Josef Triman, Evan Hollier, & Sebastian Lemm

Regression Tree Analysis of Grad Application Click Rate

Research Question

The goal with this project was to explore the relationship between certain independent variables from the Google Analytics for the Daniels College of Business and how they affect people applying to its graduate programs. The variables we chose were "Device Type", "Default Channel Grouping" (traffic source), and "Avg. Session Duration" (average time on site). This led us to create the research question, "How do traffic sources, device types, and average time on site influence the likelihood of users clicking the "Apply" button on DU's grad websites?"

Data Source and Definitions/Variables

The data for our project was obtained from Google Analytics results for the Daniels College of Business. Josef works there as a student website manager/data analyst. One of his responsibilities is analyzing and reporting on data obtained from Google Analytics of the Daniels College of Business websites. Google Analytics is a platform that collects data from websites or apps to create reports that provide insights for the businesses running it.

The variables that we were interested in conducting our model on came from two data tables within Google Analytics. The first data table is centered around the variable "Default Channel Grouping" (see Figure 1). This term refers to a categorization system that consolidates incoming site traffic from a variety of sources into distinct, commonly-defined channels. This just means that "Default Channel Grouping" provides us with a comprehensive overview of where the website traffic is originating. The "Default Channel Grouping" variable encompasses eight distinct subcategories, each associated with different ways visitors access the website:

1. Organic Search: When someone uses a search engine to get to the website.

2. Direct: Entering the URL directly into the browser.
3. Referral: Traffic that comes from other websites through referral links.
4. Social: Traffic generated from social media platforms.
5. Paid Search: Traffic resulting from paid advertisements in search engines.
6. Display: Traffic from display/banner ads.
7. Email: Traffic from email marketing campaigns.
8. Other: Any other sources not falling into the above categories. This data was combined with Direct traffic as per the guidance of Josef's boss.

The second data table is where we get the variable "Device Type" (see Figure 2). "Device Type" is how Google Analytics tracks what device people are using to access websites. "Device Type" contains three subcategories for what device people are using to access the Daniels websites (Desktop, Mobile, Tablet).

Both data tables contain the variable "Avg. Session Duration" which represents the average time on site based on traffic source or device type. They additionally contain the variable "Link Click – Apply" which is the number of people clicking "Apply" based on either traffic source or device type. The data we collected from these tables ranges from October 2021 to May 2023 and is broken up semi-monthly, resulting in 38 periods of data.

Regression Tree introduced

A regression tree is a type of decision tree used for predicting continuous numerical values. Each node in the tree represents a proportion of the data confined above or below certain thresholds of a variable. A node contains the predicted value of the outcome variable in addition to the percentage of data represented. Each node is then split into two sub-nodes based on a particular true/false question about the data. This results in a tree structure that functions like a flowchart, lending regression trees to be very visual.

There are several reasons to choose a regression tree for predictive modeling. Firstly, regression trees are highly interpretable, making them easy to understand. They can handle both numerical and categorical data, making them versatile. Furthermore, regression trees naturally account for non-linear relationships and interactions between variables, which is beneficial when dealing with complex data patterns. Another advantage is that regression trees have very lax requirements of the data. Finally, by pruning the tree, removing less important branches, the model's complexity can be controlled, striking a balance between accuracy and simplicity as well as avoiding overfitting and underfitting.

Regression Tree explained in principle

Regression trees operate by evaluating each variable at every possible threshold to determine the splits that minimize the sum of squared residuals (SSR). Excessive splits may indicate overfitting, as well as making the model more challenging to interpret. This is where pruning comes in. Since an unpruned regression tree already has the lowest possible SSR, modifying it will increase the SSR. Therefore, SSR alone cannot be relied upon for tree pruning. To address this, a penalty is applied for every terminal node or "leaf." Tweaking the intensity of this penalty results in bigger or smaller trees.

Regression Tree Applied and Interpreted

After cleaning and merging the data in both Python and R, we ran three regression trees and pruned one of them to analyze how the variables predicted the number of people clicking "Apply" to Daniels graduate programs.

