



Customer Churn Analysis

04.29.2023

April Kim, Boran Sheu, Kai Zhang, Olivia Lee, Sreekar Lanka



Executive Overview

This project aimed to analyze the customer data of YETI.com and identify the key drivers of churn to identify new strategies for customer acquisition and churn prevention. This will enable YETI to save approximately 2.2M dollars annually and strengthen the company's customer base and brand.

To achieve our goal, we tried to answer three business questions in this project:

- 1) What is the appropriate definition of customer churn
- 2) What are the key drivers of customer churn
- 3) Do these drivers differ by customer segment?

We were given access to the customer data of US e-commerce YETI.com stored in GCP, and we only analyzed purchases from users with two or more purchases. The project was divided into two steps: defining churn and analyzing churn drivers.

To define churn, we used unsupervised learning techniques with a parallel structure. They ran K-means twice, once with only the frequency of recency to determine whether a customer had churned or not, and then compared the distribution of the mean frequency and mean recency to define the target variable, churn or no churn. On the other end, we used the other side of the K-means to segment the customer base into eight different clusters based on purchase history, age categories, preferences, and other features. We reduced the 21 features to seven principal components for feature-reduction purposes. The result of the clustering technique was eight different segments, and we interpreted the customer profile for each segment.

We then utilized XGboost as the main model to analyze the key drivers of churn. We analyzed the feature importance for each cluster and determined the actual effect of these drivers on churn propensity, and made four major findings that can be seen in the figure below.

	<u>Finding</u>	<u>Action</u>	<u>Expected Return</u> *Assumption: success rate = 20%
1	Customization has an increasing impact on churn propensity	Target gift givers	Lowers churn propensity by 6% and generates business value of \$1,744,604.
2	Interest in more categories increase churn propensity for Accessory Fans but decrease for Outdoors/Cargo Fans.	Losing accessory fans to other channels?	Lowers churn propensity by 1.3% and generates business value of \$251,349.
3	Dedicated hard cooler fans display more loyalty.	Targeted 'not-so-pure' hard cooler fans	Lowers churn propensity by 3% and generates business value of \$26,243.
4	Lifetime decreases churn propensity at earlier stage for Hard Cooler Fans, compared to Super Fans	Targeted 40-50 month old hard cooler fan	Lowers churn propensity by 10% and generates business value of \$97,788.

Our findings identified different actions that could save YETI approximately 2.2M dollars. For example, for one of the clusters, we found that the customers were primarily one-time purchasers who had not interacted with the brand after the initial purchase and suggested that YETI could offer a loyalty program or a post-purchase follow-up to retain these customers.

Overall, this project provides valuable insights into customer churn and helps YETI acquire new strategies to improve its business performance. By understanding the key drivers of churn, YETI can implement targeted actions to retain customers and strengthen its brand.



Business Outcomes

Business Context

The client, YETI, operates in the outdoor lifestyle industry and specializes in outdoor products, such as coolers, drinkware, and bags which leads to \$1B in annual revenue. This data science project was launched to understand the key drivers of churn to provide different perspectives on customer churning. In a world where competition for customers is ever-increasing, understanding key drivers of churn will allow us to increase our retention rates by preventing churn and identifying new acquisition strategies. Both will, in turn, strengthen YETI's customer base and, thus, the brand. Furthermore, a churn model could highlight certain operational inefficiencies YETI can tackle as a business.

Business Questions


We aimed to answer three business questions: 1) What is the appropriate definition of customer churn based on the timeframe? 2) What are the key drivers of customer churn, and 3) Do these drivers differ by customer segment? Understanding different key drivers of churn for different customer segments will enable us to provide YETI with more efficient strategies that target specific customers to reduce churn and reinforce revenue.

Business Outcomes

From this project, the client can expect to identify key drivers of customer churn for YETI products over a specified time, measured by retention rate and customer lifetime value, and quantify dollar value for an increase in revenue opportunity by preventing churn.

Our team explored the key drivers of churn across various customer segments and extracted four significant insights that could provide YETI with fresh perspectives to retain customers. After analyzing the impact of these insights, we calculated the expected business value of preventing customer churn. This was done by adding the retention rate and new acquisition amount and multiplying it with the customer lifetime value. We assumed that the success rate of preventing churn would be 20%. By using this approach, YETI could understand the potential financial benefits of our suggested strategies for preventing customer churn.

First, we discovered that customization greatly impacted churn propensity across three customer segments: Drinkware Fans, Accessory Fans, and High-spending Customers. We



suggest targeting gift-givers by identifying them through their behavior and creating gift baskets and gift subscriptions to encourage diversity in the products received. This would lower churn propensity by 6% and generate a business value of \$1,744,604 annually.

Second, we found that accessory fans are more likely to churn as they have more options outside of YETI, while outdoor cargo fans are more loyal. We suggest YETI develop strategies to prevent losing accessory fans to other channels, such as creating YETI-exclusive items and promoting them more heavily to entice accessory fans and increase their loyalty. This would lower churn propensity by 1.3% and generate a business value of \$251,349 per year.

