Report for Case Competition Concerning Donation Behavior

Advisor: Professor James Albert

Xiang Cao* Haitao Liu* Liwen Tong* Yang Yu^

*Department of Applied Statistics & Operations Research

^Department of Statistics

Bowling Green State University

Bowling Green, OH 43403

1. Summary

Over 52 million hits will pop out if you do the Google search on the phrase "fundraising analysis", it would be outdated to think that non-profit donation is an absolute random individual behavior which beyond prediction and over-optimistic to expect a growth of outcome simply by increasing the coverage of market promotion. Colleges, institutions and other non-profit organizations amass great deal of information about the people they serve, with effective fundraising analysis such as data mining and predictive modeling, we can yield significant benefits in cost saving and more productive donor contact.

In this project, 95412 records basically about the customers' demographics information, socio-economic status and their history of donation were collected to build our model and the model performance will be tested on the prediction of customer's donation behavior on another data set. Several analytical approaches were exploited such as logistic regression, Lasso and ridge regression, random forest, Naïve Bayes and K nearest neighborhood to build our candidates of model. As we are facing an imbalanced problem in this project (class of interest is less than 5%), several evaluation metrics such as cross-validation, F1 score, AUC(ROC) and LIFT were discussed to build a more appropriate criterion of our model. Lasso regression shows the best performance and used to predict the response variables.

2. Exploratory data analysis

The variables can be classified into three categories, including customer's general information, neighborhood's demographics and socio-economic status, and the information of customer's donation history. We picked up some variables in these three categories to do exploratory analysis to show some possible correlations between predictors and response variable, which would be helpful for fitting models.

2.1 State

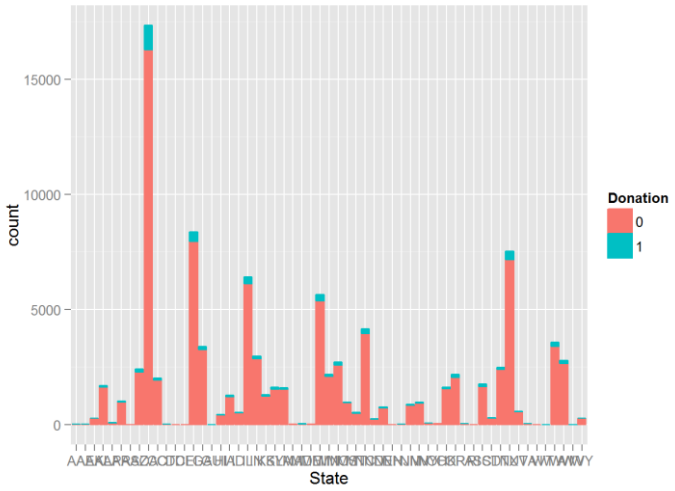

Fig. 1 Histogram of State

The number of samples vary dramatically in different State. The CA has the largest 16284 observations while there are nine states in which samples are less than ten. As a result, the ratio of donation=1 to donation=0 could be influenced by incidents in states with small number of samples. Another problem is that the states in training data and test data are inconsistent. Therefore, we don't consider this feature at the starting point.

2.2 Gender, Age and Income

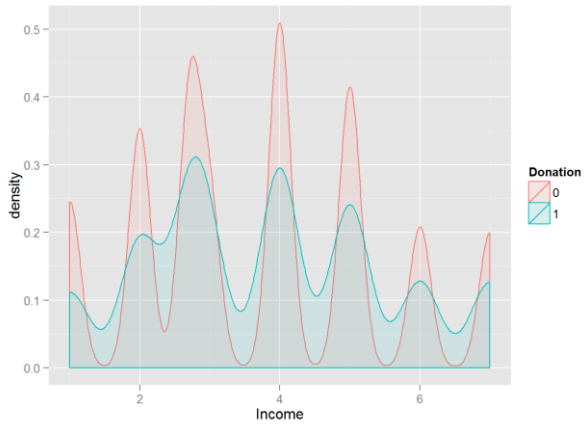

Fig. 2 Density plot of Income

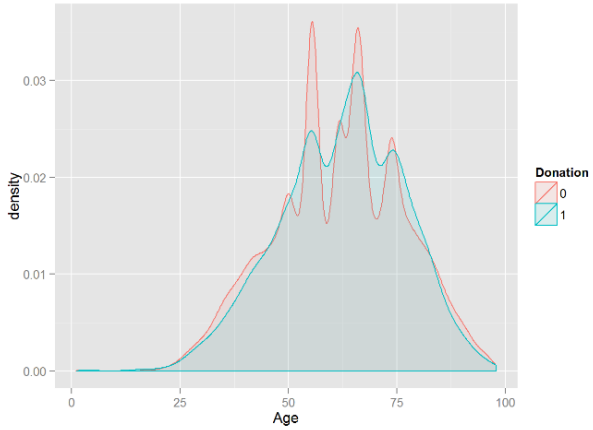

Fig. 3 Density plot of Age

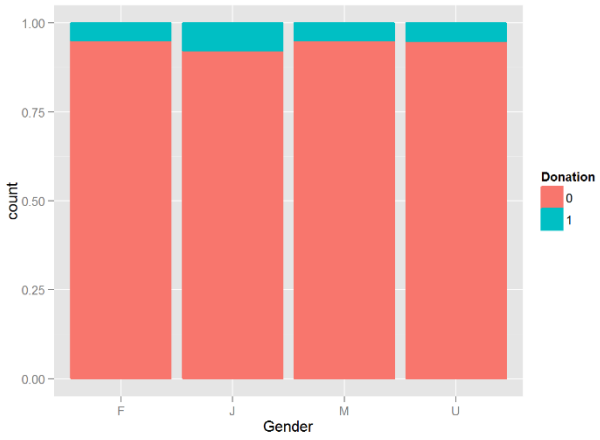

Fig. 4 Bar plot of Gender

We predicted the missing values in Income and Age first and then plot their density graphs. No obvious different pattern between donation=0 and donation=1 appears in the plot. To Gender, there is no different donation preference between female and male. However, for Joint account (J), a higher probability of donation is displayed in the plot.

