Case Background

Mattson Nutrition is one of the premier nutrition retailers in the United States.[1] Mattson has a mission to help its customers live healthier lifestyles by supplying high-quality vitamins, supplements, proteins, superfoods, and other nutritional products. All of Mattson's products may be purchased over the counter without a prescription. Mattson Nutrition was founded in 2008 with retail operations strategically located in cities and gyms around the United States. Those retail storefronts make up roughly 30% of its revenue. The remainder of its revenue comes from its online operations primarily in partnership with Amazon, eBay, and other specialized online platforms. Customers have the ability to make purchases online via their website, their mobile application, or their online retail platform partners. Staffing constraints along with increasing rents has hurt their profit margins with its traditional storefront locations, which has led the executive team to exercise caution when investigating new potential brick and mortar locations. They are also investigating closing certain storefronts that are in areas with low access to staff and high rents. As a result, Mattson wants to focus on its online and mobile sales channels because those require less staff and less high-priced retail space, which should help their overall margins in the longer term.

Mattson has focused on its online presence with social media platforms such as Facebook, Reddit (and sub-Reddits) and Instagram. Engaging potential customers on these social media platforms by sharing nutritional advice and healthy-living content has been extremely useful at attaining and retaining its customers. They currently have hundreds of thousands of followers on both platforms. Mattson uses this platform to share valuable information about different nutritional products and how to live healthier lifestyles in whatever manner their customers define healthier lifestyles. They have made a conscious effort to have content on their social media pages that appeals to all different types of healthy lifestyles and body types. Doing so has helped Mattson differentiate itself from other fitness and nutrition social media accounts who only focus on one singular definition of "healthy." Their social media presence has also allowed Mattson to gather valuable data on its customers that they have used in their machine learning related projects to predict customer purchase patterns.

Mattson Nutrition recently partnered with a large-scale health system in the United States. This partnership has allowed Mattson's team of nutrition experts to work with family physicians to provide educational resources related to healthy living and general well-being for select patients. Individuals who see their family physician regularly generally care about their health, which makes them ideal target customers for Mattson Nutrition. The nutritional experts from Mattson also regularly hold daily or nightly meetings to discuss the latest trends in nutritional products. These events are held at the health systems facilities and have generally been very well attended. Individuals who attend their in-person discussion forums and educational sessions typically start following Mattson Nutrition on social media and eventually become customers for a variety of products. Partnering with health systems has also allowed Mattson to get additional data about these potential customers. These data have been valuable to construct of a variety of predictive models using different machine learning algorithms.

Another creative idea that the marketing team at Mattson Nutrition implemented was offering customers discounts if they share their "fitbit" data. They typically give customers 15 to 20% off their purchases for sharing these data each month. Essentially, Mattson is paying customers for these data because these data help them segment customers to better market to them with customized coupons and other offers. The

[1] This example is completely fictitious. Any similarities to actual companies is coincidental.

marketing team has not yet incorporated these data into any of their machine learning models, but they have plans to do so. As a result, Mattson Nutrition has hired you to use your machine learning expertise to train and test different models using a variety of different machine learning algorithms to predict customer purchase patterns.

To start your training process, the IT department has provided you with a data file that includes a variety of customer attributes, data from the customer's family physician (with proper permissions and anonymized to be HIPPA compliant), prior purchases, and "fitbit" (or other similar wearable devices) activity data (averaged across the period). The data file they provided is named "Mattson_nutrition_customers.json". The json data file contains many sub-documents, because the IT department constructed the json file from several different data sources. After several rounds of discussion with the IT department, they informed you that this format is the best format that they will be able to provide given the complexities that they had gathering and extracting the data. Therefore, you have a fair bit of clean up and preparation work needed to get these data ready to use to train and test machine learning models.

You decide to use the "sales" variable as your known target for your supervised machine learning models. You observe that not all of the features may be used to train the models because certain features are not observed before the sales target, which would result in a target leakage problem.

Tasks

1. Use all of the machine learning algorithms you have learned to train and test different models.
2. Select the best machine learning model that balances bias and variance.
3. Use that model to make predictions with the data contained in the "sample_implementation.txt" file. Output those predictions with only a few of the import fields to a MS Excel file. Make sure to include a tab with the details concerning your training and testing process.
4. When making those predictions, perform some what-if analyses (i.e., what happens to expected sales if the customer's stress increases or decreases by 50%). For these what-if analyses, ask the user what field to use to perform these what-if analyses from the following: Likes, Shares, WebsiteVisits, MobileAppLogins, Steps, REM Sleep, Deep Sleep, Light Sleep, and Stress.
5. Based on the selection, calculate the predictions assuming a 50% reduction and a 50% increase in that value. For instance, if the user selects steps and they had steps of 1000, then calculate the predicted sales assuming they had steps of 500 and 1500.

