Regression Tree Analysis of Grad Application Click Rate

Josef Triman, Evan Hollier, Sebastian Lemm

Research Question

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

- Google Analytics: Platform that collects data from your websites and apps to create reports that provide insights into your business.
    - Data: Daniel’s College of Business Websites

Google Analytics

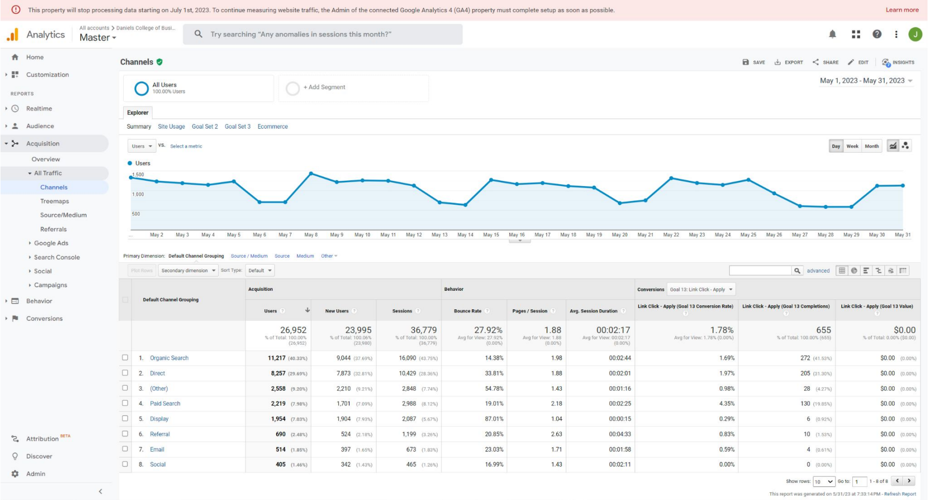
Definitions within data source: Report 1

Variable - Default Channel Grouping: aggregates traffic that comes to your site from various sources into known channels, based on common definitions.

1. Organic Search
2. Direct / Other
3. Social
4. Email
5. Referral
6. Paid Search
7. Display

Values:

1. Average Session Duration
2. Conversion Rate



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.

May 1, 2023 - May 31, 2023

Primary Dimension: Default Channel Grouping

Sort Type: Default

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%)

Show rows: 10 Go to: 1 1 - 8 of 8

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

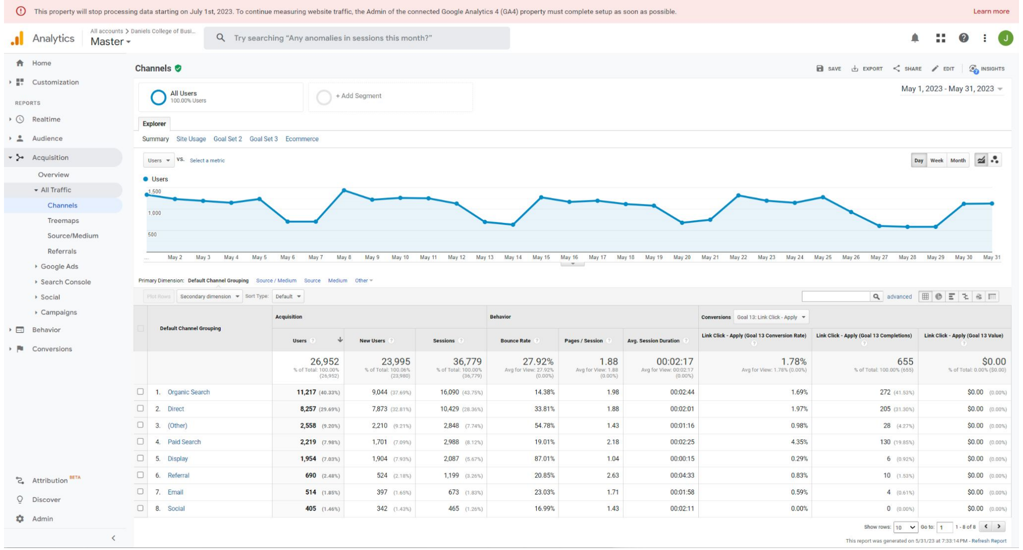
Definitions within data source: Report 2

Variable 2: Device Category: What device people are using to access the websites

1. Desktop
2. Mobile
3. Tablet

Values:

1. Average Session Duration
2. Conversion Rate



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.

