Wyatt Rasmussen

DSC 680

Customer Propensity Modeling

Business Problem

In order to increase the revenue at our online shopping website we want to increase the sales of our products by looking into customer purchasing habits. By recording how users are going through the buying process we can look for actions that indicate potential high propensity to purchase.

Background/History

Propensity modeling is often used as a method that will determine the likelihood of a user will complete a purchase based on previous actions users have taken. For example, if we find that historically users that look at the cart tend to complete the buying process we can begin to model for future users who look at the cart. This can give us a model that takes into account all of the users actions and give us a probability that a particular user will complete the buying process. Once we have this information we can target users who went through the process, but didn't complete it based on the likelihood they would have purchased.

Data Explanation

For this project we have two datasets. One is a training dataset of historical data. This has over 450,000 instances of user shopping experiences complete with whether the user completed the checkout process and bought items. It also contains 25 total features including, but not limited to: user sign in, device type, user view checkout, etc.

There is also a testing dataset that is over 150,000 instances of yesterday's users shopping experiences. This dataset has 100% of users not completing the buying process in able to apply a probability that a user will purchase the items they were shopping for. This dataset also include all of the features that were included in the training dataset as well.

Methods

There are a few methods that go into building a customer propensity model. The first method that I followed was creating an exploratory data analysis. In doing this since all of our data is true/false data I created a series of count plots that count the number of responses for each variable broken up with the number of each response that ordered.

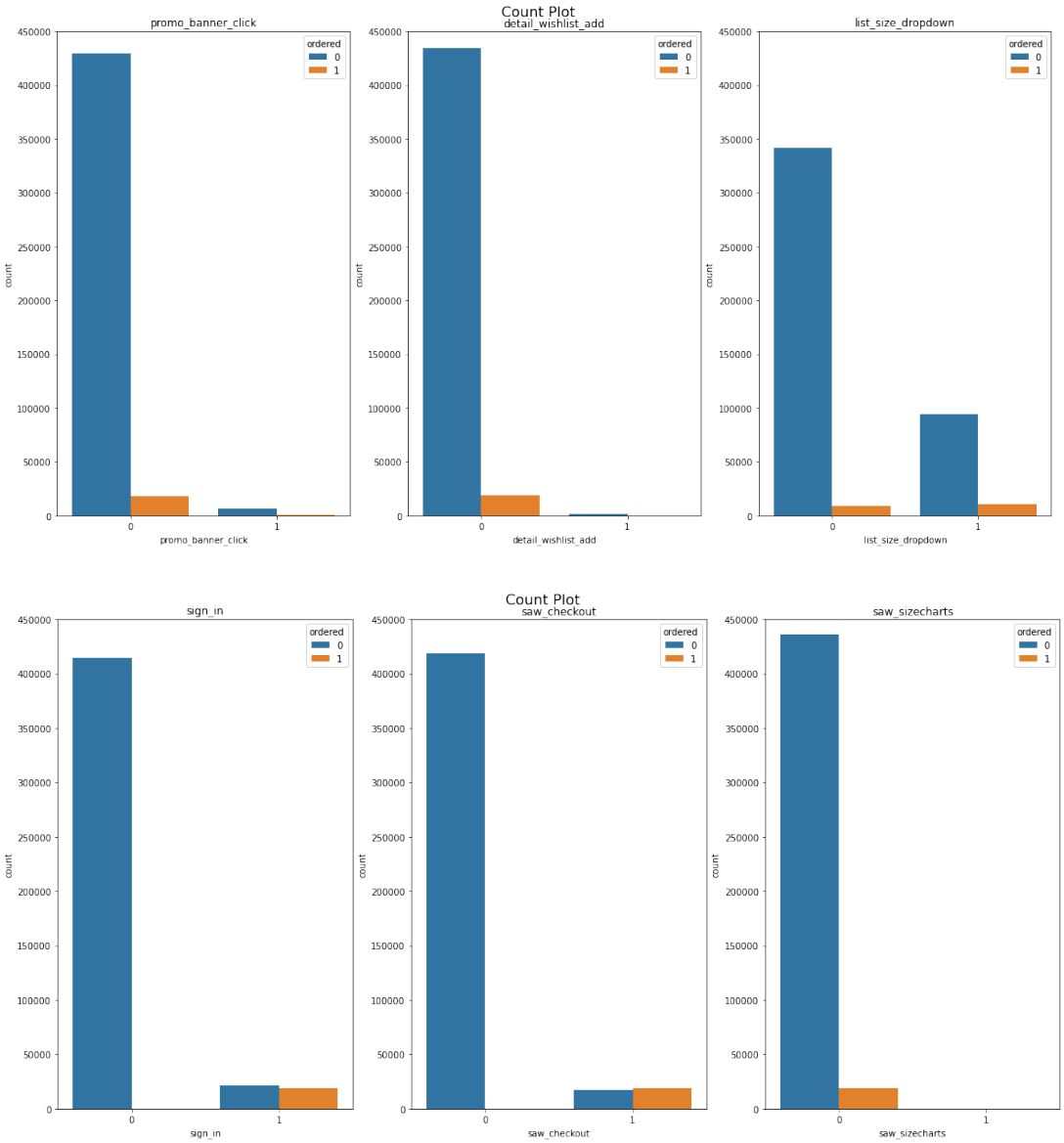
The next primary method was selecting the features that had the highest correlation to buying in the end. To do this I created a correlation heatmap and found the correlations of each of the different features. There correlations of the features would vary quite a bit allowing me to take two methods of feature selection, using features over 0.15 correlation and using features over 0.02 correlation.

