KDD CUP 1998: DIRECT MARKETING FOR PROFIT OPTIMIZATION

RESULTS

Rajasekhar Battula

September 17, 2020

TOPICS

Objective

Modelling Approach

Data Understanding

Preprocessing

Variable selection & Modelling

Model Evaluation

Results

OBJECTIVE

To understand how to best target people who are most likely to give to charity.

The tasks are a classification problem where goal is to predict which people are more likely to donate to a charity and a regression problem where the goal is to estimate the return from a direct mailing in order to maximize donation profits.

MODELLING APPROACH

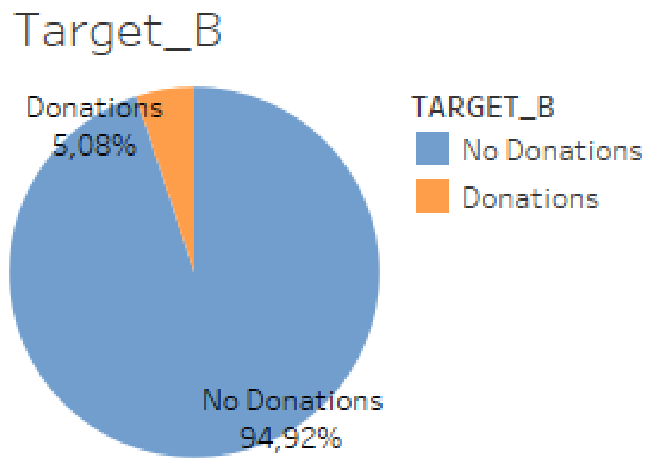
The general process used for addressing the problem

Data Understanding	Preprocessing	Model Development	Model Evaluation	Post-Analytics
Exploring the data in terms of the dimensions, cardinality, datatypes, distributions and data definitions	Perform data cleaning, Identify the feature correlations, feature selection, important data features and their influence on predicting the donations	Prepare data for modelling, data standardization, model selection for both classification and regression, training, testing and valuation along with Hyper-parameter tuning	Evaluate various models based on the metrics selected for the use case like recall, precision scores for classification and mean absolute score for regression	Make observations and recommendations based on the findings from model output

DATA UNDERSTANDING

- The dataset is an imbalanced dataset related to TARGET_B (Response to 97 NK mail) as shown below
- Dataset doesn’t have any duplicate mails or duplicate identifier CONTROLN
- There are 3 columns with 99.5% null values
- Generated a Pandas profile report to which explains the distributions and statistics of each column in the data and observed the dispersion and variance of the columns
- Dataset has high cardinality variables, noisy data

Attributes	**Training data**	**Validation data**
Number of observations	95412	96367
Number of columns	481	479
Number of categorical variables	74	74
Number of Numerical variables	407	405
Target variables	TARGET_B, TARGET_D	



DATA CLEANING & FEATURE ENGINEERING

- Truncated suffix '-' for some values in column ZIP
- Gender column has values A,U,J,C which has been replaced with nulls and imputed using the column TCODE which is the donor title.
- 3 columns with greater than 99.5% null values has been removed.
- 1 column with variance in its data less than 0.1% has been removed.
- Converted FISTDATE, LASTDATE columns to date format.
- Some of the binary fields (RECINHSE, RECP3,RECPGVG,RECSWEEP,MAJOR,PEPSTRFL, NOEXCH) has values (<space>, X) are replaced with (0,1)
- Created new columns for every columns with missing data. For e.g: if column RAMNT_8 has missing data then RAMNT_8_was_missing column is created and for each corresponding row of missing value 1 is updated else 0.
- After creation of above missing value columns, SimpleImputer() is used to impute null values in existing columns
- 92 columns with high correlations among themselves are removed to solve the problem with multi-collinearity (this problem will reduce the predictive power of the models)

DATA ANALYSIS

RATE OF DONATION BY DEMOGRAPHIC FEATURES

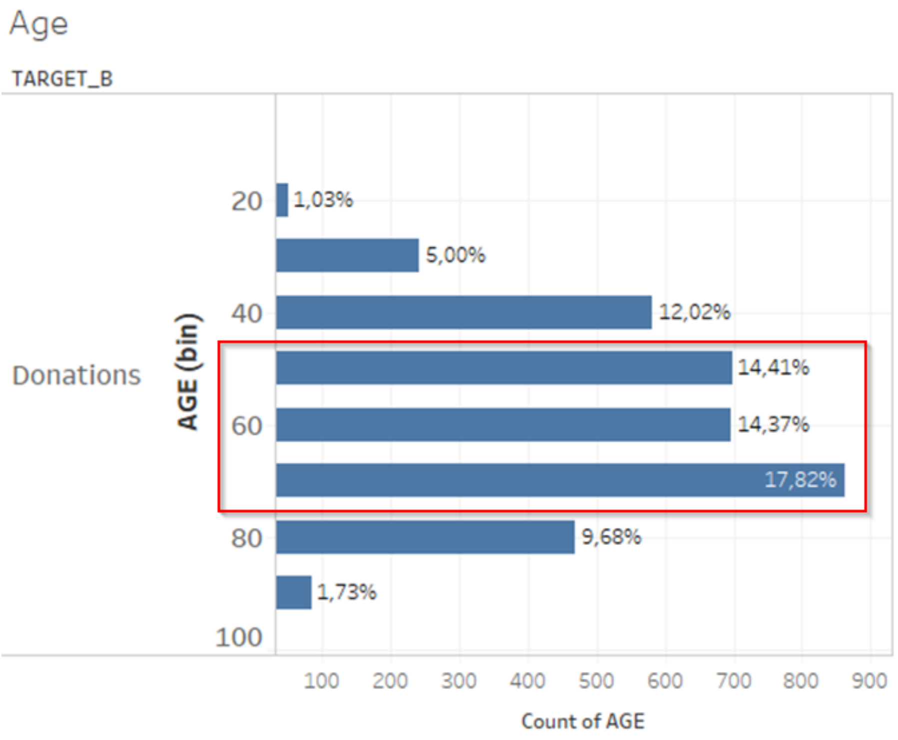
- NUMCHILD: As the number of children increases it is evident that donation rate decreases
- People in the age group 50-70 years make more donations

