**Data Audit Report**

Prepared by

Jason Noble

In support of

Predictive model to identify the more likely donors for the AnyState VFW

Requested by

AnyState Veterans of Foreign Wars

September 11, 2024

**Introduction**

The analytics team has been asked by the AnyState VP of Marketing to build a predictive model and score their marketing database. The goal is to spend resources more wisely by targeting those most likely to donate and improve the effectiveness of their solicitation campaigns,

It is the analytics team’s understanding that the target variable segments for this model are:

1. Yes customer has a 1 response to the 97NK.
2. No customer has a 0 response to the 97NK.

The supplied datafiles are intended to support the development of a model differentiating the donation characteristics between these two segments. The estimated model will be applied to another, much larger bank customer file, supplied later to the analytics team for scoring. The final delivery will be this scored “marketing” file.

The purpose of this data audit is to ensure the following:

1. The analytics team has received all datafiles intended for this project.
2. The analytics team understands the content, layout, and format of these files.
3. The data in these files are of sufficient integrity and quantity to support the model development.

This data audit consists of 4 sections:

1. Datafile Summary: A list and description of all datafiles received.
2. Datafile Detail: For each datafile, tables showing all data fields received, their values, summary statistics, and distributions. Data fields are categorized into one of 4 types of analytical variables:
   * Categorical - data fields with distinct levels or values which represent categories; can be a number or a label, nominal or ordinal.
   * Date - data fields that are identified as calendar dates.
   * Numeric - data fields that are continuous numeric data.
   * Character - data fields whose values are characters and are not otherwise classified as categorical.
3. Modeling Sample – After merging all supplied datafiles, a determination is made as to whether there is adequate sample size for each target sample to support model development.
4. Questions – The auditing process will uncover data integrity issues. This section lists what the analytics team has found in this regard. This section also poses specific questions on data field definitions, field coding, and interpretation, answers to which will facilitate the team’s model development effort.

**Datafile Summary**

The analytics team has received 4 datafiles from AnyState Veterans of Foreign Wars as listed in Table 1.

Table 1. Datafiles Received

|  |  |  |  |
| --- | --- | --- | --- |
| Filename | File Type | # Of Records | File Contents |
| Donor\_census2 | SAS | 19,372 | Income info, Vietnam Vet, Control # |
| Donor\_survey2 | csv | 58,116 | Control #, Survey Q/A vars |
| Donor\_profile2 | csv | 19,372 | Control#, Age, Gender, relational vars |
| Donor\_history2 | csv | 19,372 | Data on promotions and gift, tracking vars |

**Datafile Detail**

Each datafile contains the analytic data fields as shown in the following tables. Note that the data fields have been classified based on their potential analytical usage.

|  |
| --- |
| ***Datafile #1:*** Donor\_census2  ***File Analytic Contents:***  Numeric Fields (6): Wealth\_Rating Median\_Home\_Value Median\_Household\_Income  Per\_Capita\_Income PCT\_Owner\_Occupied PCT\_Vietnam\_Veteran  Categorical Fields (0):  Character Fields (1): Control\_Number  Date Fields (0):  ***Records:*** 19,372  ***Columns:*** 7 |
| ***Notes:*** The data field Control\_Number appears to be a row ID. PCT\_Vietnam\_Vets could be categorical, I do not understand the values for the Vietnam\_Vet variable. |

Table 1. Donor\_census2 - Numeric Data

|  |  |
| --- | --- |
| *Analytic numeric data fields* |  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |

Table 2. Donor\_census2- Character Data

|  |  |
| --- | --- |
| *Analytic character data fields* |  |
|  |  |
|  |

|  |
| --- |
| ***Datafile #2:*** Donor\_survey2  ***File Analytic Contents:***  Numeric Fields (0):  Categorical Fields (1): Survey\_Value  Character Fields (2): Control\_Number Survey\_questions  Date Fields (0):  ***Records:*** 58,116 (3x19,372)  ***Columns:*** 3 |
| ***Notes:*** The data field Control\_Number appears to be a row ID. |

Table 1. Donor\_survey2 - Categorical Data

|  |  |
| --- | --- |
| *Analytic categorical data fields* |  |
|  |  |
|  |

Table 2. Donor\_survey2 - Character Data

|  |  |
| --- | --- |
| *Analytic character data fields* |  |
|  |  |
|  |
|  |  |
|  |

|  |
| --- |
| ***Datafile #3:*** Donor\_profile2  ***File Analytic Contents:***  Numeric Fields (2): Donar\_Age Recent\_Star\_Status  Categorical Fields (5): Income\_Group Donor\_Gender Home\_Owner SES URBANICITY  Character Fields (2): Control\_Number Cluster\_Code  Date Fields (0):  ***Records:*** 19,372  ***Columns:*** 9 |
| ***Notes:*** The data field Control\_Number appears to be a row ID. Cluster\_Code could be Categorical if data is defined different than initial assessment. |

Table 1. Donor\_profile2 - Numeric Data

|  |  |
| --- | --- |
| *Analytic numeric data fields* |  |
|  |  |
|  |
|  |  |
|  |

Table 2. Donor\_profile2 - Categorical Data

|  |  |
| --- | --- |
| *Analytic categorical data fields* |  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |

