Column Name	Description	Comment
Odate	Date of donor's first gift to PVA	97 - First two digits of ODATEDW, to get the years since first gift
Osource	Code indicating which mailing list the donor was originally acquired from	Excluding this data as it won't impact any of the target variables.
State	State abbreviation	
Zip Code	Zipcode	Excluded as this feature won't give us any major improvement in modelling
Target_B	determines whether given candidate is donor or not	**Target Leak for predicting Target_D.** Excluded it while predicting the regression model, as it is highly co-related.
Target_D	determines the amount paid by the predicted donor	**Target Leak for predicting Target_B.** Excluded it while computing the classification model
MAILCODE	Mail Code	Excluded rows with B value as mail code as the mailer would never reach to the customer and thus no impact on the target variables. After removing the above mentioned rows this column has only one unique value so excluding the column.
PVASTATE	Indicates whether the donor lives in a state served by the organization's EPVA chapter	Created separate column for each unique value in this column and then recoded the columns
DOB	Date of birth	Excluded as the correctly calculated age is already present as a separate column
NOEXCH	Do Not Exchange Flag	Converted blanks to 1 and X to 0 and then Recoded the column as categorical
RECINHSE	In House File Flag	Converted X to 1 and blanks to 0 and then Recoded the column as categorical

RECP3	P3 File Flag	Converted X to 1 and blanks to 0 and then Recoded the column as categorical
RECPGVG	Planned Giving File Flag	Converted X to 1 and blanks to 0 and then Recoded the column as categorical
RECSWEEP	Sweepstakes file flag	Converted X to 1 and blanks to 0 and then Recoded the column as

		categorical
MDMAUD	The Major Donor Matrix code	Excluded as we already have three separate columns which contains the same value of RFA which are present in this column as a complex type
CLUSTER	Code indicating which cluster group the donor falls into	Replaced missing values with 0 and then Recoded this column as categoric
Age	Overlay age	Mean imputed
Ageflag	Ageflag	Excluded as this won't have any impact on the target variables
HomeOwner	HomeOwner flag	Excluded as too many blank values
Child03	Presence of Children age 0-3	Replace blank with N and recoded as categoric
CHILD07	Presence of Children age 4-7	Replace blank with N and recoded as categoric
CHILD12	Presence of Children age 8-12	Replace blank with N and recoded as categoric
CHILD18	Presence of Children age 13-18	Replace blank with N and recoded as categoric
NUMCHLD	NUMBER OF CHILDREN	Mean imputed with average value
INCOME	HOUSEHOLD INCOME	Mean imputed with average value

GENDER	GENDER	Created new columns for each unique value M, F, J, U. For A and C there were only four records. Out of those 4, two were having Tcode 1, so gender would be M, other two were not that straight forward so used U. In addition to this, recoded each column as categoric.
Wealth1	Wealth Rating	Imputed the missing numerical value with the mean value and then recoded this column as categoric as each number corresponds to a separate segment.
HIT	Indicates total number of known times the donor has responded to a mail order offer other than PVA's.	

MBCRAFT	Buy Craft Hobby	Mean imputed with average value
MBGARDEN	Buy Gardening	Mean imputed with average value
MBBOOKS	Buy Books	Mean imputed with average value
MBCOLECT	Buy Collectables	Mean imputed with average value
MAGFAML	Buy General Family Mags	Mean imputed with average value
MAGFEM	Buy Female Mags	Mean imputed with average value
MAGMALE	Buy Sports Mags	Mean imputed with average value
PUBGARDN	Gardening Pubs	Mean imputed with average value
PUBCULIN	Culinary Pubs	Mean imputed with average value
PUBHLTH	Health Pubs	Mean imputed with average value
PUBDOITY	Do It Yourself Pubs	Mean imputed with average value

