DATA EXPLORATION & CLASSIFICATION MODEL

Thuy Nguyen

Task

1. *What % of 'email' companies signed up on a Saturday or Sunday?* Answer: **32.36%**

Steps:

- Extract the day of the week from 'free_trial_signup_date' column by using the formula: = TEXT(B2,"ddd")
- Create a 'sign up on weekend' column with Boolean value:
  0 = not sign up on weekend
  1 = sign up on weekend

I used the if(or ()) function: =IF(OR([@[Day of the week]]="sun",[@[Day of the week]]="sat"),1, 0)

company_id	free_trial_signup_date	google_analytics_medium	Day of the week	Sign up on weekend
17711836	02/11/2020	cpc	Mon	0
17712950	08/11/2020	cpc	Sun	1
17723760	12/12/2020	(none)	Sat	1
17713380	09/11/2020	cpc	Mon	0
17724228	13/12/2020	email	Sun	1

- Create a pivot table:
  - Drag the google_analytics_medium column to Rows
  - Add count of "Company ID" and sum of "Sign up on weekend" to the Values box

Medium	Sum of Sign up on weekend	Total signup
(none)	233	745
cpc	188	619
email	169	522
organic	1249	4167
social	11	34
Grand Total	**1850**	**6087**

We'll get a table that provide us with the info needed:
Here there are 169 "email" companies that sign up on the weekend over the total of 522 "email" companies.
So 169/522*100% = **32.36%** "email" companies sign up on the weekend.

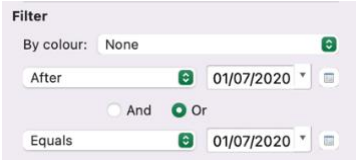
2. *Are app users or non app users more likely to currently be a subscriber?* **App users are more likely to currently be a subscriber.**
   Using pivot table again, we get:

App User	Count of subscription_status
FALSE	**4050**
Cancelled	1435
Subscribed	2615
(blank)	
TRUE	**1910**
Cancelled	268
Subscribed	1642
(blank)	
Grand Total	**5960**

The percentage of app users who are currently subscribers: 1642/1910*100% = 85.97%
The percentage of non_app users who are currently subscribers: 2615/4050*100% = 64.57%

3. *Of companies that signed up on or after 1st July 2020, which medium had the highest average minutes activity on any device over the last 7 days?* ***Organic***

Here, I used MATCH & INDEX to take the column 'Minutes_activity_last_7_days_all_devices" from the signup file to the company_information file. The I changed the free_trial_signup_date filter to *equal to or after* 01/07/2020:



medium	Average of minutes_activity_last_7_days_all_devices
(none)	4.324832215
organic	3.861930295
cpc	3.168012924
email	2.084291188
social	0.147058824
Grand Total	**3.528135991**

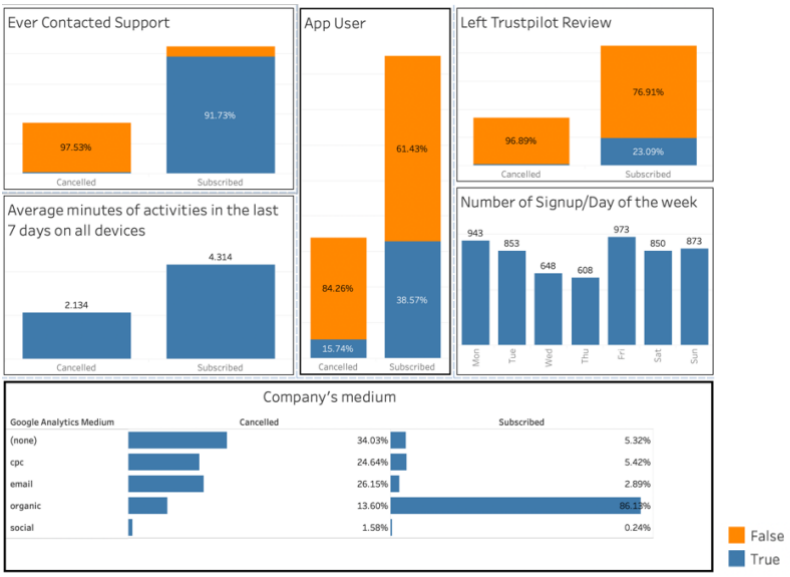
I created a pivot table and sort the values from largest to smallest. Here, undetected medium (none) had the highest average minutes activity on all devices over the last 7 days. Among the detected medium, **organic** had the highest average minutes activity on all devices over the last 7 days.

