I’m Exhausted

Modelling fundraising data to target Donors and Donation Size

Problem Identification

- **Client:** Post-Secondary Foundation
- **Dataset:** Alumni database containing biographical features and donation history
- **Problem:** How can this data be leveraged to increase donations and decrease costs of raising a dollar (customer acquisition).
- **Solution:**
  - **Classifier:** Whether a constituent has/will donated or not
  - **Regression:** What amount have/will they donated

Data Description

- Using the sample dataset from COOL DATA, a how-to-guide for predictive modeling for higher education available for free
- The dataset can be divided into 4 sections:
  - 12 boolean columns that describe biographical features of an alumni
  - 1 float column aggregating an alumni’s current total donation with the client
  - 1 categorical column describing the alumni’s current marital status
  - 1 date year column indicating the alumni’s graduation year

Data Description continued

- The two predictor variables:
  - **Regression:** Cumulative Donations, available in dataset
  - **Classification:** Has the alumni donated, generated from Cumulative Donations greater than $0.00.

Predictive Modelling Lifecycle

```
# 0th bit [-1]:         0 - logistic regression
#                       1 - linear regression
# 1st bit [-2]:         0 - grad_year int
#                       1 - grad_year binned
# 2nd bit [-3]:         0 - cum_donation float
#                       1 - cum_donation binned
# 3rd-5th bit [-6:-3]:  000 - no automatic feature selection
#                       001 - chi square filtering (chi)
#                       010 - Random Forest Importance (rfi)
#                       011 - Recursive Feature Elimination Cross Validation (rfe)
#                       100 - Forward Feature Elimination (ffe)
# 6th-7th bit [-8:-6]:  00 - unscaled
#                       01 - MinMaxScaler
#                       10 - StandardScaler
#                       11 - RobustScaler
# 8th bit [-9]:         0 - Cross Fold Validation
#                       1 - Stacking
```

Predictive Modelling Lifecycle

- Automatic Feature Selection produced a wide number of total columns to drive model development. Models had as little as 1 predictor column all the way up to 43.

