Case Competition for Fundraising Project

Team Members:

Xiang Cao

Haitao Liu

Liwen Tong

Yang Yu

Content

Introduction

Data Description

- More than 100 variables
- Over 90,000 observations

Feature Categories

- General Information of Donators
- Neighborhood Information
- Donation History

Goal

- Prediction
- Segmentation(Marketing Recommendation)

Data Overview

Data cleaning

Convert Date of First Gift into a new numerical variable

- e.g. 8901 → 107

Convert Zip Code

- convert 5% negative into positive
- add zero back: e.g. 1002 → 01002

Neighborhood Code

- create a new factor for 2.5% of missing Neighborhood Code

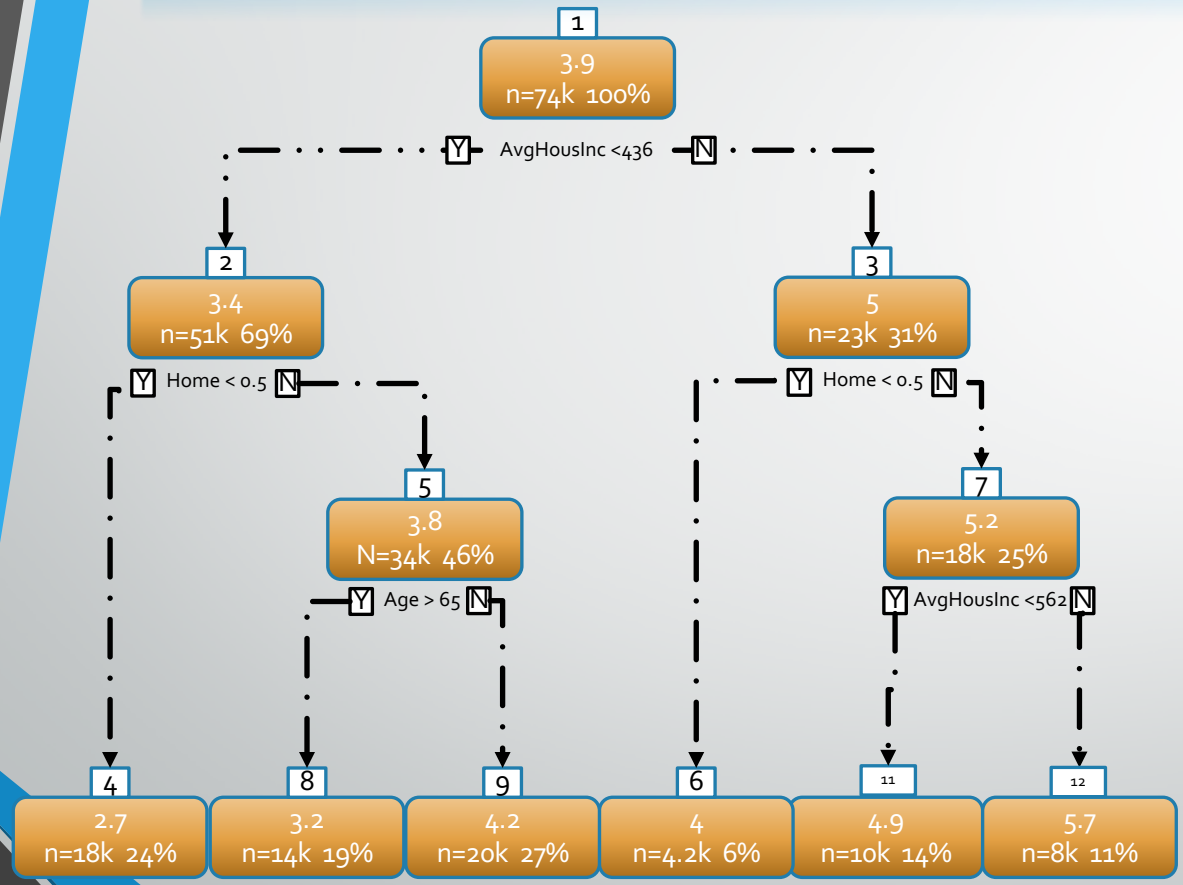
Gender

- Convert A & C into the dominated level Female