DATA EXPLORATION & CLASSIFICATION MODEL

Thuy Nguyen

Investigation

I'd like to create a classification model that can predict whether a company will subscribe to Company A using data from the datasets, such as whether they've ever contacted support, left a Trustpilot review, their minutes of activity, their free trial registration date, and so on. Before I do that, I'd like to look into the relationship between the target (subscription status) and the available independent variables.



1. Exploratory Data Analysis

The dashboard suggests that the independent variables are highly correlated to the target (subscribe/cancel). High engagement suggests subscription. Subscribers spend more than twice as much time on Company A than cancellers (4 minutes vs 2 minutes). Customer support engagement is likewise substantially connected, with 91.73 percent of subscribers contacting support and 97.5 percent of cancellers never contacting support. The percentage of subscribers using app is more than double that of cancellers (38% vs 16%). Subscribers are 20 times more likely than non-subscribers to leave a Trustpilot review. Subscribers discover Company A mostly through their own searches, with 86 percent coming from organic sources. As for cancellers, the arrived medium is equally distributed. The most popular day to sign up for free trial is Friday.

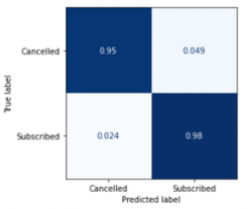
2. Data Pre-processing

I used Python on Jupiter Notebook to work on the model. 339 missing values in the column *subscription_status* (the target) was removed, and the merged dataset was split into training and testing set by stratified sampling based on subscription_status. The categorical variable, *Google Analytics Medium, is encoded into dummy variables, Minutes_activity_last_7_days_mobile* and *Minutes_activity_last_7_days_desktop* are not considered because they will be correlated with another all-including variable, *Minutes_activity_last_7_days_all_devices.*

3. Model

Random Forest is the machine learning algorithms used to fit the data. It is chosen because of the higher accuracy through cross validation. Grid Search is used to find the best Random Forest model. Here, we can see that the best model returns 96,6% accuracy. The correlation matrix on the test set shows 98% of the subscribed companies are correctly predicted. The mean train score is only 1.6% larger than the mean test score, suggesting little overfitting.

	params	mean_train_score	mean_test_score	diff, %
4	{'max_depth': None, 'n_estimators': 200}	0.981694	0.965908	1.608030
5	{'max_depth': None, 'n_estimators': 500}	0.981697	0.965305	1.669735
3	{'max_depth': None, 'n_estimators': 100}	0.981697	0.965076	1.693112
0	{'max_depth': 5, 'n_estimators': 100}	0.963929	0.962546	0.143518
1	{'max_depth': 5, 'n_estimators': 200}	0.964726	0.961758	0.307602
2	{'max_depth': 5, 'n_estimators': 500}	0.963631	0.961161	0.256360




```
year                                        0.321397
ever_contacted_support                      0.285080
month                                       0.194452
organic                                     0.103240
day_of_week                                 0.030391
email                                       0.020023
minutes_activity_last_7_days_all_devices    0.015135
cpc                                         0.011781
left_trustpilot_review                      0.008980
app_user                                    0.008116
social                                      0.001406
```

4. Assumptions and notes

I assumed that the *Minutes_activity_last_7_days* is only for free-trial users. Otherwise, it would be obvious that subscribers would be more active on the website than cancellers. Also, the screenshot shows the list of independent variables that have been used by the model and their weights in the model. Data for different years are not evenly collected (only 1 month of data for 2021, 4 months for 2019 and the whole 12 months for 2020). In this model, I assume that the month and year observations are distributed evenly in training and testing dataset. Variable year and month might need to be removed.