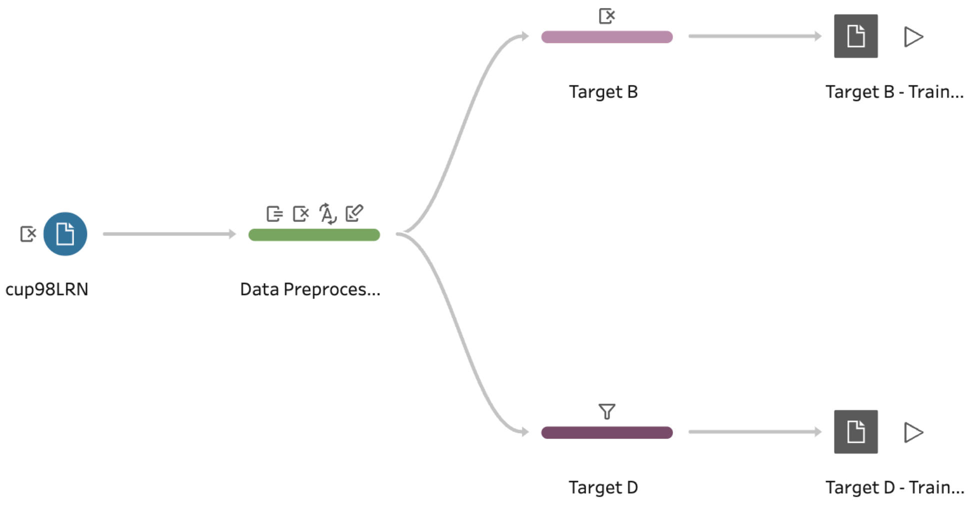
Business Case.

Paralyzed Veterans of America (PVA) is a nonprofit organization which aims to help injured veterans. They send out mailings asking for donations to their previous donors. It would be helpful to them to have a priority mailing list so that they could reach out to the right donors which are most likely to donate the most. Hence, the problem is to find the right people to target from their existing donor lists.

Data Preprocessing.

The learning dataset cup98LRN.txt contains 95,412 records and 481 fields, which was loaded into Tableau Prep Builder. There are 2 targets in this dataset that we are trying to train the models on. The first step is to clean the dataset by preprocessing it. Following are the steps taken for the final file creation for each of the targets.

- Data Preprocessing (green) - is the process where we change the variable type and remove features for DataRobot to read the data properly.
- Target B is about whether the donor donated or not and Target D is about how much the donor donated.
    - For Target B - we removed Target D as it was a target leak for B.
    - For Target D - we kept only the donors that donated for the campaign by filtering them from Target B. It was done because the percent of donors is 5% amongst all the donors that were mailed for the campaign.



Processed Data:

ODATE was converted to TENURE by using 97 - INT (LEFT(STR([ODATEDW]),2). It was the date when the donor was first added to the database, which means we could calculate the duration of the relationship between the donor and the organization.

Following features were converted from numerical to categorical features and _cat was added to them to avoid any misinterpretation. Also, the older features after converting to newer ones were removed from the dataset before model building. For instance, TCODE was dropped and TCODE_CAT remains in the dataset.

Feature Name	Code to convert	New Name
TCODE	STR ([TCODE]) + '_cat'	TCODE_CAT
ZIP	STR([ZIP]) + '_cat'	ZIP_CAT
NOEXCH	STR([NOEXCH]) + '_cat'	NOEXCH_CAT
CLUSTER	str([CLUSTER]) + '_cat'	CLUSTER_CAT
NUMCHLD	STR[NUMCHLD] + '_cat'	NUMCHLD_CAT
INCOME	STR([INCOME]) + '_cat'	INCOME_CAT
WEALTH1	STR([WEALTH1]) + '_cat'	WEALTH1_CAT
DATASRCE	str([DATASRCE])+'_cat'	DATASRCE_CAT
SOLP3	str([SOLP3])+'_cat'	SOLP3_CAT
SOLIH	str([SOLIH])+'_cat'	SOLIH_CAT
WEALTH2	str([WEALTH2])+'_cat'	WEALTH2_CAT
GEOCODE	str([GEOCODE])+'_cat'	GEOCODE_CAT
LIFESRC	str([LIFESRC])+'_cat'	LIFESRC_CAT
MSA	str([MSA])+'_cat'	MSA_CAT
ADI	str([ADI])+'_cat'	ADI_CAT
DMA	str([DMA]) + '_cat'	DMA_CAT
TARGET_B	STR([TARGET_B])+'_cat'	TARGET_B_CAT
HPHONE_D	str([HPHONE_D]) + '_cat'	HPHONE_D_CAT
RFA_2F	STR([RFA_2F]) + '_cat'	RFA_2F_CAT
CLUSTER2	str([CLUSTER2]) + '_cat'	CLUSTER2_CAT

Following dates were in the form YYMM, which was converted to YY and the MM part was dropped as overall it doesn't have a huge significance on the outcome because the campaigns are widespread across so many years. Moreover, previous features were removed after changing them to years. For instance, MAXADATE was removed and MAXA_YEAR is kept in the dataset.

