E-Commerce Subscription Study Analysis

Aastha Amin, Lucas Chin, Doan Nguyen, Xiaohui Ma, and Brady Snyder

Department of Statistics, University of California, Riverside

STAT171: General Statistical Models

Dr. Esra Kurum

20 March 2025

Table of Contents

Introduction:

The modern digital landscape is characterized by a proliferation of subscription services, a trend that has picked up in popularity since the dot-com days. (Rheude, 2025) While some services like Netflix and Disney+ are purely subscription-based, others like Spotify, Hulu, and YouTube offer both free and paid services, allowing users to opt for additional benefits like ad-free viewing and unlimited streaming. This variability in membership models raises the central question of what drives the subscription decision of customers for premium memberships. This study attempts to examine the drivers of such decisions through two central questions: What attributes have significant impacts on the behavior of a customer in subscribing for a premium membership?, and how may we effectively predict the willingness of a customer to subscribe for a premium membership? By understanding these dynamics, businesses are able to optimize customer retention, estimate revenue, and optimize growth through strategic expansion. Our primary objective is to determine the most appropriate statistical model to predict the subscription likelihood of customers based on relevant predictors.

The dataset we will be using for our analysis is the "E_commerce_subscription.csv" file given in the class page.

Dataset Description:

Below shows the description of each variable and their relevance towards the overall dataset:

Age (int) – Represents the customer's age. This helps analyze age-based trends in subscription likelihood, spending habits, and engagement with the e-commerce platform.

Income_Level (float) – The annual income of the customer. This provides insights into purchasing power, premium subscription preferences, and average cart size.

Avg_Cart_Size (float) – The average amount a customer spends per shopping session. This is a key metric to understand spending behavior and whether higher spenders are more likely to subscribe to premium services.

Website_Visit_Frequency (int) – The number of times a customer visits the website in a given period. Higher visit frequencies may indicate higher engagement and a greater likelihood of subscribing.

Marketing_Emails_Opened (int) – The number of marketing emails the customer has opened. This measures how responsive consumers are to email campaigns, helping assess the effectiveness of marketing strategies.

Device_Type (object) – The type of device used by the customer (e.g., Desktop, Mobile, Tablet). This can help edit platform optimizations and marketing efforts based on device preferences.

Loyalty_Program_Member (int) (binary) – Indicates whether the customer is part of a loyalty program (1 = Yes, 0 = No). Loyalty members may be more likely to subscribe to premium services and have higher retention rates.

Subscribed_Premium (int) (binary) – Our selected **target variable**, indicating whether a customer has subscribed to the premium membership (1 = Yes, 0 = No).

Data Preprocessing:
Below is a summary of the variables with key metrics such as the mean and standard deviation that hope to guide us in what data cleaning method we should do.

Variables	Mean/Count	Standard Deviation
Age	43.65	15.32
Income Level	82819	37114.06
Average Cart Size	253.78	143.36
Website Visit Frequency	27.04	14.01
Marketing Emails Opened	14.68	9.25
Device Type	Desktop: 138 Mobile: 136 Tablet: 26	NA (because Categorical Variable)
Loyalty Program Member	Yes: 208 No: 92	NA (because Categorical Variable)

There were no missing or NA values so data cleaning was not required.

EDA:

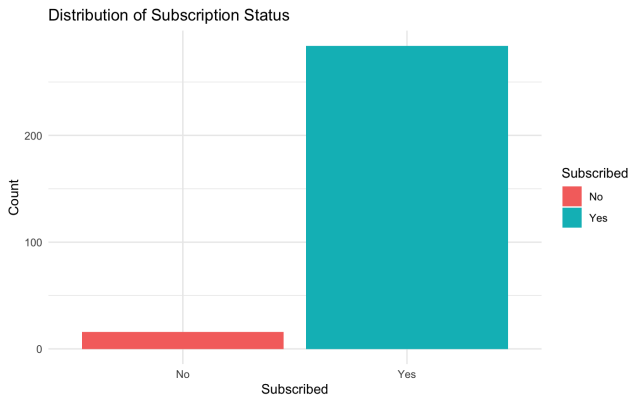
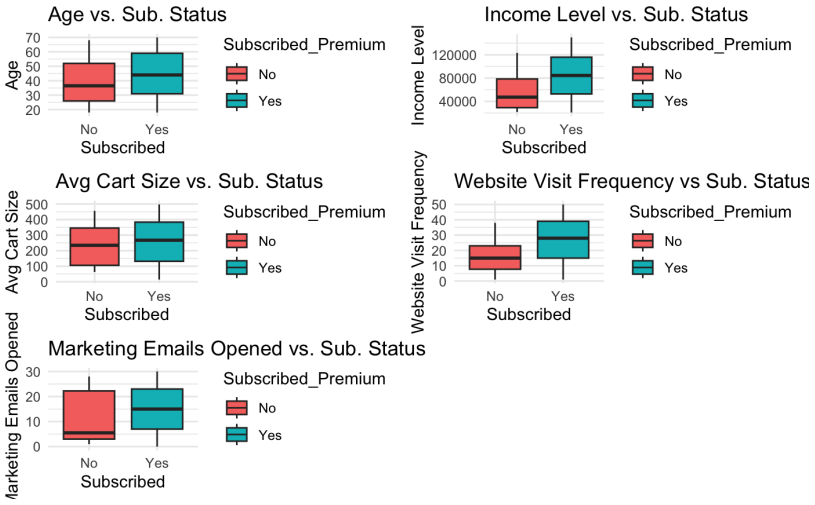