2.3 Neighborhood

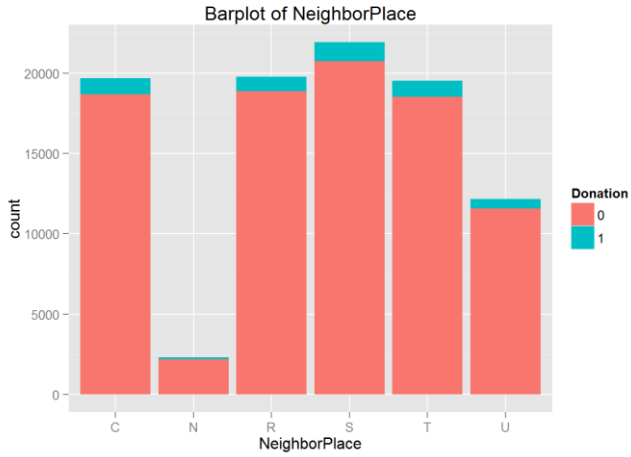

Fig. 5 Bar plot of NeighborPlace

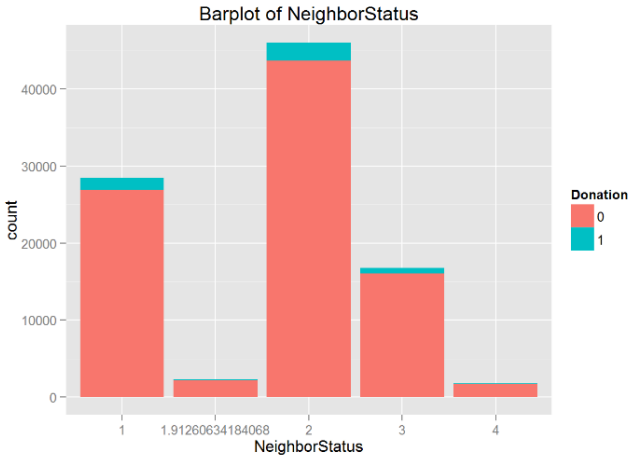

Fig. 6 Bar plot of NeighborStatus

	Donation			Donation	
NeighborPlace	0	1	**NeighborStatus**	0	1
C	0.9489	0.0511	1	0.9445	0.0555
N	0.9396	0.0604	1.912606	0.9396	0.0604
R	0.9532	0.0468	2	0.9497	0.0503
S	0.9459	0.0541	3	0.9562	0.0438
T	0.9485	0.0515	4	0.9618	0.0382
U	0.9526	0.0474			

Table 1. NeighborPlace and NeighborStatus vs. Donation

We split the variable NeighborhoodCode into two new variables, NeighborPlace and NeighborStatus. The missing values are assigned with NeighborPlace factor N and NeighborStatus value 1.926 which is the mean of all status values. From the probability of Table 1, consumer from suburban (s) and Highest status are more likely to donate, and those from urban (U) and lowest status are less likely to donate.

2.4 Donation history

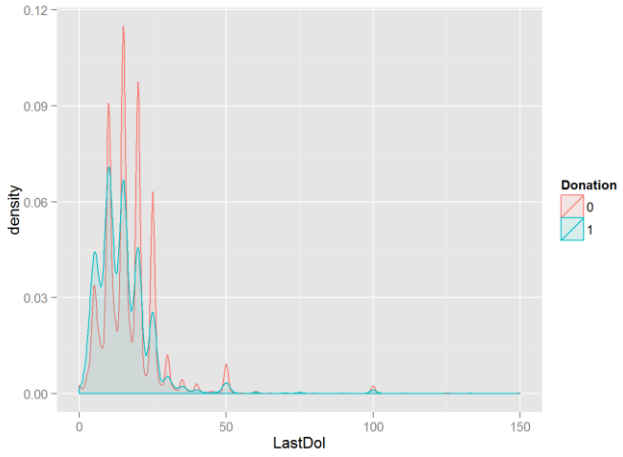

Fig 7. Density plot of LastDol

The variable LastDol represents the recent dollar amount of donation. According to the Fig 7, given a large LastDol amount like 50 or 100 dollars, a customer is less likely to donate; while given a small LastDol amount like those less than 10 dollars, a customer is more like to donate.

2.5 Correlation

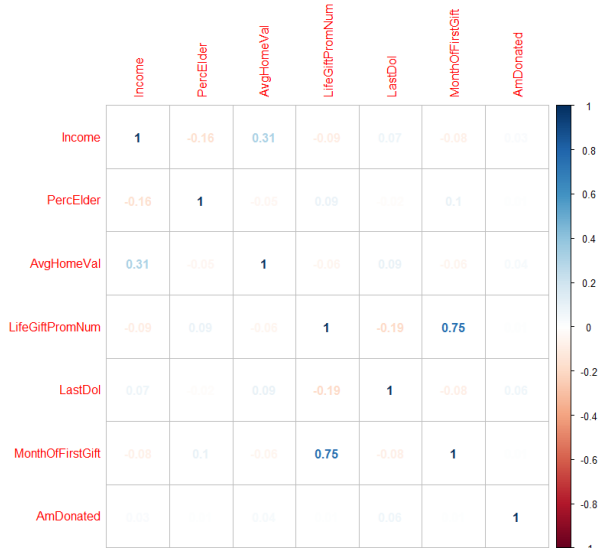

Fig 8 Correlation plot of selected variables

We picked several variable that we are interested and compute the correlation between each two of them. Some independent variables show little correlations, however, none of them are correlated with response variable.

