***Data Analysis Final Report***

**Predicting Purchase Behavior for an Online Luxury Shoe Retailer**

**Prepared by: Abigail Sandquist, Connor O’Grady & Michael Daley  
 Course: BAIS-329  
 Instructor: Dr. Lauren Laker  
 Institution: Xavier University  
 Date: May 2nd, 2025**

***Executive Summary***

We were tasked with building a model that would help predict whether customers would make a purchase for an e-commerce sneaker company. Our goal was to create an accurate model with minimum financial risk and maximum efficiency in targeting customers.

***Introduction***

**The Problem**

A luxury sneaker resale company wants to understand and predict consumer purchasing behavior based on data from 15 recent targeted online ad campaigns. Platforms like Facebook, Instagram, and podcasts were all used in the campaign for collecting data, while other data points included gender, location (American or not), and the number of transactions within the last year.

**Why it’s important**

Accurately predicting whether a customer will make a purchase allows the company to optimize its marketing strategy, allocate ad spend more effectively, and increase conversion rates to boost revenue. For a mid-size company, there is little room for error. Money can not be wasted targeting the wrong markets or using the wrong platforms.

**The role of Analytics**

Using predictive modeling, analytics can identify which customer characteristics and campaign elements are most (and least) strongly associated with purchasing behavior. This will enable data-driven decisions for future marketing efforts. Analytics equips the company with the tools to move from intuition-based marketing to an evidence-based strategy. Evidence-based strategies should lead to an effective solution to the problem and increase the company's return on investment.

***Methods***

**High-Level Summary of the Dataset**

The dataset consisted of 1600 observations, representing individual customer interactions with an online luxury shoe retailer. Key variables used in the analysis included:

* Time Spent on Website: Measuring customer engagement, this variable served as an indicator of interest in making a purchase.
* Gender: Capturing demographic differences in purchasing behavior.
* Advertisement Type: Identifying the impact of various marketing strategies in directing customers to the retailer's website.
* Geographic Data: Differentiating between customers located within the United States versus international visitors.

The dataset allowed us to examine patterns and correlations between these factors and purchasing outcomes, providing actionable insights for the retailer's marketing strategy.

**Summary of the Predictive Model**

The objective was to develop a model that could predict the likelihood of a customer making a purchase based on their interaction with the retailer's website. To align with the constraints and goals of a relatively small retailer, the team chose a model that emphasized efficiency and cost-effectiveness, focusing on eliminating spending on ineffective marketing tactics.

**In non-technical terms:**

* The model highlighted waste in marketing expenditure by identifying customer segments or advertising types that did not contribute to sales conversions.
* For example, if certain advertisements consistently directed customers who spent little time on the website and did not make purchases, those ads were flagged as low-performing. Conversely, highly effective ads that drove engaged customers were prioritized for future campaigns.
* By analyzing these variables, the model provided actionable recommendations to refine marketing strategies, ensuring that limited resources were allocated to tactics with the highest return on investment.

