Windfall Case Study: 2020 FEC Contributions

Saran Ahluwalia

December 13, 2022

Contents

0.1 Questions 1 Rationale . . . 2

0.1.1 What are some ways that we can uniquely identify individuals? . . . 2

0.1.2 What transformations on the data should we consider making to accurately group and associate transactions to individuals? . . . 3

0.2 Questions 2 Rationale . . . 11

0.2.1 For the calendar year 2020, what were the names of the top 20 committees receiving contributions, and how much did they receive? . . . 11

0.2.2 Who were the top 20 individuals making contributions and how much did they contribute? . . . 11

0.2.3 How many committees are both recipients and contributors? Do these results align with your expectations? . . . 11

0.3 Questions 3 Rationale . . . 12

0.3.1 What is the calendar effect and role of seasonality (if any) in donation activity? Does the seasonality change at all in off-year election cycles? . . . 12

0.4 What are the benefits of using time-to-event models and why is the GLM with a logit link function not always appropriate? . . . 12

0.4.1 Considering alternative modeling approaches . . . 13

1 Appendix **14**

1.1 Elaborating on using Linear Mixed-Effects when applied to voter contributions . . 14

Abstract

The following exposition and the accompanying PDF documents in the results directory detail:

1. Model selection considerations - as alternates to the generalized linear model (GLM) with a logit link especially when faced with censored observations in the voter contribution dataset. Various business implications are discussed and alternative modeling suggestions are detailed.

2. I expound on the original questions regarding grouping and transformations by providing an example characterization of contributions in the past 2020 senate election in North Carolina. I chose this election because, as a resident of North Carolina, I was curious to know more about the successful election campaign of Republican candidate Thom Tillis. Based on this analysis, it is clear that Thom Tillis was heavily supported by retirees and older demographics; conversely, the unemployed heavily favored Cal Cunningham. Finally, there was an outsized monetary investment from outside of North Carolina in both campaigns.

3. Finally, to fully appreciate the end-to-end voter identification, characterization and potential avenues for improving predictive capabilities I use 1) the American Community Survey (ACS) 2015 - 2019; the 2020 American Presidential Election precinct results; 3) North Carolina registered voter registration records, and 4) North Carolina campaign donation records to answer: *Who will donate money to a political campaign in 2020?* I characterize voters and their contribution history spanning 2012 - 2020, inclusive.
4. Finally, the aforementioned datasets are used to construct a logit model. Using an out-of-sample testing set, and a threshold of 0.166, the baseline logit model with the least absolute shrinkage and selection operator (LASSO) can correctly classify 78.8 percent of cases (i.e., 22.2 percent testing error). This is not an improvement on the no information rate, which is 0.8489. Moreover, the model only correctly identifies positive classes (i.e., 2020 donors) 16.37 percent of the time. The model correctly identifies negative classes (i.e., non - 2020 donors) at 90 percent. Additional details, model results, the approach, and caveats are discussed in the final documentation in the results and modeling directories, respectively.

0.1 Questions 1 Rationale

0.1.1 What are some ways that we can uniquely identify individuals?

1. Upon inspection of the schema, the immediate - but perhaps naive strategy - is to use the combination of the name, city, zip code, employer, and occupation (based on the following guidelines for individual reporting). However, I can hypothesize that a political party, PAC, or candidate running for office that uses this data would like a more enriched characterization of an individual. For example, race, ethnicity, and age, are also critical. There is a well-established body of research demonstrating that there this significant associations between individual behaviors and indicators of the things group members have in common support; concluding that the group context of participation influences choices to register and to vote ([5]). Moreover, determining the race and ethnicity of someone based on name alone is fraught with potential biases. There has been previous work that attempts to provide a more rigorous approach ([8]) that has been used in accurately identifying Medicare and Medicaid beneficiaries.

2. In addition, small contributions (defined as < 200) may not be represented in this dataset.

   The type of grouping - if there was not such significant wealth inequality in the United States - may have been difficult to discern based on employer and occupation as well as zip and city. It may be very well that there could be multiple "Joe Smoes" who could all be retired or homemakers - all living in the same geographical area or may have multiple addresses across states. This ambiguity is exacerbated by each state has its own State Board of Elections that maintains varying standards for collecting registrations and voting results. Name standardization and address standardization vary significantly. For example, what I discovered in the North Carolina contributions dataset were names such as: "Steven 'Steve' Abernathy", 'Amanda Abernathy Stokes', 'Amanda J Abernathy Stokes', 'JOHN TALBOT' and 'MICHAEL D TALBOT JR'.

   In the end, it may be helpful to normalize these names and use a combination of first, last, middle, abbreviation, and all of the other fields available using some open-source solution such as Human Parser [3]. Of course, if we had a standard voter identifier that is universal across states and federal agencies we would not have to resort to such tedious additional steps. Moreover, it would be more difficult to accurately identify someone's trajectory without observing the temporal nature of each individual's contribution. One can construct this using two-dimensional panel data with a year-zipcode fixed effect.

3. This brings me to the third point. These records do not include any concept of the temporal nature of filing and re-filing. I have taken for granted that this table is a snapshot of the FEC filings for the 2020 cycle, without observing any non-overlapping tumble periods where

contribution periodicity may occur. I have attempted to reconstruct - in aggregate donations in the North Carolina senate seat - in the accompanying notebook. In doing so, I observe the lack of seasonality, but a clear co-integration in the time series ([2]).