3. Modeling and Results Evaluation

3.1 Introduction of **imbalanced** classification problem and solutions.

3.1.1 More appropriate evaluation metrics

Evaluation metrics plays a critical role because they are used to guide the algorithms and results of data mining. However, commonly used evaluation metrics such as classification accuracy are defective when dealing with rare classes and rare cases. Thus several metrics that take rarity into account are used to improve our analysis.

Notations:

True Positive(TP): Donated and classified as donated.

True Negative(TN): Non-donated and classified as non-donated.

False Positive(FP): Non-donated but classified as donated.

False Negative(FN): Donated but classified as non-donated.

$$\text{Specificity} = \frac{\text{TN}}{\text{N}} = \frac{\text{TN}}{\text{TN} + \text{FP}} \qquad \text{Sensitivity} = \frac{\text{TP}}{\text{P}} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{Yrate} = \frac{\text{TP} + \text{FP}}{\text{P} + \text{N}}$$

Lift tells how much better a classifier predicts compared to a random selection:

$$\text{Lift} = \frac{\text{Precision}}{\text{Proportion of Positive}} = \frac{\text{TP}/(\text{TP} + \text{FN})}{\text{P}/(\text{P} + \text{N})} = \frac{\text{Sensitivity}}{\text{Yrate}}$$

F_1 score is a measurement of test accuracy considers both precision and sensitivity:

$$\text{F}_1 \text{ score} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$$

ROC curve can be used to assess the accuracy of our classifier independent of the threshold. Since each pair of a probabilistic classifier and threshold t could define a binary classifier, by varying the threshold t of probabilistic classifier, we can plot $\text{y} = \text{sensitivity(t)}$ against $\text{x} = 1 - \text{specificity(t)}$ to get the ROC curve, each point represent a binary classifier. The area under ROC curve (AUC) is used as a measure of quality of the probabilistic classifier.

$$AUC = \int_0^1 \frac{TD}{D} d\frac{FD}{N}$$

3.1.2 cross-sensitive learning

Since the rare cases are of primary interest, one solution is to use cost-sensitive learning which exploit the fact that the value of correctly identifying the donated (minority class) outweighs the value of correctly identifying the non-donated (majority class). We can associate a cost with each of the four outcomes notated in the former part, which refer to as $C_{TD}, C_{FD}, C_{TN}, C_{FN}$. By assign $C_{TD} = C_{TN} = 0$ and $C_{FN} > C_{FD}$, we bias our classifier desirably and improve the performance with respect to the rare class.

However, this method may not be preferred for several reasons. Firstly, there are not cost-sensitive implementations of all learning algorithms and a wrapper-based approach such as sampling is the only option. Secondly, our data is enormous and highly skewed which need to be reduced in order for learning to be feasible, under-sampling seems to be a reasonable strategy in this case. Finally, the misclassification costs unknown, we can not assign a proper ratio on C_{FN} over C_{FD}.

For all above reasons, we decide to use sampling rather than cross-sensitive learning in this report.

3.1.3 Sampling

Another common idea to deal with rare events is sampling. Through adjusting the ratio of majority-class examples and minority-class examples, we reduce or eliminate class imbalance.

3.1.3.1 Basic sampling methods

The basic sample methods include under-sampling and over-sampling. Under-sampling omits observations from majority classes in the training data, while over-sampling, contrarily, duplicate observations from rare classes. Obviously, both of the two sampling methods decrease the overall level of class imbalance, thereby making the rare class less rare. These two sampling techniques, however, do have their own disadvantages. By under-sampling, we could risk removing some of the majority class examples than can be very representative, thus discarding useful information. Over-sampling just duplicates the minority class, which may lead to overfitting to a few examples. On one hand, over-sampling introduces "new data" into consideration, which can increase the time to build a classifier. On the other hand, paradoxically, those "new data" are often an outright remake of existing minority-class observations, which means that it may provide little additional useful information. For the reason stated above, under-sampling is more often a better choice than

over-sampling. Some researches involving imbalanced data adopt hybrid approach by combining under-sampling of majority class and over-sampling the minority class, however, the improvement is nonsignificant in the lift index (Ling & Li, 1998). Next section we will introduce an adjusted combination of these two sampling methods that can overcome the drawbacks described above.

3.1.3.2 Advanced sampling methods

1) SMOTE: Synthetic Minority Over-sampling Technique

As stated before, over-sampling is just making exact copies of existing data. With the adoption of SMOTE, however, the minority class is over-sampled by creating new observations along the line segments joining the k minority class nearest neighbors instead of by over-sampling with replacement (default k=5). For example, if the amount of over-sampling needed is 300%, three neighbors from the five nearest neighbors are chosen and one observation is generated in the direction of each. The algorithm is as follows:

a) For each observation x in minority class, calculate the Euclidean distance between x and each other observation in minority class, then obtain the five minority class nearest observations.

b) Depending on the amount of over-sampling needed, for each x_i, several neighbors can be chosen from the five nearest neighbors, and denoted as $\hat{x}$.

c) For each $\hat{x}$ chosen, a new observation can be generated using the following formula.