The first Regression Tree model was constructed using two independent variables, Device ("Device Type") and Avg_Duration ("Avg. Session Duration" based on "Device Type"), to predict the dependent variable, the number of people clicking "Apply" on Daniels graduate program websites (see Figure 5). The tree starts by splitting the data into two groups: "mobile" and "tablet" versus "desktop". This suggests that "Device" type is a crucial determinant of the likelihood of clicking "Apply". In particular, users on "mobile" and "tablet" devices are generally associated with fewer "Apply" clicks compared to "desktop" users. The importance of the device type is further reinforced by its variable importance score (847,210.9), which is higher than that of Avg_Duration (760,687.4). For the mobile and tablet group, the model further differentiates between "mobile" and "tablet" users, with "tablet" users showing significantly fewer Apply clicks on average. For the desktop group, the "Avg_Duration" variable is used for the split. Specifically, "desktop" users who spend less than 167 units of time are associated with fewer "Apply" clicks than those spending 167 units of time or more. This suggests that "Avg_Duration" is also an important factor in predicting the likelihood of people applying. In summary, both "Device Type" and Avg_Duration play significant roles in predicting the number of Apply clicks, with "Device Type" having a slightly higher impact according to this model.

The second regression tree illustrates "Grad_Apply" as a function of "Source" and "Avg_Duration," segregating the data according to the traffic source (see Figure 6). For traffic originating from Display, Email, Paid Search, Referral, or Social sources, the model bifurcates further. One node represents Display, Email, Referral, and Social, and the other represents Paid Search. In the instance of Display, Email, Referral, or Social sources, the average "Grad_Apply" stands around 2.35. The model progressively splits based on "Source" and "Avg_Duration" until no significant divisions can be made. Notably, Organic Search sources contribute to the highest "Grad_Apply" count, totaling 146, while short length Paid Searches also contribute significantly, with a count of 120.

The third Regression Tree used a single variable, Avg_Duration, to predict Grad_Apply. We pruned this regression tree, resulting in one fewer split (see Figure 4). The overall expected value for Grad_Apply over a two-week period is estimated to be 61 conversions. The population is split evenly below and above 2-minute duration. The conversions for durations below 2 minutes are notably lower, at 28. The sweet spot lies between 2 and 3 minutes, where a significant increase to 115 is observed. An even sweeter spot lies between 2.4-3 minutes, representing 25% of the data and 134 conversions predicted.

We utilized the R programming libraries "rpart" and "rpart.plot" for constructing and visualizing regression trees. The regression model for our merged dataframe was built using "Grad_Apply" as the response variable and "Avg_Duration" as the predictor. An important step in our analysis was pruning the tree with a complexity parameter that minimized the cross-validation error, a process facilitated by the "rpart" package.

The complexity parameter is a value used in the construction of a decision tree or regression tree that influences the size and fit of the tree. It is utilized as a mechanism to control overfitting, which occurs when the model becomes too complex and starts to fit not only the underlying pattern in the data, but also the random noise. In the context of "rpart", the complexity parameter is a non-negative numeric scalar that dictates the minimum amount of improvement in the model needed for a split to happen. If the improvement, measured by the reduction in overall error, is less than the complexity parameter, the split is not made, and the node is considered a terminal node.

This pruning process essentially trims the tree down to a simpler version, reducing the risk of overfitting by removing branches that provide little predictive power and are likely capturing noise. This also makes the final model easier to interpret and more generalizable to new data.

For the "device_df" and "source_df" dataframes, "Grad_Apply" was used as the response variable while "Avg_Duration" along with "Device" and "Source," respectively, were employed as predictors.

Data Requirements

To ensure the robustness of our regression tree model, we adhered to several key principles. The quality of data is crucial, as regression trees - like other decision tree models - are essentially nested if-else conditions. The accuracy of these conditions, and thus the overall model, is directly impacted by the quality of the data. We ensured clean and error-free data by performing data cleaning and preprocessing steps, such as handling missing values, removing duplicates, and dealing with outliers as necessary.