NUMCHILD

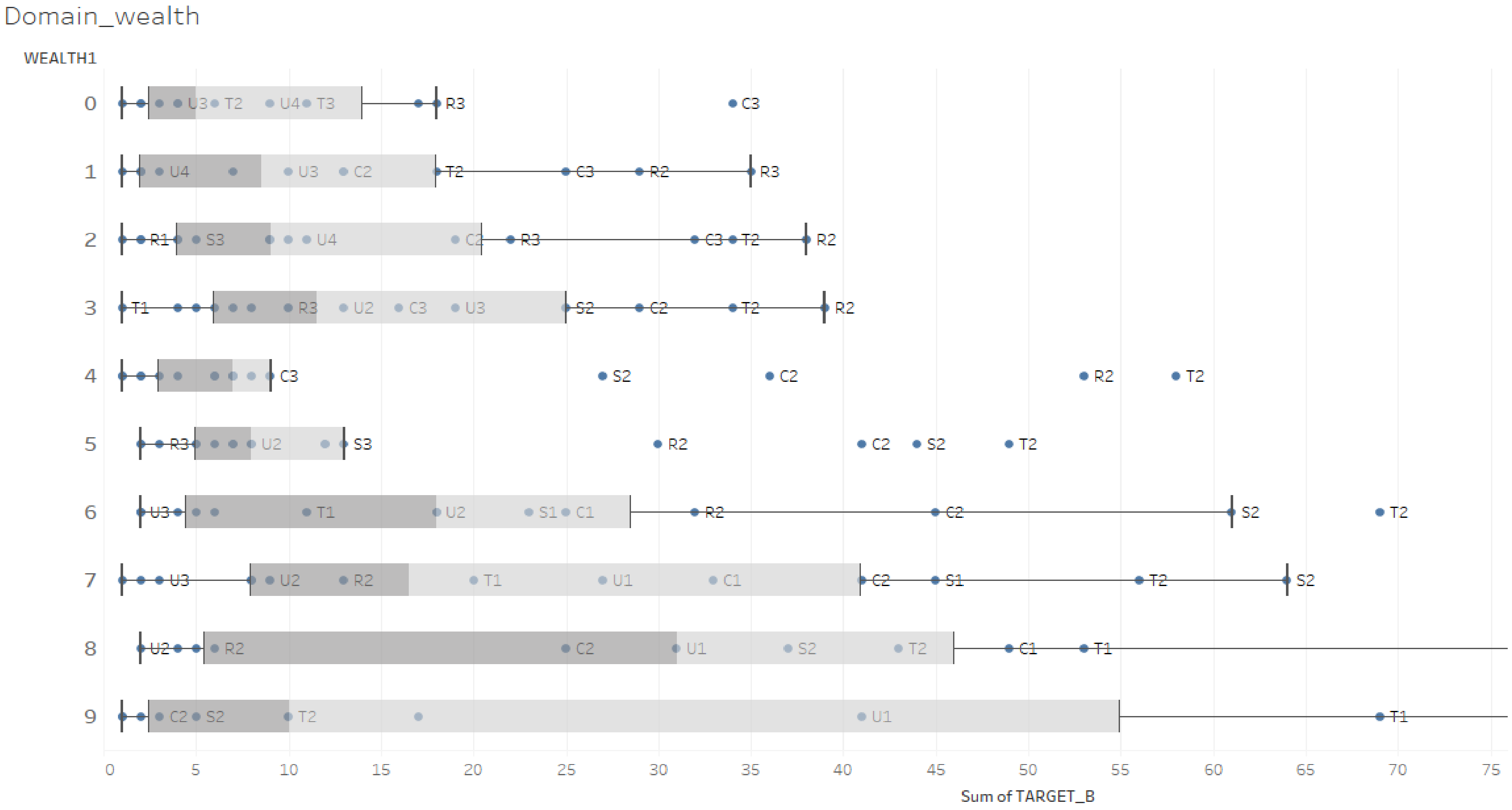
	TARGET_B	
NUMCHLD (group) 1	No Donations	Donations
	94,85%	5,15%
1	95,32%	4,68%
2	95,11%	4,89%
3, 4, 5 and 2 more	96,77%	3,23%

% of Total Count of cup..
3,23% – 96,77%

% of Total Count of cup98LRN.txt broken down by TARGET_B vs. NUMCHLD (group) 1. Color shows % of Total Count of cup98LRN.txt. The marks are labeled by % of Total Count of cup98LRN.txt.