$$x_{new} = x + rand(0,1) * (\hat{x} - x)$$

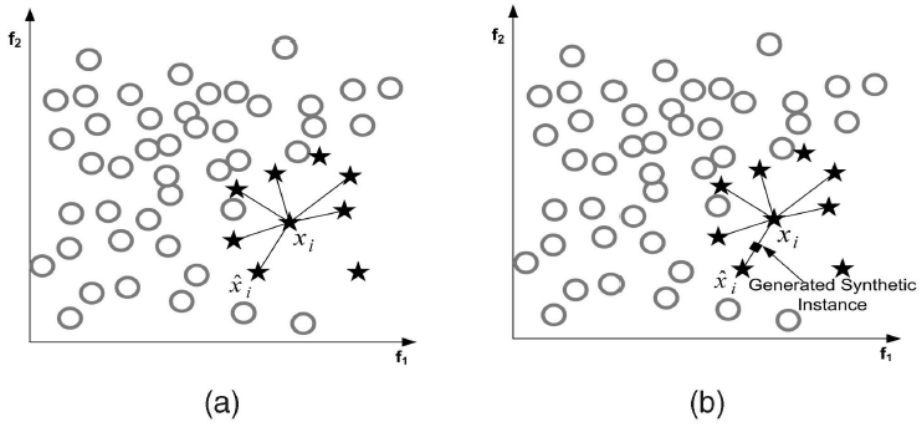

Fig 9. The figure of SMOTE algorithm

2) Under-sampling and SMOTE combination

In this report, we implement the combination method of under-sampling and SMOTE, by under-sampling the majority class and "over-sampling" the minority class using SMOTE to adjust the ratio of the observations in the two classes. Compared with other advanced sampling methods (e.g. CUBE for under-sampling), SMOTE tends to retain the completeness of useful information in the original data set and provide significant improvement to reduce the level of imbalance.

3.2 Methodology:

3.2.1 Lasso and Ridge Regression

Based on the characteristics of our data set, we applied five methods to our project, including Lasso Regression, Ridge Regression, Random Forest, Naïve Bayes, and K Nearest Neighbors.

Lasso Regression and Ridge Regression are the typical generalized linear models. The theory behind these two models is to use shrinkage approach involving all predictor variables and making estimated coefficients shrink toward zero associated with the least squares estimates.

Consequently, these two models reduce variance.

Here is the Ridge Regression Model:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \text{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2$$

From the equation above, we can see that Ridge Regression is similar to least squares except for the second term. The coefficient estimates $\hat{\beta}^R$ in Ridge Regression are the values to minimize equation above, where $\lambda \geq 0$ needs to be determined separately. Ridge Regression is the model seeks to find the coefficient estimates by making RSS as small as possible and λ to control shrinkage penalty. It is worth mentioning that when $\lambda = 0$, the second term, namely shrinkage penalty, exerts no effect in the model, obtaining the same result with the least square estimate. As λ tends to be infinite, however, the effect of shrinkage penalty increases and the coefficient estimates $\hat{\beta}^R$ approaches to zero.

Different from Ridge Regression, Lasso Regression can overcome an obvious disadvantage for Ridge Regression in which the penalty term will shrink all the predictor variables' coefficient towards zero but not exactly to zero. It will be difficult for us to interpret the model when the number of predictor variables are large using Ridge Regression. Here is the Lasso Regression Model:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \text{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|$$

Lasso Regression uses $L1$ penalty, forcing some of the predictor variables' coefficients to be exact zero when the tuning parameter λ is sufficiently large. Therefore, Lasso Regression can help us select variables. Compared to the Ridge Regression, the model from Lasso Regression is way easier to interpret than that from Ridge Regression. Cross-validation will help us select a good value of λ to minimize the error.

3.2.2 Random Forest

Random forest is an ensemble of simple tree predictors, which is induced from bootstrap samples if the training data by using random feature selection in the tree based method. For classification issues, it can vote for the most popular class, while for regression problems, random forest's responses are averaged to get an estimate of the the response variable. This method enforces each split in the tree to consider only a subset of the predictors so that other predictors will have more chances of being selected rather than only considering the strong predictor. It is designed to reduce the overall error rate, and focuses more on the accuracy for predicting the majority class.

3.2.3 Naïve Bayes

Naïve Bayes, an easy model to build, is an approximation to the Bayes classifier. It is to calculate the posterior probability. Denote π_k as the overall or prior probability that a randomly chosen observation is from the k^{th} class. In addition, denote $f_k(X) \equiv \Pr(X = x|Y = k)$ as the density function of X for an observation from the k^{th} class. Therefore, the formula for Naïve Bayes is as follows:

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

The result obtained from the Naïve Bayes classifier is usually well in large number of observations and often outperforms more other sophisticated classification models. Therefore, we prefer applying this model to our project to check the result.

3.2.3 K Nearest Neighbors

K Nearest Neighbors, one of the non-parametric methods without making any assumptions on the underlying data distribution, provides more flexible approach for classification. K points, represented by N_0, in the training data which are closest to x_0 are identified by KNN, given the the positive integer K and a test observation x_0. It estimates the conditional probability for class j as the fraction of points in N_0 whose response values equal j:

$$\Pr(Y=j|X=x_0)=\frac{1}{K}\sum_{i\in No} I(y_i = j)$$

Hence, KNN will predict the highest probability of categorical response variable belonging to which classification.

3.3 Data Preprocessing

Besides the customer ID and two target fields, there are 139 variables describe the information of customer, neighborhood demographics and socio-economic status, and history of donation. A few variables have missing values or inaccurate notation, so we clean the data first.

3.3.1 Clean data:

1) Convert variable DateOfFirstGift into a new variable.

   The date of first gift is indicated by the string containing the year and month when the first gift was send to the customer. E.g., 8901 represents the first gift was sent in January 1989. According to the variable DOB(date of birth) and Age, we knew the data was collected in around January 1998, then we compute the number of months between the date of first gift and the "current month" January 1988. A new variable MonthOfFirstGift was created which is a continuous variable to fit the later models. E.g. 8901 is converted into number of months 107.

2) Predict the missing value of Income.

   More than 22% of the samples in both training data and test data are missing. We combine the training data and test data, split the combined dataset based on whether the Income is missing or not. Then we fitted a regression tree mode using the data set which has Income to predict the missing values of Income.

