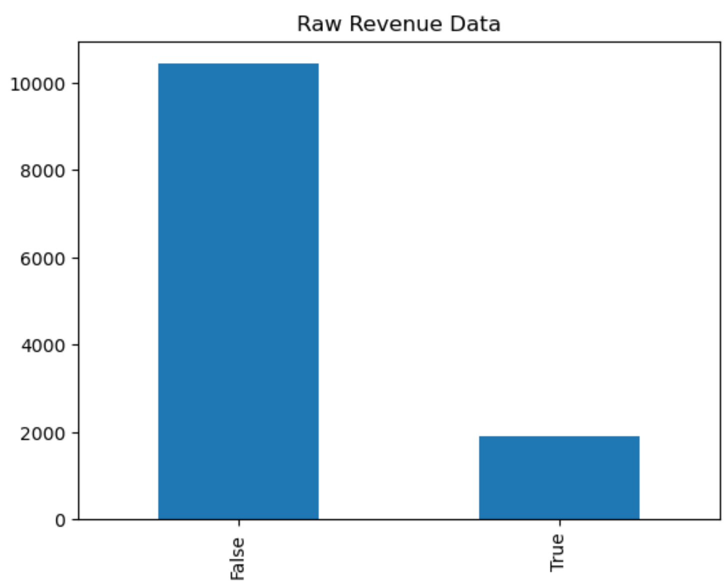
Shopping Intention Analysis

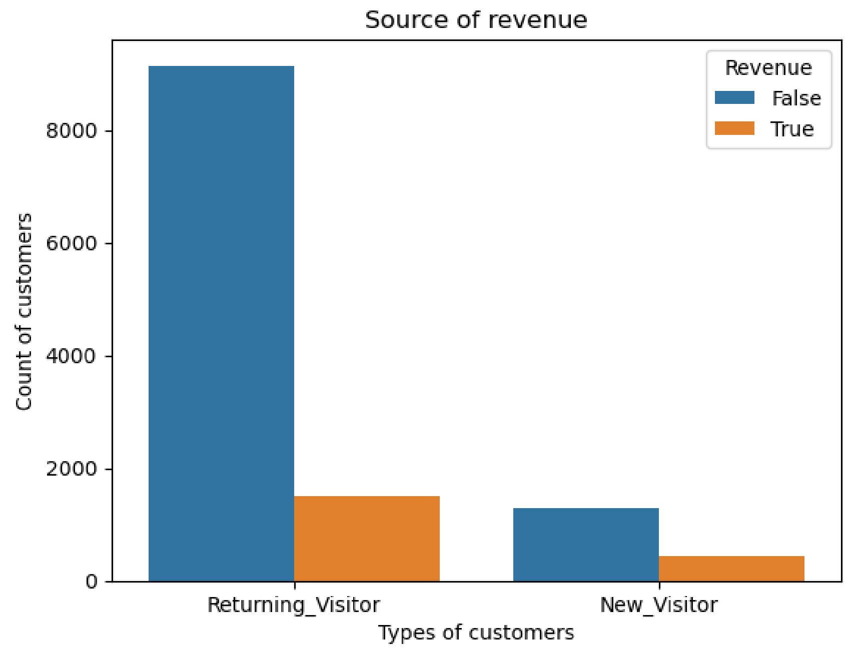
Window shopping is not a new concept - and with the rise in online shopping, online window shopping has increased as a way to pass the time. However, online window shopping by itself drives no profit. As technology continues to advance, ordering items online has become almost second nature, only helped along by the Covid pandemic. As malls start to become defunct more companies need to consider the customer experience online in order to ensure a continued steady flow of revenue. What we're trying to do here is figure out if there is a pattern to customers who do end up buying something when they come to a website and see what customers who are just 'window shopping' do when they're online. Once this information is collected we can determine the next best steps to increase revenue.

We begin with a dataset taken from Kaggle, whose current split between buyers and non-buyers is as follows:



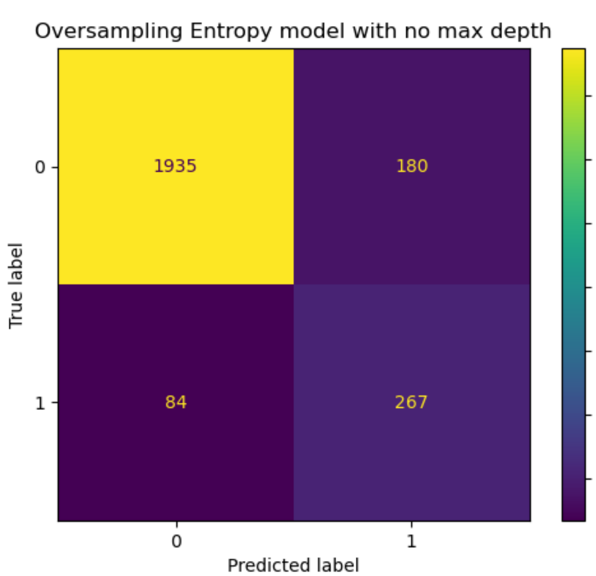
It is because of this large imbalance that I over and under sample my data when testing my models, because the 82-18 split on its own could lead to misleading results. It is also for this reason that I won't just be relying on accuracy to determine the best model. As it is now, if our model simply classified every session as not ending in a sale, it would still give us 82% accuracy - but what happens if the new data we're given actually has more sessions where revenue was earned? We would then miss out on opportunities to better advertise to these buyers and try to get them to spend more money - or at least look at ads and generate revenue for us in that way.