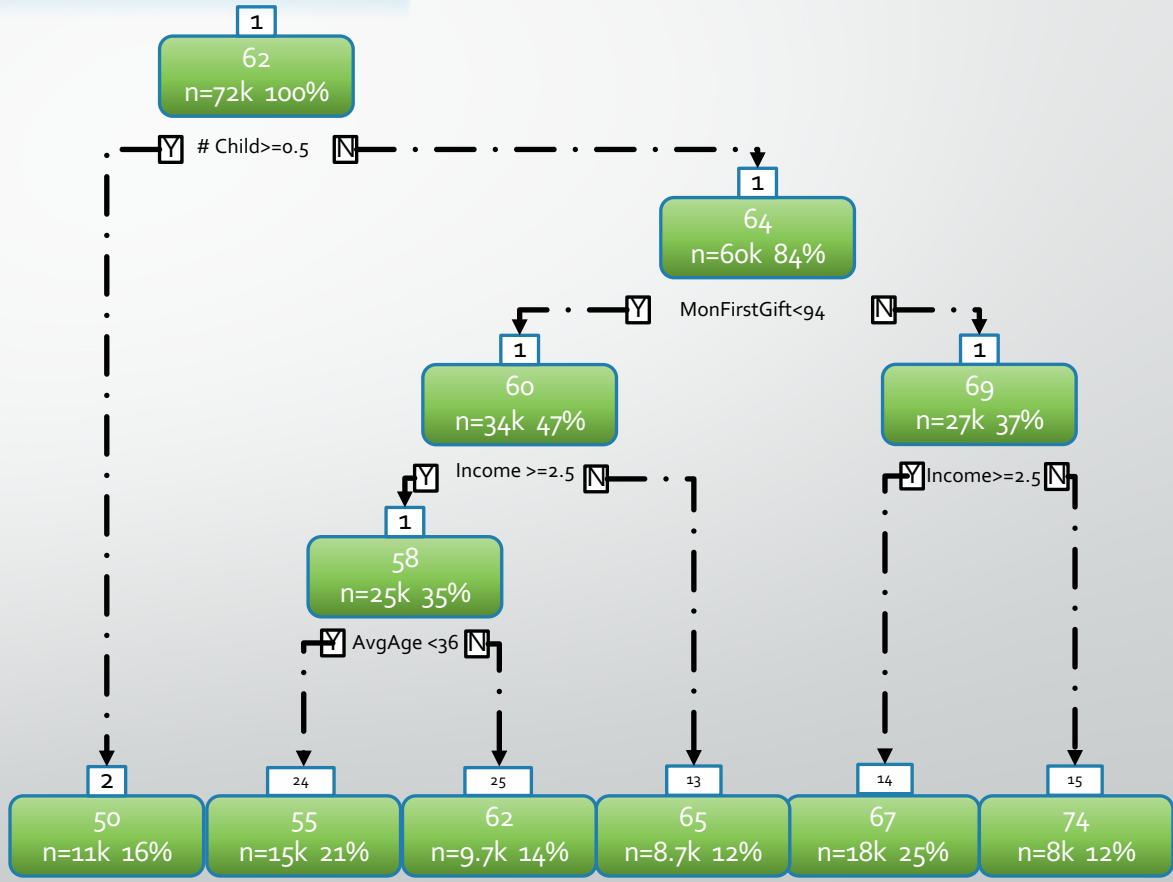
Data cleaning

Data Overview

Prediction Missing Values of Income and Age



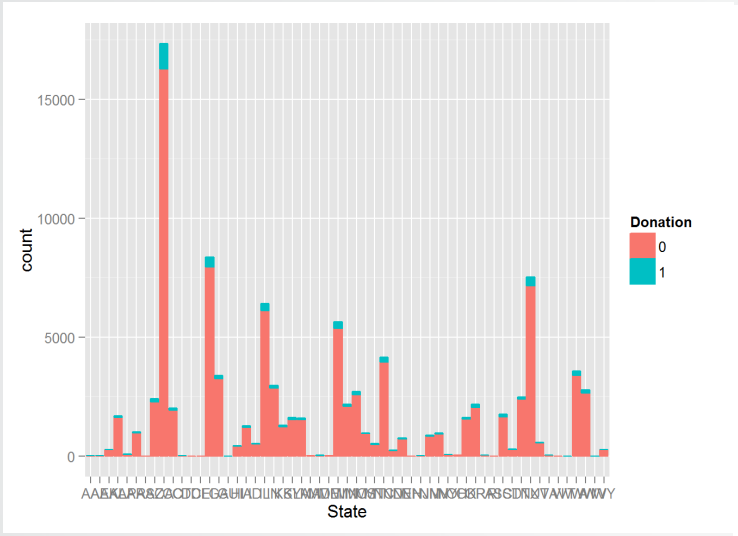
Regression Tree for Income

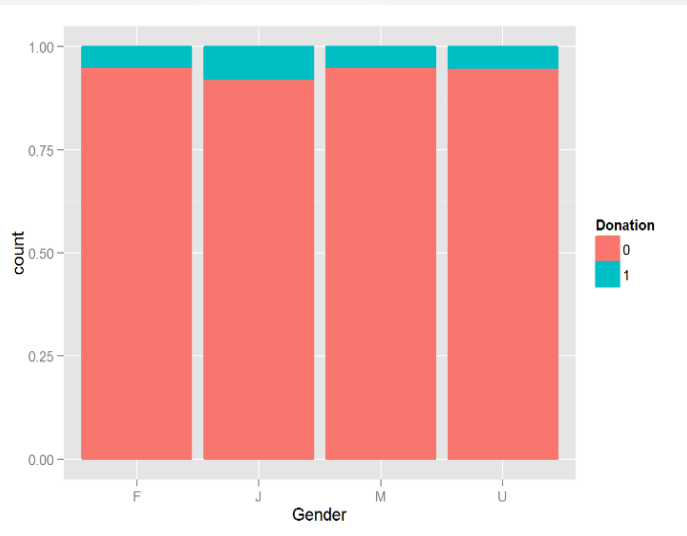


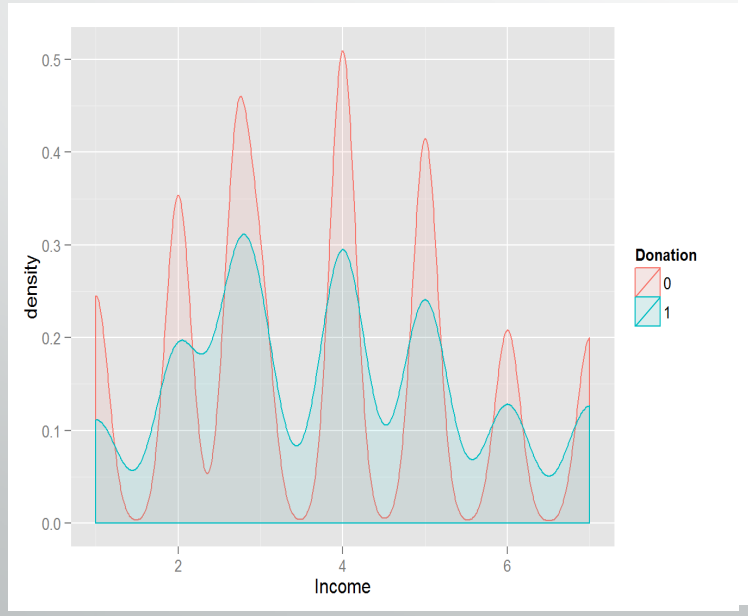
Regression Tree for Age

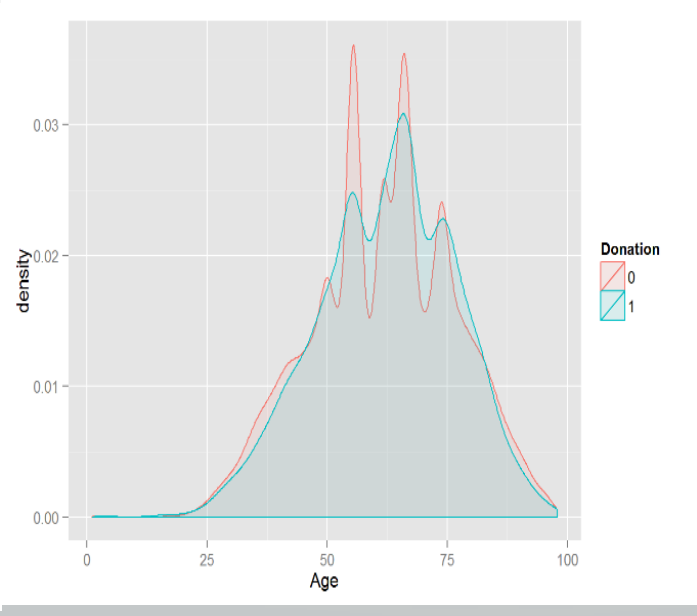
Data Overview

Descriptive statistics