PUBNEWFN	News / Finance Pubs	Mean imputed with average value
PUBPHOTO	Photography Pubs	Mean imputed with average value
PUBOPP	Opportunity Seekers Pubs	Mean imputed with average value
DATASRCE	Indicates which third-party data source the donor matched against	Excluded as it doesn't directly impact the target variables
MALEMILI	% Males active in the Military	
MALEVET	% Males Veterans	
VIETVETS	% Vietnam Vets	
WWIIVETS	% WWII Vets	
LOCALGOV	% Employed by Local Gov	
STATEGOV	% Employed by State Gov	
FEDGOV	% Employed by Fed Gov	
SOLP3	SOLICIT LIMITATION CODE P3	Excluded the value
SOLIH	SOLICITATION LIMIT CODE IN HOUSE	Excluded the value
MAJOR	Major ($$) Donor Flag	Converted X to 1 and blanks to 0 and then Recoded the column as categorical

Wealth2	Wealth Rating	Imputed the missing numerical value with the mean value and then recoded this column as categoric as each number corresponds to a separate segment.
GEOCODE	Geo Cluster Code indicating the level geography at which a record matches the census data.	Replaced blank values with 0 and then Recoded this column as categoric
COLLECT1	COLLECTABLE (Y/N)	Recoded this column as categorical
VETERANS	VETERANS (Y/N)	Recoded this column as categoric
BIBLE	BIBLE READING (Y/N)	Recoded this column as categoric
CATLG	SHOP BY CATALOG (Y/N)	Recoded this column as categoric
HOMEE	WORK FROM HOME (Y/N)	Recoded this column as categoric
PETS	HOUSEHOLD PETS (Y/N)	Recoded this column as categoric
CDPLAY	CD PLAYER OWNERS (Y/N)	Recoded this column as categoric

STEREO	STEREO/RECORDS/TAPES/CD (Y/N)	Recoded this column as categoric
PCOWNERS	HOME PC OWNERS/USERS	Recoded this column as categoric
PHOTO	PHOTOGRAPHY (Y/N)	Recoded this column as categoric
CRAFTS	CRAFTS (Y/N)	Recoded this column as categoric
FISHER	FISHING (Y/N)	Recoded this column as categoric
GARDENIN	GARDENING (Y/N)	Recoded this column as categoric
BOATS	POWER BOATING (Y/N)	Recoded this column as categoric
WALKER	WALK FOR HEALTH (Y/N)	Recoded this column as categoric
KIDSTUFF	BUYS CHILDREN'S PRODUCTS (Y/N)	Recoded this column as categoric
CARDS	STATIONARY/CARDS BUYER (Y/N)	Recoded this column as categoric
PLATES	PLATE COLLECTOR (Y/N)	Recoded this column as categoric
LIFESRC	Indicates source of the lifestyle variables	Excluded this column as it doesn't directly impact the target variables
PEPSTRFL	Indicates PEP Star RFA Status	Converted X to 1 and blanks to 0 and then Recoded the column as categorical
Variables reflecting characteristics of the donors neighborhood		
ADATE_2	Date the 97NK promotion was mailed	Calculated the time in months from when the promotion started and this data set was generated. Mean imputed the data

ADATE_3 - ADATE_24	Date of all the other promotions	Calculated the time in years since the campaign was sent and this data was generated. Mean imputed the data
RFA_2	Donor's RFA status as of 97NK promotion date	Created separate column for each unique value in this column and then recoded the columns. In addition to this, excluded RFA_2_R as it has only one unique value.
RFA_3 - RFA_24	Donor's RFA status as of all other promotion date	Created separate column for each unique value in this column and then recoded the columns

CARDPROM	Lifetime number of card promotions received to date	
MAXADATE	Date of the most recent promotion received	Calculated the time in years since the most recent promotion and the date this data was generated.
NUMPROM	Lifetime number of promotions received to date	
CARDPM12	Number of card promotions received in the last 12 months	
NUMPRM12	Number of promotions received in the last 12 months	
RDATE_3 - RDATE_24	Date the gift was received for various campaign	Calculated the time in years between the particular campaign and the date this data was generated. Mean imputed the data
RAMNT_3 - RAMNT_24	Dollar amount of the gift for various campaign	Mean imputed with average value
MAXRAMNT	Dollar amount of largest gift to date	
LASTGIFT	Dollar amount of most recent gift	
LASTDATE	Date associated with the most recent gift	Calculated the number of years between most recent gift and the date this data was generated.
FISTDATE	Date of first gift	Calculated the number of years between first gift and the date this data was generated.
NEXTDATE	Date of second gift	Calculated the number of months between second gift and the date this data was generated. In addition to this, mean imputed the missing values to improve the model
TIMELAG	Number of months between first and second gift	Mean imputed with average value
AVGGIFT	Average dollar amount of gifts to date	
CONTROLN	Control number	Identity Column
HPHONE_D	Indicator for presence of a published home phone number	Recoded this column as categoric
RFA_2R	code for RFA_2	Excluded as it has only one unique value

