FoodX

Summary of Data and Errors:
While parsing through the data provided to me, I was able to determine some of the issues that were apparent within the data frame. First, I saw that you cannot make generalizations based on the data for the menu, since the time that is given is not consistent, as times later in the day around 14:00 and higher can result in a variation that shifts your analysis a bit lower than normal as that is after peak hours. Not only that, the data for the majors may be misleading as it showcases Chemistry as the most popular major for the food orders. This is because this data heavily depends on a university that was popular and not on a popular major across all universities, as you could have Purdue University on the list but their common majors stem across technology and engineering-based degrees. Finally, I noticed that the spread of the universities was skewed heavily to the right when I graphed that particular column in a barpot. This is because I only saw Bulter University, Indiana State University, and Bell State University as the top three with the most meals, whereas other big universities like Purdue University barely had any orders. This could be an error in the data that was collected as well as depending on the source of it.

Ethical Implications:
I think that in general, whenever data is retrieved, it needs to be neutral as much as possible and not biased by outside factors. An instance of this would be that various students from a particular university already know that data collection will occur on a certain day, and because of that the results get biased since everyone gets substantially more food during that time than normal.

Business Outcome:
One implication that could occur that affects the business outcome is the actual results that are obtained. This is because the skew of results, such as the standard deviation or variability can make a significant impact on whether data analysis was done correctly or not. If a business needs to determine the number of meals that they need to have for a certain university to make a good profit, they could overachieve on that and have a loss as a result of the data analysis that I've currently done.

Technical Implications:
In general, having all of the prediction and regression models that are being employed within the actual data available, such as graphs of results that would be obtained from the data being taken. This would make it easier to spot any errors that are created.

Data Analysis I Used for Predicting Customer Order:
After doing research on the scikit-learn library that was provided, I found two links (https://scikit-learn.org/stable/modules/tree.html|https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.RandomOverSampler.html) that showcased that I could easily use the RandomOverSampler and DecisionTreeClassifer so that I could use the .predict() function that was recommended to do so to get the final analysis. I also needed to do some modifications with NumPy itself by using get_dummies, as it was a simple way to get the data formatted correctly so that I wouldn't get an error with the data types that I was converting to and from. Not only that, I was able to use pickle, as we needed to make sure that the results from the machine learning model that I used were in fact "pickable". With that, all I did with that aspect was just using with(open()), which I found out how to use here

(https://stackoverflow.com/questions/11218477/how-can-i-use-pickle-to-save-a-dict-or-any-other-python-object). This was able to effectively display the accuracy results of the model within an actual output file.

Future Objectives:

For the future, one aspect that I would do would be to ensure that there is more accuracy within the model that I used. This is because I only got 63% of the total accuracy of the model, and that could be higher next time if I use a neural network or some other machine learning process. This is something that I can definitely improve on when working with a company, as I can make sure I do research on more libraries and other neural networks like GAN and CNNs which will provide the best performance in the long run.