DONATIONS BY DEMOGRAPHIC AND SOCIO-ECONOMIC FEATURES



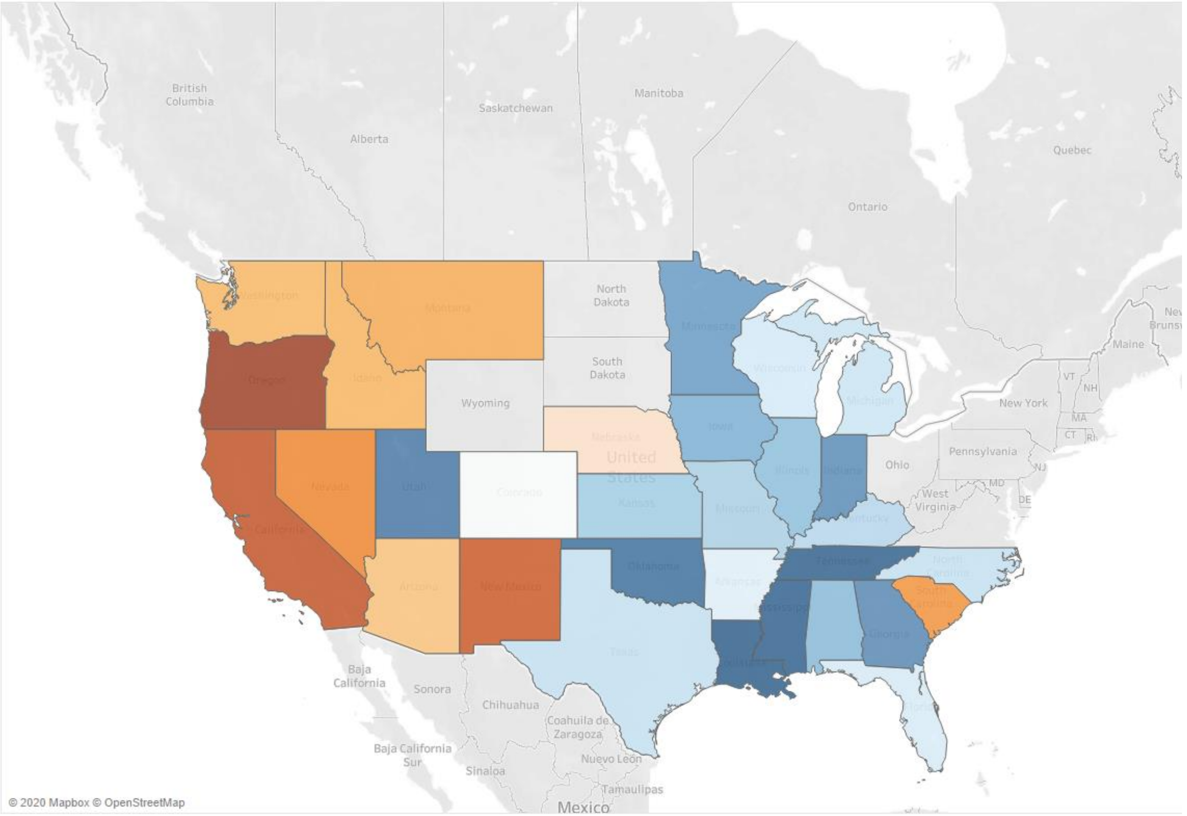
Sum of TARGET_B for each WEALTH1. The marks are labeled by DOMAIN. The data is filtered on TARGET_B, which keeps Donations. The view is filtered on WEALTH1, which excludes Null.

Donations are high in the highest, average Socio-economic status neighborhood of city, town and urban regions and among wealthy people

```
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.)
```