4. Lastly, primarily because I am a ”glass-half-empty” chap, a more extreme example of any grouping or segmentation using individual-level information does not reflect an individual. Consider organizations such as WinRed - which is listed under multiple entities such as Org, and PAC - may masquerade as individuals (in the worst case scenario). Conversely, this could imply that, for example, an individual may repeat contributions under the guise of another individual; or a political PAC or corporate entity may be represented by a large sample of individuals that are focused on influencing a local congressional seat. In that case, the grouping strategy may artificially inflate or deflate the represented contribution amount of an identified individual.

0.1.2 What transformations on the data should we consider making to accurately group and associate transactions to individuals?

The transformations conducted would vary based on the question’s level of study - or resolution. For example, are we considering studying a precinct or individual; household or PAC? Consider the following questions that are natural extensions of the original prompts for Question 2:

- *Were retirees and self-employed individuals more likely to vote for Thom Tillis in the 2020 North Carolina Senate election?* Such a question requires segmentation by geography, employer, and by type of political contribution. Using the North Carolina case study, we can easily identify here that retirees and self-employed overwhelmingly voted for Thom Tillis.

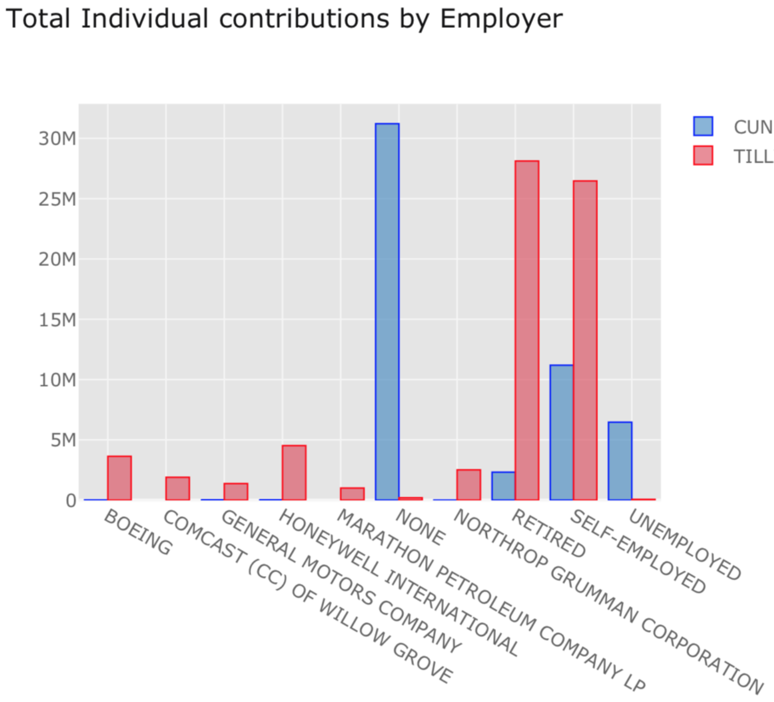

Figure 1

- *What is the frequency of contribution activity to a specific candidate*? This would leverage using window functions in conjunction with grouping to ensure that the proper timeline is captured at the level of the donor. In essence, we could partition by the voter and uniquely ascertain their contribution activity based on their zip code, employer, and occupation. Here, in aggregate - again using the North Carolina senate race and reviewing contributions from Jan 1, 2019, to May 2020 - it's not immediately clear how nuanced such questions can be from an operations standpoint (2).