Next, I created two different types of models that could predict the user behaviors. Those two models were Gaussian Naïve Bayes and Logistic Regression. These two models perform best in predicting binary outcomes which is the case here. In total 4 different models were created using the different feature selection and the different models. After looking at the 4 models I was able to determine that feature reduction using correlation produced the best results. Taking that information I was able to implement hyper parameter tuning in order to get the best parameters for the models.

After determining the most successful model, the last step is using the features on the testing dataset to determine the probability of any given user to purchase. By using sklearn's predict_proba function I was able to use the models created in the last step and implement that into my code. At this point you can deduct that any probability higher than 0.500 is more likely to purchase than not purchase. The next step is to separate the users with over 0.500 probability in a data frame of User ID and the probability to purchase and target those users.

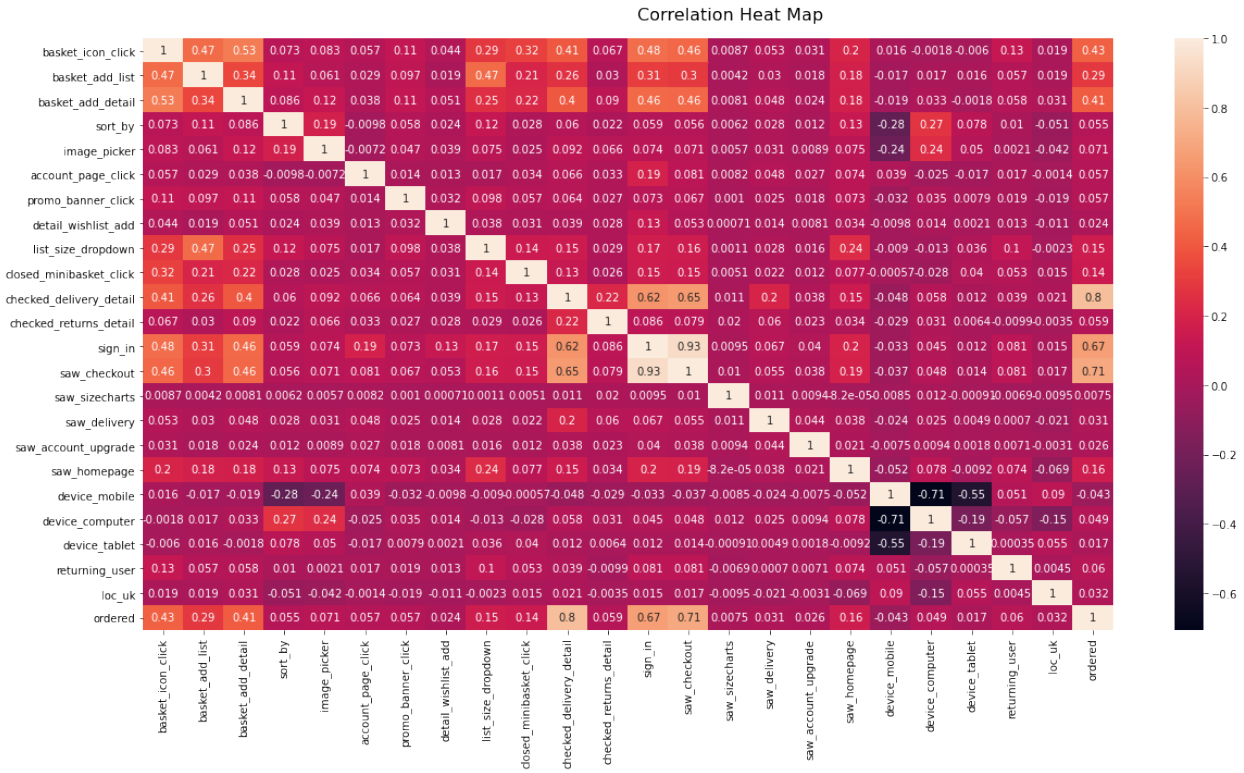
Analysis

The data visualizations on this helped a ton with understanding the relationships of each of the features with whether the user ordered or not. Below are a few of the different count plots results.



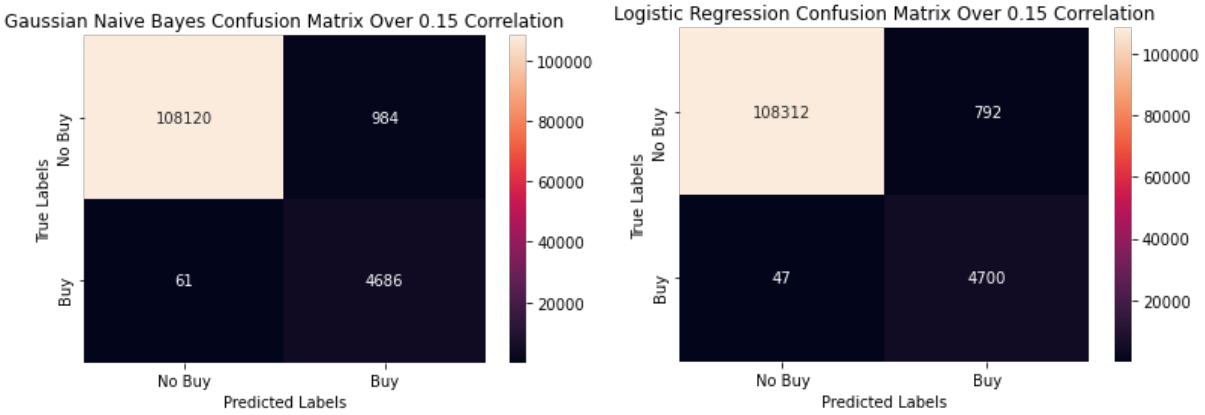
I felt that these visualizations did a good job of showing the buying habits depending on action. As you can see not many users looked at size charts or added to wishlist. This likely means that these features aren't as important when it comes to predicting the buyers habits. While the number of users who saw the checkout or signed in had about a 50% likelihood of buying the items. This is where we can start looking at positive buying habits.

The next step to start understanding the buying habits was to look at the correlation. Below is a correlation heatmap that was created to understand the relationships.



The most important part of this correlation heat map is the ordered row. Generally speaking the lighter shades are the items that have a higher correlation and thus more of a predictive relationship with whether the customer ordered.

