Propensity to Give

TOC

Problem Statement

Can we determine who will be an elite donor?

- Given past donation amounts
- Provided a set of features describing donors
- “Elite donor” being any donor likely to give more than $20,000 over 5 years.

Steps to solve

1. Obtaining the data.
2. Cleaning, organizing, and featurizing data sets.
3. Building a model to analyze data.
4. Visualizing and reporting the findings.

Project objective 1

Data Collection and Preparation

Data Collection

- Provided as two .CSVs:
  - 1st set contained historical donation amounts.
  - 2nd set, a collection of features describing donors.

Data Preparation

Step 1:

Creating a binary target from donation history.

Data Preparation

Creating a binary target from donation history:

- Dropped years prior to 2000
- Converted dates to Pandas Timeseries
- Summed donations per year, per donor

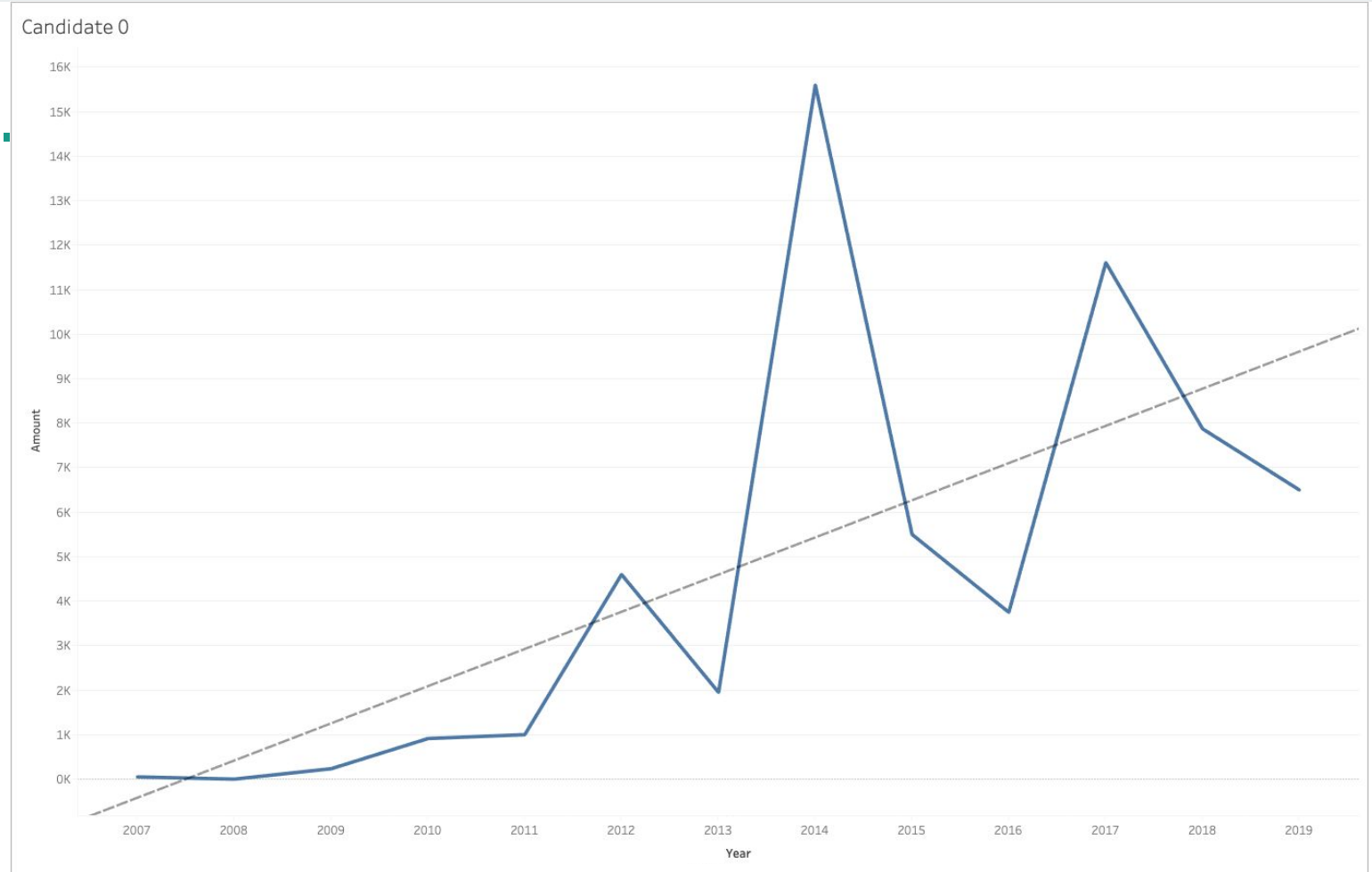
Data Preparation

		amount
cand_id	trans_date	
candidate_0	2007-12-31	50.00
	2009-12-31	234.39
	2010-12-31	912.50
	2011-12-31	1000.00
	2012-12-31	4600.00
	2013-12-31	1950.01
	2014-12-31	15600.00
	2015-12-31	5500.00
	2016-12-31	3750.00
	2017-12-31	11609.94
	2018-12-31	7874.97
	2019-12-31	6500.00