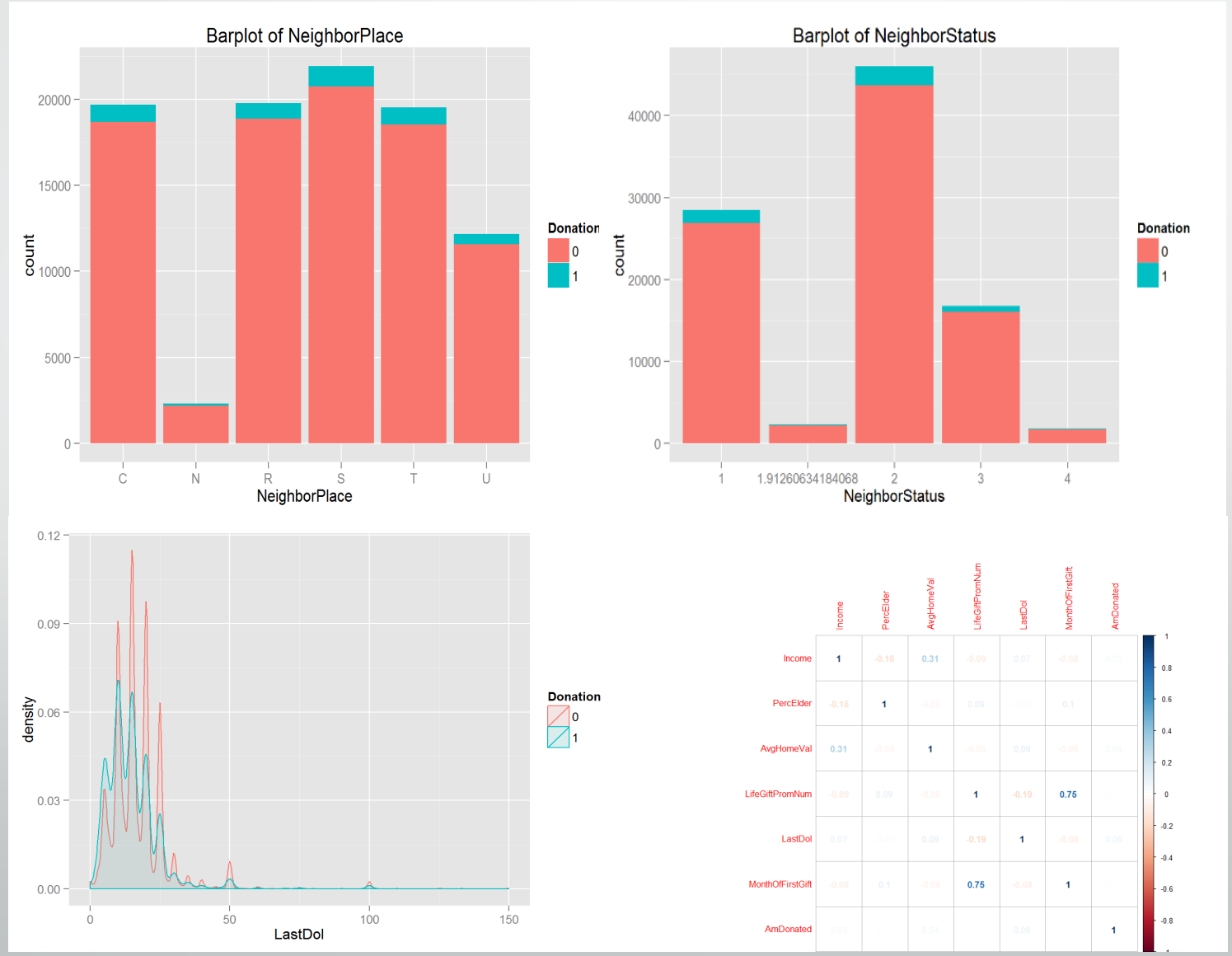






Data Overview

Descriptive statistics



Data Overview

Dealing with Rare Event: Metrics

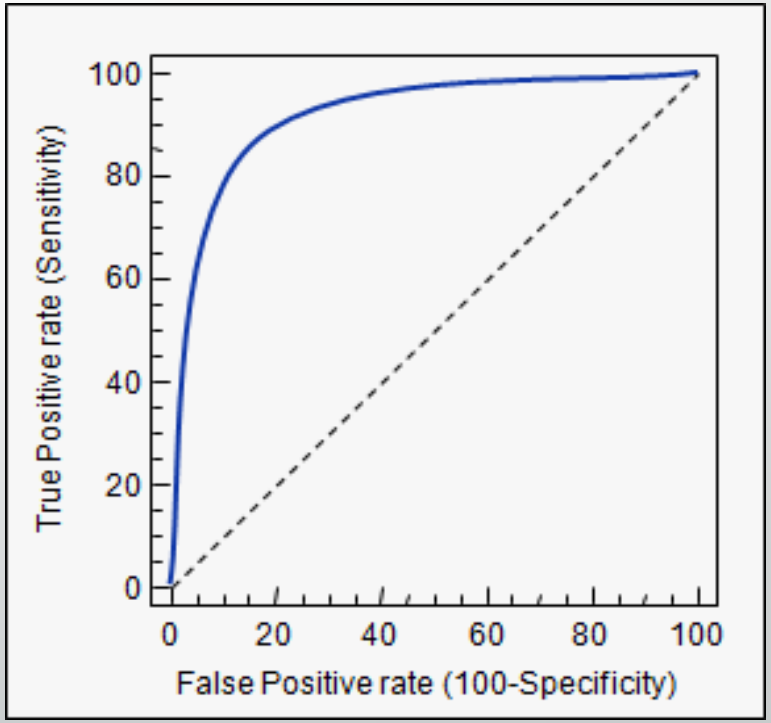
Lift & F1 Score

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

Precision

Sensitivity

AUC/ROC Curve



Data Overview

Dealing with Rare Event: Sampling

Cons for Over-sampling:

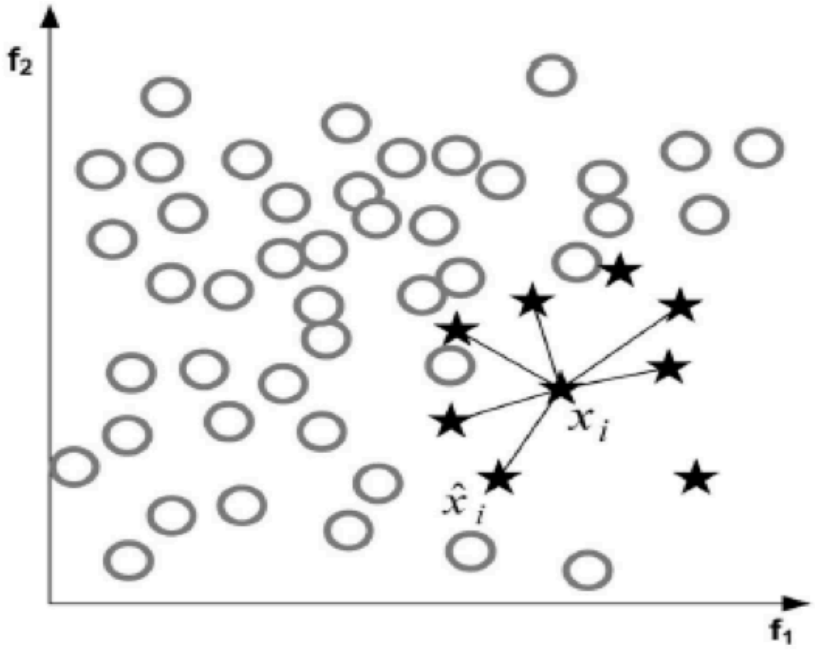
- Just duplication and no additional information
- Overfitting

Cons for Under-sampling:

- Risk of removing some representative observations

Synesthetic Minority Over-Sampling Technique (SMOTE) & Under-sampling

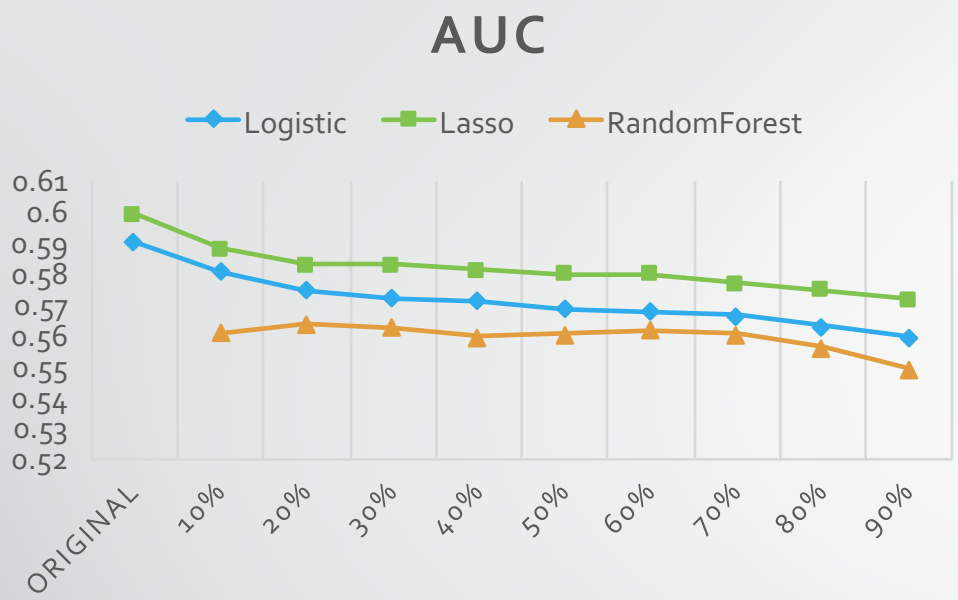
$x_{new} = x + rand(0,1)*(\hat{x}\text{-x})$

