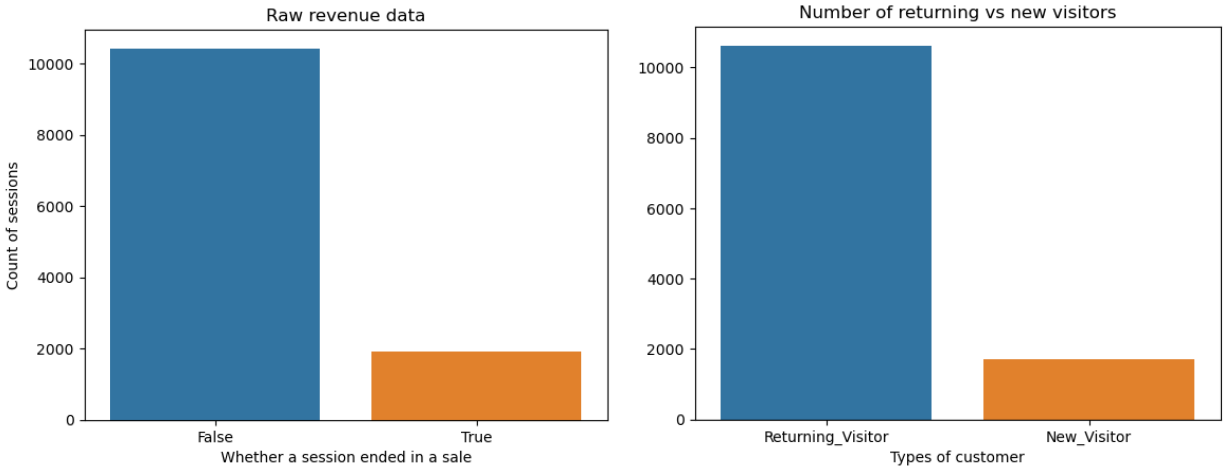
Shopping Intention and Customer Classification Analysis

As technology continues to advance and become a more prominent part of our lives, window shopping - and shopping in general - has become increasingly popular to do online. Companies need to adapt and figure out new marketing tactics since what works in store may not necessarily work online. Is it best to try and capture new customers? Is it better to try and get returning customers? With these two questions in mind, I decided to analyze browsing sessions from an online shop and test several ensemble models to see if it can correctly classify customers and their intentions based on browsing behaviors.

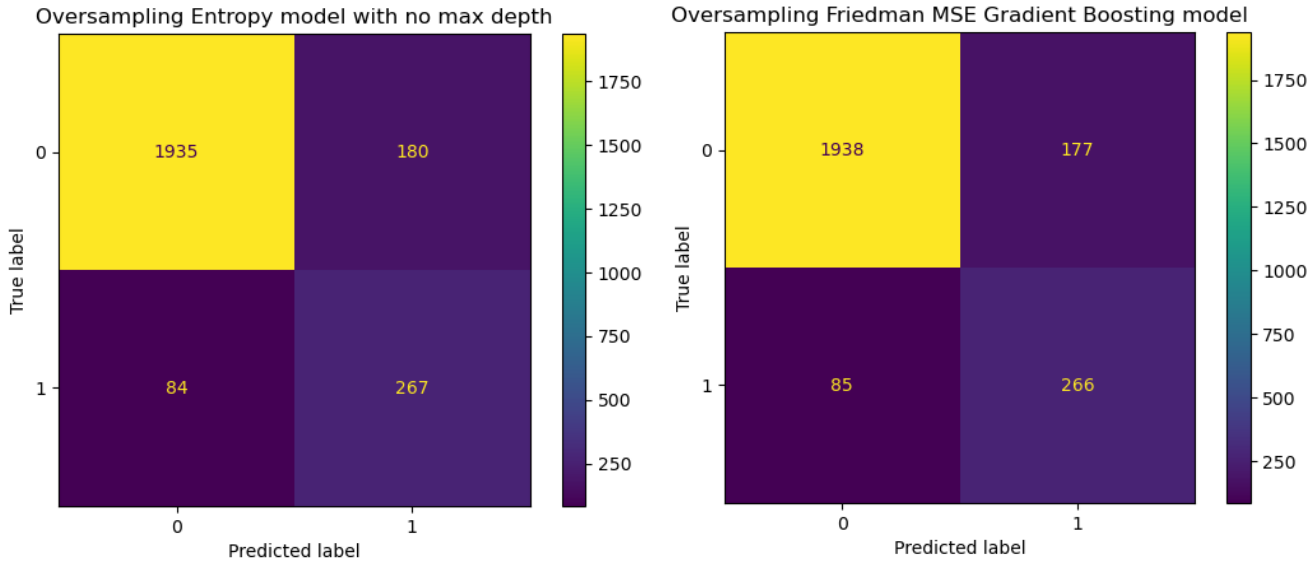
To answer this question, let's first look at the data I retrieved from Kaggle. The two things I wanted to focus on were how many sessions ended in a sale, and the amount of the two types of customer coming to browse the site. As we can see, both are heavily imbalanced which means that when evaluating a model we cannot just rely on a model's accuracy, since it could just dump every data point into the majority class and still give ~80% accuracy.



