also inflate the purchases on the next day open.