	Acquisition			Behavior			Conversions Goal 13: Link Click - Apply		
Default Channel Grouping	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 - Refresh Report

Collection of Data

Export Google Analytics data as a CSV

Split Semi-monthly for 19 months (38 periods)

Done for both Source data and Device data (76 total CSVs)

Use Jupyter Notebook to clean CSVs before exporting to R

3 Devices x 38 periods = 114 rows in Device dataframe

7 Sources x 38 periods= 266 rows in Source dataframe

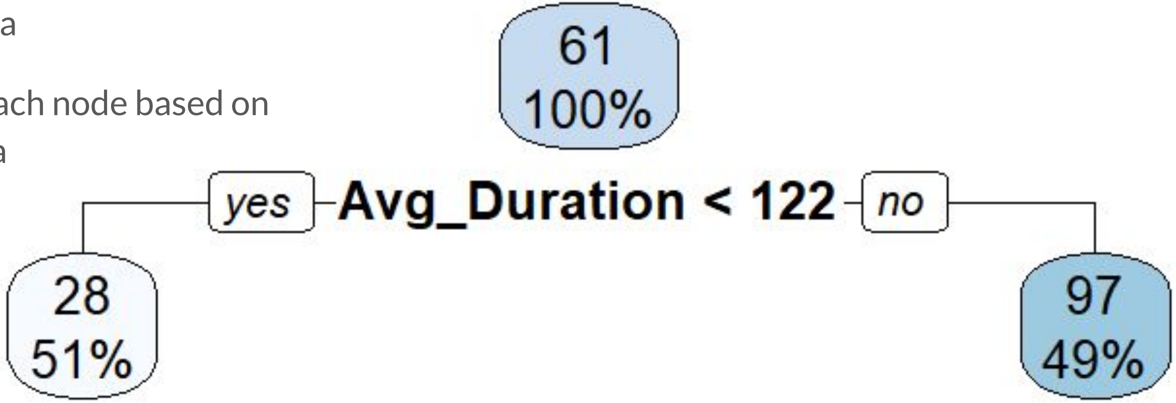
No missing values

What is a Regression Tree?

A type of Decision Tree used to predict a continuous numerical value

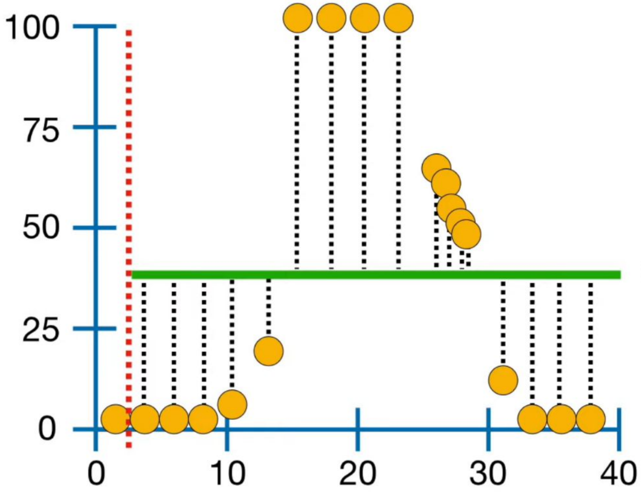
Each node contains the predicted value of outcome variable and the proportion of data

Works like a flowchart, splitting each node based on yes or no questions about the data

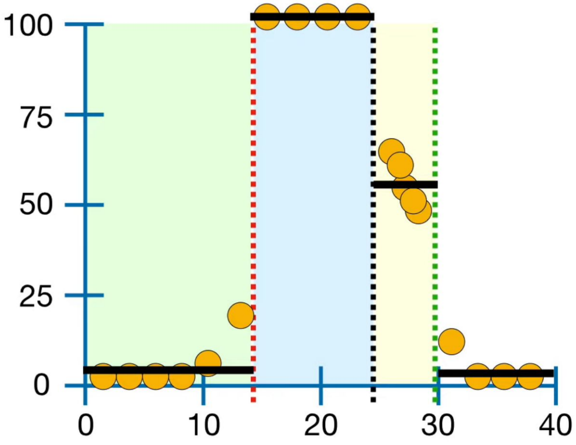


How does a Regression Tree determine splits?

For every variable and threshold, calculate sum of squared residuals and pick the lowest.



$$(0 - 0)^2 + (0 - 38.8)^2 + (0 - 38.8)^2 + (0 - 38.8)^2$$
$$+ (5 - 38.8)^2 + (20 - 38.8)^2 + (100 - 38.8)^2$$
$$+ (100 - 38.8)^2 + \ldots + (0 - 38.8)^2$$

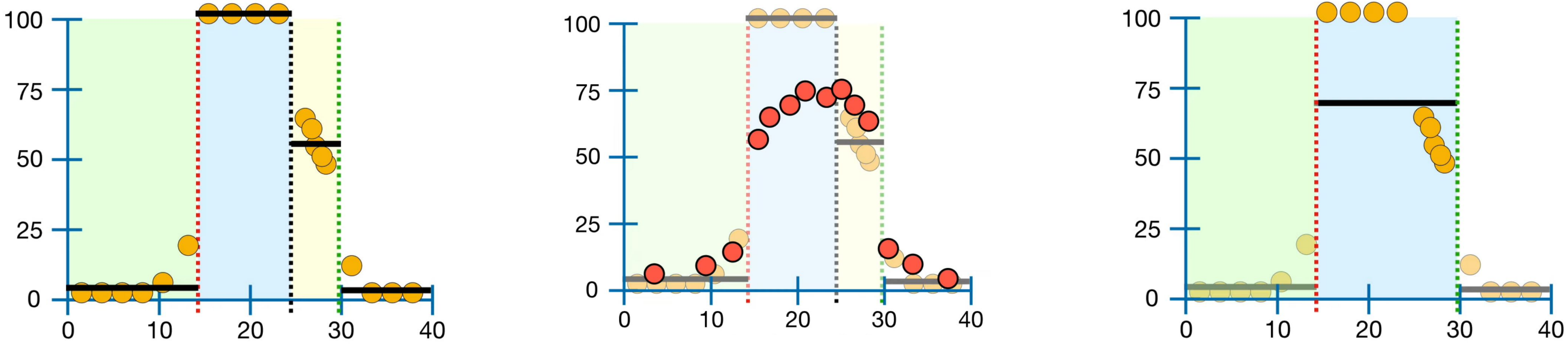


Pruning a Regression Tree

When there are many splits, it might be a sign of **overfitting**. The model is also more difficult to interpret.