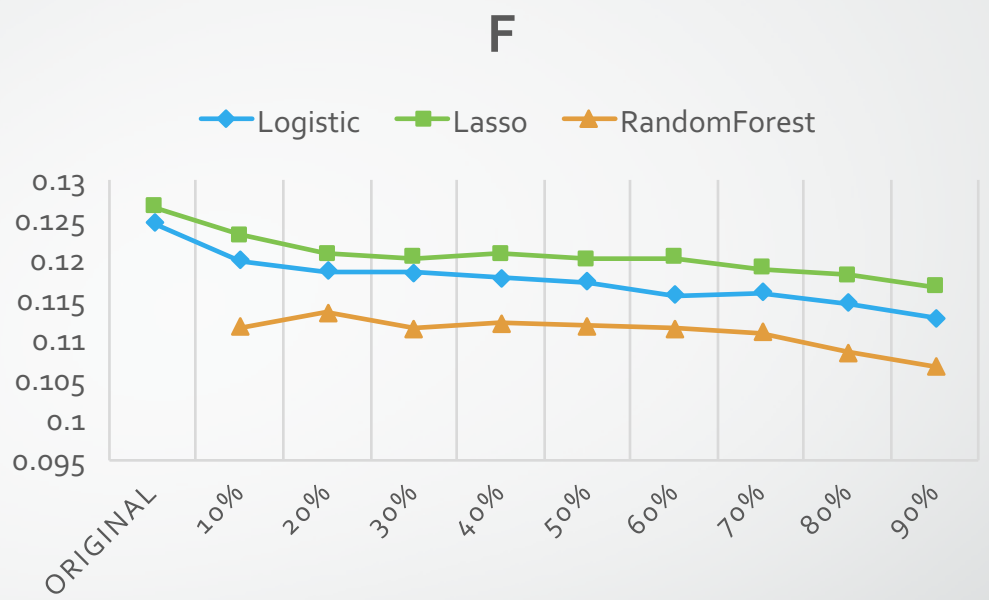


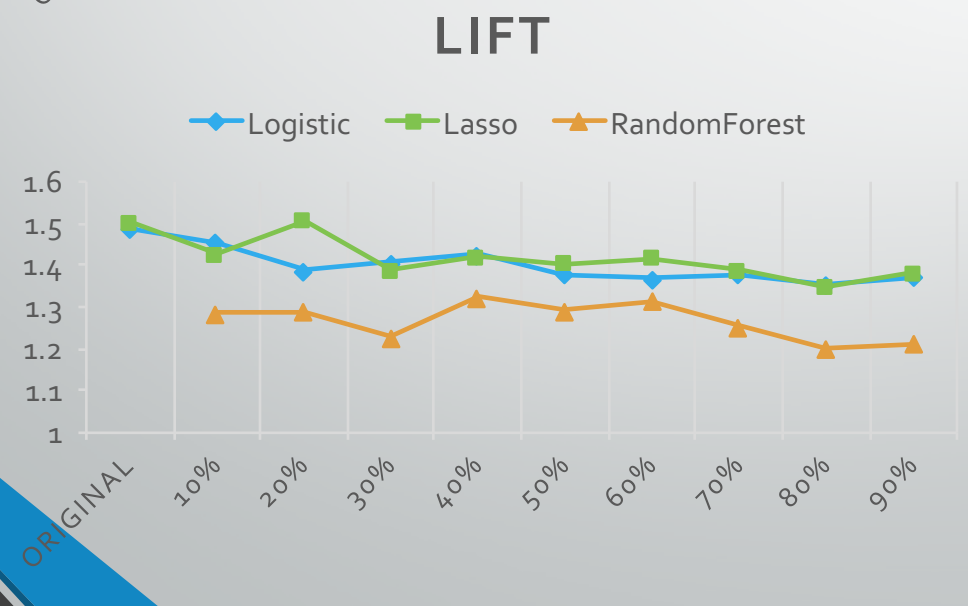
Methods Comparison

KNN	Classify a given test observation by identifying its K near neighbors
Naïve Bayes	$\Pr(Y = k \mid X = x) = \frac{\pi_k \times \Pr(X = x \mid Y = k)}{\sum_{l=1}^{K} \pi_l f_l(x)}$
Logistic	$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}$
Ridge & Lasso	$RSS + \lambda \sum_{j=1}^{p} \beta_j^2$ & $RSS + \lambda \sum_{j=1}^{p} \lvert\beta_j\rvert$
Random Forest	A tree-based classifier using a random subset of predictor at each split