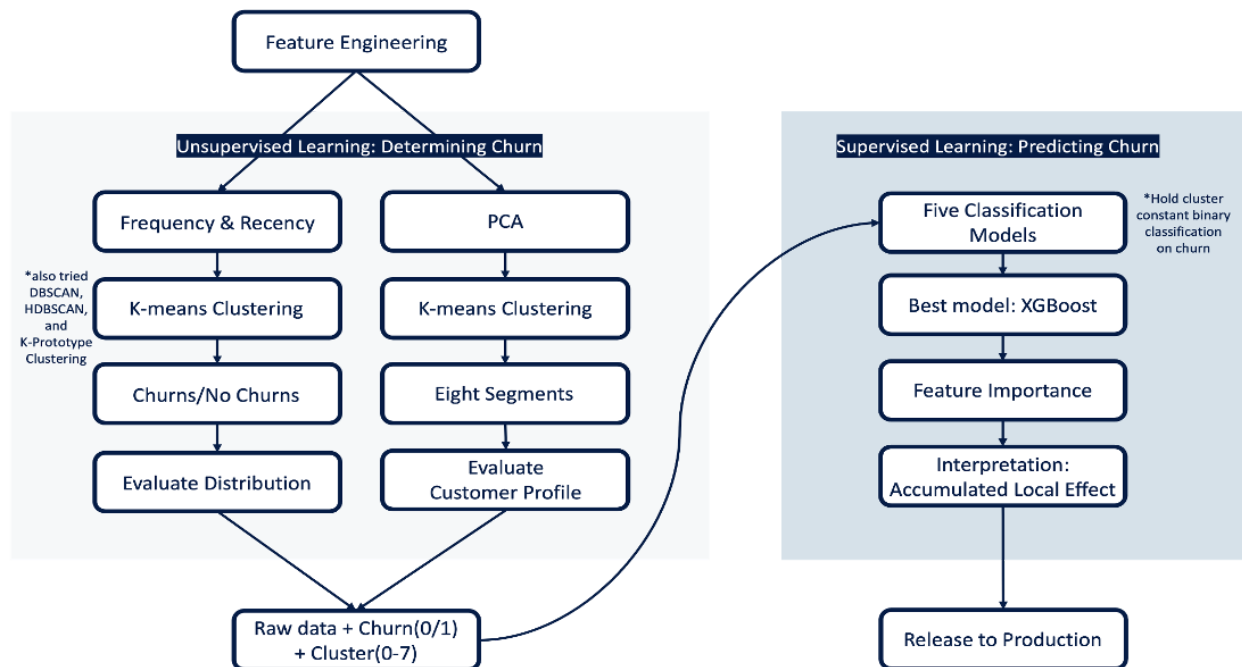
Third, we discovered that adding a small amount of drinkware to a hard cooler fan's purchase can decrease churn, but too much can increase it. YETI should encourage "pure" hard cooler fans to purchase high-value items like soft coolers or cargo to increase loyalty and decrease churn, which would lower churn propensity by 3% and generate a business value of \$26,243 per year.

Lastly, we found a difference in the threshold for churn propensity between hard cooler and soft cooler fans, with hard cooler fans displaying higher churn propensity after 50-55 months compared to soft cooler fans after 40 months. We believe that targeting customers with lifetimes falling between 40 and 50 months to intervene at an earlier stage would lower churn propensity by 10% and generate a business value of \$97,788 per year.

The execution of YETI's new strategies based on our suggestion can generate a business value of \$2,119,984 in total per year.

Explore Source Data

We were provided with the customer data of US e-commerce yeti.com stored in GCP that we could access via BigQuery. The data contained information useful to inform the business decisions under consideration because it was real online customer data of high quality, and it was well-prepared with existing keys between tables so that we could spend more time on data preparation and joining the appropriate tables to create the input file rather than cleaning low-quality data.



The first step in our analysis was defining churn using unsupervised learning techniques. We ran k-means twice, once with only frequency and recency, to define churn or no churn. With recency outside of a certain range of frequency rate, we will define this group as a churn group. We then used the other side of the k-means to separate our customer base into different segments based on purchase history, age categories, and preferences, among other features. Through this process, we obtained two new columns: "churn" and "cluster number."

To analyze the drivers of churn, we utilized XG Boost as our main model. We analyzed feature importance for each cluster, which allowed us to determine what makes a feature important for the model to make accurate decisions, as well as the actual effect of these drivers on churn propensity.