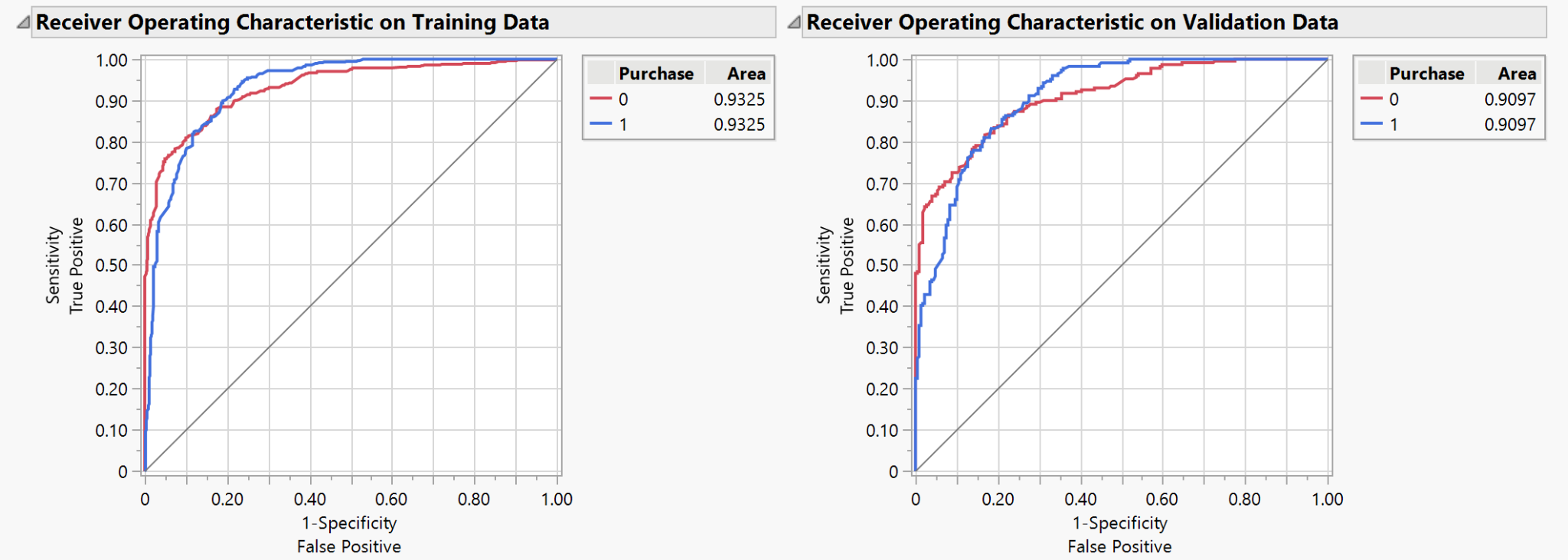
**Business Implications**

The approach acknowledges the competitive landscape of the footwear industry, particularly in e-commerce, where large brands like Nike and Adidas dominate the market. For this smaller retailer, the model's focus on cutting unnecessary spending empowers them to optimize their marketing strategies without the substantial budgets that larger brands possess.

By adopting data-driven decision-making, the retailer can channel efforts into high-performing advertisements, enhancing conversions, and driving sustainable growth.

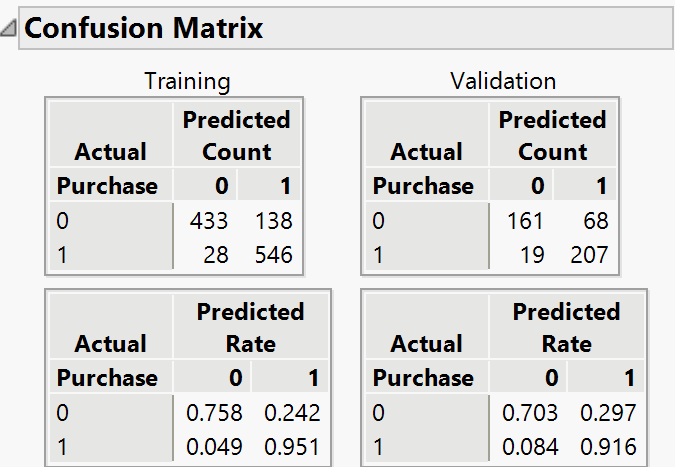
***Results***

For our final model, we chose to use a Bootstrap forest to predict customer purchasing behavior. This model was a good fit for the problem because it can handle lots of different variables at once and does great at balancing accuracy and generalizability. On the training data, it had an accuracy of 81.4% and a misclassification rate of 14.5%. On the validation set, the model had an accuracy of 75.38% and a misclassification rate of 19.12%. These values indicated that the model performs consistently well on unseen data, maintaining an acceptable balance between correctly identifying buyers and non-buyers.