Methods Comparison

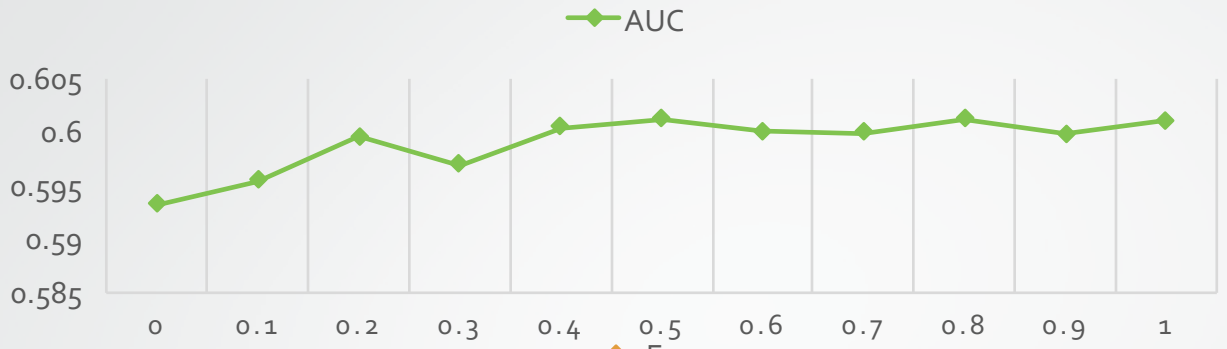


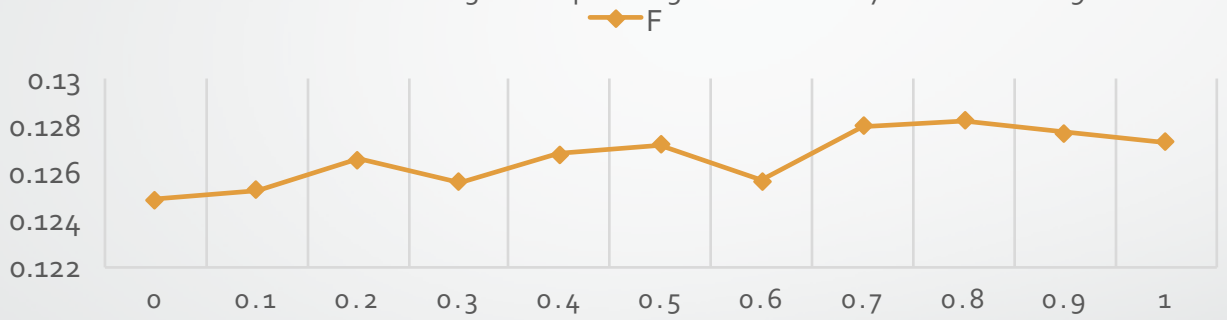


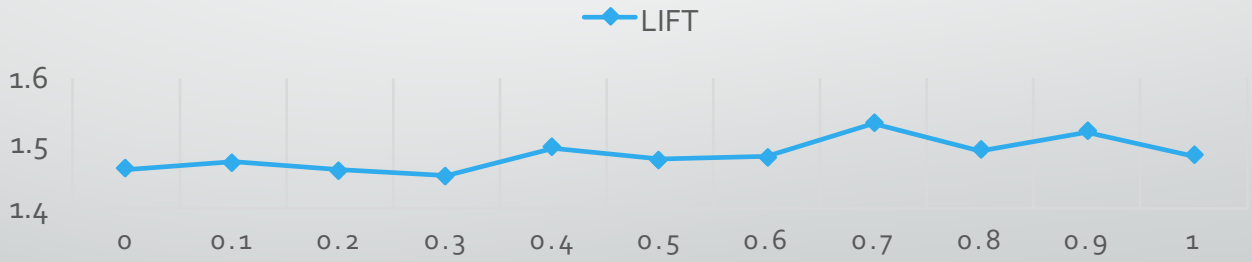


	NAÏVE BAYES	KNN(K=7)
AUC	0.5502	0.4885
F	0.1095	
LIFT	1.2112	

Methods Comparison

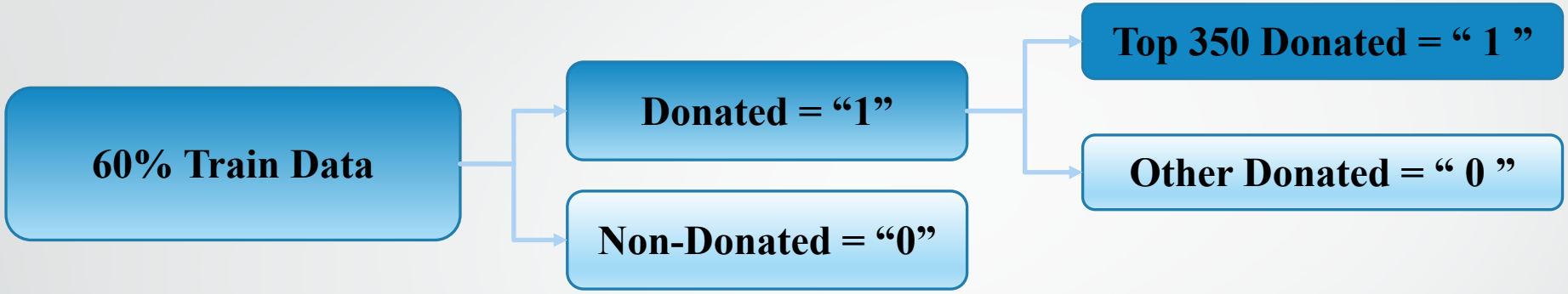






Methods Comparison

New classification for Top 350



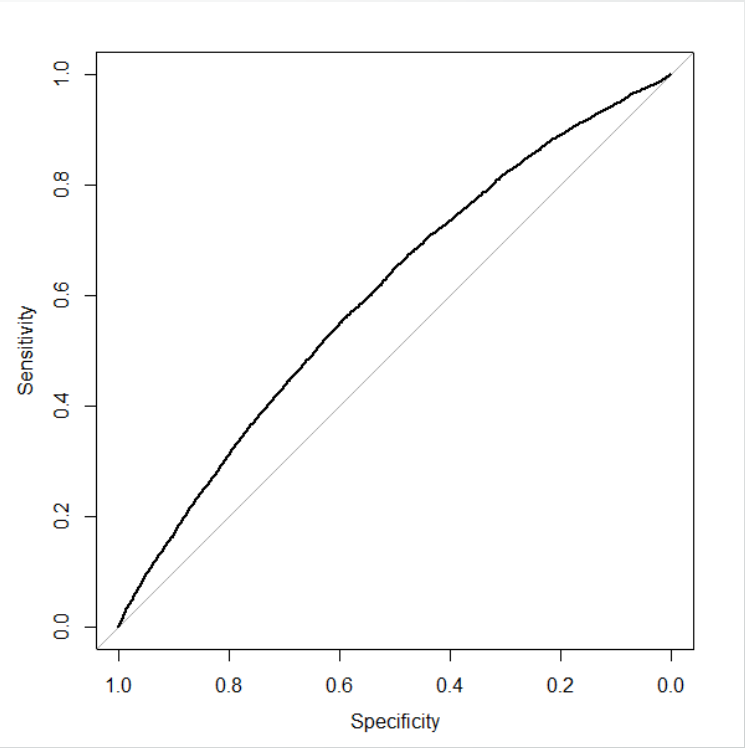
	Lasso	Regression	Overlaps 1	Random Forest	Overlaps 2
Top 50	2	2	2	0	0
Top 100	2	3	2	1	1
Top 150	5	3	3	1	1
Top 200	6	5	3	2	1
Top 250	9	8	6	4	3
Top 300	11	11	9	8	4
Top 350	14	12	9	10	5
Top 400	15	15	10	15	7
Top 450	17	16	12	15	7
Top 500	23	18	16	18	8

Result & Interpretation

Donation prediction

AUC	0.6013304
F1 SCORE	0.1249
LIFT	1.4896

		PREDICTED	
		0	1
ACTUAL	0	69594	21900
	1	3089	1784



AUC: 0.6013304

Result & Interpretation

Top 350 prediction

Custom ID	Donation	AmDonated
No.185147	1	500
No.185056	2	250