Removing a leaf will increase SSR, so it alone cannot be used to prune a tree.

Thus, a penalty is applied for every terminal node ("leaf").



Why a regression tree?

1. **Interpretability:** Regression trees are simple to understand and interpret.
2. **Handling Different Types of Data:** Regression trees can handle both numerical and categorical data
3. **Non-linearity and Interactions:** Regression trees naturally account for non-linear relationships and interactions between variables.
4. **Lack of assumptions:** Don't need to assume that the relationship between the independent and dependent variables is linear or that the errors are normally distributed.
5. **Model Complexity Control:** pruning the tree (i.e., removing less important branches), we can control the complexity of the model, balancing accuracy with simplicity to avoid overfitting and underfitting.

Data Requirements

1. Quality of Data: The data should be clean and free of errors. In our project, we ensured this by performing data cleaning and preprocessing steps before feeding the data into the regression tree. This included handling missing values, removing duplicates, and dealing with outliers if necessary.

2. Relevant Features: The features used to train the model should be relevant to the outcome we are trying to predict. In our case, we used 'Avg_Duration', 'Device', and 'Source' as predictors for 'Grad_Apply', which are all relevant to the user's decision to apply.

*3. Sufficient Data: Regression trees require a sufficient amount of data to learn effectively. Our 410 observations might be on the low side, and we decided it wasn't enough to split into training and testing sets.

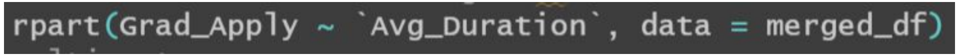
4. Independence: The observations should ideally be independent of each other. In our project, we assumed that each user's decision to apply is independent of others.

Overall, Regression Trees are very flexible because there are few requirements.

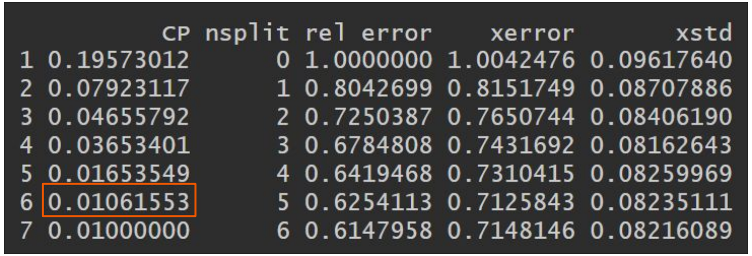
Application of Primary Method

We used the libraries, 'rpart' and 'rpart.plot', which are used for creating and visualizing regression trees in R.

For the merged dataframe, the model was fitted using 'Grad_Apply' as the response variable and 'Avg_Duration' as the predictor.

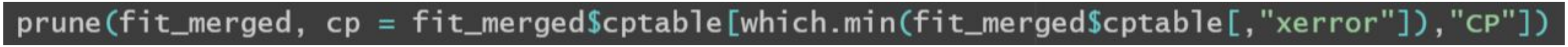


```
rpart(Grad_Apply ~ `Avg_Duration`, data = merged_df)
```



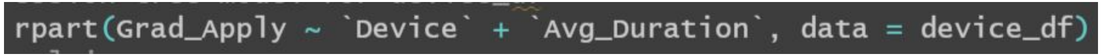
```
          CP nsplit rel error    xerror       xstd
1 0.19573012      0 1.0000000 1.0042476 0.09617640
2 0.07923117      1 0.8042699 0.8151749 0.08707886
3 0.04655792      2 0.7250387 0.7650744 0.08406190
4 0.03653401      3 0.6784808 0.7431692 0.08162643
5 0.01653549      4 0.6419468 0.7310415 0.08259969
6 0.01061553      5 0.6254113 0.7125843 0.08235111
7 0.01000000      6 0.6147958 0.7148146 0.08216089
```

We pruned the merged tree using the complexity parameter (cp) that minimized the cross-validation error. *cp was determined by the rpart package.