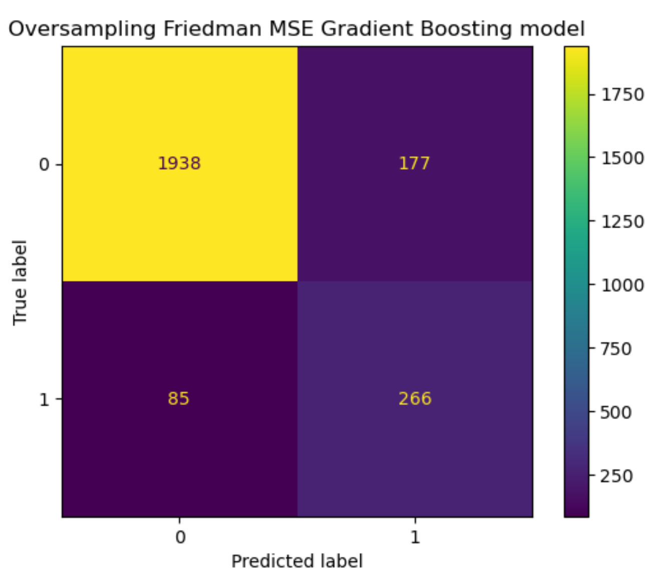
I decided to take a look and see what the division was between new and returning customers and whether their browsing session ended in a sale. This is what I found:



While this information won't necessarily help us make the model classify better, it's good information to keep in mind when thinking of next steps after we have selected a model.

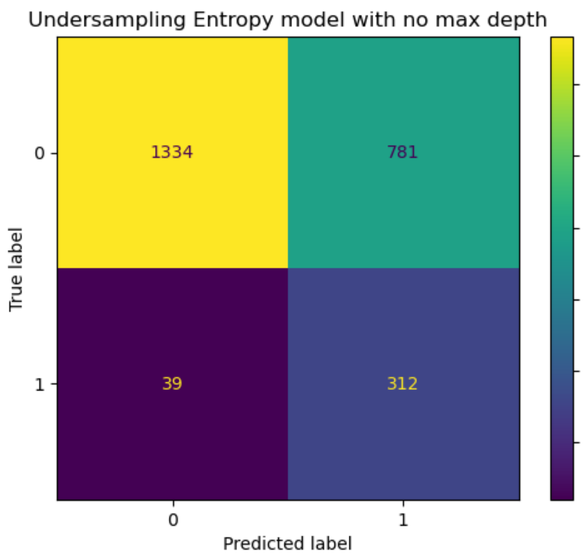
All of the ensemble learning models scored very closely to each other which makes it hard to pick which one is truly "the best." Random forests without a specified depth and gradient boosting all scored around .93 (confusion matrix for both pictured below) when calculating their AUC curve, so I pushed my focus more towards these models while I moved on to under sampling.

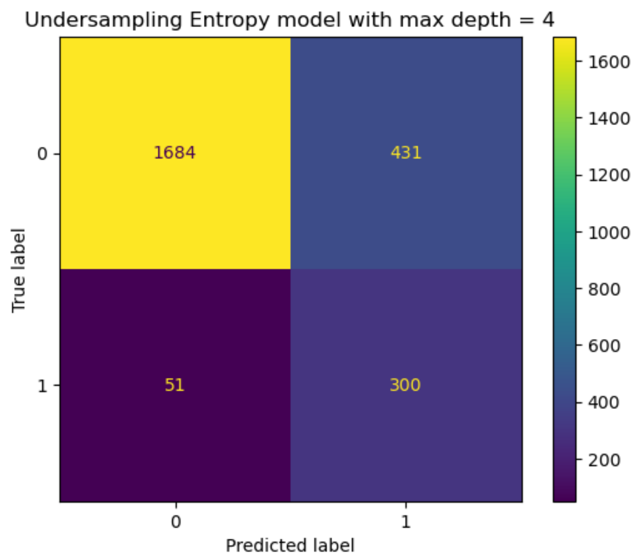




If certain models perform well when under sampling, I would consider them the most robust and my top choice for classification. While I knew the models wouldn't do as well with the under sampling, I was surprised to see how much the models misclassified false positives.

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. This makes me believe that if we had a higher max depth on the over sampling forests we would see better results.





Ultimately, the random forest models performed the best overall and I would likely choose the entropy model to move forward with. It did very well when classifying while over sampling, and did pretty decent when under sampling. It adapted the best out of all of the models that I tested here and I feel confident in its abilities moving forward.

Having this information, I would suggest the client, if they haven't already, start a kind of rewards system to entice more people to come back since returning customers are their main source of income - or find a way to reward returning customers if they refer a friend and the friend makes a purchase. They could also advertise a discount when the customer goes to checkout if they hit a certain amount of money. Alternatively, they could cater ads towards their customer base more, since not a lot of new people stick around anyway, and get affiliate click-through links from other companies if they advertise things off-site.

I would consider doing further research on how other companies run their rewards programs and incentivise customers to return, as well as look into the psychology behind people's reaction towards 'big sales' - whether it's something to pursue and take advantage of or if efforts are better put elsewhere. It might also be worth looking into whether certain website designs work better to get people to click on products and/or buy them, as well as the efficacy of advertising the store's website on other sites.