Data Preparation

		amount
cand_id	trans_date	
candidate_0	2007-12-31	50.00
	2009-12-31	234.39
	2010-12-31	912.50
	2011-12-31	1000.00
	2012-12-31	4600.00
	2013-12-31	1950.01
	2014-12-31	15600.00
	2015-12-31	5500.00
	2016-12-31	3750.00
	2017-12-31	11609.94
	2018-12-31	7874.97
	2019-12-31	6500.00

Variation year to year


Candidate 0
16K
15K
14K
13K
12K
11K
10K
9K
8K
7K
6K
5K
4K
3K
2K
1K
0K
Amount
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
Year

Data Preparation

Creating a binary target from donation history:

- Created a rolling average using Pandas Rolling method
  - Minimum years equal to one, to include new donors
  - Large window to account for variation
- Filtered for final average donation

Data Preparation

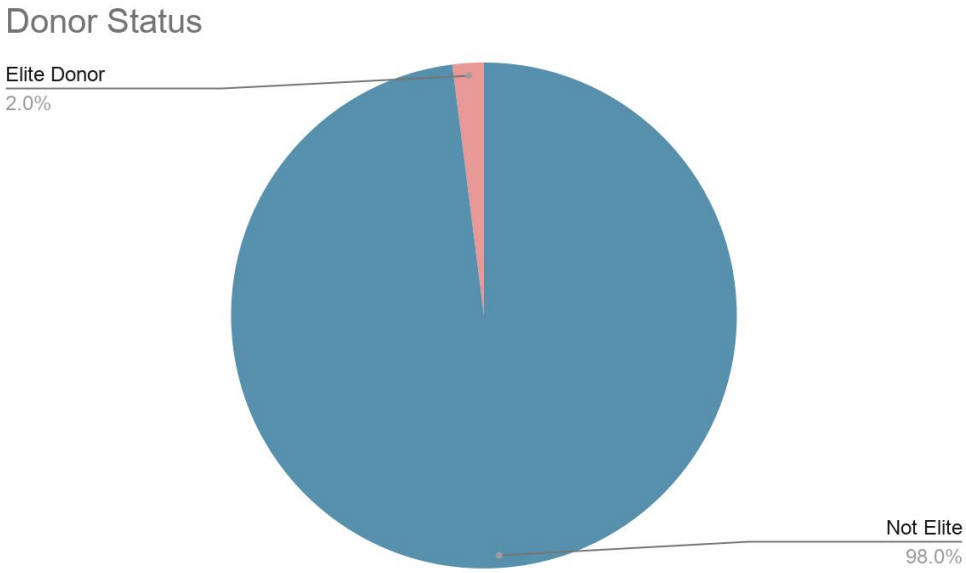
	cand_id	rolling_avg
11	candidate_0	4965.150833
12	candidate_1	25000.000000
13	candidate_10	85.000000

Data Preparation

Creating a binary target from donation history:

- Assumed a yearly donation average greater than $4,000 would be likely to donate $20,000 over 5 years
- Created binary target on the above assumption

Data Preparation



Data Preparation

- Columns containing rolling average and binary target merged with features dataset on the candidate's ID
- Any candidate missing status classification or a feature set dropped
- Final dataframe exported for analysis

Project objective 2

Model Testing

Model Testing

Determined baseline model performance

- StandardScaler for numeric features
- Logistic Regression
    - Default parameters
- No handling of target imbalance

Model Testing

Baseline model performance

- Accuracy = *97% vs 98% null model*
- Precision = 0.08
- Recall = 0.02
- F1-score = 0.04
- ROC AUC = 0.51

Model Testing

Featurization

- All features utilized
    - Some correlation noted between variables
- StandardScaler for all numeric features
- Other techniques such as PolynomialFeatures, RFE, and setting a variance threshold did not increase performance

Model Testing

Handling imbalanced data

- SMOTE & ADASYN (oversampling) from imblearn library, with a range of sampling strategies, tested with logistic regression
- Both increased recall by ~3500% (0.02 → 0.77)

Model Testing

- Several models fit with varying sampling strategies.
- Random forest classification with ADASYN highest performer

Model Testing

Best model parameters determined via GridSearchCV:

- StandardScaler
- Sampling strategy = 0.8
- Criterion = entropy
- N estimators = 1000
- Oob score = true

