GENERAL ASSEMBLY

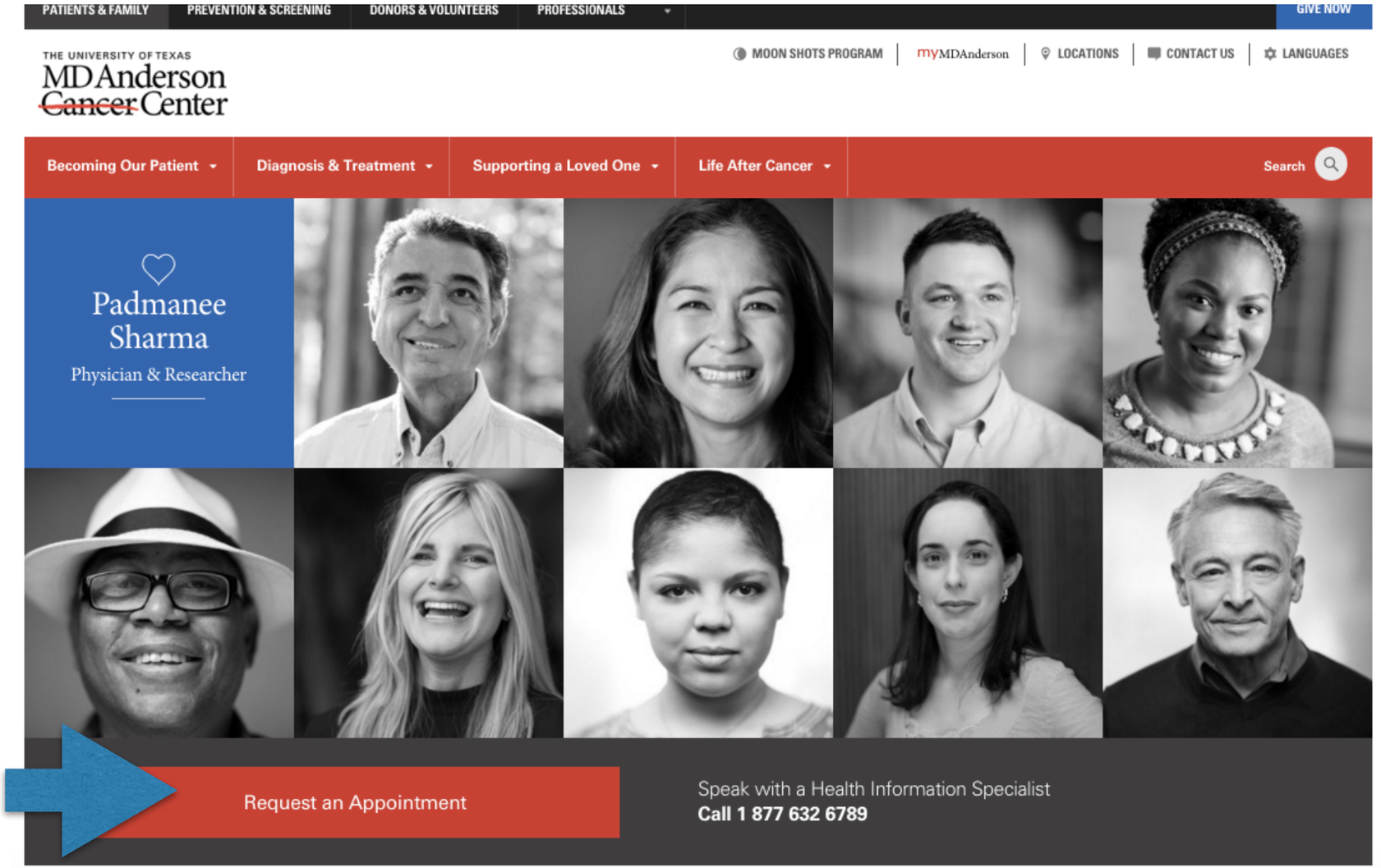
FINAL PROJECT

Nicole Baez - MD Anderson, “Predict Appointment Requests”

THE PROBLEM - PREDICTING LEAD CONVERSION 2

- MD Anderson is one of the largest cancer centers in the world. They have been working to eliminate cancer for more than six decades and drive 1.2M visitors to their global site per day, yet appointment requests are less than 1% (conversion).
- Requesting an appointment online is one of the easiest ways for patients to get ahead of early stage cancer detection and is one of the top lead conversions captured by the digital team to drive and inform fundraising and targeted media spend.
- **Goal: This project attempts to understand the relationship between request an appointment conversion online (target) and site behaviors which can positively impact lead submissions.**

THE PROBLEM - 1.2% LEAD CONVERSION



- Less than 2% of site visitors engage with the largest call to action on the homepage. By understanding the relationship with variables, a prediction model to target likely converting visitors can help with this challenge.