Figure 1: Overall, about 5.33% are not subscribed and 94.67% are subscribed to the premium membership in the dataset. So, less than 10% of our outcome data consisted of "No" which is something that may cause some issues in creating our model.

Figure 2: The average is higher in the numerical predictors for the people who have subscribed. Those with higher income levels tend to be subscribed. The average cart size for those who are and aren't subscribed appears to be close in range and mean. Also, the overall website frequency is higher for those with premium membership.



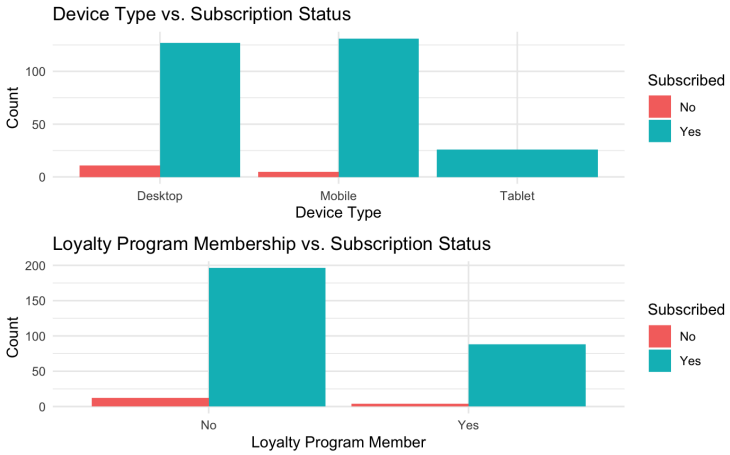
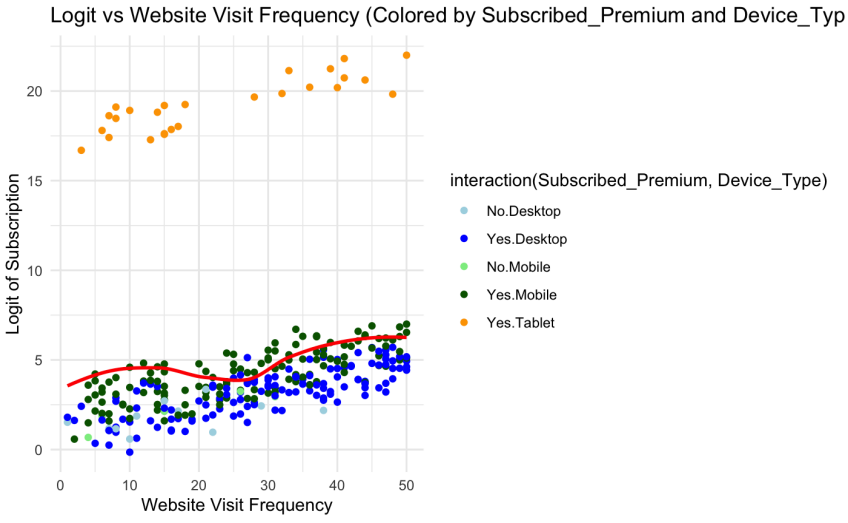


Figure 3: Those who use a tablet are all subscribed. More people who are not subscribed use a desktop compared to a mobile device. There is a higher number of users who are subscribed however are not a part of the loyalty program membership. Since, there is low data on those who are not subscribed to the outcome there may be issues in predicting the "No" response.

