Cross Sell Report

1. Overview

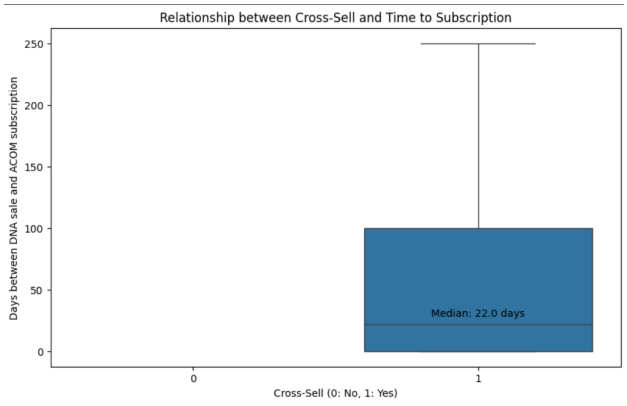
The key question in the exploratory data analysis is whether to focus on the entire dataset, including customers who did not subscribe to ACOM, or narrow it down to only those who made a new subscription (xsell_gsa equal to 1). By analyzing the entire dataset, a comprehensive understanding of customer behavior can be gained, but this introduces class imbalance, a classification task of 220,726 that were already ACOM subscribers or did not purchase the ACOM subscription and 31,216 customers that qualify as a cross-sell. Alternatively, focusing solely on new ACOM subscribers may simplify the analysis but could result in a biased view. More targeted analysis specifically geared toward cross-selling behavior can be created by narrowing down to the subset with xsell_gsa equal to 1, which can be done in one line of code (commented out in the script). The entire dataset is analyzed with a few trends and next steps identified.

2. Data Cleaning Process

Notable preprocessing steps include the 'ordercreatedate' column converted to datetime format, and 12 rows with the creation date of 1900-01-01 are imputed with the mean. Feature engineering is then applied to extract date components from the order date, and a binary indicator, 'is_weekend,' is created to signify whether the purchase occurred on a weekend. Further feature engineering involves one-hot encoding for the 'regtenure' column, using both one-hot encoding and label encoding for the 'customer_type_group' column, and defining the response variable 'cross_sell' based on the specific conditions related to 'xsell_gsa' and 'xsell_day_exact.' Activation date features, such as 'is_weekend_a' and 'activation_minus_order,' are introduced. NaN's are already mostly dropped with the creation of 'cross_sell'. In the activation date column, NaN's are imputed as '-1', so a machine learning algorithm can handle it properly.

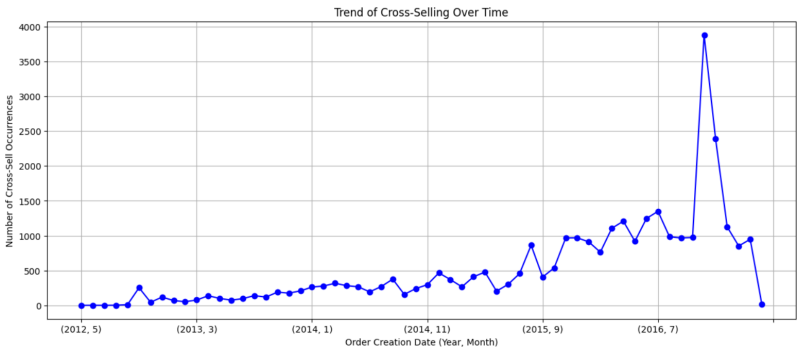
2.a. Trend 1

This box plot compares the distribution of the time it takes for customers to subscribe to ACOM after purchasing a DNA product. The median time it takes for a customer to purchase the subscription after the DNA product is 22 days.