RATE/AMOUNT OF DONATION BY STATE

States by rate of donation

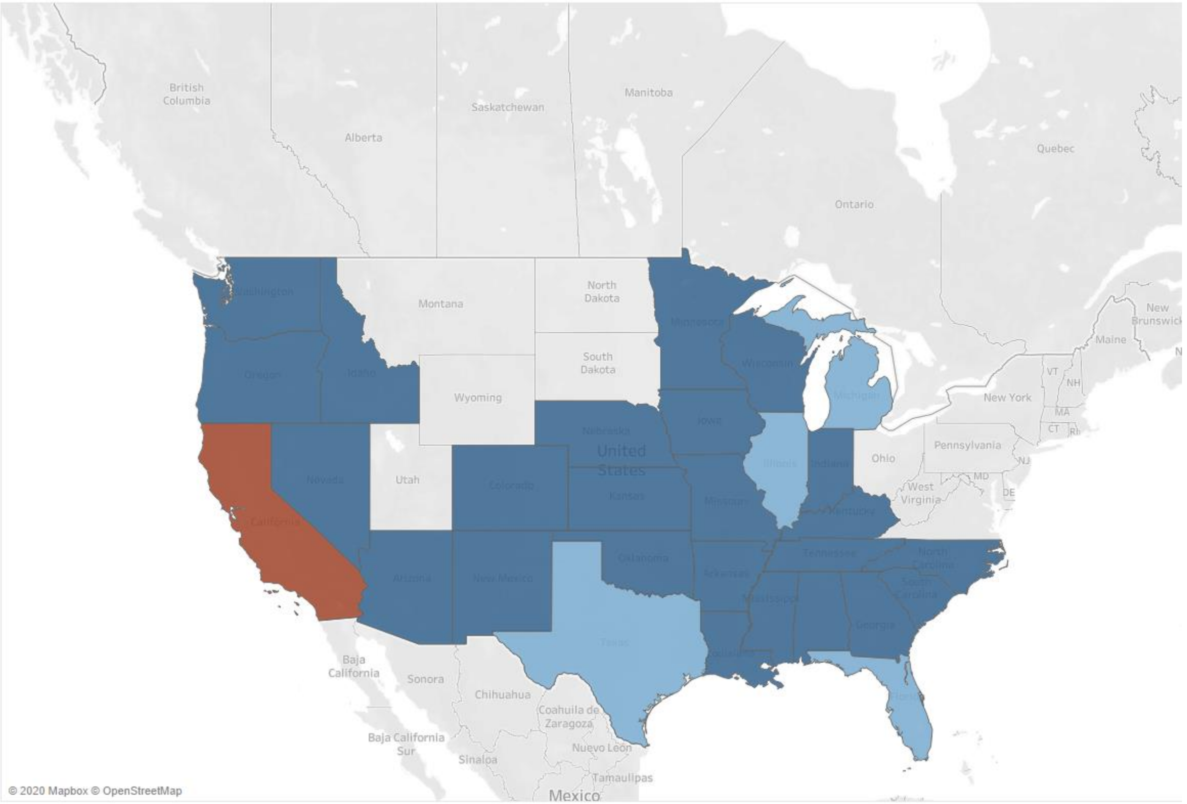


Map based on Longitude (generated) and Latitude (generated). Color shows average of TARGET_B. Details are shown for STATE. The view is filtered on STATE, which keeps 31 of 57 members.

AVG(TARGET_B): Perce..

0.03827 0.06322

States by amount of donation



Map based on Longitude (generated) and Latitude (generated). Color shows sum of TARGET_D. Details are shown for STATE. The view is filtered on sum of TARGET_D, which ranges from 500 to 18,103.

TARGET_D

513 18,103

Observations:

- West US states has more rate of donations compared to Eastern US states, esp. *California, Oregon* are the one's with the highest rate of donation.
- States with high contributions to the donations are from California, Florida, Illinois, Michigan, Texas.

RATE OF DONATION BY FEATURES

- Observed important variables in terms of rate of donation:
  - It is observed that income and wealth columns has positive correlation with donation.

Wealth

WEALTH1	TARGET_B No Donations	TARGET_B Donations
Null	95,00%	5,00%
0	95,40%	4,60%
1	95,31%	4,69%
2	95,06%	4,94%
3	94,90%	5,10%
4	95,07%	4,93%
5	95,19%	4,81%
6	94,45%	5,55%
7	94,68%	5,32%
8	94,85%	5,15%
9	94,42%	5,58%

% of Total Count of cup..
4,60% 95,40%

% of Total Count of cup98LRN.txt broken down by TARGET_B vs. WEALTH1. Color shows % of Total Count of cup98LRN.txt. The marks are labeled by % of Total Count of cup98LRN.txt.

INCOME

INCOME	TARGET_B No Donations	TARGET_B Donations
Null	94,82%	5,18%
1	95,83%	4,17%
2	95,18%	4,82%
3	95,06%	4,94%
4	94,97%	5,03%
5	94,74%	5,26%
6	94,46%	5,54%
7	94,30%	5,70%

% of Total Count of cup..
4,17% 95,83%

% of Total Count of cup98LRN.txt broken down by TARGET_B vs. INCOME. Color shows % of Total Count of cup98LRN.txt. The marks are labeled by % of Total Count of cup98LRN.txt.

RATE OF DONATION BY RFA FIELDS

RFA_2F

	TARGET_B	
RFA_2F	No Donations	Donations
1	96,24%	3,76%
2	94,86%	5,14%
3	93,44%	6,56%
4	91,68%	8,32%

% of Total Count of cup..

3,76% 96,24%

RFA_2A

	TARGET_B	
RFA_2A	No Donations	Donations
G	96,44%	3,56%
F	95,62%	4,38%
E	93,53%	6,47%
D	90,61%	9,39%