	Frequency & Recency	Purchase History	Product Category	Color Preference	Email Interaction	Web Action
		<ul style="list-style-type: none"> Customer lifetime Purchase time Lifetime spent Lifetime quantity Per order spent Per order quantity Discount frequency order Discount frequency product 	<ul style="list-style-type: none"> Distinct category count Max entry product price Outdoor equipment pct Cargoal pct Soft cooler pct Bags pct Drinkwear pct Hard cooler pct Other pct 	<ul style="list-style-type: none"> Black White Navy Seaform Stainless Charcoal Nordic Purple Harvest Red Alpine Yellow Other <p>*pct</p>	<ul style="list-style-type: none"> Email received Email open rate Unsubscribed (0/1/2) 	<ul style="list-style-type: none"> Product added Product removed Product added to wishlist Customizer started Customizer completed Cart viewed Checkout started Order cancelled <p>*count</p>
Unsupervised	✓	✓	✓	✗	✗	✗
Supervised	✗	✓	✓	✓	✓	✓

We took a top-down approach to identify relevant features, including frequency and recency, purchase history, product categories, color preferences, email interactions, and web actions. These features were used as predictors for churn propensity.

Overall, our analysis allowed us to define churn and identify drivers of churn for each customer segment so this information can be used to develop targeted strategies to reduce churn and increase customer lifetime value.

Analyses


Unsupervised Learning

Moving on to the first step of our project, which involves unsupervised learning, we explored four types of clustering methods. Eventually, we decided to use K-means due to the large size of our dataset, which includes 2.1 million unique users that have at least two purchase order. Our decision was reinforced by a graph that showed that K-means significantly lowered time complexity, especially when dealing with large data sizes. While we considered other techniques such as HDBSCAN and DBSCAN, which are density-based and centroid-based respectively, we opted for K-means as it provided a more comprehensive approach to our project. However, we may explore these other techniques in the future as they have the potential to make clusters tighter and leave out outliers.

After running K-means in two feature spaces and aggregating the results, we determined which clusters were associated with churn based on if the recency is larger than frequency plus two standard deviations of the standard deviation. Some groups were too close to call, so we left them blank. For example, if the group has a 15-month frequency with 16 months of recency, it is too close to call whether they are churn or no churn. This helped us define our target variables of 0 and 1 for churn or no churn.

cluster	avg_time_bt看_orders			recency	
	count	mean	std	mean	std
0	100877	25.07	3.99	6.63	3.65
1	301100	3.66	2.41	3.51	1.55
2	87900	15.16	4.62	27.56	3.92
3	24139	66.17	9.37	11.84	7.44
4	143364	15.26	4.21	16.00	2.85
5	261189	4.20	2.81	10.12	2.01
6	201419	12.61	2.83	5.10	2.60
7	285157	3.65	2.86	18.42	2.79
8	45532	35.22	7.10	23.74	5.55
9	42098	42.53	5.87	8.59	4.78
10	196794	2.97	2.68	29.23	3.58
11	15525	6.28	6.75	53.93	10.66





We classified eight clusters into three types: super fans, category-specific clusters, and newer customers. Super fans are the most loyal YETI fans, who make purchases across multiple categories and spend a lot. Category-specific clusters focus on specific product types such as hard coolers and drinkware. Lastly, Newer customers include discount users and big spenders which are low in size but are high in volume and purchase times. We played around with the number of clusters and found that the eight clusters created interesting separations in category preferences and spending behaviors.

- **Cluster 0: 'Super Fan': 75k** Oldest customer group, made lots of purchase, spending high across most product categories
- **Cluster 1: 'Hard Cooler Fan': 95k** Infrequent buyers, once and done, first purchase is probably a hard cooler
- **Cluster 2: 'Drinkware Fan': 418k** Infrequent buyers, low spent, doesn't buy anything but drinkware
- **Cluster 3: 'Soft Cooler/Bag Fan': 77k** Infrequent buyers, once and done, first purchase is probably a soft cooler or bag
- **Cluster 4: 'Accessory Fan': 556k** High quantity on cheap item quantity
- **Cluster 5: 'Outdoor/Cargo Fan': 30k** Dabble a bit in ~3 categories, least drinkware purchases
- **Cluster 6: 'Discount User': 250k** Newest customer group, bought a lot discounted drinkware and other items
- **Cluster 7: 'Big Spender': 14k** Newest customer group, High volume, high purchase times