MD ANDERSON SITE DATA

- This project will only use data exported from Omniture, which is click-stream data.
- Data is a .csv file with 2016 YTD daily activity.
- For the initial analysis daily data includes site visitors features segmented by Search Engine, Direct Entry, and Cancer Type (pages visited related to cancer types). The outcome (target) variable is request an appointment visits (RA).

HYPOTHESIS- CONVERSION PROBABILITY

- By using a few broad site traffic features in the data set I want to determine if there is any kind of relationship to request an appointment based on segmented traffic visitors. These features showed the highest correlation to visitors who requested an appointment.
- Success for this project means a simple, clear understanding of which traffic segment is more likely to convert and submit a lead. Additional analysis of specific traffic segments such as device, time of day, city and offline paid marketing exposure could add more context.
- Since visitors who reach the site directly or a via a search engine have the highest correlation I believe these traffic segments will glean the

DATA DESCRIPTION


```
In [3]: print data.head()
```

```
     Date  Search Engine   Direct Entry   Cancer Type   RA
0  1/1/16              16             20          110   57
1  1/2/16              20             15           90   52
2  1/3/16              22             18          118   62
3  1/4/16              77            102          353  247
4  1/5/16              78             87          350  242
```

```
In [15]: data.shape
```

```
Out[15]: (93, 5)
```

```
In [4]: print data.describe()
```

```
       Search Engine   Direct Entry   Cancer Type          RA
count       93.000000      93.000000    93.000000   93.000000
mean       175.741935     198.806452   679.193548  412.462366
std         73.638055      91.011855   273.996242  172.945679
min         16.000000      15.000000    90.000000   52.000000
25%         99.000000      99.000000   435.000000  222.000000
50%        213.000000     248.000000   779.000000  519.000000
75%        234.000000     272.000000   897.000000  555.000000
max        285.000000     312.000000  1306.000000  616.000000
```

```
In [5]: data.corr()
```

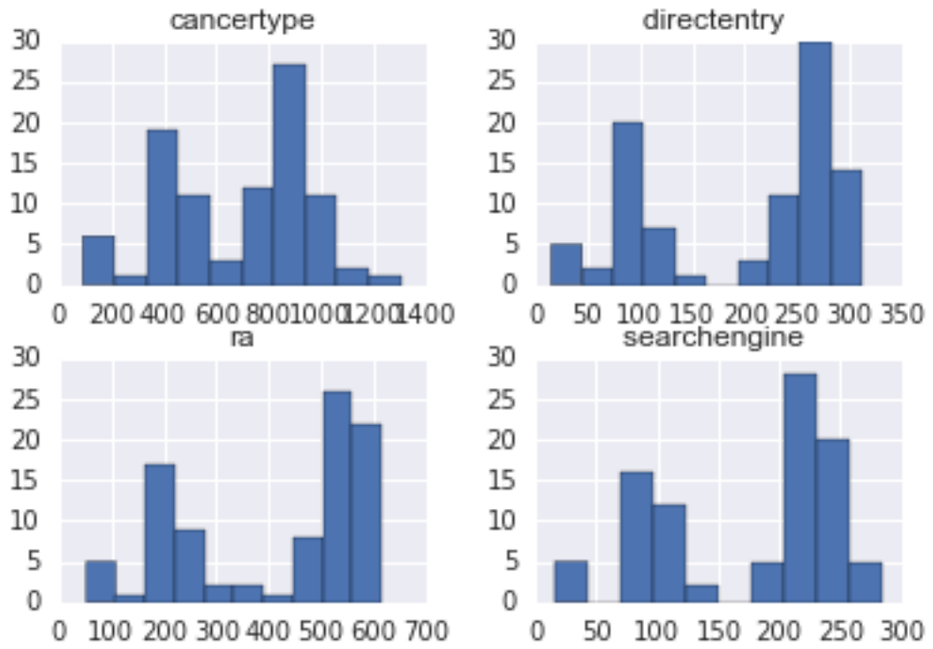
Out[5]:

	Search Engine	Direct Entry	Cancer Type	RA
Search Engine	1.000000	0.965710	0.866196	0.978897
Direct Entry	0.965710	1.000000	0.860911	0.991921
Cancer Type	0.866196	0.860911	1.000000	0.875027
RA	0.978897	0.991921	0.875027	1.000000

- 93 rows (days) and 5 columns. This data set might be too small for the analysis. Looking into exported the previous year as well.

DATA VISUALIZATION


```
In [31]: data.hist()
         plt.show()
```



> Visitors (visits)

- 93 rows (days) and 5 columns. This data set might be too small for the analysis. Looking into exported the previous year as well, but unlikely as the site tracking optimization is recent.

PREDICTION MODEL IDEAS

- Decision tree: Determining the most relevant factors for conversion probability. (Lesson 12/13)

- Logistic regression: Regressing for the probability of a categorical outcome, rather the probability of an event, in this case the event is site conversion by submitting an appointment lead. (Lesson 5/6/7)