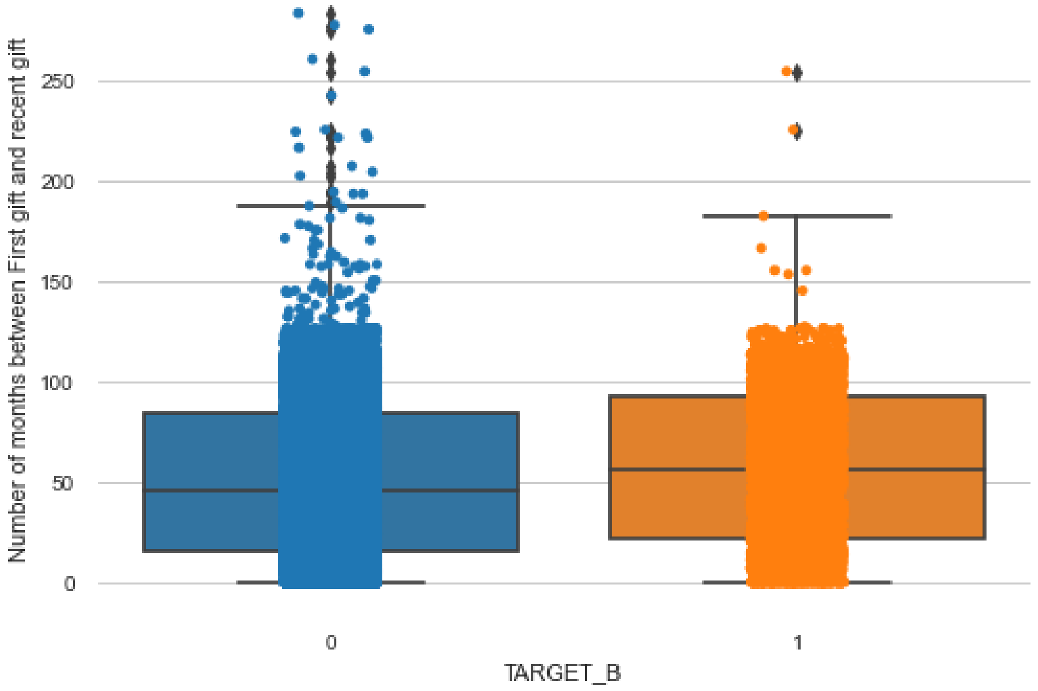
% of Total Count of cup..

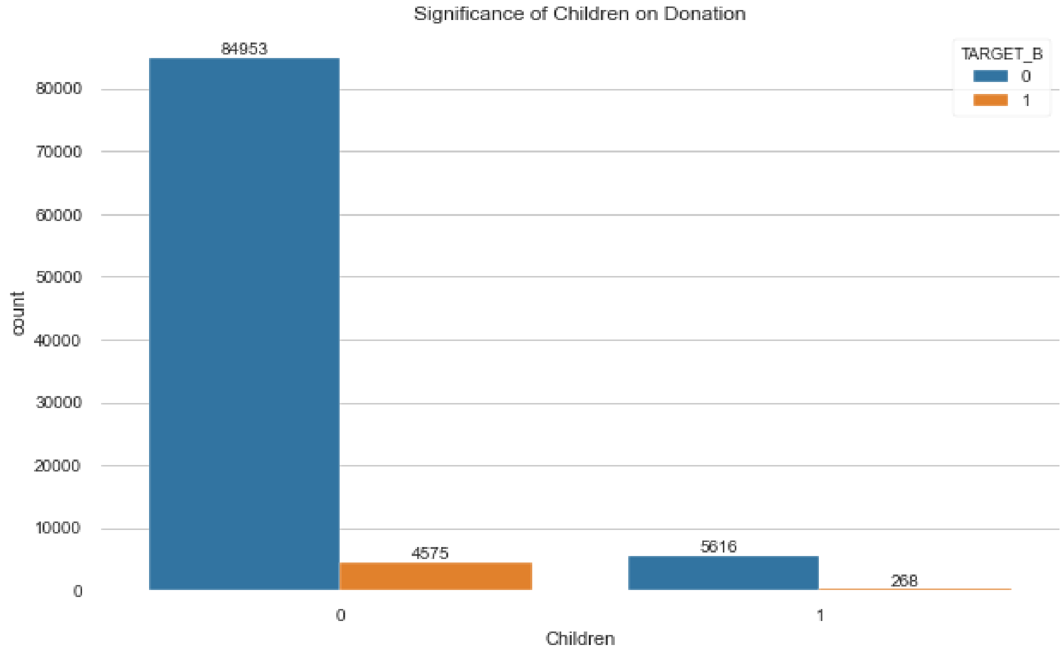
3,56% 96,44%

```
1=One gift in the period of recency
2=Two gift in the period of recency
3=Three gifts in the period of recency
4=Four or more gifts in the period of recency

Third byte of the code is the Amount of the last
gift.

A=$0.01  -  $1.99
B=$2.00  -  $2.99
C=$3.00  -  $4.99
D=$5.00  -  $9.99
E=$10.00 - $14.99
F=$15.00 - $24.99
G=$25.00 and above
```

- RFA (Recency, Frequency and Monetary) columns played an important role in the prediction of Target variables. For e.g: RFA_2 has sub columns RFA_2A, RFA_2F. The other RFA columns are also included in modelling process
  - RFA_2F: more the number of gifts the more is the donation rate. i.e frequent donations are more by such people
  - RFA_2A: more the amount donated in last gift, less is the donation rate. i.e donations are less by people who donated higher amounts recently

DISTRIBUTION OF DERIVED VARIABLES





- Observed important variables in terms of rate of donation:
  - Nb_months: It is a derived field, I.e the difference between date of first gift and most recent gift. It is observed that the average difference in nb_months is higher for people who donated vs who has not donated.
  - Children: 1 is the people with children of age 3 to 18 else 0. People with children of this age make less donations than people without children.