The relevance of features is another key consideration. The predictors we used, "Avg_Duration", "Device", and "Source," were all closely tied to the outcome we were predicting, "Grad_Apply". The regression tree works by dividing the predictor space into distinct and non-overlapping regions, making the relevance of features critical to forming these regions and predicting the mean response for every observation that falls into a region.

Additionally, having sufficient data is essential for effective learning by regression trees. Despite having 410 observations, we didn't consider it sufficient to split into training and testing sets. The quantity of data affects the model's ability to minimize the residual sum of squares (RSS), which is the objective of forming the distinct regions in the predictor space. However, there is a need to be cautious about overfitting, which can be mitigated by adjusting the tree size, a tuning parameter governing the model's complexity, and applying the cost complexity pruning algorithm.

Lastly, the assumption of independence is important in regression tree modeling. We operated under the assumption that each user's decision to apply is independent of others. Violation of this assumption can come from observations being close together in time, space, or appearing multiple times in the same data set. To ensure independence, a simple random sampling method should ideally be used when obtaining the sample from the population. This gives each individual an equal chance of being included in the sample, thereby minimizing the chances of selecting two individuals who may be closely related or in close proximity. These principles underscore the flexibility of regression trees, enhancing their effectiveness when adhered to.

Conclusion

In conclusion, we sought to investigate the independent variables device type, traffic source, and average time on site and how they affect the dependent variable of people applying to

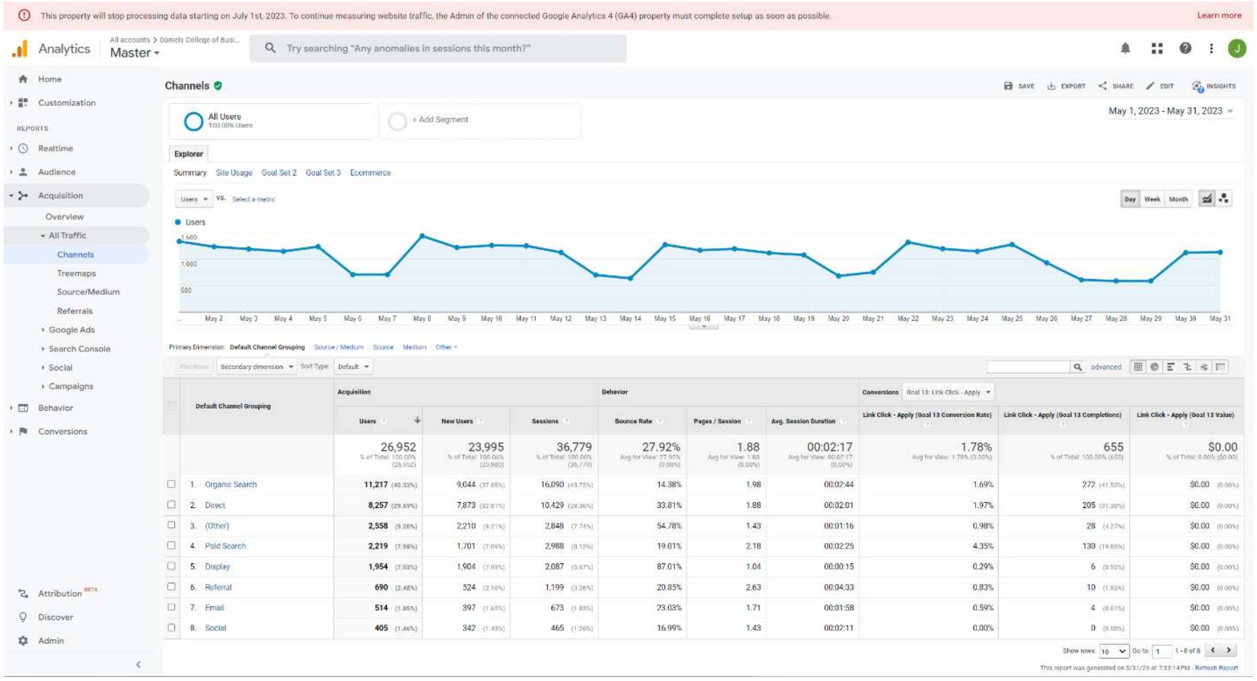
Daniels graduate programs. This prompted us to come up with the research question, "How do traffic sources, device types, and average time on site influence the likelihood of users clicking the "Apply" button on DU's Grad Websites?". We decided to answer these questions using three regression trees, with one of them being pruned further.