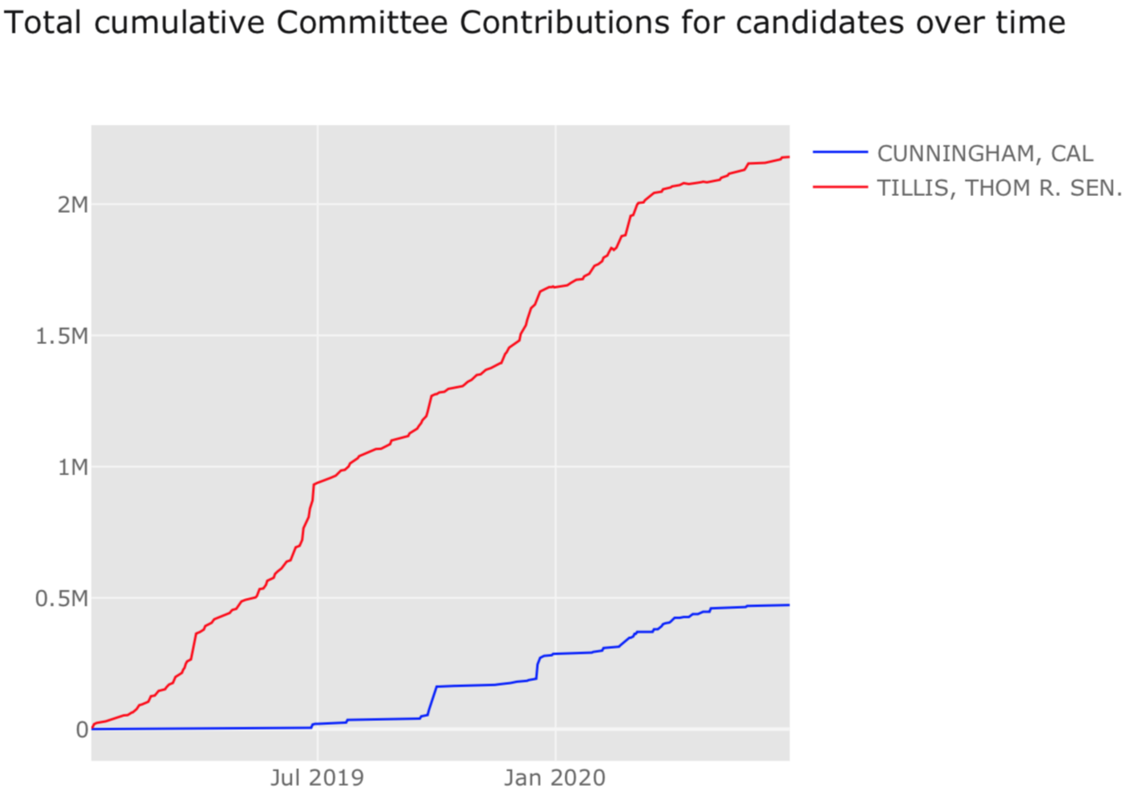

Figure 2

However, when we further refine this and identify distinct changepoints in the donations when reviewing ”for versus against” expenditures we have a more startling (and rather perplexing for me) result:

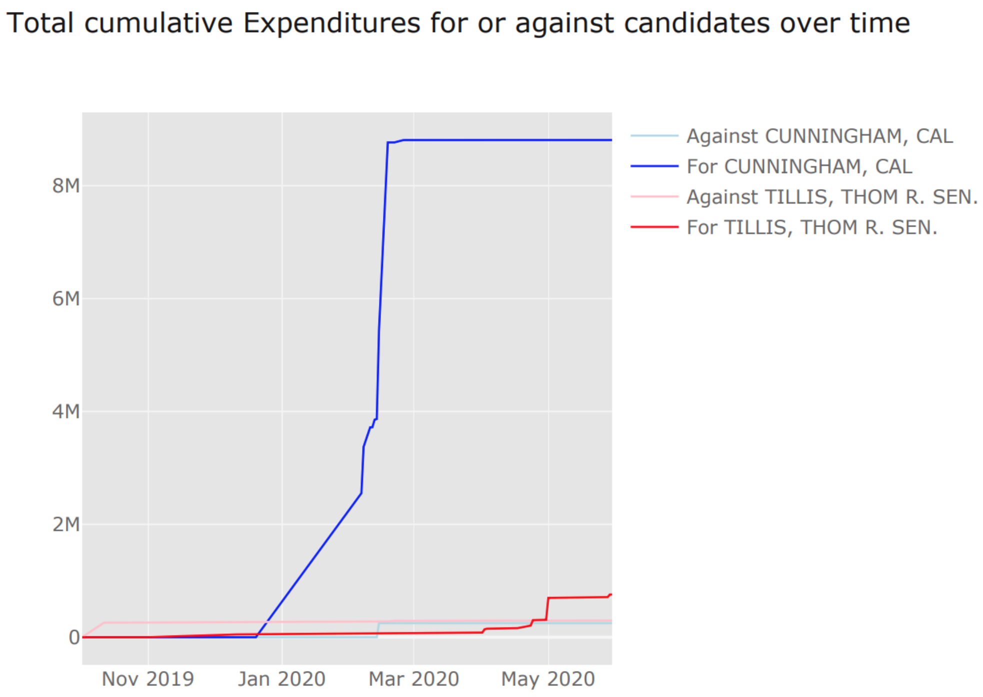

Figure 3

I elaborate on this observation in my response to Question 3.

- We could ask a more elementary question regarding how to best segment the voting bloc or stratify by income. We could ask: *does past donation signal anything about donating in 2020? Second, what does the occupation - if anything - suggest about those that did donate in 2020?*

  Consider the excerpt from the logit model study focused on the state of North Carolina. In isolation, these figures are meaningless. Together they help to better understand how different donation bands are mapped to specific occupations. These occupations in turn are associated with a wealthier or poorer census tract; indicating different priorities and affinity groups.

  For example, does the predominance of real estate occupations (4) suggest anything about local political proclivities towards increasing housing demand, furthering policies that increase population growth, or laws that expand access to businesses?

  Conversely, we have to reconsider minimum contribution amounts in a campaign email if those who donated a total amount of < 1000 did not contribute in 2020 (5). Alternative framing of the business problems is discussed in greater detail in Section 0.4.