Figure 4: In the scatterplot on the right our data with points grouped by device type and subscription status. The shape of the plot appears to be the same for all of the combinations, however there appears to be a dip in the middle which appears to show a quadratic pattern as the gap appears to be from the lack of data from the tablet.



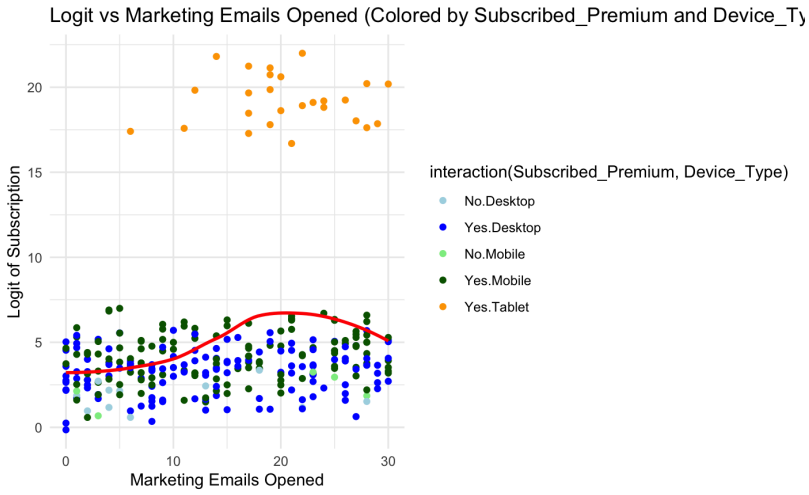


Figure 5: This scatterplot of the left is also grouped by device type and subscription status. We noticed a peak which is where the users who are subscribed and use tablets are clustered. This is another possibility for a squared term in our model. The other data points appear to follow the same pattern and using device type may skew our results if used in our model.

Methodology:

A generalized linear model with a binomial distribution and a logit link function was used to model the log-odds of subscribing to a premium membership. We ended up with the logistic regression GLM as our response variable is binary (0/1), with 1 indicating a user is subscribed and 0 indicating a user is not subscribed. This regression model provides insights into how different factors influence the log-odds of a customer subscribing to the premium service. Doing a logistic regression model ensures that the probability remains between 0 and 1 (IBM, 2024). In our model, each predictor's coefficient represents the change in log-odds for every one unit increase in that predictor, holding all other variables constant (UCLA, OARC Stats, n.d.). When we exponentiate the coefficient, that gives us the odds ratio which shows how the odds of subscribing changes when the predictor increases by one unit.

To reach our final model, we followed a systematic approach that involved multicollinearity tests, stepwise regression, and significance tests. Initially, we did a multicollinearity check with the main effects model to see if any coefficients are inflated due to correlation with other predictors. This test resulted in the variance inflation factor (VIF) values for the main effects to be all around 1, which let us start with a clean model where multicollinearity isn't a problem. We then did backwards elimination, starting with a model up to 3 way interactions along with squared values for numerical predictors. This caused perfect separation in the model–meaning one or more predictors perfectly predicted the outcome variable. We then removed all 3 way interactions and qualitative terms and their interactions. From here it was a combination of removing and adding predictors and comparing the Akaike information criterion (AIC) and Bayes information criterion (BIC) between each removal and addition of terms. Some predictors, while insignificant in the model, when removing it, caused the AIC to go up and caused other predictors that used to be significant to no longer have significance in the model. This most likely means that while the term has no significant effect in predicting the response, it still

contributes to the overall fit in some way. There were a couple strong contenders to be the final model, our analysis to see which one of our models were the best included doing a global Wald test. The global Wald test is conducted with the null hypothesis that at least one of the regression coefficients in the model is significantly different from 0. Only one of our models came back with a significant global Wald test so we will conclude with that as our final model. The last test for this model is to test for multicollinearity again. This final model has higher order and interaction terms so we will have to center each predictor by subtracting their mean and then test; centering the variables is a way to reduce structural multicollinearity while still having the interpretations of coefficients remain the same (Frost, 2017). The multicollinearity check concluded that all predictors had VIF values lower than 5 when centered.