Data Dictionary

The variables without a prefix are not time dependent. The variables with the "delta1" prefix are from the previous period (quarters). The variables with the "delta2" prefix are from two periods ago.

CustomerID – The unique identifier of each customer.
Extract_Date – This field is the date of the current period. The "delta1" period starts three months prior to this date and ends on that date. For instance, if the extract date is 10/31/2022, then the delta1 period starts on 7/31/2022 and ends on 10/31/2022. The "delta2" period is the period (quarter) prior to "delta1." Therefore, the delta2 period would start on 4/30/2022 and end on 7/31/2022 if the extract date was 10/31/2022.
dob – This field represents the date of birth of each customer.

NOTE1 about Customer Age: To calculate age in a period, use the last day of the period (quarter) when subtracting the period date and the extract date. Unfortunately, there are many customers who do not provide their date of birth to Mattson Nutrition. It is a bit unfortunate that these data are missing because customers in different age ranges have different purchasing patterns of nutritional products.

gender – This field represents the gender of each customer.

NOTE about Gender: Similar to age, it is common for customers to not provide their gender to Mattson Nutrition. These missing data are significant omissions because male and female customers tend to have different nutritional consumption patterns.

Big5_Conscientiousness – This field represents the customers score on the conscientiousness portion of the Big 5 personality test.
Big5_Openness – This field represents the customers score on the openness portion of the Big 5 personality test.
Big5_Extroversion – This field represents the customers score on the extroversion portion of the Big 5 personality test.
Big5_Agreeableness – This field represents the customers score on the agreeableness portion of the Big 5 personality test.
Big5_Neuroticism – This field represents the customers score on the neuroticism portion of the Big 5 personality test.
FamilyHistory_Diabetes – This field represents whether the customer's family has a history of diabetes.
FamilyHistory_HeartDisease – This field represents whether the customer's family has a history of heart disease.
FamilyHistory_Cancer – This field represents whether the customer's family has a history of cancer.
FamilyHistory_Crohns – This field represents whether the customer's family has a history of Crohns disease.
FamilyHistory_Alzheimer – This field represents whether the customer's family has a history of alzheimer.
FamilyHistory_Parkinsons – This field represents whether the customer's family has a history of Parkinsons disease.
FamilyHistory_Depression – This field represents whether the customer's family has a history of depression.
FamilyHistory_Other – This field represents whether the customer's family has a history of other disease not included in any of the other FamilyHistory variables.

The following fields are either for the delta1 period or the delta2 period depending on the prefix.

Sales – This field represents the sales in US dollars for the period.
OfficeVisits – This field represents the number of times a customer saw a doctor during the period.
BMI – This field represents the customers body mass index measured during a doctor's appointment during the period. The following are the general rules of thumb for BMIs

Underweight: BMI is less than 18.5
Normal weight: BMI is 18.5 to 24.9

Overweight: BMI is 25 to 29.9
Obese: BMI is 30 or more

Bloodpressure – This field represents the customers blood pressure measured during a doctor's appointment during the period. The possible values are null, normal, low, and high.
Smoke – This field represents whether the customer identified as a smoker during a doctor's appointment during the period.
Drink – This field represents whether the customer identified as a drinker during a doctor's appointment during the period.
Likes – This field represents how many posts on Mattson Nutrition's social media accounts the customer liked during the period.

Shares – This field represents how many posts on Mattson Nutrition's social media accounts the customer shared during the period.
WebsiteVisits – This field represents the total number of website visits the customer made during the period.
MobileAppLogins – This field represents the total number of mobile application logins the customer made during the period.
Steps – This field represents the average number of steps per day the customer took during the period.
Deep Sleep – This field represents the average deep sleep score the customer had during the period.
Light Sleep – This field represents the average light sleep score the customer had during the period.
REM Sleep – This field represents the average REM score the customer had during the period.
HeartRate – This field represents the average HeartRate score the customer had during the period.
Stress – This field represents the average Stress score the customer had during the period.