FEATURE SELECTION

- Various methods are used for feature selection:
- RandomForestClassifier – BorutaPy Technique
  - Data scaling is performed on the Train dataset before giving it as input.
  - Shape of the dataset is 95412 rows and 388 columns.
  - 70 important features has been selected by the model
- Manual Inspection:
  - 41 features has been selected as important by manual inspection of the features
  - 20 of these features has also been selected by the RandomForestClassifier feature selection algorithm mentioned above
- Together selected 91 features, as columns selected by RandomForestClassifier are giving better results they are only used in modelling process
- Tried and tested other feature selection techniques like Recursive Feature elimination, Kbest etc. But Ensemble method like Random forest has gave the better results

MODEL DEVELOPMENT

- Multiple models (from classical Machine Learning) has been built for both prediction of TARGET_B (Classification output), prediction of TARGET_D (Regression output)

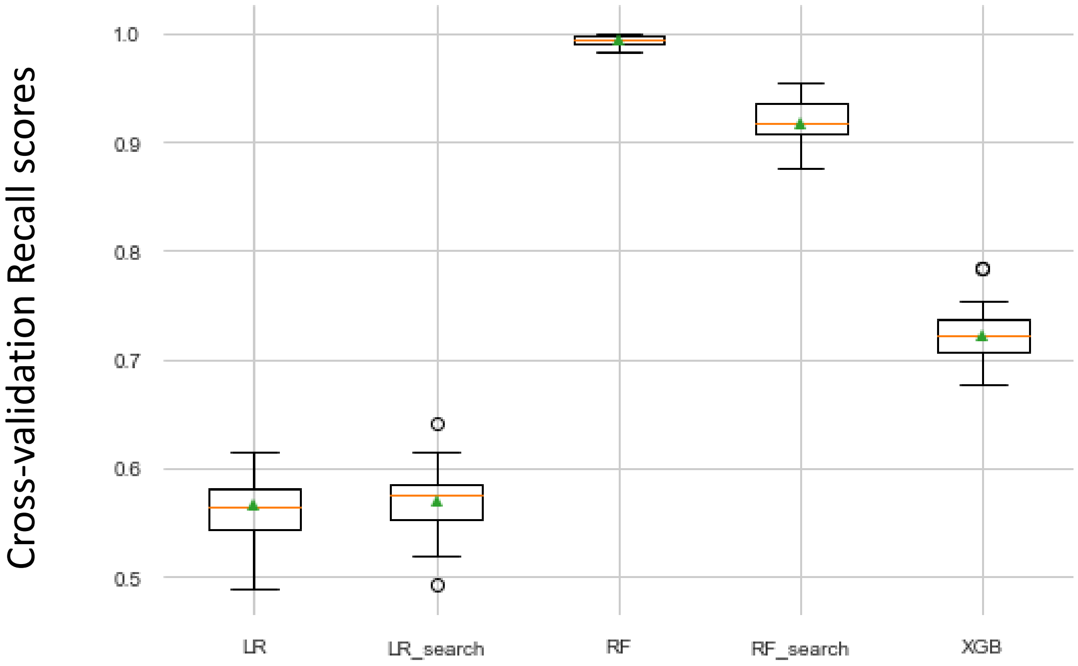
- Prediction of TARGET_B:
  - Logistic Regression with and without Hyperparameter Tuning
  - Random Forest Classifier with and without Hyperparameter Tuning
  - XGBoost Classifier without Hyperparameter Tuning
  - Finally a voting ensemble model has been built from the above 3 model outputs
- Prediction of TARGET_D:
  - Ridge Regression with and without Hyperparameter Tuning
  - Random Forest Regressor without Hyperparameter Tuning
  - Decision Tree Regressor without Hyperparameter Tuning
  - Finally an average ensemble model has been built from the above 4 model outputs

MODEL EVALUATION METRICS

- Below are the metrics considered as important for this use case and the models are trained and tested to optimize the same
  - **Recall and Precision** for Classification i.e prediction of TARGET_B variable
    - As it is important to understand how many donations are correctly predicted by model among all the actual donations
    - Also, it is important to understand how many are the correct donations among all the predicted donations
  - **Mean absolute error (MAE)** for Regression i.e prediction of TARGET_D variable