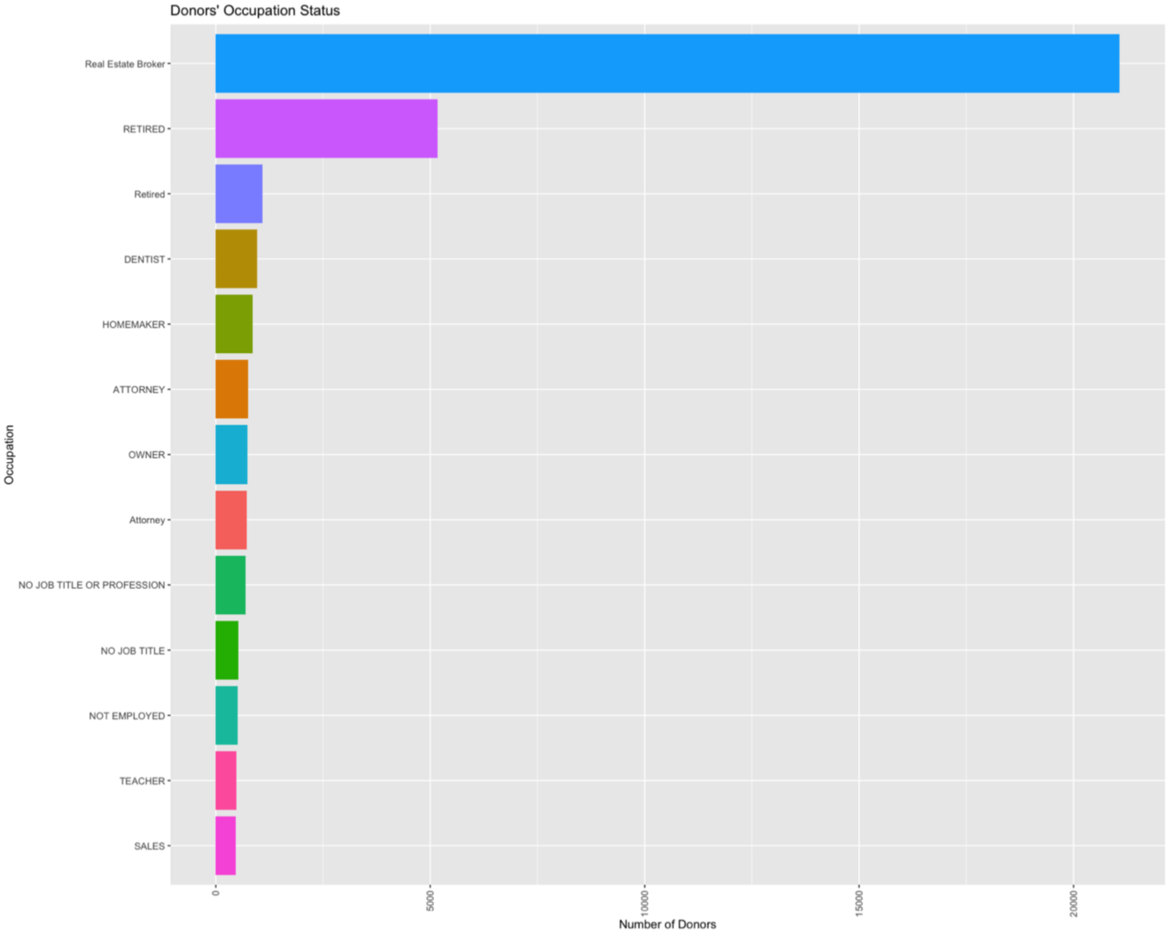

Figure 4: Occupations of registered voters in North Carolina (Source: Experiment 1 Part 1 in the modeling notebooks directory)

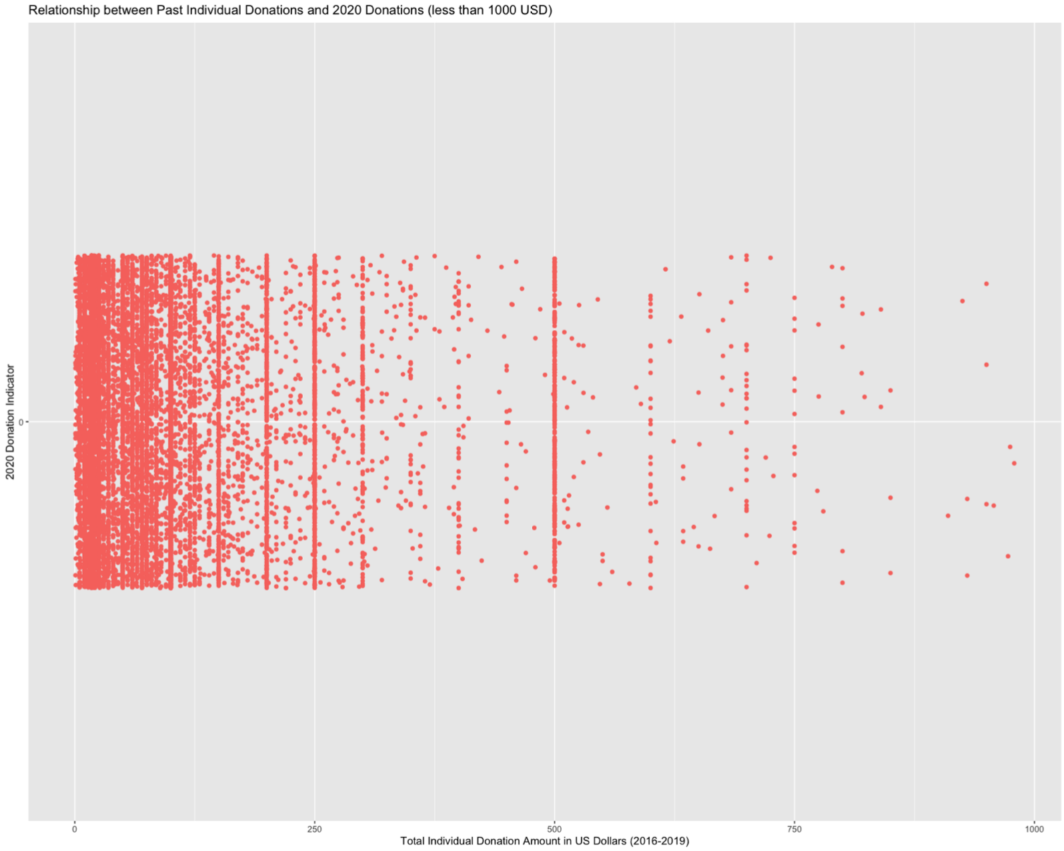

Figure 5: Relationship between past individual donations (> 1000) and not donating in 2020 - North Carolina's registered voters (Source: Experiment 1 Part 1 in the modeling notebooks directory)