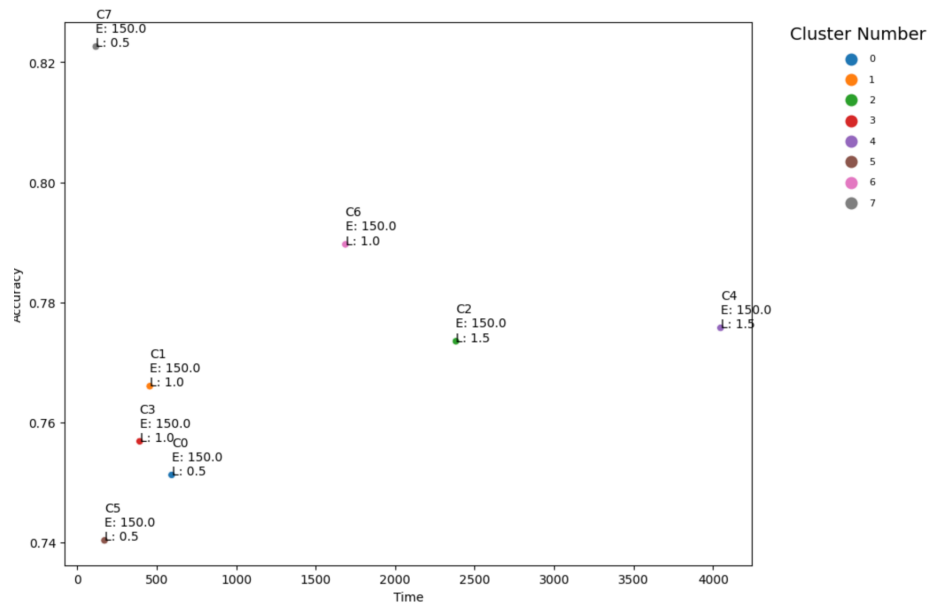
Overall, K-means helped us identify clusters associated with churn and provided interesting insights into our customer base. The clusters vary in size and churn ratios across the board, and we can use this information to tailor our marketing and retention strategies.

Supervised Learning

Moving on to the topic of supervised learning, we defined our target as a binary classification problem where the objective was to predict whether a customer would churn or not. Our approach to this binary classification problem was to go from simple models to more complex ones. The models we tried are 1) Logistic regression (Baseline Model), 2) KNN, 3) Decision Tree Classifier, 4) Random Forest, 5) AdaBoost, and 6) XGBoost.

Due to its simplicity and interpretability, logistic regression was the first model we tried in the hope of setting it as our baseline model. But logistic regression could have been better at dealing with non-linear trend data. Then, we tried KNN, but it did not give moderately accurate results on our dataset, with low accuracies ranging from 60-70%, especially compared to other ensemble methods, and also not providing us with feature importances. We tried running AdaBoost, thinking that it is good at dealing with binary classification problems, so we moved on to tuning hyperparameters, hoping to get more insights from

the model. The parameters we adjusted were the “Estimators,” “Learning rate,” and the depths of the trees, and the results are shown below.



However, the feature importance was too similar within clusters.

Cluster	0	1	2	3	4	5	6	7	Average
Feature									
months_elapsed	0.08	0.20	0.19	0.20	0.15	0.15	0.17	0.08	0.17
cart_viewed_count	0.11	0.10	0.09	0.08	0.09	0.09	0.09	0.08	0.10
lifetime_spent	0.02	0.08	0.18	0.07	0.16	0.08	0.05	0.06	0.10
per_order_spent	0.03	0.09	0.11	0.10	0.13	0.08	0.06	0.05	0.09
purchase_times	0.05	0.06	0.04	0.05	0.07	0.06	0.06	0.04	0.06
nordic_purple_pct	0.07	0.04	0.03	0.04	0.04	0.03	0.07	0.04	0.05
product_added_count	0.04	0.04	0.05	0.03	0.04	0.05	0.05	0.04	0.05
charcoal_pct	0.05	0.01	0.04	0.02	0.05	0.01	0.09	0.04	0.04
checkout_started_count	0.05	0.03	0.03	0.03	0.01	0.04	0.02	0.04	0.04
alpine_yellow_pct	0.03	0.03	0.03	0.02	0.02	0.03	0.04	0.03	0.03
discount_frequency_order	0.04	0.01	0.01	0.02	0.01	0.02	0.03	0.04	0.03
email_received	0.04	0.02	0.02	0.03	0.01	0.03	0.02	0.02	0.03
discount_frequency_product	0.03	0.03	0.02	0.03	0.01	0.02	0.01	0.01	0.02
product_removed_count	0.03	0.02	0.01	0.03	0.01	0.02	0.02	0.02	0.02
customizer_started_count	0.02	0.02	0.01	0.00	0.02	0.02	0.02	0.04	0.02
drinkware_pct	0.01	0.03	0.00	0.02	0.01	0.02	0.02	0.03	0.02