MODEL EVALUATION RESULTS - CLASSIFICATION

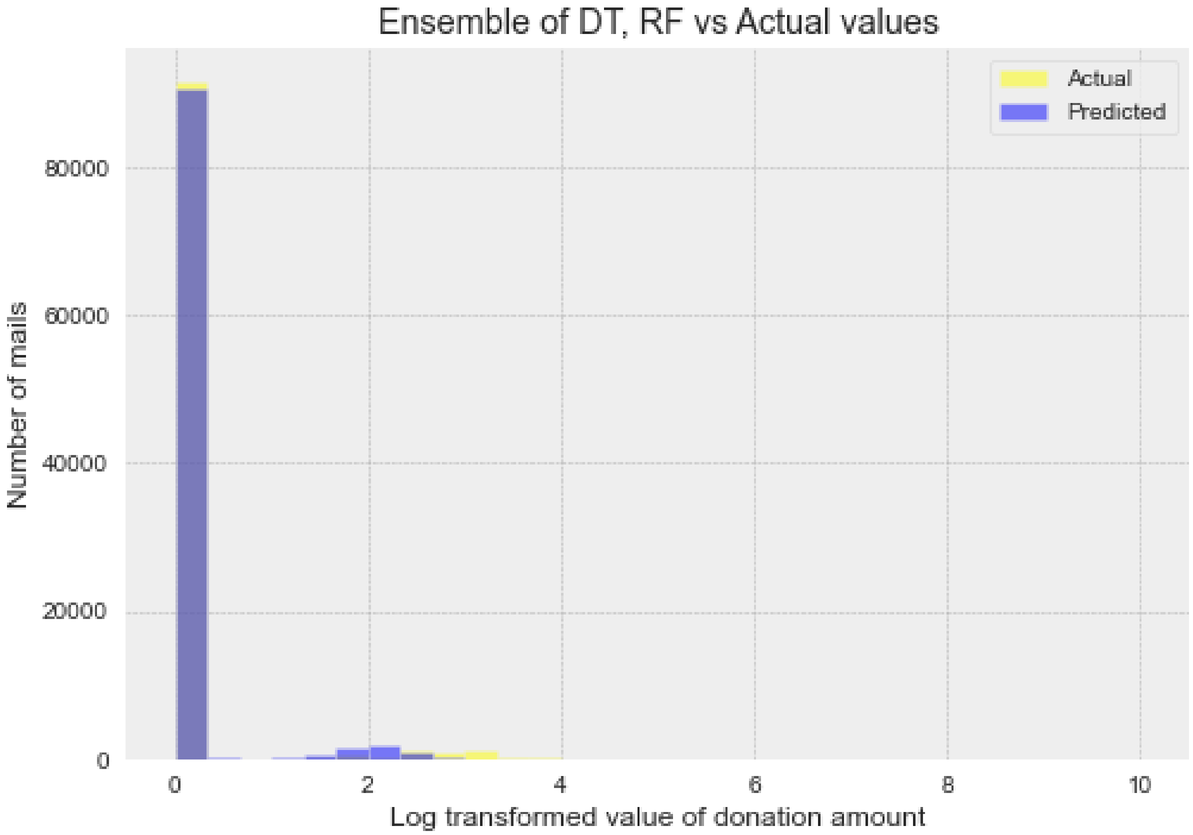
- Tree models are overfitted to the data as you can see in the below plot the train cross-validation scores for Random forest is high compared to Logistic regression and Xgboost.
- **Voting classifier** shows better weighted average Recall and Precision scores among all the classifiers, It shows a balance between Recall of individual classes



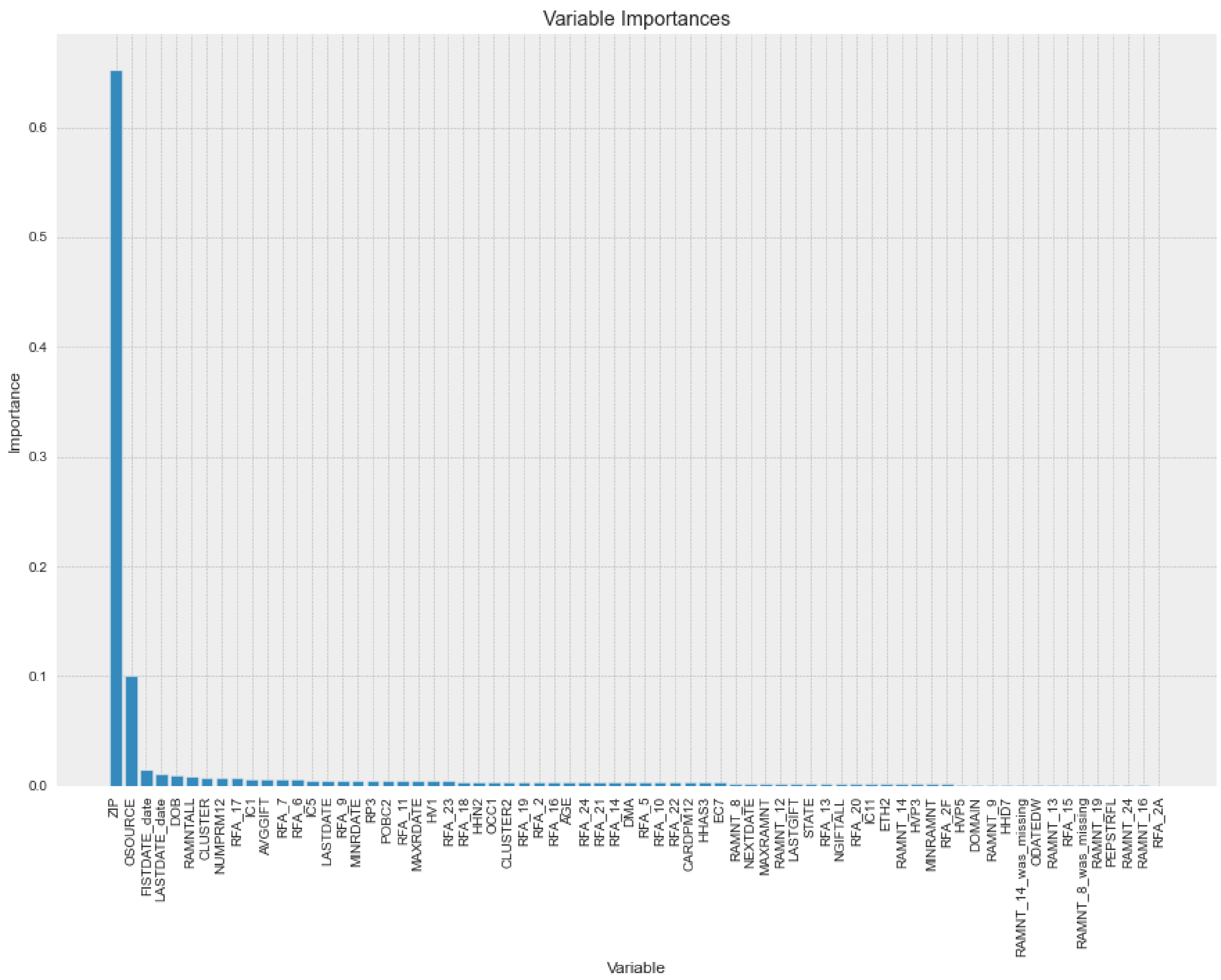
Model	Train Recall	Test Recall	Validation data Recall (Weighted avg)	Validation data Precision (Weighted avg)
LogisticRegression_default	0.564	0.569	0.57	0.92
LogisticRegression_search	0.568	0.567	0.56	0.91
RandomForestClassifier_default	0.990	0.995	0.50	0.91
RandomForestClassifier_search	0.910	0.928	0.59	0.91
XGBClassifier_default	0.641	0.754	0.68	0.91
VotingClassifier_default	***0.810***	***0.882***	***0.62***	***0.91***