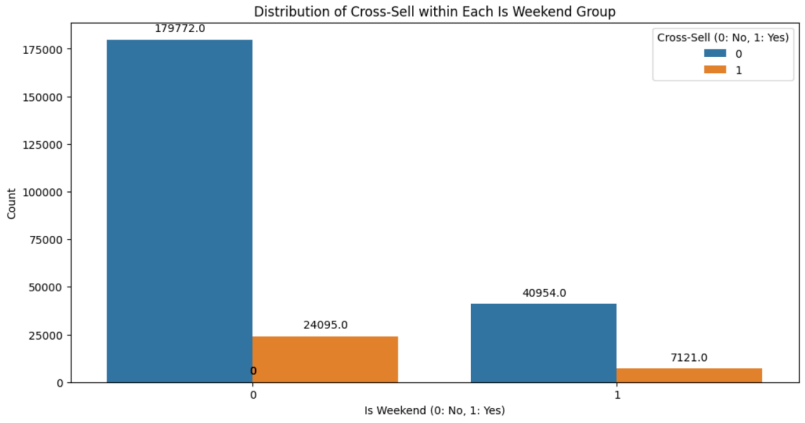
2.b. Trend 2

This line graph counts the occurrences of cross-selling for each month, and visualizes the trend over the dataset period. This time series gives insight into when the number of cross-sell occurrences were the highest. In addition, from this graph, it is apparent that there has been a general uptrend of cross-sell occurrences.



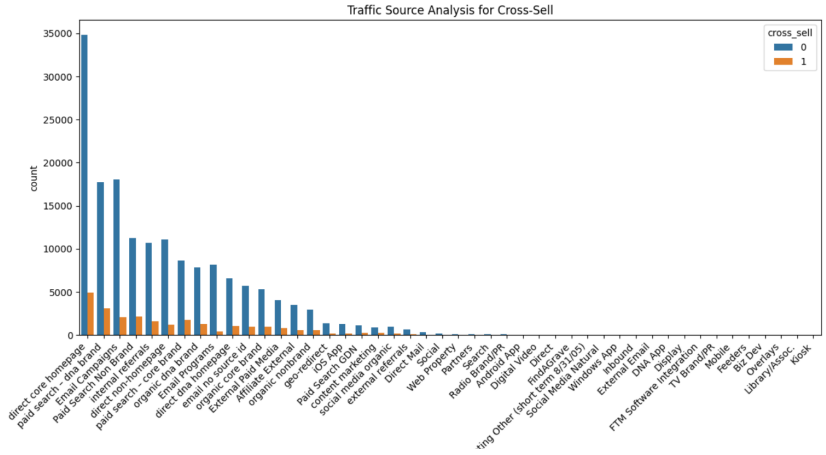
2.c. Trend 3

This bar chat visualizes the distribution of cross-sell within different groups based on the 'is_weekend_a' feature. This feature represents whether an activation occurred on a weekend (1) or not (0). The cross-sell rate is 17% on the weekend and 13% when it is not, providing insight into time based sell rates which can range from seasonality to weekly.



2.d. Trend 4

This graph shows distribution of cross-sell occurrences across different traffic sources. Most traffic, and therefore cross-sell occurrences, happen from the direct core homepage, while around half of the categories receive little to no traffic.



3. Next Steps

With this initial exploratory analysis, a few general trends are created, which can provide more insight into where and when cross-sell occurrences occur. Whether it is looking for monthly patterns, or traffic-based condition patterns, these initial trends can be the starting point for further analysis. The fraction of the observed customers cross-sell to subscription is 12.39%. The dataset is properly preprocessed for machine learning and analysis. The dataset exhibits a notable scarcity of instances where the target variable is equal to 1, indicating a cross-sell occurrence. Given this imbalance, it becomes imperative to address class imbalance to enhance predictive model performance. Class imbalance techniques are essential to mitigate the impact of skewed class distribution. For such scenarios, machine learning algorithms designed to handle imbalanced datasets prove beneficial. A logistic regression can be used initially because it serves as a fundamental and effective method due to its simplicity and robust performance. A Support Vector Classifier (SVC) can be employed, leveraging its ability to handle class imbalance. Furthermore, ensemble methods such as XGBoost can be used in addition to oversampling or adjusting the class weights to produce an effective model, addressing the relationship for cross-sell rates.