The first tree, which used the variables "Device" and "Avg_Duration" (average time on site based on device type), indicated that device type significantly affects the predicted graduate applications, with desktop users yielding the highest count greater than 167 seconds of "Avg_Duration" (262), followed by mobile users (90), and the lowest being tablet users (3.3) (see Figure 5).

The second tree, which used the variables "Source" (traffic source) and "Avg_Duration" (average time on site based on traffic source) to predict the number of users clicking apply, suggests that the source of traffic greatly influences the predicted graduate applications. Users from Organic Search and Direct sources show a significantly higher predicted application count (146 and 95 respectively), while users from Paid Search display considerable variability based on Avg_Duration, and users from Display, Email, Referral, and Social predict the lowest number of applications (2.4) (see Figure 6).

Finally, the last tree, which used only the variable "Average_Duration" of both device type and traffic source to predict the number of people clicking apply to Daniels graduate programs, suggests that there is a peak between 2 to 3 minutes at 115, which constitutes 37% of the population. Interestingly, this sweet spot becomes even more pronounced between 2.4 to 3 minutes, yielding 134 applications from 25% of the users. There is a noted drop off in applications (42) for the 12% of users spending more than 3 minutes on the site.

These were our main findings for how the variables Traffic Source, "Device Type", and Average Time on Site affect the number of users applying to the Daniels College of Business graduate programs.



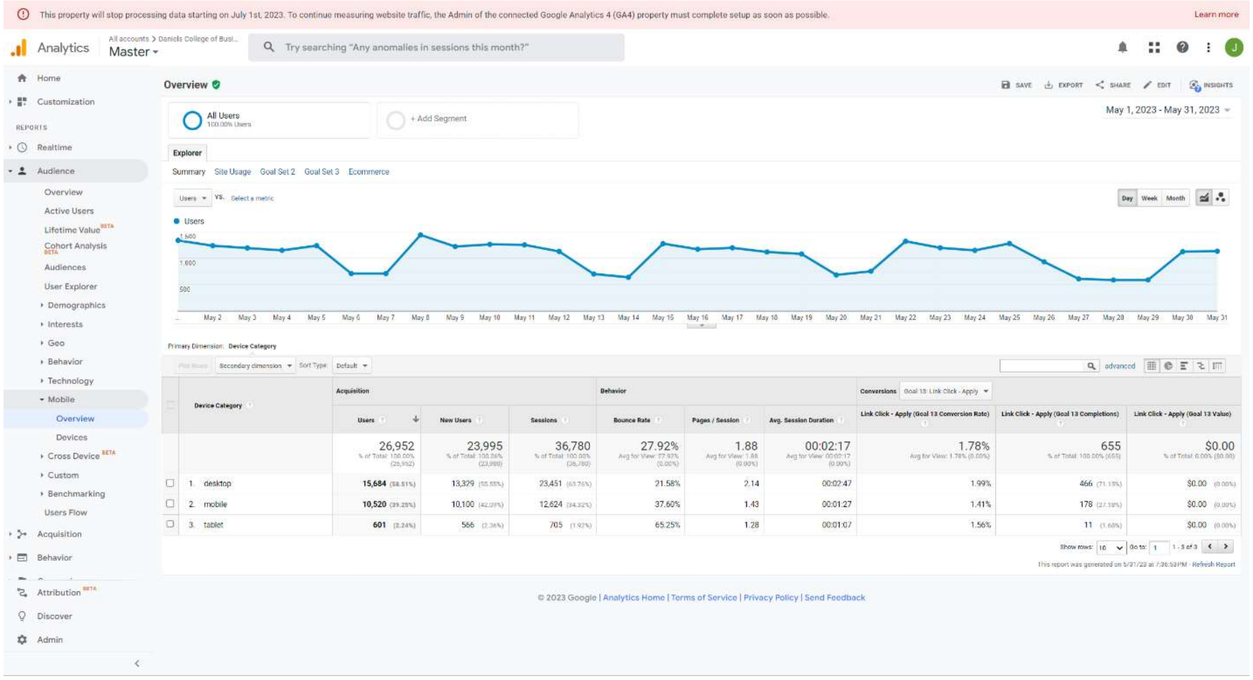
This property will stop processing data starting on July 1st, 2023. To continue measuring website traffic, the Admin of the connected Google Analytics 4 (GA4) property must complete setup as soon as possible.