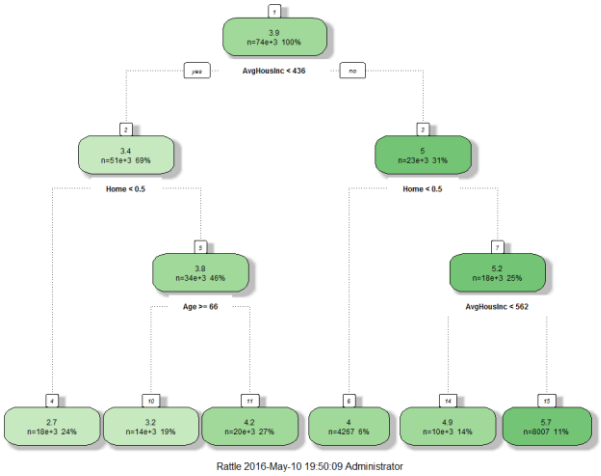

Fig 10 Regression tree for Income

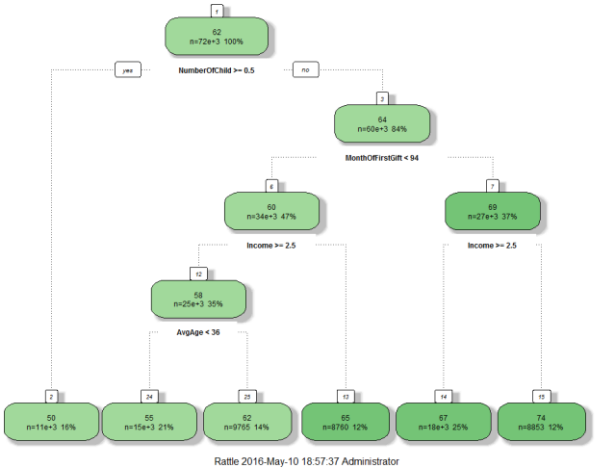

Fig 11 Regression tree for Age

3) Predict the missing value of Age.

The same method of predicting missing Income is used to predict Age.

4) Zip Code. Around 5% of zip code values are negative. According to the state the customer belongs to, we think the zip code is correct and convert the negative ones into positive.
   Also, some zip codes have only three or four digits, e.g. 778, 1002. I think that's because the value is recorded in numeric format, therefore any zero at the beginning of zip code was dropped by the system. So we add the zero back to the zip code and convert it into a string. E.g. convert 1002 into 01002, which is consistent with the State information.
5) Gender. In the combined data set of training and test dataset, there are two samples with label A and two samples with label C. Since A and C aren't defined before, we simply convert A and C into the dominated level Female.
6) NeighborhoodCode. Around 2.5% of NeighborhoodCode are missing, we create a new factor 'N' for these values.
7) All the continuous variables are centered and scaled, by subtracting the mean of the variable and dividing by its standard deviation.

3.3.2 Feature selection:

Generally, there are three ways to do feature selection. In order to reduce overfitting and improve the generalization of models, we usually do feature selection before fitting models.

1) Univariate selection.
   Examine each feature individually to determine the strength of the relationship between the feature and response variable. Criteria like Pearson Correlation, R-square, P-value are used to determine whether a variable should be included or excluded. The drawback of this method is that it is unable to remove redundant features.
2) Linear models with regularization
   As we introduced before, the Lasso regression (L1) can be used as a way to select variables. Lasso regression can force the coefficients of unimportant variables as low as zero, thereby we can select the variables whose coefficients aren't zero. While the Ridge regression (L2) forces the coefficients to be spread out, the coefficients can be very small but never be zero.
   Due to the different form of regularization, to a group of correlated variables, Ridge regression is more stable than Lasso regression. Namely, each time Lasso regression isn't guaranteed to select the same variable from the correlative one, while Ridge regression can get more stable coefficients for each of them. Although the instability of Lasso, we still implement it in this

model since it did cross validation to choose the parameter to minimize the error, so we reckon the different groups of features can achieve the same performance.

3) Tree base models.

The tree based models like decision tree and random forest will return the importance of variables based on the mean decreased impurity contributed by each variable. However, since our dataset is highly imbalanced, in Balanced Random Forest the criteria like decreased accuracy for ranking variable importance are inappropriate.

In this model, we fitted lasso regression and can select the features whose coefficients are greater than zero. However, compared to the data size, the number of variable isn't large, therefore, we didn't arbitrarily select these variables to fit all other models. Conversely, we use all variables to fit Random Forest, Naïve Bayes and K Nearest Neighbors.

3.3.3 Results: Classification

1. SMOTE sampling: Lasso, Logistic regression, and Random Forest

		Original	10	20	30	40	50	60	70	80	90
Lasso Regression	AUC	**0.5994**	0.5883	0.5828	0.5828	0.5811	0.5801	0.5797	0.5768	0.5746	0.5719
	F	0.1266	0.123	0.1207	0.1202	0.1208	0.1201	0.1202	0.1189	0.1181	0.1166
	Lift	1.5023	1.4266	1.5061	1.3869	1.4188	1.4053	1.4164	1.3869	1.3486	1.382
Logistic Regression	AUC	**0.5902**	0.5808	0.5749	0.5723	0.5713	0.5688	0.5677	0.5666	0.5629	0.5597
	F	0.1246	0.1199	0.1186	0.1184	0.1177	0.1173	0.1156	0.116	0.1146	0.1128
	Lift	1.4878	1.4564	1.3892	1.406	1.4278	1.3785	1.3707	1.3782	1.3548	1.3722
Random Forest	AUC		0.5611	**0.5638**	0.5626	0.5595	0.5607	0.5617	0.5607	0.5561	0.5494
	F		0.1116	0.1135	0.1115	0.1121	0.1118	0.1115	0.111	0.1084	0.1067
	Lift		1.2877	1.2886	1.2298	1.3259	1.2932	1.3164	1.2535	1.2021	1.2146

Table 2. Model Comparison

We resample the data with SMOTE algorithm. For each resampled data set, we set the total data amount to be 60000, over sample the rare class and down sample the dominating class. Besides the original percentage of rare class, nine percentages from 10% to 90% are implemented. 5-folds cross-validation is performed to compare different models. Namely, each time we resample a 60000-row new data set from 80% (4 out of 5-folds) of the original data, fit the model on the resampled data, and then evaluate the performance on the test data, which is 20% (1 out of 5-folds) of the original data.