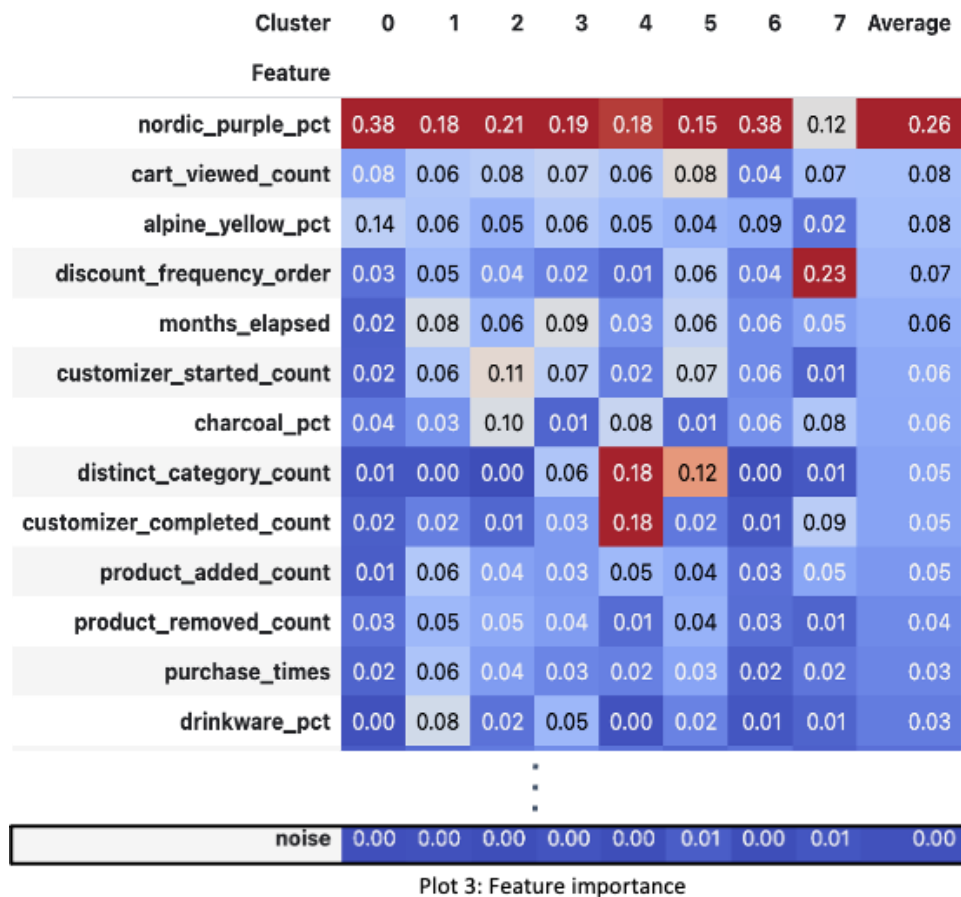
The plot above shows that all clusters take the “months elapsed” and “cart_viewed count” main features. This result provided us with the idea that it could be reasonable to consider a customer's lifetime as the critical decision factor of churning. However, we wanted to investigate further for more key insights.

After running several experimental models introduced briefly above, we concluded that XG Boost was our best model as it was the most effective in dealing with large-scale datasets and provided equalized feature importance.

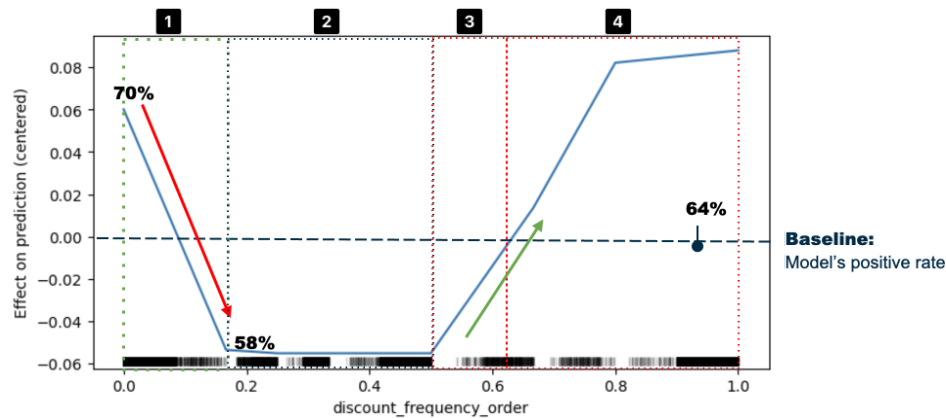
Model	Accuracy	Model Type
Logistic Regression (Base model)	74%	Probabilistic Binary classifier
KNN	70%	Distance-based model
Decision Tree Classifier	77%	Tree-based classifier
Random Forest	79%	Tree-based Ensemble
Gradient Boosting Classifier	81%	Tree-based Ensemble
XgBoost Classifier	82%	Boosted Tree-based Ensemble

Next, we performed hyperparameter tuning for the number of estimators, learning rate, and max depths, to achieve high accuracy rates and lower computing times. One of the characteristics of max depths is it gives us more equalized features. However, the deeper the depth might cause the overfitting of the data. After running the model with hyperparameters, we identified the optimal hyperparameters to be 150 estimators, a learning rate of 0.5, and a max depth of 2. These parameters yielded satisfactory accuracy rates while maintaining efficient computation time and also without any sign of overfitting.