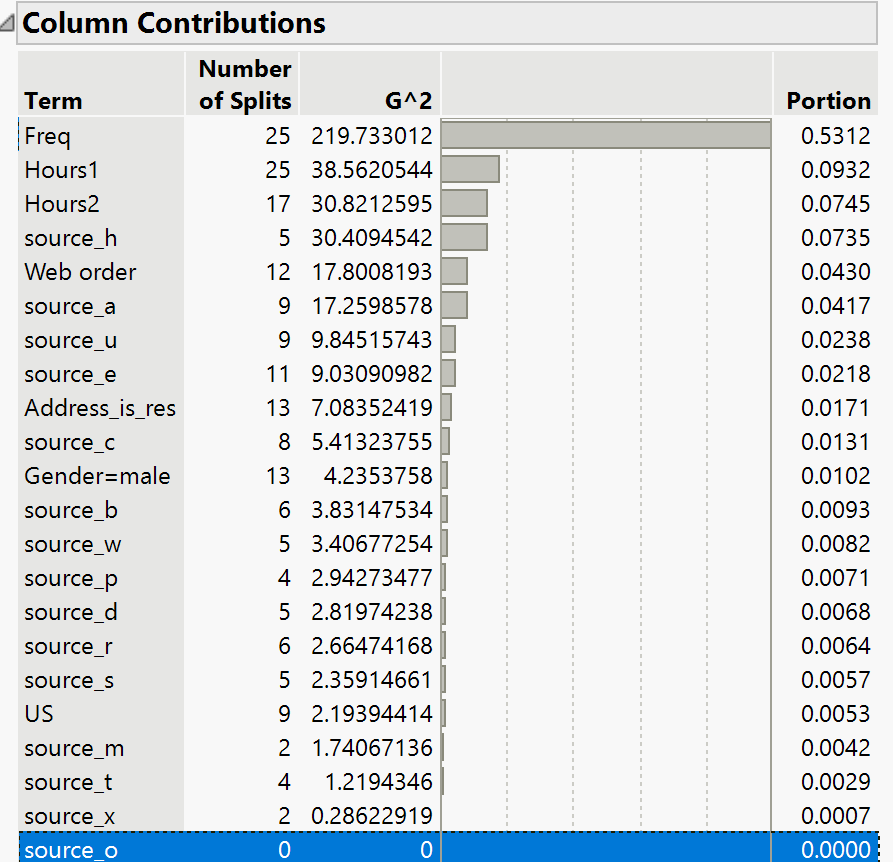


The Area Under the Curve, or AUC (seen above) , was 0.9325 on the training set and 0.9097 on the validation set, demonstrating that the model is really good at ranking customers based on how likely they are to buy, even if it doesn’t always predict perfectly.

One thing to note is the model’s high specificity of 0.9282 on the training data and 0.9258 on the validation set. Specificity measures the model’s ability to correctly identify customers who are not likely to buy, and high values here are especially important in the context of this business. Their marketing strategy involves spending real time and money on customer outreach, especially for high-end footwear. The decision threshold for this model is 0.6626. This was chosen to prioritize a high specificity while maintaining a good accuracy of the model. By prioritizing specificity, the model minimizes false positives, or cases where resources might be wasted targeting non-buyers, reducing financial risk and improving marketing efficiency. At the same time, the model still maintains reasonable accuracy, ensuring likely buyers are still captured and providing a solid balance between caution and opportunity. We can see this in the confusion matrix below:



In analyzing the column contributions from the final Bootstrap Forest model, it is clear that the variable Freq played the most significant role in predicting purchasing behavior. Freq was used in 25 splits across the ensemble and accounted for approximately 53.12% of the model’s total explanatory power, as measured by the G² statistic (219.73). This suggests that how frequently a customer interacts with the platform, whether through visits, views, or engagement, is the strongest indicator of their likelihood to purchase. Following Freq, Hours 1 and Hours 2 were the next most influential variables, contributing 9.32% and 7.45% to the model’s decisions, respectively. Despite being split multiple times, these variables had much smaller G² values (38.56 and 30.82) compared to Freq, indicating that while they play a role, their predictive power is more limited.



These insights suggest that future marketing efforts should strongly consider customer engagement frequency when targeting high-value leads.