We compute the average AUC, F1 score and Lift of the 5-folds cross validation. Since the AUC has nothing to do with probability threshold, we mainly use average AUC to evaluate the performance of the models.

From the Table 2, Lasso regression achieves the highest AUC at original rare class ratio, which means resampling doesn't bring any improvement. The result is a little bit weird since for rare event classification problem, over-sampling and down-sampling usually have a positive effect in classification. Probably in this data set, resampling doesn't amplify the actual potential pattern between the features and response variable (although it's not guaranteed if any significant pattern exists).

The other two models fitted with SMOTE resampled data are logistic regression and random forest. As expected, lasso regression outperforms logistic regression at each sampling percentage, which verified the regularization term in lasso regression can reduce the overfitting and improve performance in test data. Within different sampling percentage, random forest performs best at around 20% and 30%. The result is intuitive since balanced random forest use Gini and Information gain as criteria to measure impurity, which can't handle imbalanced classes problem. Therefore, unlike lasso and logistic regression, random forest doesn't perform good at original percentage and 10% percentage. However, the outcome of random forest unexpected, which is generally worse than lasso regression. Random forest performs perfectly in training data, but fail in testing data. More future work need to be done to check how does random forest over fitted on resampled data, but fails to extract the actual pattern (if it exists).

2. Lasso regression, Ridge regression, Elastic Net

	Ridge	Elastic Net									Lasso
Alpha	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
AUC	0.5932	0.5954	0.5996	0.5969	0.6004	**0.6012**	0.6001	0.6	**0.6012**	0.5998	**0.6011**
F	0.1249	0.1253	0.1266	0.1256	0.1268	0.1272	0.1257	0.128	0.1282	0.1277	0.1273
Lift	1.4608	1.4718	1.4583	1.4504	1.4928	1.4739	1.4785	1.5304	1.4884	1.5184	1.4817

Table 3. Comparison of Ridge, Elastic net and Lasso

We continue to tune the parameter of penalty in elastic net. Elastic net combines both the regularization term of lasso and ridge. The parameter α adjust the weights of lasso penalty and ridge penalty in elastic net. If α equals 0, elastic net is the same as ridge, and if α equals 1, elastic net is the same as lasso, while If α is the value between 0 and 1, it contains both lasso and ridge penalty.

Since Lasso regression has the best performance in original data without sampling, we assume so are Ridge regression and Elastic net. We fit the Elastic net model on the original data with 11 α values from 0 to 1 with step 0.1. Also, 5-folds cross validation are performed.

From the outcome, lasso performs at the best level. The elastic net also performs good when α equals 0.8 and 0.9. We went further to do 10-folds cross validation and found there is no obvious difference between the α 0.5, 0.8 and 0.9. So we conclude Lasso regression and Elastic net with α close to 1 performs a little bit than Ridge regression.

3. Other models: Naïve Bayes and K Nearest Neighbors

We assume the original ration of rare event is also good for other models. We fit Naïve Bayes and K Nearest Neighbors on the original data.

	Naïve Bayes	KNN (k=7)
AUC	0.5502	0.4885
F	0.1095	
Lift	1.2112	

Table 4. Outcome of Naïve Bayes and KNN

Since the AUC of Naïve Bayes and KNN are significantly smaller than that of Lasso regression, we didn't continue working on these two models.

3.3.4 Results: predict continuous variables

1. Discussion

Another task is to predict top 350 people who donates the most, which is a regression to predict the continuous value of donation amount. For regression, we usually use MSE (mean square error) to evaluate how good a model is. However, like the criterion Accuracy in Classification, we think MSE isn't a good evaluation metric for rare events in regression if we care mainly on the exact value of continuous response variable, since one model can simply predict all donation to be zero and get a possible low MSE. In other words, using MSE as metric for rare event assign equals weight on all samples no matter whether one donates and how much one donates. But actually we care more about the people who donate the most, and probably there are some methods to sacrifice MSE to predict the donation amount of top ones more accurately.

However, if we only have to predict top 350 people and don't care the exact donation value, the problem is actually another classification case in which the top 350 can be labeled and Positive and the rest are Negative. Therefore, the new classification task becomes more imbalanced, as only 0.3668% (=350/95412) of all samples are positive. It should be more difficult to predict than previous task in which we only predict whether a person donate.

2. Lasso, Linear Regression and Random Forest

We split the original training data into two parts, 60% used as training data and the rest 40% used as validation data. We fit the model on training data and predict the response amount of validation data. We pick the top K number of predicted values where K is from 50 to 1000 with a step 50, then compare the predicted value to actual value and compute how many top K customers were predicted correctly. (Due to the long computation time of random forest, no cross-validation is performed here)

	LASSO	LINEAR REGRESSION	RANDOM FOREST	OVERLAPS 1	OVERLAPS 2
TOP 50	2	2	0	2	0
TOP 100	2	3	1	2	1
TOP 150	**5**	**3**	1	3	1
TOP 200	6	5	2	3	1
TOP 250	9	8	4	6	3
TOP 300	11	11	8	9	4
TOP 350	14	12	10	9	5
TOP 400	15	15	15	10	7
TOP 450	17	16	15	12	7
TOP 500	23	18	18	16	8
TOP 550	25	19	19	16	8
TOP 600	28	21	20	17	9
TOP 650	31	25	21	20	9
TOP 700	33	29	21	22	9
TOP 750	35	33	25	23	9
TOP 800	39	36	28	23	10
TOP 850	43	37	30	24	10
TOP 900	44	40	30	26	10
TOP 950	46	43	34	27	12
TOP 1000	48	46	34	29	13

Table 5. Compare Top K customer prediction of three models

According to the output, for almost all the top K customers, Lasso outperformed Linear regression and Random Forest. Since our goal is to predict Top 350 customers in test data and the test data has almost the same rows as original training data, roughly we need predict Top 140 customers in the validation data (350*0.4=150, where 0.4 is the percentage of validation data). From the table, for the Top 150 customers in validation set, Lasso can predict 5 while regression can only predict.