MODEL EVALUATION RESULTS - REGRESSION

- **Ensemble of Decision Tree regressor, Random Forest Regressor** has resulted in better prediction than all other models. So, in this case of regression tree-based models has worked better than the linear models.



Model	Test MAE	Validation data MAE
DecisionTreeRegressor_default	0.240	0.264
RandomForestRegressor_default	0.237	0.267
RandomForestRegressor_search	0.221	0.247
RidgeRegressor_default	0.248	0.273
RidgeRegressor_search	0.244	0.269
EnsembleClassifier_DT_RF		***0.260***



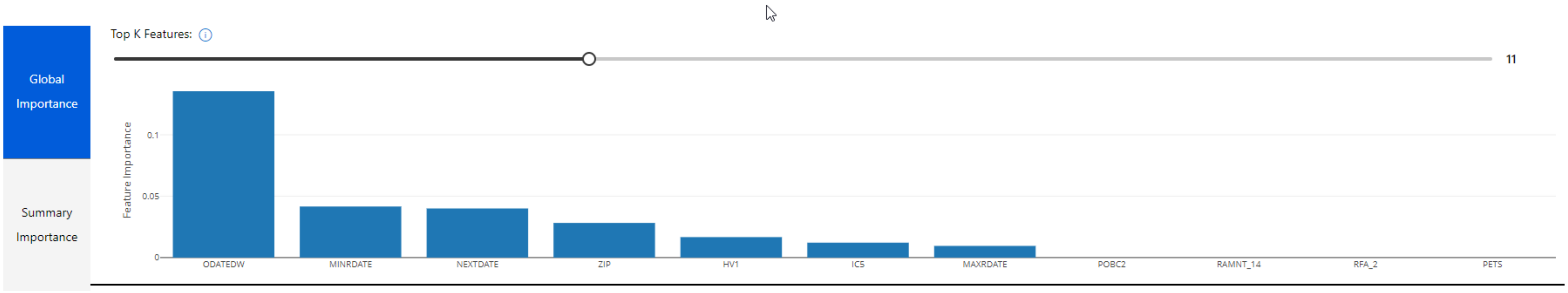
VARIABLE IMPORTANCE BY RANDOM FOREST REGRESSOR

COST MATRIX ANALYSIS

How many people we should target if each envelope was costing us $5 → 3680 rows
What would happen if each envelope only cost $1? → 5368 rows

Cost of Envolope	Number of people to target	Amount of money the charity will be able to collect
$1	5425	37818.06
$5	3760	31576.46

FEATURE IMPORTANCE BY AZURE ML



- Performed Automated ML experiment in Microsoft Azure and observed the Global importance values for the features.
- The selected model by Azure is a VotedEnsemble model and Average Recall metric for the model is 0.67

WHAT'S NEXT…

Feature Engineering improvements

- Apply date difference between 06/1997 (Date of mails) and all the date columns
- Split the bytes of RFA fields and aggregate the R, F, A bytes of all the RFA_xx columns
- Create binary variable whether the person belongs to US West or US East states
- Impute the age of the person based on the age of the children
- Better Imputation of null values in categorical columns instead of using SimpleImputer()

Model development improvements

- Apply SMOTE instead of using class_weight parameters
- Try hyper-parameter tuning for gradient boosting algorithms
- Work with probabilities from the models to try and adjust the thresholds which results in better predictions (Esp. for Logistic Regression)

THANK YOU