Conversions: Goal 13: Link Click - Apply

	Default Channel Grouping	Acquisition			Behavior			Conversions		
		Users	New Users	Sessions	Bounce Rate	Pages / Session	Avg. Session Duration	Link Click - Apply (Goal 13 Conversion Rate)	Link Click - Apply (Goal 13 Completions)	Link Click - Apply (Goal 13 Value)
		26,952 % of Total: 100.00% (26,952)	23,995 % of Total: 100.06% (23,980)	36,779 % of Total: 100.00% (36,779)	27.92% Avg for View: 27.92% (0.00%)	1.88 Avg for View: 1.88 (0.00%)	00:02:17 Avg for View: 00:02:17 (0.00%)	1.78% Avg for View: 1.78% (0.00%)	655 % of Total: 100.00% (655)	$0.00 % of Total: 0.00% ($0.00)
1.	Organic Search	11,217 (40.33%)	9,044 (37.69%)	16,090 (43.75%)	14.38%	1.98	00:02:44	1.69%	272 (41.53%)	$0.00 (0.00%)
2.	Direct	8,257 (29.69%)	7,873 (32.81%)	10,429 (28.36%)	33.81%	1.88	00:02:01	1.97%	205 (31.30%)	$0.00 (0.00%)
3.	(Other)	2,558 (9.20%)	2,210 (9.21%)	2,848 (7.74%)	54.78%	1.43	00:01:16	0.98%	28 (4.27%)	$0.00 (0.00%)
4.	Paid Search	2,219 (7.98%)	1,701 (7.09%)	2,988 (8.12%)	19.01%	2.18	00:02:25	4.35%	130 (19.85%)	$0.00 (0.00%)
5.	Display	1,954 (7.03%)	1,904 (7.93%)	2,087 (5.67%)	87.01%	1.04	00:00:15	0.29%	6 (0.92%)	$0.00 (0.00%)
6.	Referral	690 (2.48%)	524 (2.18%)	1,199 (3.26%)	20.85%	2.63	00:04:33	0.83%	10 (1.53%)	$0.00 (0.00%)
7.	Email	514 (1.85%)	397 (1.65%)	673 (1.83%)	23.03%	1.71	00:01:58	0.59%	4 (0.61%)	$0.00 (0.00%)
8.	Social	405 (1.46%)	342 (1.43%)	465 (1.26%)	16.99%	1.43	00:02:11	0.00%	0 (0.00%)	$0.00 (0.00%)

This report was generated on 5/31/23 at 7:33:14 PM

*Figure 1***.** Data table within Google Analytics showcasing the traffic source categories (Default Channel Grouping) along with data for each one. "Avg. Session Duration" (time spent on site) and "Conversion Rate (Link Click – Apply)" (clicks on "Apply") are shown as well for each traffic source.



This property will stop processing data starting on July 1st, 2023. To continue measuring website traffic, the Admin of the connected Google Analytics 4 (GA4) property must complete setup as soon as possible. Learn more