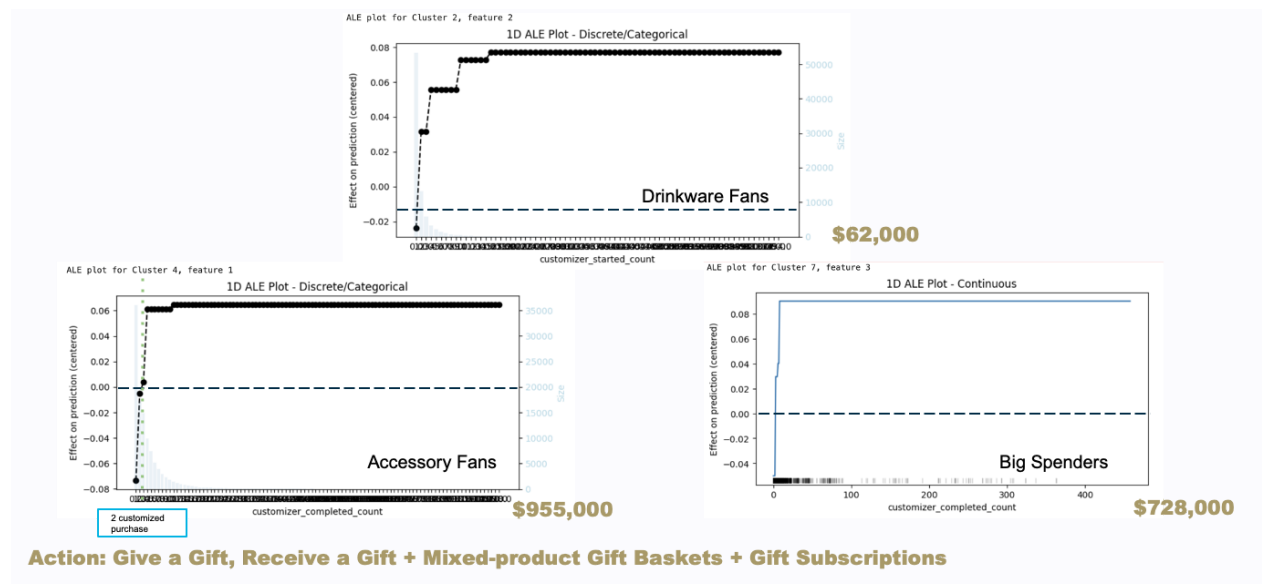
To determine the feature importance, we introduced some noise into the data and were surprised with the result. We expected month_elapsed would be the most important feature, but it turned out that the nordic_purple_pct scores the highest in the model and had the most significant impact on churn prediction. We predicted it could be because Nordic Purple was the only discounted color item from the YETI.



Then, we used the ALE (Accumulated Local Effects) plot to measure the importance of the features and found that changes in the discount frequency percentage had a significant impact on churn propensity. We provided an example of discount frequency as a feature in our ALE plot analysis. The exponent shows the baseline churn prediction of 64%, which is the zero on the graph. The y-axis represents the marginal effect on churn propensity, while the X-axis represents the unit change of the feature (discount frequency). At 0% discount frequency, the churn propensity is 70%, 6% higher than the baseline. However, as the discount frequency increases to 20%, the churn propensity decreases to 58%, which is 6% lower than the baseline. The churn propensity stabilizes between phase two and phase four. The ALE plot provides a visualization of how the churn propensity changes with continuous changes in the discount frequency feature.

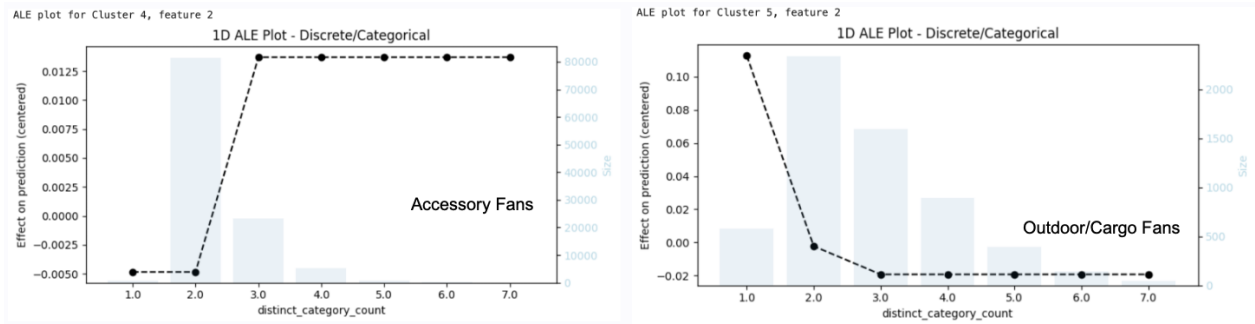


In summary, we adopted a rigorous methodology to ensure our supervised learning model was efficient, accurate, and not overfitting. We identified the optimal hyperparameters and feature importance using the ALE plot, and our findings allowed us to make informed decisions about predicting customer churn. We would like to introduce our four significant results below.