Data Exploration….I used Pandas

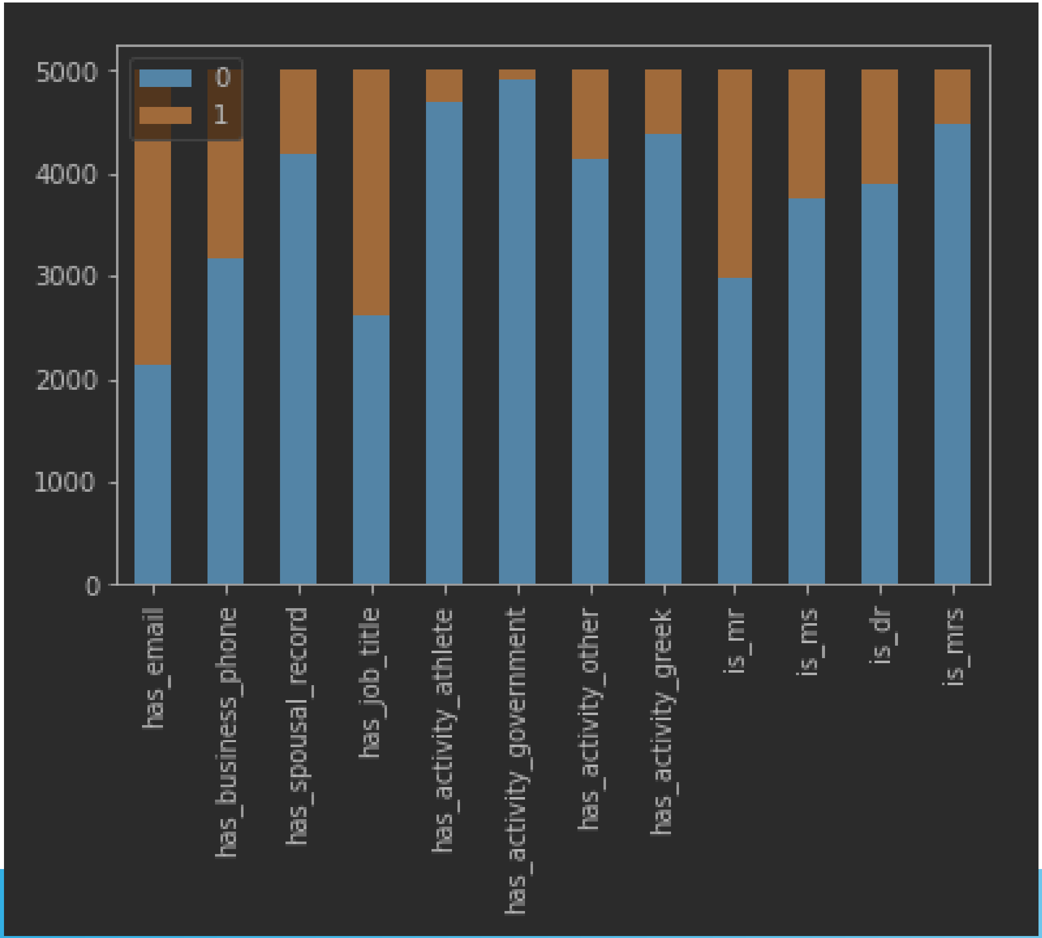
- In general

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 18 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   id                       5000 non-null   object
 1   cum_donation             5000 non-null   float64
 2   has_email                5000 non-null   int64
 3   has_business_phone       5000 non-null   int64
 4   grad_year                5000 non-null   int64
 5   marital_status           4965 non-null   object
 6   has_spousal_record       5000 non-null   int64
 7   has_job_title            5000 non-null   int64
 8   has_activity_athlete     5000 non-null   int64
 9   has_activity_government  5000 non-null   int64
 10  has_activity_other       5000 non-null   int64
 11  has_activity_greek       5000 non-null   int64
 12  is_mr                    5000 non-null   int64
 13  is_ms                    5000 non-null   int64
 14  is_dr                    5000 non-null   int64
 15  is_mrs                   5000 non-null   int64
 16  grad_decade              5000 non-null   category
 17  cum_range                5000 non-null   category
dtypes: category(2), float64(1), int64(13), object(2)
memory usage: 635.9+ KB
```

Data Exploration: Boolean Columns

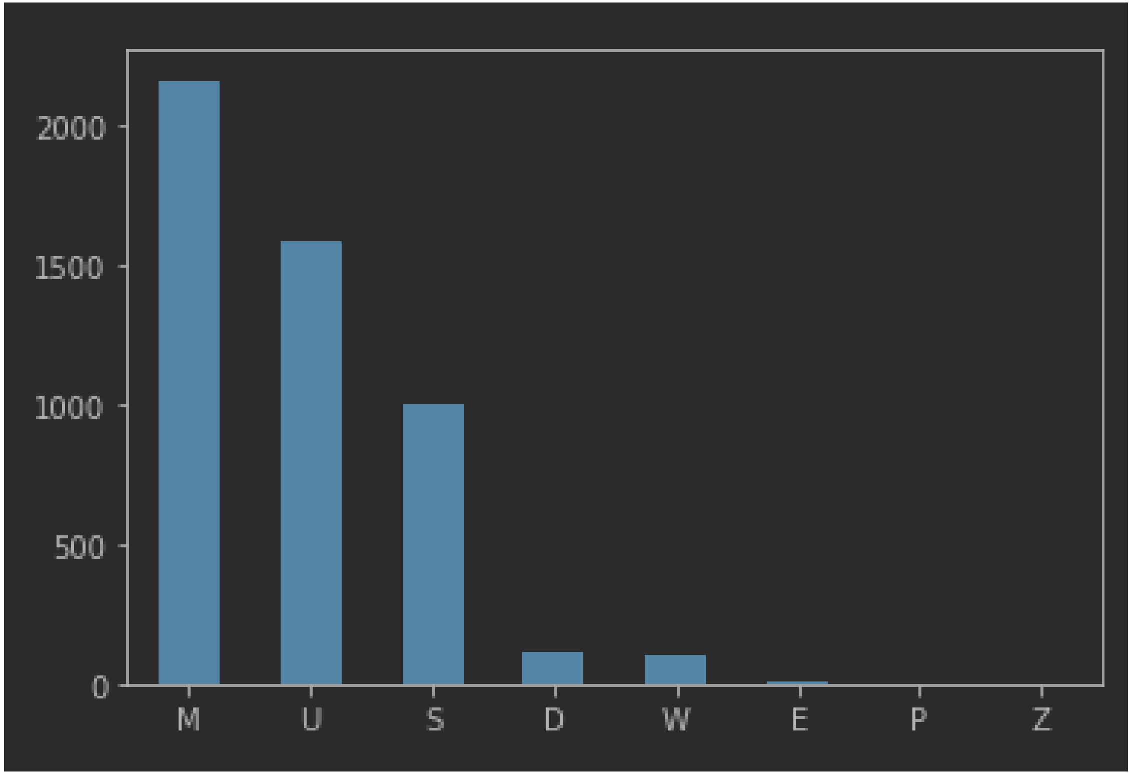
```
has_email                    5000
has_business_phone           5000
has_spousal_record           5000
has_job_title                5000
has_activity_athlete         5000
has_activity_government      5000