Results:

Final Model: $\text{logit}(Y) = \beta_0 + \beta_1(\text{Age}) + \beta_2(\text{Income Level}) + \beta_3(\text{Avg. Cart Size}) + \beta_4(\text{Website Visit Frequency}) + \beta_5(\text{Website Visit Frequency})^2 + \beta_6(\text{Marketing Emails Opened}) + \beta_7(\text{Age*Website Visit Frequency}) + \beta_8(\text{Avg. Cart Size*Website Visit Frequency}) + \beta_9(\text{Avg. Cart Size*Marketing Emails Opened}) + \beta_{10}(\text{Website Visit Frequency*Marketing Emails Opened})$

Summary Table of Coefficients

Predictor	**Estimate**	**Std. Error**	**z value**	**p-value**
(Intercept)	3.021	2.438	1.239	0.215
Age	-0.063	0.042	-1.477	0.140
Income Level	0.00002	0.000	2.580	0.010*
Avg. Cart Size	-0.002	0.005	-0.474	0.636
Website Visit Frequency	-0.415	0.191	-2.166	0.030*
(Website Visit Frequency)²	0.005	0.003	1.708	0.088
Marketing Emails Opened	0.252	0.113	2.224	0.026*
Age*Website Visit Frequency	0.005	0.003	1.910	0.056
Avg. Cart Size*Website Visit Frequency	0.001	0.0003	2.116	0.034*
Avg. Cart Size*Marketing Emails Opened	-0.001	0.0002	-2.303	0.021*
Website Visit Frequency*Marketing Emails Opened	-0.003	0.004	-0.934	0.350

*Income Level, Website Visit Frequency, Marketing Emails Opened, the interaction between Avg. Cart Size & Website Visit Frequency, the interaction between Avg. Cart Size and Marketing Emails Opened are statistically significant ($p < 0.05$). The quadratic term and interaction between Age & Website Visit Frequency are marginally significant ($p < 0.1$) however, these terms suggest some non-linear effects.

Wald Test Summary

Test	Chi-Square	df	p-value	Conclusion
Wald Test	19.0	10	0.04	Significant ($p < 0.05$)

The Wald test shows whether all of the predictor coefficients are jointly significant and we found that our final model collectively contributes to the models explanatory power.

Variance Inflation Factors of Final Model (Mean-Centered)

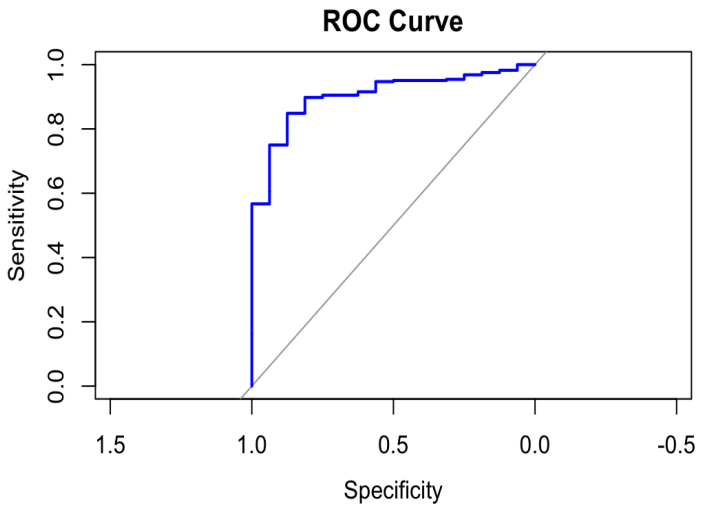
Predictor	**VIF**
Age	1.024697
Income Level	1.044246
Avg. Cart Size	1.044354
Website Visit Frequency	1.027295
(Website Visit Frequency)²	1.030409
Marketing Emails Opened	1.026866
Age*Website Visit Frequency	1.021409
Avg. Cart Size*Website Visit Frequency	1.060709
Avg. Cart Size*Marketing Emails Opened	1.048818
Website Visit Frequency*Marketing Emails Opened	1.019978

These are the VIF values of the predictors of our final model centered by subtracting the mean of each predictor; centering is done due to higher order terms and interactions. All values are around 1 which means that each predictor provides unique information without redundancy and our final model does not suffer from multicollinearity.

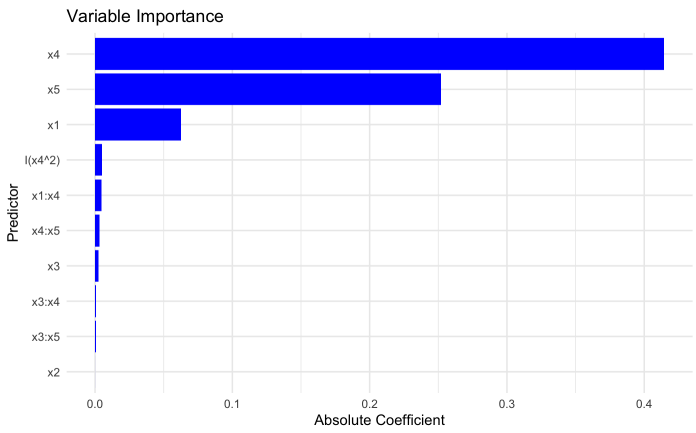
Model Performance Metrics

Confusion Matrix	**Actual: No**	**Actual: Yes**
Predicted: No	1	3
Predicted: Yes	15	281