- Time segmentation - using partitions in conjunction with the lag operator applied to the date ensures proper granularity. One can artificially ”tag” different partitions to structure the time between contribution events. The level of resolution is at the donor level. I provide such an example here:

```


-- Suppose we have the following scenario:
-- Active voter: a voter is considered active on any day where they have at least one donation
 --in the prior 28 days. For example, if the following records are present in events, the voter
 -- "Bob" is considered active on 2017-04-01 through 2017-04-28 (inclusive) even
 --if no further activity is detected.
-- Churned voter: a voter is considered to be a churned voter during the 28 days following their
--last being considered active.
-- A voter is no longer a churned voter if they become active again.

DROP TABLE `seo-project-349214.mydataset.donor_intervals_enriched`;
CREATE TABLE `seo-project-349214.mydataset.donor_intervals_enriched` AS (
  SELECT name,
    city,
    state,
    zip_code,
    amndt_ind,
    entity_tp,
    transaction_tp,
    employer,
    occupation,
    transaction_amt,
    CAST(transaction_dt AS TIMESTAMP) as timestamped,
    transaction_dt - MAX(CASE WHEN transaction_amt > 0 THEN transaction_dt END) OVER (
              PARTITION BY
              name,
              city,
              state,
              zip_code,
              employer,
              occupation
              ORDER BY transaction_dt,
              transaction_dt ROWS BETWEEN UNBOUNDED PRECEDING AND 1 PRECEDING
              ) AS days_since_last_donation
  FROM `bigquery-public-data.fec.indiv20`
  where  transaction_amt >= 1
      AND amndt_ind IN ("N")
      AND entity_tp = 'IND'
      AND transaction_tp IN ('15E',
        '15'
      )
  AND EXTRACT(YEAR FROM transaction_dt) = 2020 AND employer IS NOT NULL AND state is not NULL
  AND employer not in ('NOT EMPLOYED', 'RETIRED', 'SELF-EMPLOYED')
  ORDER BY name,
              city,
              state,
              zip_code,
              employer,
              occupation, timestamped DESC
);

--Based on the value from the previous step, we can determine which donors belong to
--some time interval and which does not occur in any. The condition is as follows:
--if the amount donated has value at least of one dollar or is no more than 27 days later
--than the previous donation, then it belongs to the same interval.

--Besides, we will also mark the interval starting donations
--which is going to be useful for the period sequence calculation.
--We know which donations starts the interval because the value must be more than
--or equal to 1 donation and the previous qualifying donations must be no more than 27 days ago.
-- here days_since_last_donation is an interval object
DROP Table `seo-project-349214.mydataset.donations_tagged`;
CREATE TABLE `seo-project-349214.mydataset.donations_tagged` AS (
      SELECT `seo-project-349214.mydataset.donor_intervals_enriched`.*
      -- does belong to active some interval flag
      ,CASE WHEN (EXTRACT(DAY FROM days_since_last_donation) <= 28)
            AND transaction_amt >= 1 THEN 1
      END AS active_interval
      -- first start of a churned flag
      ,CASE WHEN (EXTRACT(DAY FROM days_since_last_donation)) > 28
            AND transaction_amt < 1 THEN 1
      END AS churned
      -- churned to active start
      ,CASE WHEN (EXTRACT(DAY FROM days_since_last_donation)) > 28
```

```
            AND transaction_amt >= 1 THEN 1
      END AS start_of_churned_to_active_status
      FROM `seo-project-349214.mydataset.donor_intervals_enriched`
      ORDER BY name, timestamped  DESC
);
-- now we tag out partitions for every donor
DROP TABLE `seo-project-349214.mydataset.donors_tagged`;
CREATE TABLE `seo-project-349214.mydataset.donors_tagged` AS (
  SELECT
     `seo-project-349214.mydataset.donations_tagged`.*
        -- 1. Donations: sequence number of the interval calculated as sum of start_of_churned_to_active_status
      ,SUM(start_of_churned_to_active_status) OVER (
         PARTITION BY
              name,
              city,
              state,
              zip_code,
              employer,
              occupation
              ORDER BY timestamped
              ROWS BETWEEN UNBOUNDED PRECEDING AND CURRENT ROW) AS interval_seq
        -- 2. end of the period computed as qualifying donation criterion plus 27 days
      ,MAX(CASE WHEN transaction_amt >= 1 THEN EXTRACT(DAY FROM timestamped ) + 27 END)
              OVER (
              PARTITION BY
              name,
              city,
              state,
              zip_code,
              employer,
              occupation
              ORDER BY timestamped
              ROWS BETWEEN UNBOUNDED PRECEDING AND CURRENT ROW) AS interval_end
  FROM  `seo-project-349214.mydataset.donations_tagged`
  --leaving out the out-of-interval donations
  --WHERE active_interval = 1
 ORDER BY name, timestamped  DESC
);
/*
The last step is to group the donations by the donor and tagged partition (interval_seq field).
We will extract a start of an interval as the earliest donation date in the interval.
The end of the interval is calculated as the latest date from the column interval_end.
The reason is that the cumulative nature of the window functions has not allowed
us to discover the end of the interval at the first row and we were updating it until the real
end of the interval. The last column tells us the desired sum of all donations over the whole interval before a
concomitant churn or end of the donor's engagement with our organization

This derived end table allows us to create indicator variables based on the number of active intervals
(say they have donated at least 3 times) or only once. Moreover, we can now figure out their total days
before they churned. This enables us to ascertain an endpoint for when they drop off  - hence the
possible use of a time-to-event analysis.

We can also remove those who donated once and process the more "active"
```

```
    contributors within our campaign.

. */
DROP TABLE `seo-project-349214.mydataset.grouped_final`;
CREATE TABLE `seo-project-349214.mydataset.grouped_final` AS (
SELECT
      name,
      city,
      state,
      zip_code,
      employer,
      occupation,
      interval_seq
     ,MIN(timestamped) AS interval_start
     ,MAX(interval_end) AS interval_end
     ,SUM(transaction_amt) AS total_transactions
FROM `seo-project-349214.mydataset.donors_tagged`
GROUP BY
      name,
      city,
      state,
      zip_code,
      employer,
      occupation,
      interval_seq
ORDER BY name, interval_start, interval_end DESC
);
```