Primary Dimension: Device Category

	Device Category	Acquisition			Behavior			Conversions Goal 13: Link Click - Apply		
		Users	New Users	Sessions	Bounce Rate	Pages / Session	Avg. Session Duration	Link Click - Apply (Goal 13 Conversion Rate)	Link Click - Apply (Goal 13 Completions)	Link Click - Apply (Goal 13 Value)
		26,952 % of Total: 100.00% (26,952)	23,995 % of Total: 100.06% (23,980)	36,780 % of Total: 100.00% (36,780)	27.92% Avg for View: 27.92% (0.00%)	1.88 Avg for View: 1.88 (0.00%)	00:02:17 Avg for View: 00:02:17 (0.00%)	1.78% Avg for View: 1.78% (0.00%)	655 % of Total: 100.00% (655)	$0.00 % of Total: 0.00% ($0.00)
□	1. desktop	15,684 (58.51%)	13,329 (55.55%)	23,451 (63.76%)	21.58%	2.14	00:02:47	1.99%	466 (71.15%)	$0.00 (0.00%)
□	2. mobile	10,520 (39.25%)	10,100 (42.09%)	12,624 (34.32%)	37.60%	1.43	00:01:27	1.41%	178 (27.18%)	$0.00 (0.00%)
□	3. tablet	601 (2.24%)	566 (2.36%)	705 (1.92%)	65.25%	1.28	00:01:07	1.56%	11 (1.68%)	$0.00 (0.00%)

This report was generated on 5/31/23 at 7:36:53 PM · Refresh Report

Figure 2. Data table within Google Analytics showcasing the device categories ("Device Type") along with data for each one. "Avg. Session Duration" (time spent on site) and Conversion Rate (Link Click – Apply) (clicks on "Apply") are shown as well for each device type.

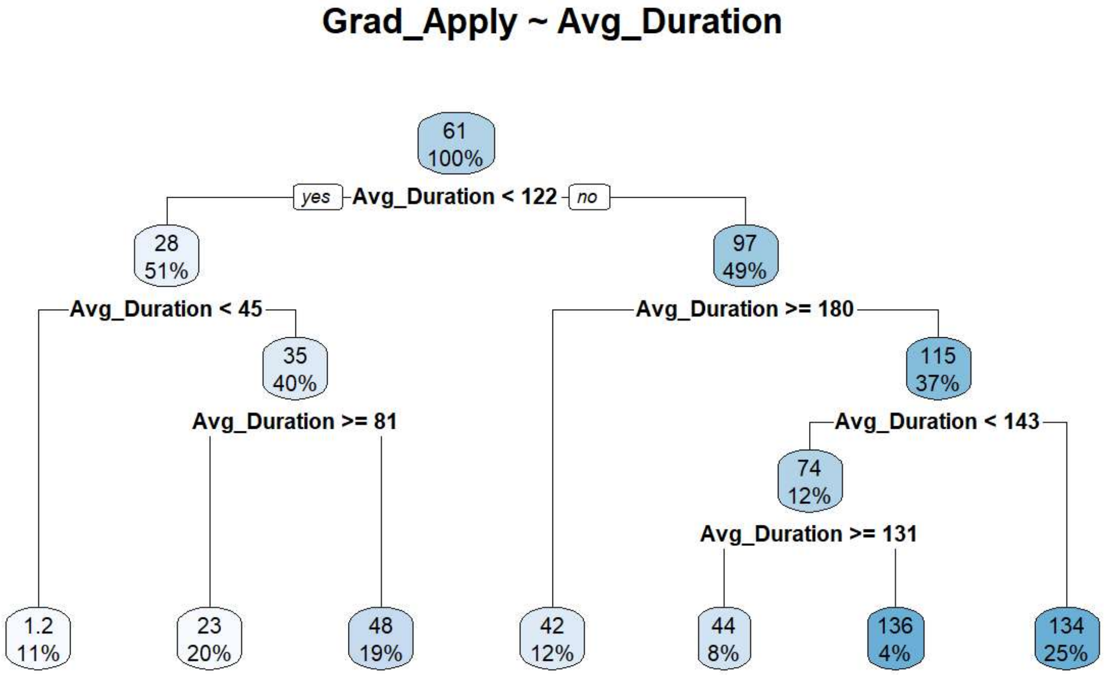


Figure 3. Regression Tree showcasing how the variable "Avg_Duration" ("Avg. Session Duration," representing time spent on site) from both the traffic source ("source_df") and device type ("device_df") tables predict the number of people clicking "Apply".