RFA_2F	Frequency code for RFA_2	Recoded this column as categoric
RFA_2A	Donation Amount code for RFA_2	
MDMAUD_R	code for MDMAUD	Excluded as it has only one unique value
MDMAUD_F	Frequecy code for MDMAUD	Excluded as it has only one unique value
MDMAUD_A	Donation Amount code for MDMAUD	Excluded as it has only one unique value
CLUSTER2	Classic Cluster Code	Replaced missing values with 0 and then Recoded this column as categoric
GEOCODE2	County Size Code	Created separate column for each unique value in this column and then recoded the columns

Summary of Classification Model

	Sensitivity	F1	Accuracy	Precision	AUC	FN	FP	Total Loss
Logistic Regression	0.001	0.0)2	0.948	0.050	0.578	976	1	$292,800

Boosted Trees Model	0.041	0.0 i1	0.936	0.041	0.589	937	297	$281,397
Decision Forest l Mode	0.00	0.00	0.949	1.000	0.584	977	0	$293,100
Decision Jungle Mode l	0.00	0.00	0.949	1.000	0.570	977	0	$293,100

Ans: On Observing the data, I realize that average cost per donor comes to be around $300 per person, while average cost of mailing to a particular candidate is around $.49 along with the punitive cost of dealing with the post office. So we realize that net cost of mailing a person comes to be around $1. So reducing false negative has a huge impact in preventing loss of opportunity cost compared to False Positive as it only results in loss of $1 per person against loss of $300 if we miss out on a suitable candidate. So, On calculating using the provided value, I realize that I lose the least amount when I'm using Boosted Tree Model in terms of opportunity cost and cost of mailing a wrong candidate, which is depicted from the table above.

Another predictor that could be used to predict the right model is the value of F1, which is another determinant used to predict the no of False positive and False negative. More the value of F1, lesser is the error. However I have come to realization that in the given data set the FN far supersedes FP and thus FN would be primary determinant.

Summary of Regression Model

	MAE	RMSE	Rel Squared error	Rel. Abs. Err
Boosted Trees	1.51	4.3	1.04	1.01
Decision Forest	1.49	4.22	0.99	1.00
Neural Networ :s	5.414	6.06	2.06	3.62

Ans: I would always prefer a model which would have the least MAE, as it would help me prevent my prediction to fluctuate the least. I would also prefer a model that would provide me with the least RMSE as that would highly help me prevent my prediction to deviate from the mean value, there by my successfully predicting the amount that I would receive per donor.

SR.No	CONTROLN	Target_B	Target_D	Donation Forecast
1	71565	0.830796421	17.69175835	14.69824951
2	12442	0.925866187	15.07478669	13.95723526
3	37875	0.873683095	15.58724953	13.61831641
4	125925	0.914245486	12.68406158	11.59634605
5	190498	0.237258375	47.1754986	11.19278212
6	5660	0.50577265	18.97765679	9.598379769
7	127398	0.722077012	13.25856286	9.573703453
8	185091	0.700227797	11.89755932	8.331001751
9	188436	0.960914135	7.956221114	7.64524533
10	68588	0.953393877	7.665957807	7.308677231
11	7999	0.871982396	8.201547067	7.151604659
12	189304	0.936090291	7.620102462	7.133103928
13	176291	0.9486624	7.447264967	7.064940259
14	1691	0.921402454	7.536958793	6.94457233
15	142666	0.780493319	8.802490242	6.870284824
16	109001	0.474645823	14.17574687	6.728459044
17	164798	0.731887639	9.190361435	6.726311928
18	162458	0.661901534	10.04630779	6.649666533
19	185077	0.141295105	46.95153265	6.634021746
20	165129	0.80880785	7.993513929	6.465216814