Metric	**Value**
Accuracy	0.94
Sensitivity	0.989
Specificity	0.062
AUC	0.904
AIC: Main Effects AIC: Final	115.49 109.28
BIC: Main Effects AIC: Final	148.82 150.019



The model correctly predicts about 94% of the cases, while this is a high accuracy there may be a strong variation between TPR and TNR. So, 98.9% of the subscribers are captured by the model which is the true positive rate. Only 6.2% of the non-subscribers are correctly identified which is the true negative rate. This suggests that our model is biased towards predicting subscription as it may overpredict “Subscribed”, leading to a high number of false positives. The ROC (Receiver Operating Characteristic) curve represents the model’s capability to distinguish between Subscribed vs Not Subscribed. Given that the AUC (Area Under Curve) is 0.904, we can see that the ROC curve has a steep increase on the left which means that the model captures most of the subscribers with a few false positives. Also, the curve flattens out near the top illustrating the declining gains as there are more false positives. The AUC is close to 1 so the model is highly effective at classification however, the specificity is too low which suggests that it struggles to correctly identify the non-subscribers. The final model has a lower AIC than the main effects model which suggests that the final model provides a better fit to the data while balancing model complexity. The final model has a slightly better BIC than the main effects model, since the BIC penalizes complexity more heavily. This suggests that the main effects model may be more preferable in terms of simplicity so, there are potentially many unnecessary interactions in the model. Since, the AIC improvement is stronger than the BIC increase, the final model is most likely the best choice overall.



We found the variable importance based on the highest coefficients. The predictors that most heavily influence the model are Website Visit Frequency, Marketing Emails Opened, and Age. The Income Level variable provides very low importance so, it can be considered that we remove it from the model. Overall, the Website Visit Frequency variables seems to have a strong impact on the model as it has several interaction terms and a squared term.

Discussion & Conclusion:

This study aims to understand the key factors that influence a customer's decision to subscribe to a premium e-commerce membership and to develop a predictive model for subscription likelihood. In our analysis, we use a logistic regression model to fit our data, and we get high accuracy (94%) in prediction, particularly in predicting subscribers.

Finally, our findings indicate that the variables of Website Visit Frequency play the most substantial role. The coefficient of "Website Visit Frequency"(-0.415) suggests that for each additional visit, the log-odds of subscription decrease by 0.415, and the odd of subscription decrease by approximately 34% ([e^(-0.415)-1]*100%). It means that the more frequently a customer visits the website, the less likelihood to subscribe than not. Some other factors, including "Marketing Emails Opened" and "Age," also play an important role in predicting subscriptions. The results shows that for each additional marketing emails opened, the odd of subscription increase by 29%([(e^0.252)-1] *100%), which positively impacts subscriptions, and for each increase in age, the odd of subscription decrease by 6.1% ([(e^-0.063)-1]*100%), which negatively impacts subscriptions. Therefore, according to our result, if companies want to increase the number of subscribers, they might consider enhancing website engagement strategies, optimizing email marketing campaigns, and targeting younger audiences.

However, despite the model's strong prediction ability, several limitations shouldn't be omitted. We have found that the true negative rate is only 6.2%, which shows inaccurate predictions for non-subscribers. According to our EDA, the majority of users were subscribers, which might skew the model's ability to predict non-subscribers. Therefore, to balance the dataset, future research should consider oversampling the minority class or undersampling the majority class when collecting data. Last but not least, while key demographics were included, additional

behavioral variables such as “Customer Satisfaction Scores” might enhance predictive power and could be added in the future.

References:

Frost, J. (2017, April 2). *Multicollinearity in regression analysis: Problems, detection, and solutions.* Statistics By Jim. https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

Home. OARC Stats. (n.d.). *FAQ: How do I Interpret Odds Ratios in Logistic Regression?*.UCLA https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/

Ibm. (2024, December 19). *What is logistic regression?*. IBM. https://www.ibm.com/think/topics/logistic-regression

Rheude, J. (2025, January 21). *The history of Ecommerce*. Red Stag Fulfillment. https://redstagfulfillment.com/history-of-ecommerce/