3. Explanation

Another interesting point is in Top 100, Lasso predicted one less than Linear regression. We think that's because the Lasso is a shrinkage method, which forces the coefficients to decrease to avoid

overfitting. So Lasso may not perform good in predicting the observations with extreme value. While linear regression doesn't have this problem. Anyway, in the following top K customers, Lasso wins with no doubts.

The last two columns represent overlaps between the prediction of Lasso and Linear regression, Lass and Random forest separately. We found Lasso and Linear regression have a large overlap, while that between Lasso and Random forest is small. The result indicates, although Random forest performs worse, it figured out some other patterns different from Lasso. In the future work, we could ensemble Lasso, Random forest and other models to improve the predicting performance.

3.3.5 Final model recommendation:

According to the models we fitted, Lasso regression performs better both in predicting whether a customer donates and predicting the Top K customers who donated the highest amount.

4. Interpretation

4.1 Interpretation of Lasso coefficients.

The neighborhood code which stands for the types of human settlements and socio-economic status is very influential to the donation behaviors. We expect that people in higher socio-economic status are more likely to donate. The coefficients of the model generally agree with this assumption. To be more specific, the suburban average and the town average are positively related while all lower status has negative effect on the donation behavior. It is also interesting to find out that the relative social status within the settlement types may be more important than the "absolute social status", people from the suburban lowest should be wealthier than the town average which expected to be more likely to donate while the result shows the opposite. Secondly, the human settlements type is also informative, people from the suburban and town seems to be more willing to donate than those from other areas. We think the life pace and amount of messages received by individuals could play a possible role here. People from the urban and city areas are always in an intensive living style and receiving tons of messages like ads, notices and invitations, the solicitation are more likely to be neglected by them. The situation comes from the opposite in the rural area. People may think too much and be reluctant to donate while they seldom receive this kind of messages. Thus, we conclude that people with modest life pace and familiarity but not fed up with the promotion message might response more actively. However, we need more direct information such as working load, monthly mail/email received to prove this point.

The number of child is negatively related and the income is positively related as we expected. They, however, are not highly correlated with donation as it is shown in Fig 8. After we consider the donation amount with no more than 200 dollars, this is rather a random kindness than a discreet economic behavior. We think the individual economic capability will surely have a somewhat positive effect on the donation behavior but not that decisive.

We unanimously agree that information of neighborhoods is vital in helping us launch donation activities afterwards. Before we start our analysis, it is reasonable for us to assume that the level of the community in which the donors live can be represented by the neighbors of the donors. Among the results of all estimated coefficients, we can notice something interesting about the percentage of neighbors from different nationalities. First of all, it seems that most nationalities turn out to be insignificant. Secondly, the estimated coefficient of *Percent of Black* is -0.027, indicating a negative effectiveness of black neighbors, which may be explained by the level of income of black people. Another interesting result is about Chinese neighbors, of which the estimated coefficient is -0.027. One possible reason can be related with the fashion arising gradually in Chinese people of studying and working in America. With a high level of mobility, people may not tend to denote to a country in which they would not stay for a long time.

Another significant result is about the *Percent Households with Person 65+ Living Alone*, with an estimated coefficient equal to 0.051, indicating a positive influence of old neighbors. With a high percent of old people in neighborhood who live alone, it is possible for the donors to be willing to accompany and help them or help each other in his/her daily life, and thus more likely to denote to help other strangers.

The estimated coefficient of AverageHomeValue is 0.102, showing a great positive significance. Since the home value and the level of incomes are closely related, it is reasonable to expect the average home value to be very important. In addition to the nationalities of neighbors, we can also obtain some information through analyzing the estimated coefficients of the characteristics of the level of education. Specifically, the estimated coefficients of *Median Years of School Completed by Adults 25+, Edu1, Edu3* are -0.029, -0.01, and -0.038, separately, suggesting that people are less likely to denote with neighbors poorly educated, which is also related to the level of average income of the community. The last two significant variables are *Percent Foreign Born* and *Percent Other Language Speaking,* of which the estimated coefficients are -0.042 and -0.018. These two

variables can be related to the nationalities as we discussed before. Again, non-citizens may not be willing to denote to the country they would leave sooner or later.

All in all, the correlations between donation and the donor's neighborhood/community are mainly reflected in three aspects. Firstly, non-citizens seem to be less likely to denote, concerning the high mobility of those people. Secondly, donors living in the community with highly educated neighbors are more willing to denote. This is actually related to the level of average income of the community. Lastly, if there are many old people living alone in neighborhood, donors would be more likely to denote to help others.

It should be highlighted that the data were collected in 1998, which was 18 years ago. There may be somewhat close relationship among neighbors at that time, however, communities gradually play a decreasingly important role in people's life concerning radical change that our society undergoes in recent years. Hence, to obtain more accurate conclusions and predictions, researchers should update the features which are being more concerned and collect more recent data.