That said, there are still a few things to be aware of before putting this model into production. First, even though the model is strong overall, its accuracy and AUC do drop between the training and validation sets, which could be a sign of slight overfitting. There's also the risk of missing some potential buyers because the model leans heavily on avoiding false positives. While that’s useful for saving money, it could mean missing out on certain customers who would’ve purchased with a bit more marketing attention. Finally, the model is only as good as the data it's trained on; if the customer behavior changes over time or if there are issues in how the data was collected, the model’s performance could suffer. Keeping the model updated and monitoring it over time will be important if the company wants to use it long-term.

***Conclusion***

**Summary:**

Our model is a conservative and accurate model to help maximize the limited resources of this e-commerce company. It is designed to define a highly specific target market, which will limit wasted marketing dollars. The model highlights purchase frequency, Facebook ads, blog posts, and podcasts as positive indicators towards a purchase. While Bing Ads and Reddit Ads are negative indicators towards a purchase. There is an adaptable threshold that would allow the company to adjust the model if they would like to be more aggressive.

This model meets our goal of making an accurate model with minimum financial risk for the company. There are clearly defined markets that the company should continue to target and strong vehicles to reach them.

**Final Results:**

With an AUC of 0.9097 and a specificity of 92.58%, the final results of our model predicted that there would be 139 purchases out of the 400 customer dataset.

**Long Term Concerns:**

There are a couple of long-term concerns. The first is that this model is only a short-term problem solver. If this model is successful early on, the company may want to take on a more aggressive approach going forward. There may be more money to spend, and the goal may shift to total revenue rather than the efficiency of a marketing campaign. Since it is a conservative model, this model would not work in a higher-risk situation.

The second is the potential of overfitting. Maybe our model was heavily skewed to the data set provided and is not an accurate predictor of future data. When tens of thousands (or more) of customers are included, how will it perform? The larger the company gets, the more customers there are, which means it’s more likely for our model to be overfit.

**Future Steps**

Future steps could include adjusting the threshold to a lower number if the company wanted to be more aggressive with its approach. I don’t think it’s necessary to build a new model because there is data-driven evidence to back up this model. However, with a high threshold, the model has a lower sensitivity, and that could be corrected by switching to a low threshold. Another future step would be finding new platforms to place advertisements and running different types of ad campaigns on the sources that were ineffective in the previous campaign.

***Appendix***

**Non-selected models**

**Logistic Regression:** While logistic regression offered a reasonably effective predictive framework, the model faced challenges in achieving an AUC of > 0.83 without introducing overfitting. This issue arose due to the complexity of the dataset and the limitations of logistic regression in capturing certain non-linear relationships. As a result, while the model provided valuable insights, it ultimately did not meet the criteria necessary for our objective, leading us to pursue a more suitable alternative.

**Decision Tree:**  
 The decision tree model encountered difficulties in effectively incorporating certain key variables, particularly frequency-related data. This limitation led to inconsistencies in predictive accuracy, preventing the model from reliably identifying patterns associated with sales conversions. Due to these constraints, the decision tree failed to serve as a viable solution for achieving our primary goal of sales prediction.

**Boosted Tree:**

The Boosted Tree model showed promise in identifying subtle patterns within customer interactions, but several challenges arose during implementation. One major issue was its sensitivity to noise, which made the model overly reactive to minor variations in the dataset. This resulted in the amplification of insignificant patterns rather than the most relevant ones, raising concerns about overfitting to specific customer behaviors instead of generalizing effectively across the dataset. The computational demands of boosting required significant tuning to balance predictive performance with efficiency. Given the retailer's resource constraints, fully optimizing hyperparameters became challenging without dedicating additional processing time. While the Boosted Tree model demonstrated strong predictive capability, these drawbacks ultimately made it less suitable for the retailer’s goal of streamlining marketing efforts and reducing wasted ad spending. Given the need to balance accuracy, interpretability, and efficiency, the boosted tree did not meet the business objectives.