After feature selection, the next step was to create the models that were mentioned in the methods section. As mentioned in the methods I got a simple baseline of which of which features were more successful. By using hyperparameter tuning I was able to get the best parameters to use on the data and to produce the models with the best scores. Those models are below using the confusion matrix in order to show accuracy of the guesses.



Both of these models performed very well in the end with the Gaussian Naïve Bayes producing an accuracy of 0.9908 and the Logistic Regression producing an accuracy of 0.9926.

After getting these models built, the next step is propensity modeling. For propensity modeling I continued using both of the models above and limited the testing dataset of over 150,000 to use the 8 features that were selected above. Using predict_proba we were able to attach a probability to each of the Users on the probability that they would purchase. For the Gaussian Naïve Bayes model, we were able to identify 1458 users that didn't purchase that had over 0.500 probability to complete the purchase. For the Logistic Regression model, we identified 1182 users that didn't purchase. In both of these cases if all were successfully transitioned to purchases that would just under 1% of these visits which would see a significant increase in revenue for our ecommerce site.

Conclusion

Using the methods that we used above we can begin to determine the purchasing habits of the users at out online retail store. By using feature reduction to just 8 features and using effective modeling, we can get a pretty good grasp of whether a user will buy a product or not. We can use this data to build a propensity model to target the users that had over 0.500

probability to purchase. Based on our results from looking at yesterday's data we could theoretically target 1% of our daily site visitors who don't purchase to increase revenue. That would lead the company to see a significant increase in total sales after a whole year.

Assumptions

There are a few assumptions that are made here that should be addressed. Not all users are the same and thus generalizing every user to determine whether they will purchase is not always accurate. Another assumption we make here is that every user is buying the same products, some products likely have a higher probability to purchase when they are shopping because they may be on sale or an essential item to every day life. If we could narrow down to certain products this would likely be more accurate.

Limitations

As mentioned in the previous section I think the biggest limitation here is that we are generalizing for all products rather than focusing on a singular product. This would likely lead to varied results across different product lines.

Challenges

I think the biggest challenges in this problem is the highly unbalanced data that is being used here. In total for the training data there was a significant difference in the users who bought items and the users who didn't buy items.

Future Uses/Additional Applications

This could be used in the future for a company to target the users that had a higher likelihood of purchasing with sales on the items they added to their cart or could be used to determine which users are likely to spend more money at our store. The possibilities for

increasing revenue are endless provided we have more data such as product type or maybe even time of day for shopping.

Recommendations

My recommendations would be to use this modeling technique in more areas of the business in order to identify consumer habits and increase the revenues of our company. By introducing and finding more features we can only continue to strengthen the modeling here.

Implementation Plan

The implementation plan for this project should be to continue to refine this model with more features and begin to target the users who are over XX probability to purchase an item. This will increase the revenue of the company and help push product out the door to the users.

Ethical Assessment

Ethically, the only concern here would be that users data is being used in order to drive the revenue for a company. If users weren't aware that this was happening it could potentially scare users away from using our platform since some users can be really private with their data. Targeting those users would likely lead to a loss of sale.

References

AltexSoft. (2021, August 10). *Predictive lead scoring: Discovering best-fit prospects with machine learning*. AltexSoft. Retrieved April 3, 2022, from https://www.altexsoft.com/blog/predictive-lead-scoring/

Mullin, S. (2020, September 25). *Propensity modeling: Using data (and expertise) to predict behavior*. CXL. Retrieved April 3, 2022, from https://cxl.com/blog/propensity-modeling/

P, B., & Das, P. (2018, June 1). *Customer propensity to purchase dataset*. Kaggle. Retrieved April 3, 2022, from https://www.kaggle.com/benpowis/customer-propensity-to-purchase-data?select=testing_sample.csv

P, B., & Das, P. (2018, June 1). *Customer propensity to purchase dataset*. Kaggle. Retrieved April 3, 2022, from https://www.kaggle.com/benpowis/customer-propensity-to-purchase-data?select=training_sample.csv