Listing 1: Example for Partitioning and checking voter drop-off based on example operational needs

- Is the contributor submitting payments that need amendments or are associated with violating the FEC regulations? This may be appropriate to address certain professions or employers that we want to analyze, that require one to still group by an employer but aggregate the unique fields that comprise an ”individual”? For example, another nuanced take on Question 2 could be the following. Notice here that this is a twist on the original question. Here I return the Top 20 names by state. The level of resolution is reduced compared to the original question’s query and here I don’t break ties between contributors that have the same donation within the state.

```

with temp_aggregation_table as (
  SELECT
    name,
    state,
    amndt_ind,
    entity_tp,
    transaction_tp,
    SUM(transaction_amt) as total_contributions
    FROM
        `bigquery-public-data.fec.indiv20`
 GROUP BY name,
    state,
    amndt_ind,
    entity_tp,
    transaction_tp
  HAVING total_contributions > 0
    AND amndt_ind = "N"
    AND entity_tp = 'IND'
    AND transaction_tp IN ('15E',
      '15')
    AND EXTRACT(ISOYEAR
    FROM
```

```
      MIN(transaction_dt)) = 2020
    AND EXTRACT(ISOYEAR
    FROM
      MAX(transaction_dt)) = 2020

)

SELECT
  *
FROM
  (
    select name, state, total_contributions,
    RANK() OVER (PARTITION BY
   state
 ORDER BY total_contributions DESC) AS contribution_rank,
    from temp_aggregation_table
  ) AS table_top_20
WHERE
  contribution_rank <= 20 AND state is not NULL
order by state, contribution_rank;
```

- Segmentation or grouping aggregations by committee, or the specific type of race (Congressional, Senate, Primary), will yield different compositions of voters. For example, in local elections or for Congressional seats there may be more contributions for individuals as there may be greater group affinity or may ”serve as a positive affirmation of identity group membership or as an expression of group solidarity and support, both of which convey psychological benefits” ([9]).

There are myriad other ways to slice and dice this; I show some of the alternatives in my modeling notebooks.

0.2 Questions 2 Rationale

0.2.1 For the calendar year 2020, what were the names of the top 20 committees receiving contributions, and how much did they receive?

Please see the results in the accompanying directory titled *results*. The query for this starts on line number 180 (assuming you open the queries.sql file in Sublime).

0.2.2 Who were the top 20 individuals making contributions and how much did they contribute?

Please see the results in the accompanying directory: *results*. The query for this starts on line number 63 (assuming you open the *queries.sql* file in Sublime).

0.2.3 How many committees are both recipients and contributors? Do these results align with your expectations?

I computed the number to be 1978. The query for this result starts on line number 220 (assuming you open the *queries.sql* file in Sublime). I had no expectations of what number I would surface - which is the fun part of being in this vocation!

What I do find discomfiting, however, is that this is an aggregate number with no specificity to space, time, campaign, the context of the Federal election, or the other outstanding questions that are beyond individual contributions - between committee expenditures and loans.

0.3 Questions 3 Rationale

0.3.1 What is the calendar effect and role of seasonality (if any) in donation activity? Does the seasonality change at all in off-year election cycles?

As prefaced earlier, regardless of the specific Federal race I noted a lack of seasonality, but did notice a clear co-integration in the time series - particularly as the election day becomes closer. Citing the "for and against" funding for the North Carolina Senate race, I did notice that was a significant surge for Cunningham mid-way through the year. Whether or not this was an artifact of when the funding was reported or when it occurred is not immediately clear.

0.4 What are the benefits of using time-to-event models and why is the GLM with a logit link function not always appropriate?

Although a generalized linear model with a logit link may be suitable it may be more favorable to choose a model that can provide contemporary updates on potential voters' behavior. For example, consider a marketing campaign or fundraising campaign that measures the cycle in terms of one quarter - defined as 90 days.

Based on this, it may not be such a bad thing after all if we just modeled a binary outcome of contributing within the window or not. However, the reality is that someone could submit a contribution after the 90-day window.

In light of the possibility of contributions at any time from the near to very distant future and the benefit of receiving contributions further along in the time horizon, it's important to frame contribution modeling as such. In light of this, consider the following construction:

- As suggested by the phrase "at an instantaneous time" survival analysis models an outcome along a time continuum whereas a logit model is applicable in how likely an *outcome occurs in a certain period of time*, rather than a point in time [7]. This is where the survival model could be useful to a canvasser's or a campaign manager's anticipation of a voter not continuing to provide donations. There could be a direct email, phone call, or "nudge" as the probability of contributing decreases up to a predetermined threshold.