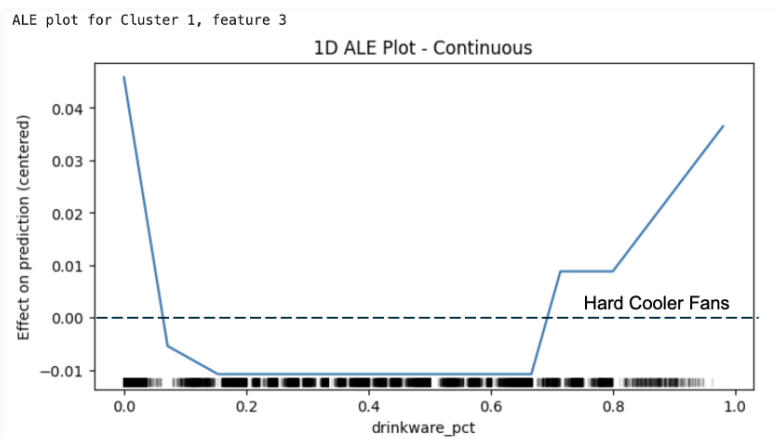
Our first insight into the data revealed the high impact of customization on churn propensity across the three customer segments: Drinkware Fans, Accessory Fans, and High-spending Customers. As shown in the ALE plots, churn propensity shoots up by 8% and continues to increase by 8% for every increase in the customization count. While

customization is a unique feature that customers want, it is interesting to note that it may lead to higher churn rates.



Action: Losing accessory customers to other channels? Push new categories(freebies?) + create YETI.com exclusive items \$250,000

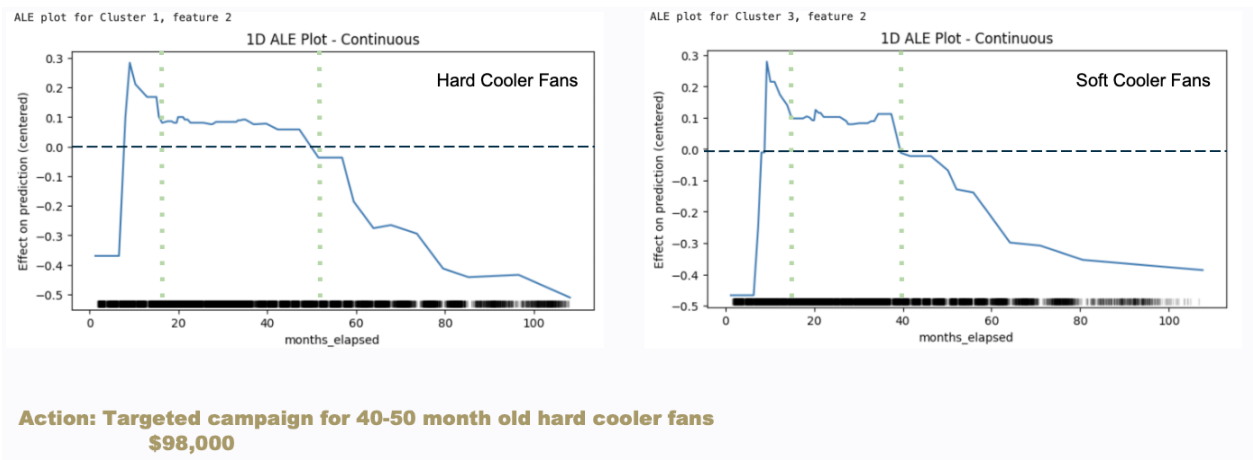
The second insight focused on the distinct category accounts and the relationship between customer exposure and churn propensity. We found that the more exposure accessory fans receive, the more likely they are to churn. In contrast, the more exposure outdoor cargo fans receive, the less likely they are to churn. We hypothesized that this could be because outdoor cargo fans are more likely to buy products from yeti.com instead of other channels, while accessory fans have more options.



Action: Keep hard cooler fans interested in more hard coolers + Targeted campaign on 'not-so-pure' hard cooler fans \$26,000

Insight three focused on the relationship between hard cooler fans and their drinkware percentage or how much they deviate from being a "pure" hard cooler fan. We found that if

hard cooler fans purchase even a small amount of drinkware, their churn propensity drops significantly. However, there is a certain threshold where the more drinkware hard cooler fans purchase, the more likely they are to churn. We cannot make the same interpretation for customers who started with hard coolers as their first purchase, as their behavior may differ.



Our final insight is related to customer lifetime, an essential feature that predicts the future churn propensity of a YETI customer. We analyzed the churn propensity between hard cooler and soft cooler fans, and we observed that both customer segments follow an identical trend throughout their lifetime, indicating a decrease in churn propensity. However, we noticed a difference in the threshold for hardcore fans, which is around 50 to 55 months, whereas, for soft cooler fans, it could be as low as 40 months. Based on our analysis, we assume that the trend for decreasing churn propensity is similar for both customer segments, and we can lower the threshold for hard cooler fans to match that of soft cooler fans.