Feature Name	Code to convert	New Name
MAXADATE	'19' + LEFT(STR([MAXADATE]),2)	MAXA_YEAR
MINRDATE	'19' + LEFT(STR([MINRDATE]),2)	MINR_YEAR
MAXRDATE	'19' + LEFT(STR([MAXRDATE]),2)	MAXR_YEAR
LASTDATE	'19' + LEFT(STR([LASTDATE]),2)	LASTYEAR
FIRSTDATE	'19'+ LEFT(STR([FISTDATE]),2)	FIRSTYEAR
NEXTDATE	'19'+ LEFT(STR([NEXTDATE]),2)	NEXTYEAR

DOMAIN was split into 2 different features as it had data which was consolidated in the same feature coded as shown below:

1st byte = Urbanicity level of the donor's neighborhood
U=Urban
C=City
S=Suburban
T=Town
R=Rural

2nd byte = Socio-Economic status of the neighborhood
1 = Highest SES
2 = Average SES
3 = Lowest SES (except for Urban communities, where
    1 = Highest SES, 2 = Above average SES,
    3 = Below average SES, 4 = Lowest SES.)

Thus, it was split into:

- URBANICITY_LEVEL using LEFT ([DOMAIN],1)
- SOCIOECONOMIC_STATUS using RIGHT ([DOMAIN],1) + '_cat'

The following features were removed:

- **CONTROLN**
    - It was removed as it is a unique identifier and hence, has no value.
- **DOB**
    - It was removed as AGE is already present as a feature in the dataset. Generally it is preferred to keep the DOB in the database but for model building it won't be useful in our case.
- **MDMAUD (The Major Donor Matrix code)**
    - The codes describe frequency and amount of giving for donors who have given a $100+ gift at any time in their giving history.
    - The individual bytes could separately be used as fields and refer to the following:
        - First byte: Recency of Giving
            - C=Current Donor
            - L=Lapsed Donor
            - I=Inactive Donor
            - D=Dormant Donor
        - 2nd byte: Frequency of Giving
            - 1=One gift in the period of recency
            - 2=Two-Four gifts in the period of recency
            - 5=Five+ gifts in the period of recency
        - 3rd byte: Amount of Giving
            - L=Less than $100(Low Dollar)
            - C=$100-499(Core)

            - M=$500-999(Major)
            - T=$1,000+(Top)
        - 4th byte: Blank/meaningless/filler
    - 'X' indicates that the donor is not a major donor.
    - Also, this feature is already extracted in the dataset as MDMAUD_R, MDMAUD_F, and MDMAUD_A which makes keeping this column redundant.
    - Hence, we have removed this feature.
- **RFA_2**
    - RFA_2 is split into RFA_2R, RFA_2F, and RFA_2A in the dataset already. It is redundant and RFA_2 is consolidated with the columns together hence RFA_2 has been removed.
- **RFA_3 to RFA_24**
    - RFA (recency/frequency/amount) of past 23 campaigns are present in the dataset, but considering that RFA_2 is the most recent and the most meaningful one for the current scenario, RFA_3 to RFA_24 have been dropped from the dataset.
- **ADATE_2 to ADATE_24 and RDATE_3 to RDATE_24**
    - ADATE had the dates of when the promotions were mailed to the donors, whereas, RDATE has the dates of when the gift was received from the donors.
    - Using these dates, response time of the donor can be calculated and observed.
    - Considering the time limitations, it isn't possible to preprocess all these dates. Hence, these features have been excluded from the dataset.

For the Targets:

- **Target B**
    - For this cleaning step, only Target D is removed as it is a target leak for Target B.
    - This dataset is further saved into .csv format to be used as the training dataset for Target B.
- **Target D**
    - For this cleaning step, Target B was filtered and only the donors that donated were kept in the dataset.
    - Doing this was necessary otherwise the data would be extremely imbalanced because 95% haven't donated.
    - Hence, for Target D (amount of donation) only filtered values of Target B was kept and saved into a .csv file to be used as the training dataset for Target D.