Project objective 3

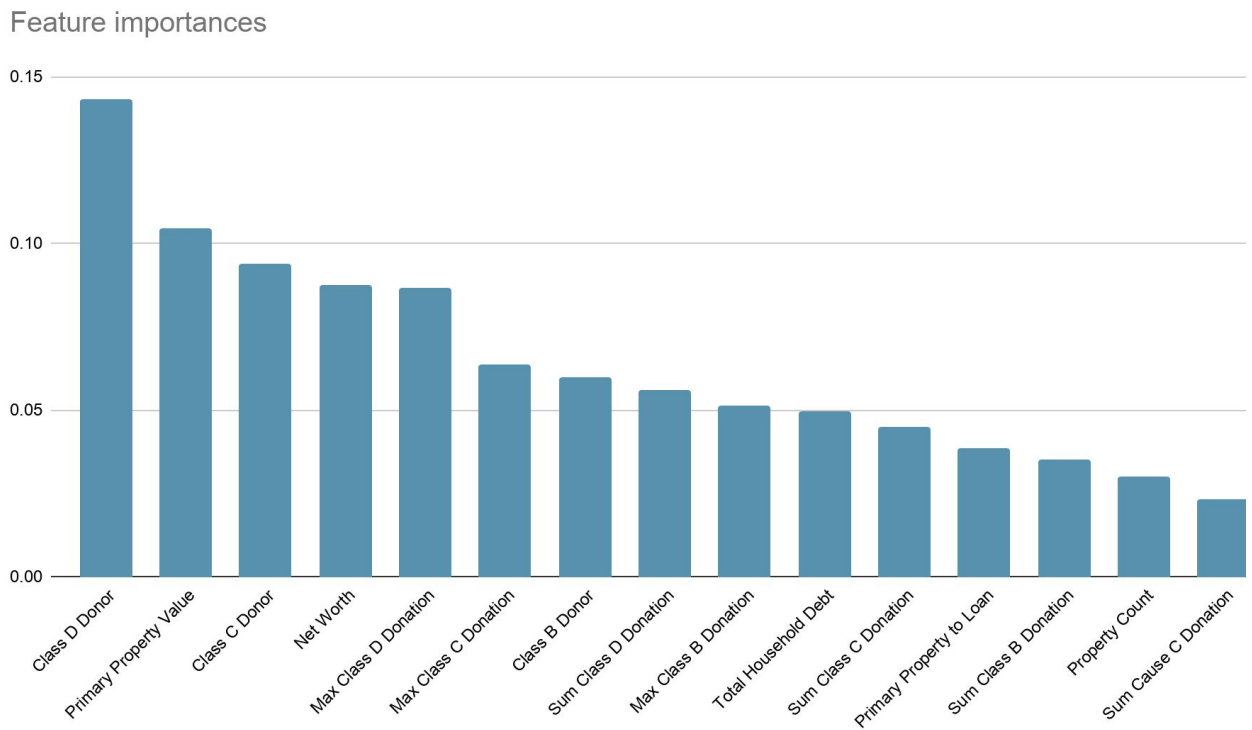
Classification and Evaluation

Model Performance

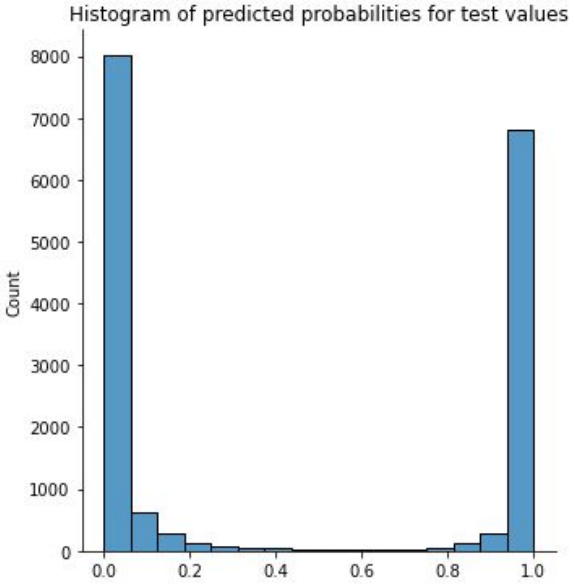
- Accuracy = 99% *vs 55% null model*
- Precision = 0.99
- Recall = 0.99
- F1-score = 0.99
- ROC AUC = 0.99

Confusion matrix shows a slight tendency toward false positives.

Feature Extraction



Probabilities



Probabilities calculated for each donor could be used to add additional individuals to a marketing campaign.

Project objective 4

Deploy a Functional Script

Deploying Script

- User inputs 2 .CSV files
- App creates target from rolling average, models data, and outputs metrics, predictions, probabilities, and feature importances, and an optional EDA report.


Variables

Conclusion

- Donor targets can be created using Pandas Timeseries and Rolling methods
- Under-represented target variables can be over-sampled using imblearn's ADASYN method
- "Elite" status donors can be classified with 0.99 recall
- The most important features and prediction probabilities can be extracted for greater marketing abilities
- The work above can be deployed in a user-friendly application to streamline analysis

Any Questions?

All photos courtesy of pixabay.com

Thanks!

Corey J Sinnott

Data Scientist

CoreyJSinnott.squarespace.com

Sinnott.CJ@gmail.com