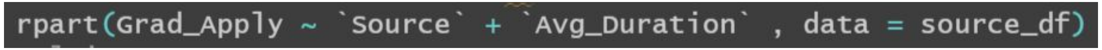


```
prune(fit_merged, cp = fit_merged$cptable[which.min(fit_merged$cptable[,"xerror"]),"CP"])
```

For device_df, the model was fitted using 'Grad_Apply' as the response variable and 'Device' and 'Avg_Duration' as predictors. For source_df, the model was fitted using 'Grad_Apply' as the response variable and 'Source' and 'Avg_Duration' as predictors.



```
rpart(Grad_Apply ~ `Device` + `Avg_Duration`, data = device_df)
```

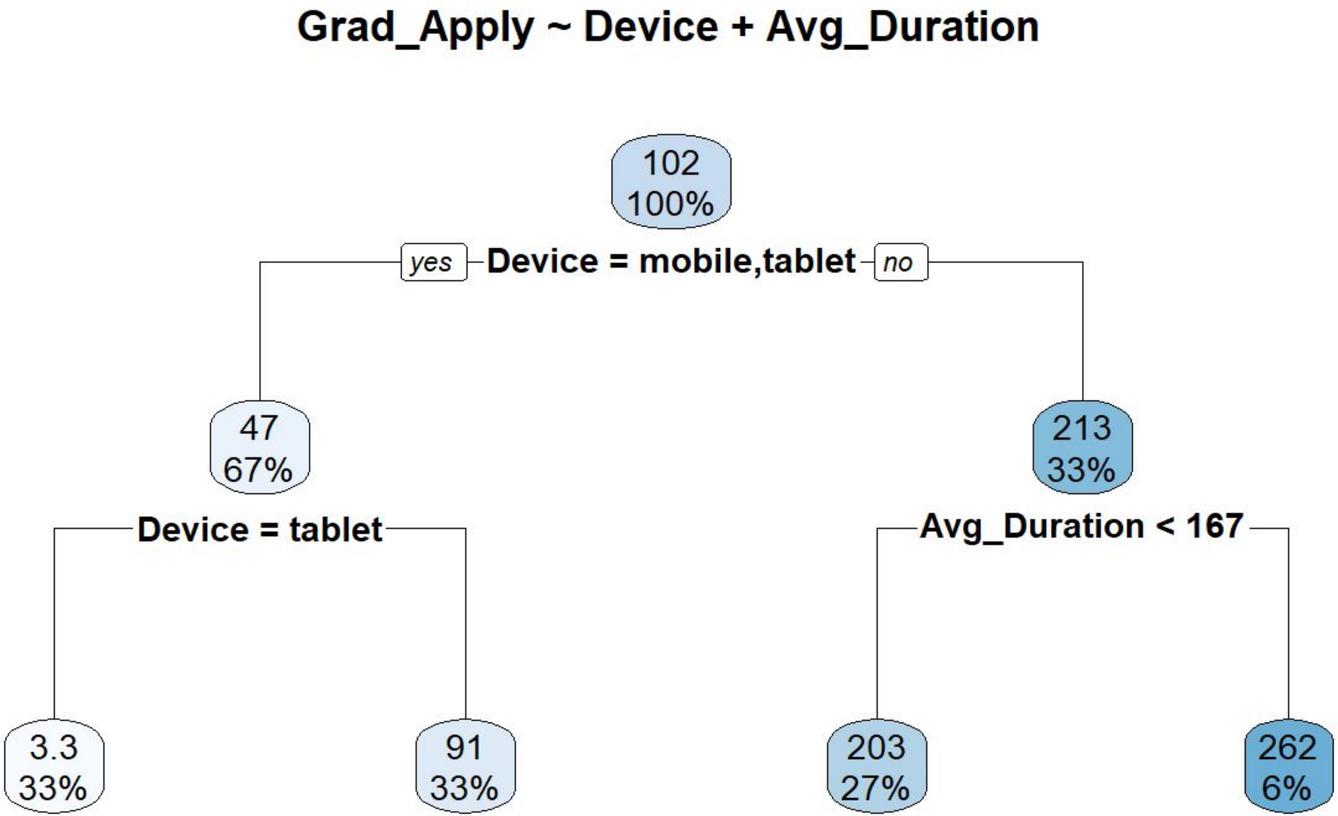


```
rpart(Grad_Apply ~ `Source` + `Avg_Duration` , data = source_df)
```

Variable Importance:

Device 847,210.9

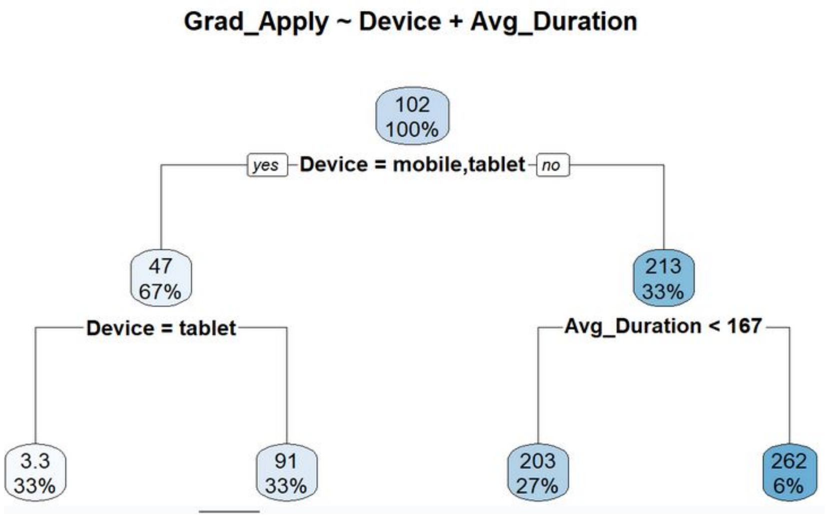
Avg_Duration 760,687.4



Device Tree Interpretation

Grad_Apply as a function of Device and Avg_Duration:
The tree looks at whether the device is mobile or tablet then splits the data based on the type of device used. If the device is a mobile or tablet, it further splits into two nodes. For instance, if the device is a tablet, the average 'Grad_Apply' is approximately 3.26. If the device is a desktop, it further splits based on 'Avg_Duration'. Essentially, the model is saying that if a user is on Desktop, that is the top indicator of clicking the ‘apply’ button. If a user is on a mobile or tablet, we can expect 3.26 users to apply against the