Table 3. Donor\_profile2 - Character Data

|  |  |
| --- | --- |
| *Analytic character data fields* |  |
|  |  |
|  |
|  |  |
|  |

|  |
| --- |
| ***Datafile #4:*** Donor\_history2  ***File Analytic Contents:***  Numeric Fields (16): Frequency\_status\_97nk Last\_gift\_amt Lifetime\_avg\_gift\_amt  Lifetime\_card\_prom Lifetime\_gift\_amount Lifetime\_gift\_count  Lifetime\_gift\_range Lifetime\_prom Months\_since\_first\_gift Months\_since\_last\_gift Months\_since\_last\_prom\_resp Months\_since\_origin Pep\_star Recent\_avg\_gift\_amt Recent\_card\_response\_count Recent\_response\_count  Categorical Fields (3): In\_House Target\_B Regency\_Status\_96NK  Character Fields (1): Control\_Number  Date Fields (0):  ***Records:*** 19,372  ***Columns:*** 20 |
| ***Notes:*** The data field Control\_Number appears to be a row ID. |

Table 1. Donor\_history2 - Numeric Data

|  |  |
| --- | --- |
| *Analytic numeric data fields* |  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |

Table 2. Donor\_history2 - Categorical Data

|  |  |
| --- | --- |
| *Analytic categorical data fields* |  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |

Table 3. Donor\_history2 - Character Data

|  |  |
| --- | --- |
| *Analytic character data fields* |  |
|  |  |
|  |

**Modeling Sample**

|  |  |
| --- | --- |
| **Segment** | **Count** |
| **Target\_B (1) sample** | 4,853 |
| (could be yes, need confirmation) |  |
| **Available non-event (0) sample** | 14,529 |
| (could be no, need confirmation) |  |
| **Total (target) sample** | 19, 372 |
|  |  |
| **Total records in datafile** | 19, 372 |

Sample sizes in the target segments appear to be of sufficient size to support the predictive model as both target segments have over 1,000 observations.

**Questions**

1. Does the above information appear to be correct? Specifically:
   * Does the analytics team have all the data that was meant to be sent?
   * Is the team interpreting the data correctly?
   * Do the data appear to have reasonable values?
2. Here is a list of the data integrity issues the analytics team uncovered. Please review:

* WEALTH\_RATING (8810 missing values)
* MEDIAN\_HOME\_VALUE (218 valued at 0)
* MEDIAN\_HOUSEHOLD\_INCOME (174 valued at 0)
  + - * + PCT\_OWNER\_OCCUPIED(218 valued at 0, does that mean renter?)
        + CLUSTER\_CODE (454 missing values)
        + DONOR\_AGE (4795 missing, there are many donors from 0 to 18, are minors donating?)
        + INCOME\_GROUP (4392 missing values)
        + SES (454 missing values, linked to Cluster\_Code)
        + URBANICITY (454 missing, linked to SES?)
        + MONTHS\_SINCE\_LAST\_PROM\_RESP (246 missing, also has negative values)

1. The following are specific questions the analytics team has about the data. Please review:

Confirm these fields are in dollar amounts:

* LAST\_GIFT\_AMT
* LIFETIME\_GIFT\_RANGE
  + - * + LIFETIME\_GIFT\_AMOUNT
        + LIFETIME\_AVG\_GIFT\_AMT
        + LIFETIME\_MIN\_GIFT\_AMT
        + LIFETIME\_MAX\_GIFT\_AMT
        + RECENT\_AVG\_GIFT\_AMT

Confirm MEDIAN\_HOME\_VALUE are displayed in 1,000's of dollars.

Confirm CONTROL\_NUMBER is just a character identification, and does not have any computation meaning.

Define the relationship between the WEALTH\_RATING scores?

Define PCT\_OWNER\_OCCUPIED definition and check low values, including 0.

Define PCT\_VIETNAM\_VETERANS, it seems it should be a binary choice.

PER\_CAPITA\_INCOME and MEDIAN\_HOUSEHOLD\_INCOME have a difference between their 0 values (173:174).

Clarify Survey\_Question: CAUSES\_DONATED\_TO\_LAST\_YEAR, is it strictly numerical?

What does Socio-Economic CLUSTER\_CODE define? should it be categorical or character?

DONOR\_AGE has values from 0-18: It does not make sense that babies and minors are donating.

Need to know the connection between INCOME\_GROUP, MEDIAN\_INCOME, PER\_CAPITA\_INCOME, and WEALTH\_RATING.

Define DONOR\_GENDER values and how answers A and U should be categorized.

Define HOME\_OWNER, the category labels.

Define how SES relates to CLUSTER\_CODE and URBANICITY.

Confirm LAST\_GIFT\_AMT does have donations with decimals.

MONTHS\_SINCE\_LAST\_PROM\_RESP has negative (-) numbers, that does not make sense.

Confirm TARGET\_B values are 0=No, and 1=Yes, that customer donated for that campaign

How does TARGET\_B relate to FREQUENCY\_STATUS\_97NK?

How does MONTHS\_SINCE\_ORIGIN have a smaller range than MONTHS\_SINCE\_FIRST\_GIFT?