These insights became valuable information for us to make recommendations to help YETI make informed decisions to reduce churn and increase revenue. By addressing the issues identified through these insights, which will be further explained in detail in the next section, YETI could retain more customers and increase customer loyalty, ultimately leading to higher revenue and profitability.



Recommendations

Our first finding was that customization has an increasing impact on churn propensity for three customer segments: Drinkware Fans, Accessory Fans, and High-spending Customers. We suggest targeting gift-givers by identifying them through their behavior and creating gift baskets and gift subscriptions to encourage diversity in the products received. This strategy could help reduce churn rates by offering more diverse product options to customers introduced to YETI just once for customized gifts.


Our second finding was that having an interest in more than two product types has an increasing impact on churn propensity for Accessory Fans but a decreasing impact for Outdoors/Cargo Fans. To address this, we propose creating YETI-exclusive items and pushing more diverse marketing materials to entice accessory fans and offer them unique products they cannot find elsewhere. This approach could help reduce churn rates among accessory fans and increase their loyalty to YETI.

We also found hard cooler fans who are also interested in drinkware exhibit a lower propensity for churn, while those who are purely dedicated to hard coolers tend to display higher levels of loyalty. To keep “pure” hard cooler fans, YETI should encourage them to purchase high-value items by soft coolers and or cargo. Even though hard cooler fans who are also interested in drinkware exhibit a lower propensity for churn, those who are pure hard cooler fans tend to display higher levels of loyalty. By focusing on these areas, we can potentially decrease churn and increase customer lifetime value.

Last but not least, our last finding was that customer lifetime has a decreasing impact on churn at an earlier stage for Hard Cooler Fans(40 months), compared to Soft Cooler Fans(50 months), which tend to show a similar effect to. Our recommendation is to target customers whose lifetime falls between 40 and 50 months, as we believe that there is something important happening during that timeline, and we can intervene at an earlier stage to reduce the drop in churn propensity.

The following are the recommendations for the next steps for potential operational executive plans YETI could follow to utilize our findings while addressing some limitations.

Firstly, improving the feature space could be the first step the company should take for potential improvement. There may be additional features, other than what we used to



determine and predict churn in this project, that could be added to increase the accuracy of churn prediction and enable better identification of churn. A new metric could be formulated by incorporating more features into the model, providing a deeper understanding of the reasons behind customer churn.

Secondly, more clustering methods could be utilized. Although we used K-means due to its computational efficiency and time complexity, there are other methods available, such as DBSCAN, density-based methods, and other hierarchical methods. Although these methods may be slightly more computationally intensive, they may yield more accurate results in certain circumstances.

Thirdly, feature effects could be examined in greater detail. We looked at a couple of features to see how they impact the financial propensity metric. However, this approach could be scaled up to include all available features. This would enable identifying more interesting features that could explain why a particular customer is churning.

The fourth area for improvement is explainability or interpretability. Other than ALE, different measures to understand the features, such as LIME/SHAP, can be utilized, as they are more model agnostic and may yield more accurate results.

Finally, we recommend YETI examine repeat purchase probability. In this project, we were only considering customers that have placed at least two orders, which represents 40% of the entire set of customers. However, by adding more features and trying those features with more unsupervised methods, the company could determine the repeat purchase probability for customers that have placed only one order. This would enable them to predict the probability that a customer would make another order, providing valuable insights into customer behavior.



Conclusion

The project aimed to help YETI improve customer retention and identify key drivers of churn to increase revenue and strengthen its brand. Using unsupervised and supervised learning techniques, we could cluster customers based on purchase history and predict customer churn. The team identified four significant insights and suggested strategies to prevent churn for different customer segments. The insights gained from the clustering and predictive models can be used to tailor marketing and retention strategies. We suggest further improvements, including utilizing more clustering methods, examining feature effects in greater detail, improving explainability or interpretability, and examining repeat purchase probability. These recommendations could help YETI decrease churn and increase customer lifetime value. Overall, the strategies proposed by the team are expected to generate significant business value for YETI, with a projected value of \$2,119,984 per year.

References

1. GeeksforGeeks. (2023, April 22). *Principal component analysis with python*. GeeksforGeeks. Retrieved April 30, 2023, from <https://www.geeksforgeeks.org/principal-component-analysis-with-python/>
2. Molnar, C. (2023, March 2). *Interpretable machine learning*. 8.2 Accumulated Local Effects (ALE) Plot. Retrieved April 30, 2023, from <https://christophm.github.io/interpretable-ml-book/ale.html>
3. XGBoost documentation. XGBoost Documentation - xgboost 1.7.5 documentation. (n.d.). Retrieved April 30, 2023, from <https://xgboost.readthedocs.io/en/stable/>