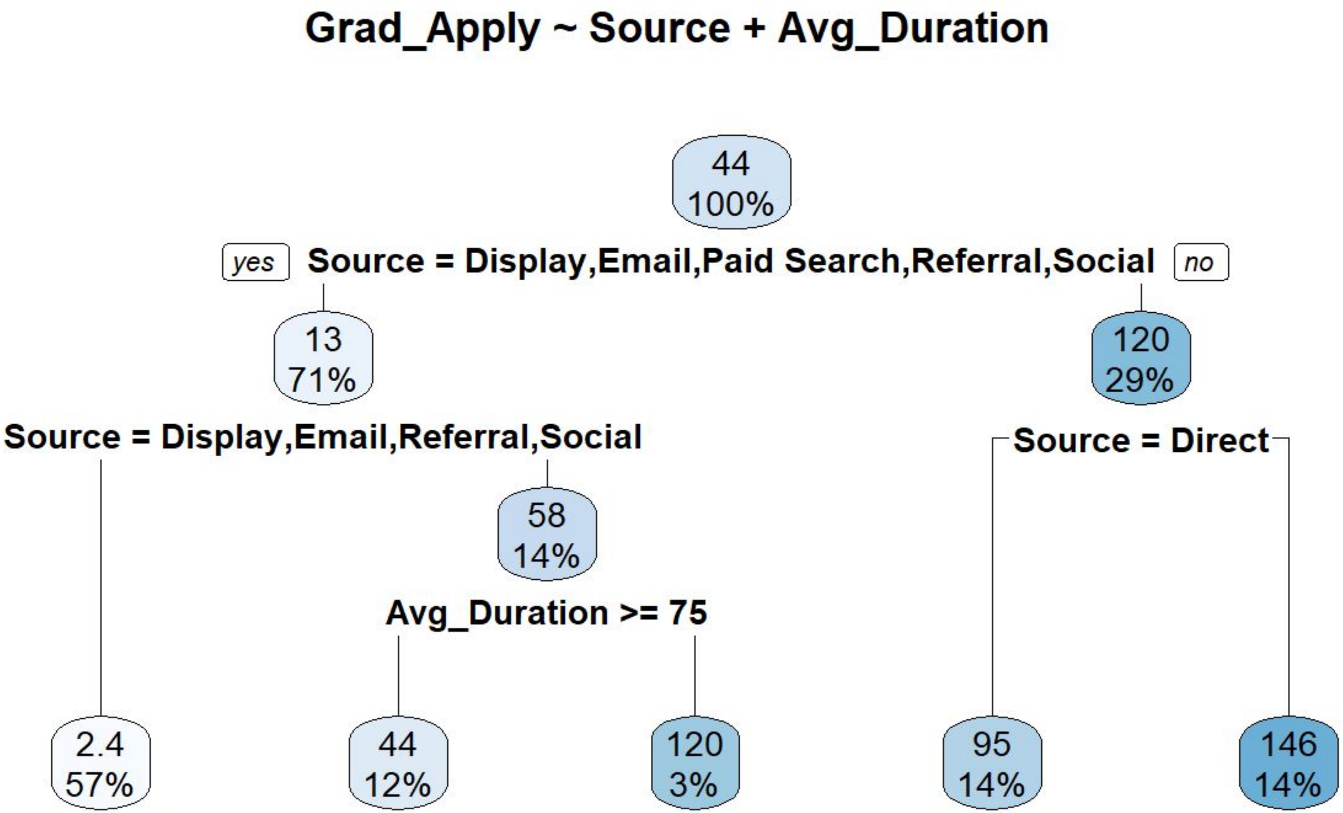
This visual shows that a desktop user is far more likely to hit the apply button within the Grad Apply and Device Model.



Variable Importance:

Source 761,325.1

Avg_Duration 283,633.8

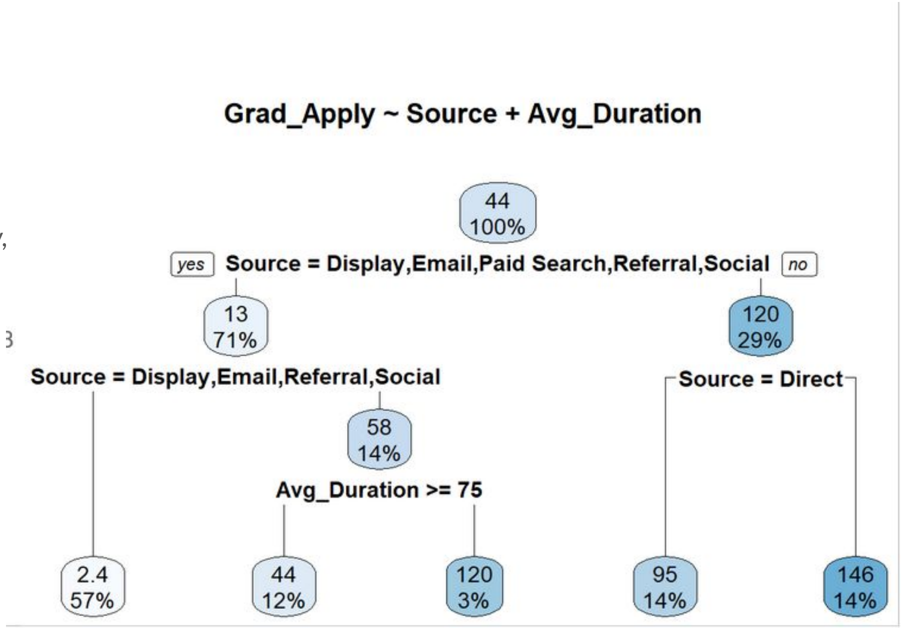

Grad_Apply ~ Source + Avg_Duration
44
100%
yes
Source = Display,Email,Paid Search,Referral,Social
no
13
71%
120
29%
Source = Display,Email,Referral,Social
58
14%
Avg_Duration >= 75
Source = Direct
2.4
57%
44
12%
120
3%
95
14%
146
14%

Source Tree Interpretation

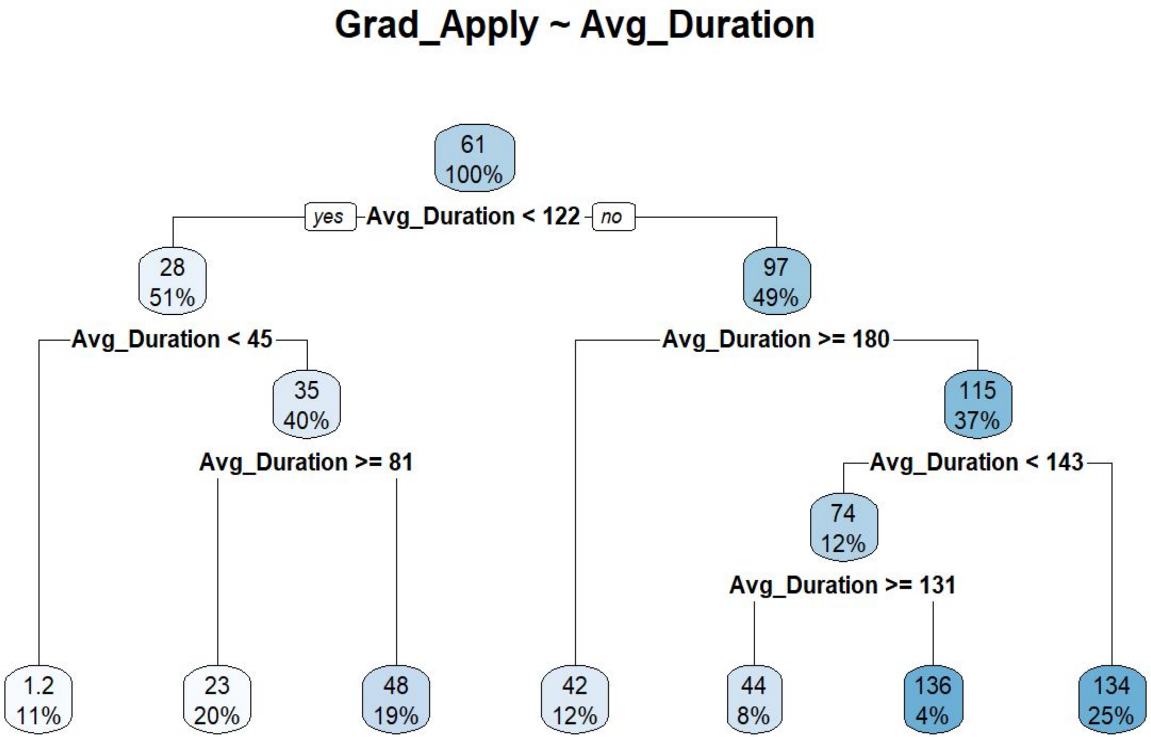
Grad_Apply as a function of Device and Avg_Duration:

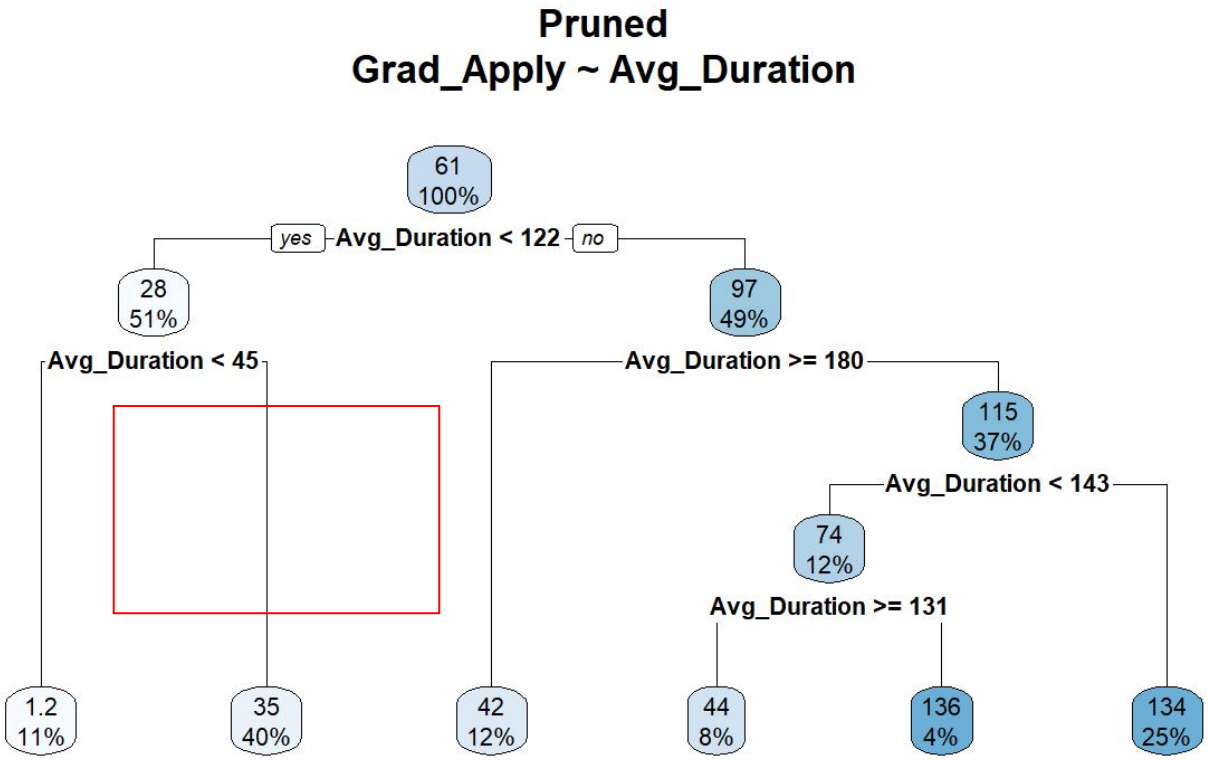
This tree splits the data based on the source of the traffic. If the source is Display, Email, Paid Search, Referral, or Social, it further splits into two nodes. One for Display, Email, Referral, and Social, and the other for Paid Search. For instance, if the source is Display, Email, Referral, or Social, the average 'Grad_Apply' is approximately 2.35. The tree continues to split based on 'Source' and 'Avg_Duration' until no further significant splits can be made.

Organic Search has the most, at 146. Short length Paid Searches also was a sweet spot, at 120.



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
Source = Direct
58
14%
Avg_Duration >= 75
2.4
57%
44
12%
120
3%
95
14%
146
14%


Grad_Apply ~ Avg_Duration
61
100%
yes
Avg_Duration < 122
no
28
51%
97
49%
Avg_Duration < 45
35
40%
Avg_Duration >= 180
115
37%
Avg_Duration >= 81
Avg_Duration < 143
74
12%
Avg_Duration >= 131
1.2
11%
23
20%
48
19%
42
12%
44
8%
136
4%
134
25%



Pruned
Grad_Apply ~ Avg_Duration

Avg_Duration only Tree Interpretation

Grad_Apply as a function of Avg_Duration:

Overall expected value is 61 grad applications for two weeks.

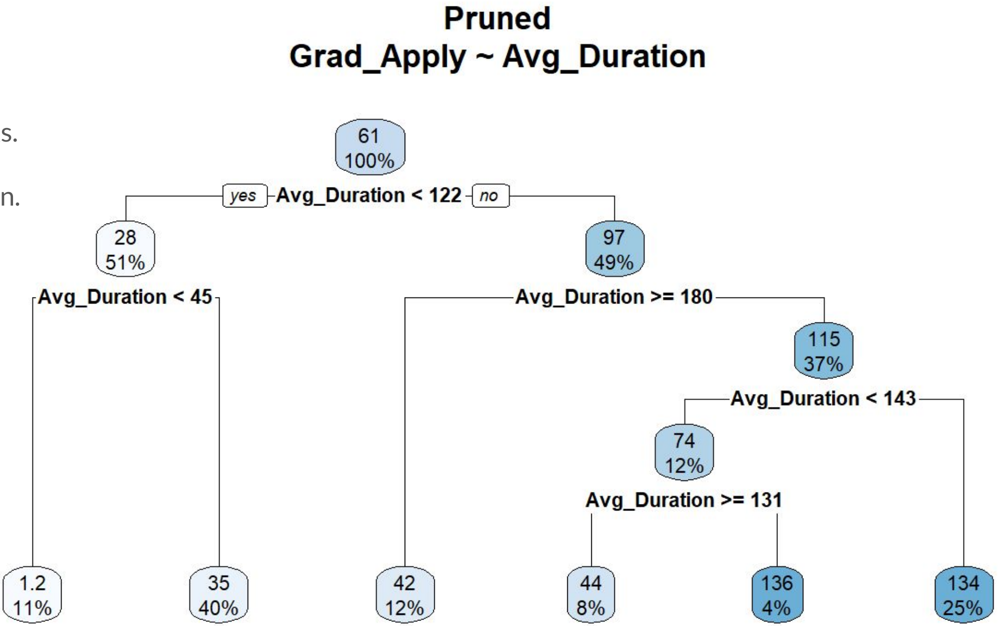
Population is split evenly below and above 2 minute duration.