- The previous point yields a subtle - but critical - distinguishing point between a logit model and survival analysis. While survival analysis accounts for censoring, logit models do not. Consider the case of right censoring - instances when a participant (in our case voter or potential campaign donor) is "lost to follow-up" (i.e. dropped out of a study or some other observational setup before the end of the observational period, or did not have the event before the end of the follow-up period) or simply cannot be observed anymore due to death or some other event that precludes the event of interest.

- A crucial ramification of accounting for censoring versus not is that accounting for censoring can increase the p-values of independent variables. This is essential because logit models count a censored event like any other non-outcome.

- Statistical significance aside, another very practical upshot of using survival analysis rather than logistic regression is, when we're modeling an outcome such as not contributing to a committee (similar to the B2B SaaS unsubscribing use case or a Medicaid beneficiary disenrolling), we don't have to pre-define what length of time we want to observe "churn", especially as this length of time could vary greatly at different times. Instead, every single moment in time is considered regarding how likely it is that "churn" (whatever your end-stage) will occur. This has the advantage of preventing many iterations of retraining whenever the typical time for churn changes quite a bit or whenever stakeholders change their mind about what length of time of 'no-show' constitutes no contributions in the election cycle.

0.4.1 Considering alternative modeling approaches

Augmenting the original construction, from above, here are additional suggestions for modeling:

- Regular Cox-PH survival analysis The most intuitive one here is where you're answering the following question: at a given time t, what is the probability that the individual will contribute at time t_1, where t_1 could be any point in time after time t? This framing affords greater flexibility than the logit model as you don't have to pre-define a window of time in which a contribution occurs.

- Because most contributors are likely to be donating before the deadline, you'll probably have a log-normal survival curve.

- A cure model can be used if it's expected that your survival curve will have a long flat tail, i.e. if you have a lot of voters who decide to not contribute due to an economic recession.

- An accelerated failure time (AFT) model is a type of survival analysis model whose outcome variable is simply the time to event. An accelerated failure time is more flexible than a Cox-PH model because it's semi-parametric. You'll probably see a log-normal distribution of time to contribute since there will most likely be an early peak before the election.

- The question of individual contributions in full may not even be the most pertinent. Voters may often submit contributions as fractions of their total contribution at different times. Hence, what you might want to model here is the percentage of their total contribution they will pay for, say, each month after they initially contribute. This is very similar to predicting sales over time. In such cases, a linear mixed-effects model may be more suitable (1.1).

There is ubiquitous work that compares and contrasts these aforementioned methodologies. A fantastic review is cited here ([6]).

References

[1] Ghahramani, Z. : *Introduction to Graphical Models*. August 2007

[2] Granger, C. W.: Time Series Analysis, Cointegration, and Applications. In: *American Economic Review* 94 (2004), June, Nr. 3, 421-425. http://dx.doi.org/10.1257/0002828041464669. – DOI 10.1257/0002828041464669

[3] Gulbranson, D. : *Human Name Parser*. https://github.com/derek73/python-nameparser. Version: 2010

[4] Jones, K. : *Do multilevel models ever give different results?* http://www.bristol.ac.uk/cmm/learning/multilevel-models/. Version: February 2019

[5] Logan, J. R. ; Darrah, J. ; Oh, S. : The Impact of Race and Ethnicity, Immigration and Political Context on Participation in American Electoral Politics. In: *Social Forces* 90 (2012), 03, Nr. 3, 993-1022. http://dx.doi.org/10.1093/sf/sor024. – DOI 10.1093/sf/sor024. – ISSN 0037–7732

[6] Ngwa, J. S. ; Cabral, H. J. ; Cheng, D. M. ; Pencina, M. J. ; Gagnon, D. R. ; LaValley, M. P. ; Cupples, L. A.: A comparison of time dependent Cox regression, pooled logistic regression and cross sectional pooling with simulations and an application to the Framingham Heart Study. In: *BMC Med. Res. Methodol.* 16 (2016), Dez., Nr. 1

[7] Rodríguez, G. : *Lecture Notes on Generalized Linear Models: Survival Models*. 2007

[8] Suriyan Laohaprapanon, G. S. ; Naji, B. : *ethnicolr*. https://github.com/appeler/ethnicolr. Version: 2017

[9] VALENZUELA, A. A. ; MICHELSON, M. R.: Turnout, Status, and Identity: Mobilizing Latinos to Vote with Group Appeals. In: *American Political Science Review* 110 (2016), Nr. 4, S. 615–630. http://dx.doi.org/10.1017/S000305541600040X. – DOI 10.1017/S000305541600040X

1 Appendix

1.1 Elaborating on using Linear Mixed-Effects when applied to voter contributions

Assume that we could select any arbitrary county or precinct (a more refined spatial aerial unit may be necessary). The initial setup is to assume that we can take any given citizen's average contribution and model a contribution using the following normal linear model.

$$y_d \sim N(\mu_d, \sigma^2), \quad \mu_d = \beta_0 + \beta_1 x_d, \quad \text{for } d \in 1 \ldots n,$$

where y_d represents the voter's contribution on their dth observation (d represents days), and $x_d \in \{0, 2, \ldots n = 9\}$ indicates the day when this observation happened.

Using 1) $\beta = [\beta_0, \beta_1]$ and 2) "plate" notation that symbolizes repeating nodes within a bounding plate according to an index ([1]) - which in this case is $d \in 1 \ldots n$ - we can provide a similar model for each voter in the experiment - indexed by $m \in 1 \ldots M$, this would lead to M independent normal linear models.