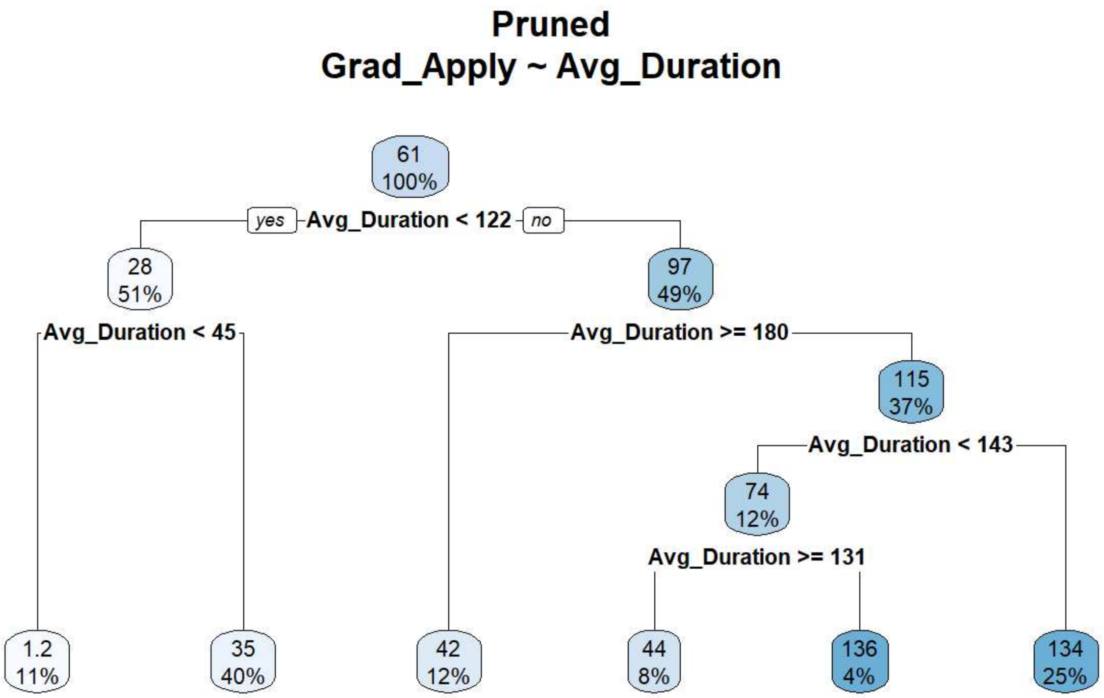


Figure 4. Pruned Regression Tree with the same formula as Figure 3.