We are also interested in the relations between donators' history profiles and amount donated. Donators' history profiles can provide informative and accurate analysis to help our market teams to initiate their strategic marketing plans. Firstly, *LifeGiftNum*, the number of lifetime gifts to date, is positively correlated to the amount of donation, indicating that the donators who give more gifts are more likely to donate in the future. In this sense, in order to retain those donators, we suggest our market team keep their updated contact information and distribute our pamphlets to them periodically. Secondly, *LifeGiftPromNum*, the number of lifetime gifts to promotions to date, is highly correlated to our target variable. In another word, the larger of number of lifetime gifts to promotions to date, the larger amount of donation will be donated. Overall, the people with higher frequency of donation are more likely to donate in the future. There are several possible reasons to explain this. One might be that these people are philanthropic and prefer to donate in a fixed time. Therefore, keeping tracking of records for our previous donators is of great importance so as for our market teams to narrow down the potential future donators in the limited marketing budgets with higher probability of getting donations. Thirdly, *MinDol*, the dollar amount of smallest gift to date, is negative correlated to our response, which makes sense for the reason that the larger lower bound for the people to donate before, the fewer chances for those people to donate later. We don't have to exert much effort and invest much time in persuading those people whose smallest amount of donation is relatively larger than others to donate. Fourthly, the dollar amount

of most recent gift plays a vital role in predicting the amount of future donation. The larger amount of donation that people gave as most recent gift, the less likely for those donators to donate in the future. Perhaps, in out future design of survey, we can create some new questionnaires to dig the reasons why those people are not prone to donate and come up with solutions to address this issue. Moreover, *AvgDol*, the average dollar amount of gift to date, is also in a negative relation with the amount donation, which is in accordance with *MinDol* and *LastDol*. There is a weak negative relationship between Month of First Gift and the amount of donations, indicating that there is no significant difference between the old donators and new donators in amount donations. From this variable, we can conclude that our marketing team should not hesitate to distribute promotions and contact the old donators because they exert similar effects on the response variable compared to our new donators.

4.2 Segmentation and marketing standpoint.

In this section, we choose several most significant variables and split the observations into different group depending upon the features of the variables. Specifically, we divide the variables into three categories, social status involving variables Neighborhood Code, economic capability involving Income and Average Home Value, and historical donation behavior involving Dollar Amount of Most Recent Gift and Number of Lifetime Gifts to Promotions to Date.

Following calculations are used as criterion for the segmentation:

% Donated: Proportion of donor number in the whole data

% Within: Proportion of donor number within the subgroup, showing the probability of people to donate

Amount/Total: Average donation amount per person, a criterion to reflect the rate of return from a business standpoint

Amount/Donated: Average donation amount per donor, showcasing the importance of donors in each level

4.2.1 Settlement type and socio status

The rural area has the lowest donation proportion in total but highest in all other three criteria, which indicate the rural area might be a potential segment for fundraising. The other three area have not much difference.

Socio status has a clear positive relationship both in total or within each area as we expected, and the donation proportion, average donation within area differs more significant than in total. This result seconded our lasso model coefficient very well.

Thus, we recommend a different fundraising strategy to different level of socio status within each settlement area might be a good idea and starting to pay more attention on the rural area than others is also necessary in future promotion plan.

4.2.2 Economic capability

For the variables Income and Average Home Value, we use original segments to set up the groups, and some obvious patterns can be found from the table. As described above, % within and Amount/Donated separately reflect the probabilities of donating and donation amount for people in each group. For Income, the results of % within and Amount/Donated are both increasing along with the increase of levels, indicating that people making higher income are more likely to denote and denote more on average. For Average Home Value, the results are similar with what of Income, as the average value of homes in neighborhood goes up, which also offers another perspective on the income level of the donors, the probability and amount of donation tend to increase accordingly. The results of the segmentation analysis are consistent with the interpretation for our model.

4.2.3 Historical donation behavior

Based on four criteria above and characteristics of LifeGiftPromNum, we divided our data into 10 groups with the range of 4 in each interval. As for percentage donated, there is an obvious decreasing trend. We suggest our marketing team send emails and update our activities or campaigns to those people whose lift gift promotion number is less than 24 frequently. Moreover, our main target should be the groups with interval (0,4), (4,8), and (8,12). But those people who donate the largest amount because of our promotions, should also be informed when we launch fundraising activities. Once they donate, their amount of donation will be larger than others and the rate of return from these group is the highest

On the basis of LastDol, we are interested in 7 groups. Amount/Donated witnessed a descending trend from group 1 to group 7, indicating that most people hesitate to donate large amount of money or gifts. Therefore, we can cover more people to donate to increase our total donation amount rather than targeting fewer people who donate a lot.

References

1. Bahn, G. D., and Massenburg, R. (2008) Deal with Excess Zeros in the Discrete Dependent Variable, the Number of Homicide in Chicago Census Tract. Joint Statistical Meetings, Social Statistics Section.
2. Burez, J., and Poel, D .V . D. (2009). Handling Class Imbalance in Customer Churn Prediction. Expert Systems with Applications, 36, 4626–4636.
3. Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002) SMOTE: synthetic minority over-sampling technique, Journal of Artificial Intelligence Research, v.16 n.1, p.321-357.
4. Chen, C., Liaw, A., Breiman, L. (2004). Using random forest to learn imbalanced data. (Tech. Report No. 666). Retrieved from Department of Statistics, University of California, Berkeley. website: http://www.stat.berkeley.edu/tech-reports/666.pdf
5. Diving into data. (n.d.). Retrieved May 27, 2016, from http://blog.datadive.net/selecting-good-features-part-i-univariate-selection.
6. Diving into data. (n.d.). Retrieved May 27, 2016, from http://blog.datadive.net/selecting-good-features-part-ii-linear-models-and-regularization.
7. Diving into data. (n.d.). Retrieved May 27, 2016, from http://blog.datadive.net/selecting-good-features-part-iii-random-forests.
8. Huang, P. J. (2015) Classification of imbalanced data using synthetic over-sampling techniques. University of California, Los Angeles.
9. Muchlinski, D., Siroky, D., He, J., and Kocher, M. (2015). Comparing Random Forest with Logistic Regression for Predicting Class- Imbalanced Civil War Onset Data. Political Analysis, pp. 1-17.
10. Ridout, M., Demetrio, C. G. B., and Hinde, J. (1998). Models for count data with many zeros. Invited paper presented at the Nineteenth International Biometric Conference, Capetown, South Africa.
11. Weiss, G. M. (2004) Mining with rarity: a unifying framework, ACM SIGKDD Explorations Newsletter, v.6 n.1.