Moreover, suppose we denote the average contribution on observation d for voter m by y_{md}, this set of models is as follows.

$$\begin{aligned} y_{md} &\sim N(\mu_{md}, \sigma_m^2), \\ \mu_{md} &= \beta_{m0} + \beta_{m1} x_{md}, \quad \text{for } m \in 1 \ldots M, \text{ for } d \in 1 \ldots n_m. \end{aligned}$$

Assume there is a shared residual standard deviation term σ. This replaces the individual standard deviations for each of the M voters. Notice that this model is identical to a varying intercept and varying slope linear model - originally discussed in our course.

Elaborating on this model we have:

$$\begin{aligned} y_{md} &\sim N(\mu_{md}, \sigma^2), \\ \mu_{md} &= \beta_{m0} + \beta_{m1} x_{md}, \quad \text{for } m \in 1 \ldots M, \text{ for } d \in 1 \ldots n_m, \end{aligned}$$

Notice that in the aforementioned construction there is no indication of levels. To introduce the concept of levels let us now consider the multivariate representation ([4]): $\beta_m = [\beta_{m0}, \beta_{m1}]$ is drawn from a multivariate Normal distribution with mean vector β and covariance matrix Σ. This model can be written as follows:

$$\begin{aligned} y_{md} &\sim N(\mu_{md}, \sigma), \\ \mu_{md} &= \beta_{m0} + \beta_{m1} x_{md}, \quad \text{for } m \in 1 \ldots M, \text{ for } d \in 1 \ldots n_m, \\ \vec{\beta}_m &\sim N(\beta, \Sigma) \quad \text{for } m \in 1 \ldots M \end{aligned}$$

The Bayesian plate diagram for this model is shown below in Figure (6). It should be noted that this notation is more ubiquitous in the topic modeling and "unsupervised" learning literature. However, the Bayesian setup is still applicable, and, I feel more conceptually accessible when communicating with a non-domain expert.

Here each β_m are modeled as functions of β and Σ.

Notice here that we can rewrite the multilevel model as:

$$\begin{aligned} \text{for } i \in 1 \ldots n, \quad y_i &\sim N(\mu_i, \sigma^2), \\ \mu_i &= \beta_{[s_i]0} + \beta_{[s_i]1} x_i, \\ \text{for } m \in 1 \ldots M, \quad \vec{\beta}_m &\sim N((\beta, \Sigma). \end{aligned}$$

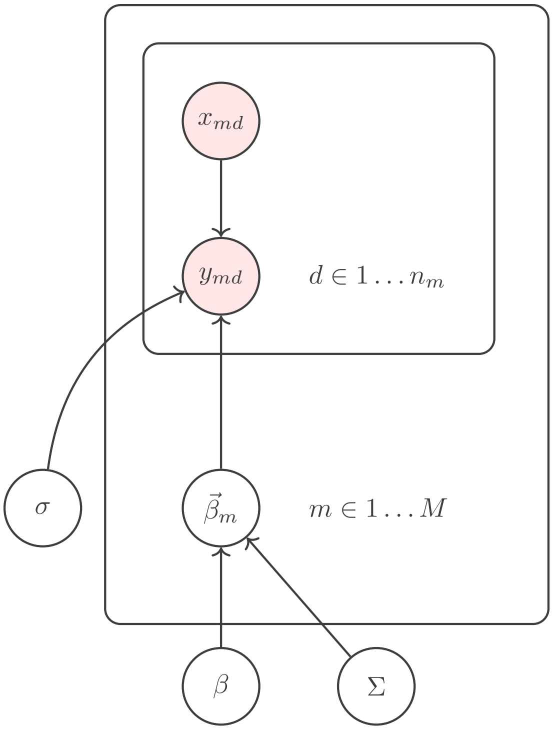

Figure 6: Plate diagram showing varying slopes and intercepts in a linear model

Here i indexes all voters i.e. $i \in 1, 2 \ldots n$, and each $s_i \in 1, 2 \ldots M$ is an indicator variable representing the voter's observed characteristics and response, respectively across $\forall i$.

Given this construction we can substitute $\vec{\beta}_m \sim N(\beta, \Sigma)$, and write $\vec{\beta}_m$ as $\vec{\beta}_m = (\beta + \vec{\delta}_m)$ where $\vec{\delta}_m \sim N(0, \Sigma)$.

Substituting $\beta + \delta_m$ for $\vec{\beta}$, and thus substituting $\beta_0 + \delta_{m0}$ and $\beta_1 + \delta_{m1}$ for β_{m0} and β_{m1}, respectively, we have the following model.

$$\begin{aligned}
\text{for } i \in 1 \ldots n, \quad y_i &\sim N(\mu_i, \sigma^2), \\
\mu_i &= \underbrace{\beta_0 + \beta_1 x_i}_{\text{fixed effects}} + \underbrace{\delta_{[s_i]0} + \delta_{[s_i]1} x_i}_{\text{random effects}}, \\
\text{for } m \in 1 \ldots M, \quad \vec{\delta}_m &\sim N(0, \Sigma).
\end{aligned}$$

As we can see from this, a multilevel normal linear model is equivalent to a non-multilevel model (the *fixed effects* models) plus a normally distributed random variation to the intercept and slope for each voter (the *random effects*). The fixed effects apply to all observations. For example, we can "fix" the county, time of the year, occupation, and employer. Hence, fixed effects yield the average effects in the *population*.

The random effects, on the other hand, vary across each different value of the grouping variable, which in this example is an individual voter or campaign contributor in the natural experiment. Therefore, individual variation around this average is given by the random effects. Together fixed and random effects are synergistic in that they capture heterogeneity in the population of the study.