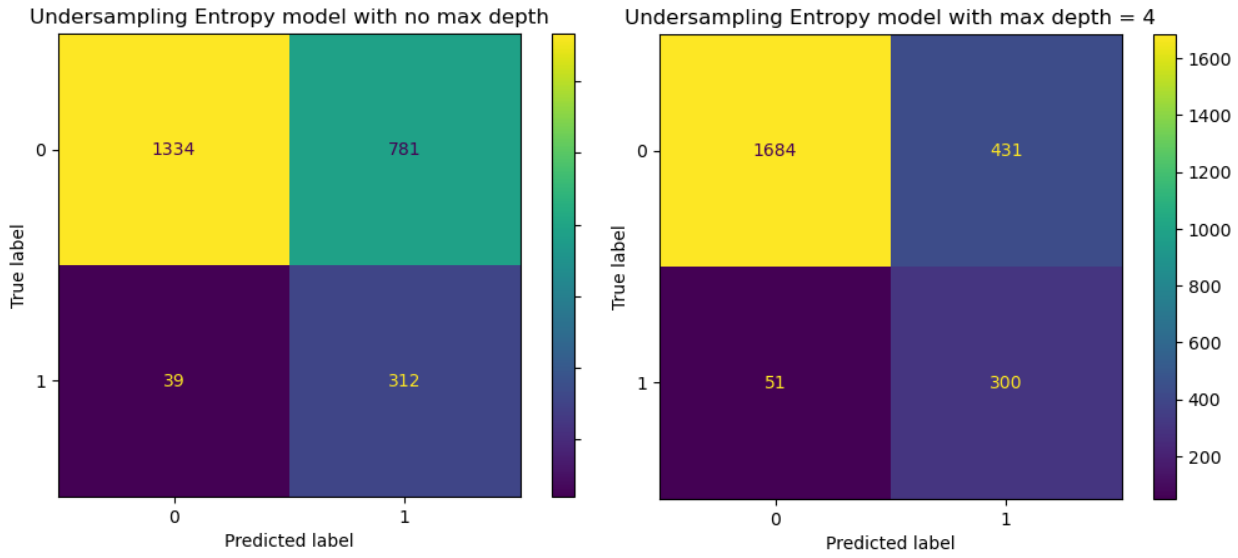
We want to focus on a well rounded model when classifying revenue, and a model that has a minimal amount of false positives (classifying new customers as returning) for visitor types. The hope would be to be able to use the same classification algorithm for both with only minor tweaks in the code to keep things simplified.

I first focused on finding the best model to classify revenue data. While over sampling, Random Forests and Gradient Boosting performed the best overall, with F1 scores of .90. I figured something like this would happen, which is why I then go on to test the same models

with under sampled data. If either of the two models perform better with less data, then that is the one I would move forward with because it is more adaptable to new data, or lack thereof.



The specified depth Random Forests did much better in handling the false positives without much sacrifice to the true positives. It makes sense that having a specified depth would improve the score in this case - the model was most likely overfitting on the small amount of data it had before, and by trimming the trees we let the model be a little more robust to new data it encounters. The specified max depth cut down on false positives by around 300 while also doing better on the true positives and false negatives. Gradient Boosting struggled in a similar fashion to the Random Forests without specified depth, which makes choosing the right model for this problem an easy one. With a little more tweaking on the depth we could further refine the Entropy model to give us better results.



Next we can shift our focus to classification of customers - whether someone is new to the site or a returning customer. All of my testing was pretty similar to what I did with the revenue classification, except this time when evaluating models I was mostly concerned with false positives - which in this case means the model classifying a new customer as a returning one. Because I don't need a well rounded model, it made going through the algorithms a lot simpler. Much to my delight, the Random Forests performed well while over and under sampling, meaning that we can use the same algorithm for revenue classification and customer type classification.

Moving forward, I would look into what it takes to push little pop-ups to new customers to entice them to sign up for a rewards program or offer them a special discount when they make their first purchase, as well as pop-ups for customers who are likely to make a purchase to do some kind of 'bundle deal' so they spend a little more while also thinking they're getting a really good deal. We could also utilize the people that are just window shopping and push more on site ads to the pages that they click on and spend time looking at.