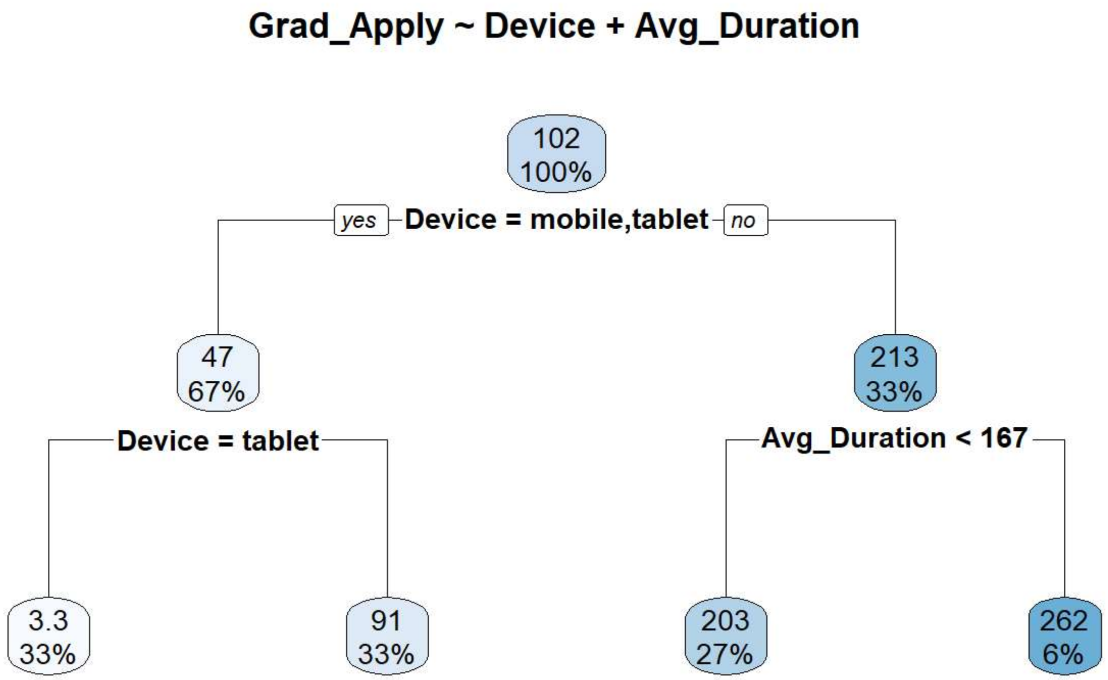


Figure 5. Regression Tree showcasing how "Device" ("Device Type") and "Avg_Duration" (time spent on site based on "Device Type") predict the number of people clicking "Apply".

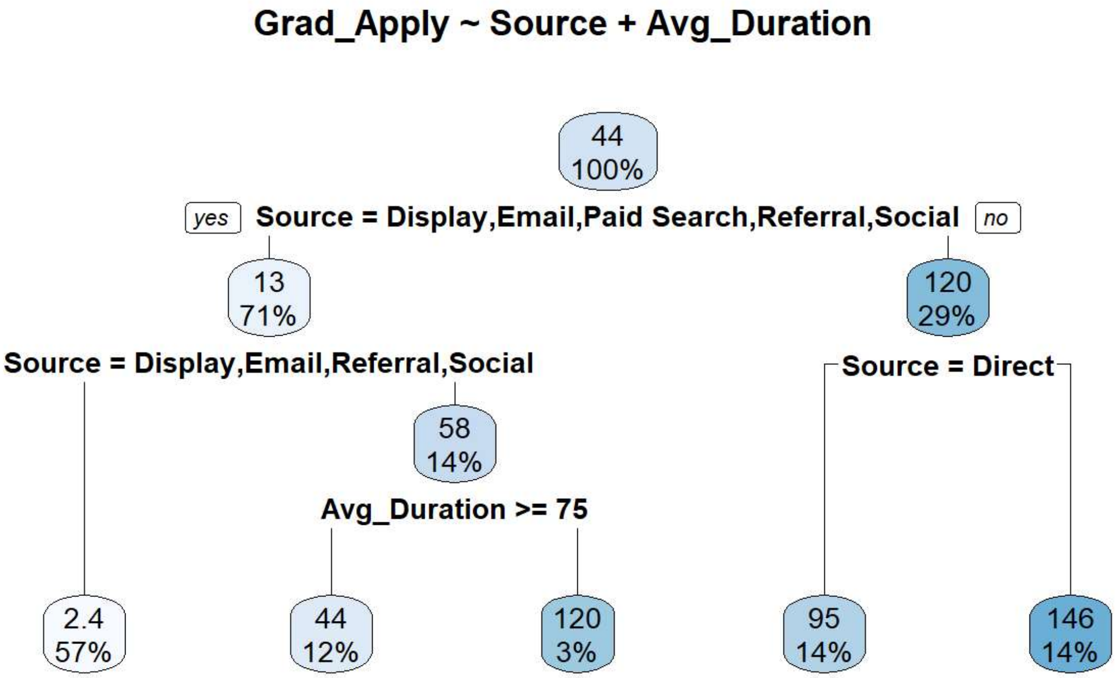

Grad_Apply ~ Source + Avg_Duration
44
100%
yes
Source = Display,Email,Paid Search,Referral,Social
no
13
71%
120
29%
Source = Display,Email,Referral,Social
58
14%
Avg_Duration >= 75
Source = Direct
2.4
57%
44
12%
120
3%
95
14%
146
14%

Figure 6. Regression Tree showcasing how "Source" ("Default Channel Grouping," representing the traffic sources) and "Avg_Duration" (time spent on site based on "Default Channel Grouping") predict the number of people clicking "Apply".