has_activity_other           5000
has_activity_greek           5000
is_mr                        5000
is_ms                        5000
is_dr                        5000
is_mrs                       5000
dtype: int64
```

	0	1
has_email	2125	2875
has_business_phone	3171	1829
has_spousal_record	4184	816
has_job_title	2626	2374
has_activity_athlete	4703	297
has_activity_government	4902	98
has_activity_other	4145	855
has_activity_greek	4380	620
is_mr	2968	2032
is_ms	3752	1248
is_dr	3908	1092
is_mrs	4484	516



Data Exploration: Marital Status

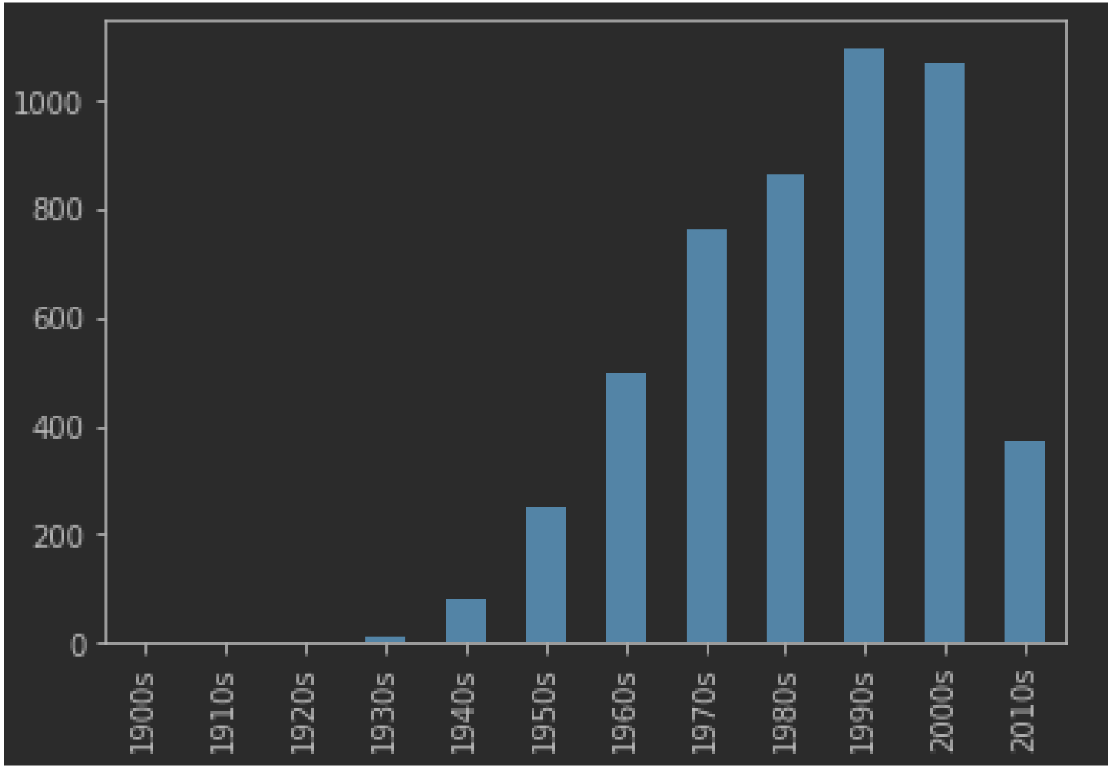
```
null count 35
value count 4965

M    2160
U    1586
S     996
D     110
W     106
E       4
P       2
Z       1
Name: marital_status, dtype: int64
```



Data Exploration: Grad Year

```
null count 0
value count 5000
min val 1911
max val 2013
```

```
1900s       0
1910s       1
1920s       0
1930s      12
1940s      78
1950s     249
1960s     497
1970s     761
1980s     864
1990s    1096
2000s    1068
2010s     374
Name: grad_decade, dtype: int64
```



Data Exploration: Cumulative Donations

```
null count 0
value count 5000
min val 0.0
max val 11187224.58
```

```
$0                   2555
$1-$999.99           1843
$1K-$9.99K            518
$10K-$24.99K           46
$25K-$49.99K           14
$50K-$99.99K            7
$100K-$249.99K         10
$250K-$499.99K          4
$500K-$999.99K          0
$1M-$2.49M              2
$2.5M-$4.99M            0
$5M-$9.99M              0
$10M-$14.99M            1
Name: cum_range, dtype: int64
```