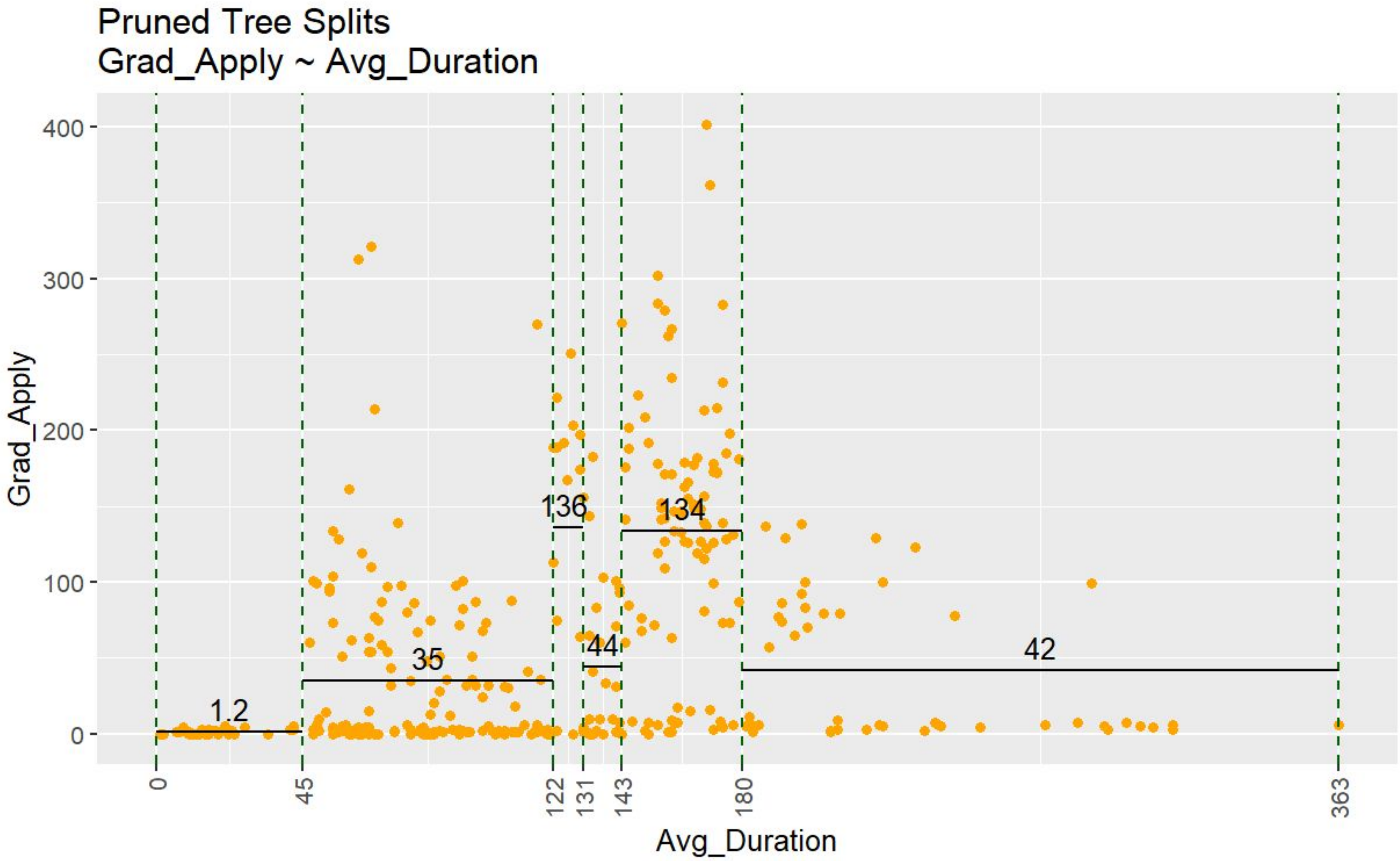
Conversions below 2 minutes are much lower (28).

Between 2 and 3 minutes is the sweet spot (37% at 115)

12% are over 3 minutes, which is predicted to drop off (42).

Even sweeter spot between 2.4-3 minutes (25% at 134)




Pruned Tree Splits
Grad_Apply ~ Avg_Duration
Grad_Apply
400
300
200
100
0
1.2
35
136
44
134
42
0
45
122
131
143
180
363
Avg_Duration

Conclusion

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?

- Regression with pruning involved
- Traffic Source
  - Organic Search has highest
  - Paid Search & Less than 75 seconds has second highest
- Device Types
  - Desktop with Duration greater than 3 minutes
- Average Time on Site
  - Sweet spot between 2 - 3 minutes