The business question: For retail firms involved in e-commerce, how can we identify and reduce the number of online orders that are likely to be cancelled?

The data set used for this project included orders from US customers. It ranges full year 2015, 2016, and includes orders for the first 9 month of 2017. The data itself is one row per order and includes details about the order, buyer, and seller. It's clean and ready to use for the most part. Only the zip-code attribute included NA's, although that attribute was not used for this analysis.

At the beginning of the project, I intended on answering the question of 'How can this company make more profit?'. However, the data description file for the dataset did not indicate whether the currencies shown were in USD or local. Further investigations showed that although I was able to limit a sub set of the data to include only US buyers, I could not confirm currencies were in USD: I picked some items from the data set and compared their final price with an equivalent price from a similar product I Googled and found instances where the prices where outside a reasonable price range. After failing to receive an email confirmation from the publisher of the dataset, I decided to alter the business question independent of any currencies. Because the data set includes attributes indicating geographical locations and timestamps of when the orders where placed, I first mapped out the data by each one of these attributes separately. The geographic display of the data showed orders concentrated most in major metropolitan cities, with less frequent clusters toward central US. Although one would expect a uniform placement of orders throughout the US given that we are dealing with e-commerce, it is important to remember that our data was limited to one retailer, in the business of selling apparel, electronics, and sports items. We should expect a much fuller map of orders if, for example, we were working with Amazon data or a much bigger company. The data showed no time-dependent patterns at the hour, month, quarter, or yearly levels, except for day of week. Graphing orders by day of week for each year, showed a recurrent pattern: many orders placed on Monday, Thursday, and Friday. Because of the various products sold, I broke each item into three major buckets: electronics, sports, and apparel. This greatly simplified the attribute for item-type when reading the model. To simplify the geographical variable, I condensed all the geo-locations into 9 major areas: pacific, west-south-central, west-north-central, south-atlantic, mountain, new-england, middle-atlantic, east-south-central, and east-north-central. Bucketing the geographical coordinates into the 9 major regions allows for a more digestible interpretation of the model while retaining the uniqueness and behavior of each geographical region. Although each of these attributes were key in identifying features that could make an order more likely to cancel, it was necessary to include interactions between some of these variables that could further explain the nature of what makes orders more likely to cancel. For example, after finding that more orders were placed toward the start of the week and again as the weekend approached, I was able to introduce into the model an interaction attribute that helped explain the psychological behavior in a customer's desire to order more items as the weekend approached than at the start. Or an interaction variable that accounts for weekly and geographical locations, capturing locations where the retailer might be more established and thus indicate whether orders from these places are less likely to cancel.

I decided to use logistic classification for this dataset for two major reasons. First, the interpretation of the model allows the business to identify which attribute of an order leads to the highest likelihood of cancellation. For example, after finding that 3 items orders are more likely to be cancelled than 1 item orders, the retailer can thus proceed in further investigating what is it about 3-item orders that increase its chances of being cancelled. Questions such as: is our 3-item product bundle combination deal not working as we expected? Is there a glitch in our online platform when a customer places a 3-item order? The interpretation of the logistic classifying model gives us further insight into how each attribute contributes to the prediction of a cancelled order. Although we do have other classification models that could've been used, such as Support Vector Machines and Regression Trees, they do not provide the unique insight the logistic regression model provides. For example, to use Regression Trees as our predicting model would have us making definitive decisions at each step, making each attribute of an order dependent on the previous attribute when making the prediction. Furthermore, it would not indicate how each attribute contributes to the likelihood of a cancellation, the model would just indicate the conditional requirements for the order to meet to be classified as likely to cancel. Alternatively, the logistic classification model attaches the likelihood of the cancellation of an order to each attribute, and as a result, provides us insight into the contribution of each orders' attribute on the likelihood of cancellation.

As mentioned earlier, the logistic classification model provides the retailer information about what makes an order more likely to cancel, in addition to the actual prediction. This modeling approach provides a unique perspective on an order's characteristics, such as finding that 3-item orders are more likely to cancel than 1-item orders. These findings allow the company to identify possible areas of improvement and implement changes that could later decrease the chances an order being cancelled. I also found that electronic orders placed in the north central region are just over 25% more likely to cancel when compared to order on apparel items placed by pacific-region customers. This would prove useful to the company because it opens the discussion of replicating the better business strategy from the pacific region over to the north central region.

The retailer would feed the model data every 2-3 hours, receive the prediction within minutes, identify those orders at risk of cancellation, and kick-start a promotional offer to those customers, ideally before kickstarting the shipping cycle of the product. Effectively, this incentivizes the customer to not cancel, and thus safeguards the retailer from costs associated with cancelled orders midway through the shipping chain. Additionally, it prevents the retailer from introducing new costs to customer in the form a shipping & handling price increase. This price increase will likely deter customers from placing an order in the first place.

The testing suite is mostly straightforward. Because the model's inputs are all categorical, the testing suite examines whether the user made valid entries. For example, the client input only accepts entries of type 'customer', 'business', and 'home office' entries, any entry outside of this would not allow the model to make a prediction. Similarly, the product, region and quantity inputs have a set of predefined entries that are accepted. To simplify things, the customer, product, and region entries are accepted in all lower case, all capitalized, or a mix of both. Quantity only accepts integers between 1 and 5, mostly due to data containing this range—perhaps the retailer's online platform does not allow for orders beyond 5 or some other external explanation. The date entry when the order was placed is accepted in the "MM/DD/YYYY HH:MM" format. The testing suite examines for this type, an order outside of this format will not be accepted. The time stamp is required here in the for where the retailer decides to supply us with additional order information. For example, in the future when have data on the time or cancellation, we could easily begin to classify orders who were cancelled shortly after placement compared to those cancelled hours after placement.

After working the data, there are additional pieces of information that could be used to make a much more profound analysis and thus a more robust model. For example, if the data had more information about the cancellation—such as, how long after was the order cancelled—it would have allowed for more interaction effects in the model, explaining more of the reason why the order was cancelled initially. An immediately cancelled order compared to those cancelled a few hours after placement provides additional insight: Did the customer cancel the order accidentally? Did the customer find the same product from an alternative retailer at a much cheaper price? Answers to these questions would be addressed in our model had we this information.