No.4006	4	200
No.12595	6	200
No.185094	7	112
No.8391	8	100
No.12333	11	100
No.5720	12	100
No.5050	13	100
No.11835	15	100
No.12592	17	100
No.137965	19	100
No.4206	21	100
No.8844	26	100
No.5660	30	100
No.14463	42	55
No.12367	73	50
No.6136	84	50
No.12284	109	50

$$\#Ratio = \frac{19}{350} = 5.4\%$$

$$Ammout\ Ratio = 15.2\%$$

Interpretation

Selected features with coefficients of Lasso Model

Customer Information	
City Low	-0.0545
Rural Average	-0.0757
Suburban Average	0.1177
Suburban Low	-0.0978
Town Average	0.0510
Town Low	-0.1520
Urban Average	-0.0912
Urban Below Average	-0.0285
Urban Low	-0.0265
Age	-0.0193
NumberOfChild	-0.0421
Income	0.0686
GenderJ	0.2939

Donation History	
LifeGiftNum	0.0264
LifeGiftPromNum	0.1519
MinDol	-0.0221
LastDol	-0.2246
AvgDol	-0.0154
MonthOfFirstGift	-0.0060

Census Information of Neighborhood			
PercEmpState	0.0120	IC3H	-0.0080
PercEmpFed	-0.0154	IC6H	0.0104
Persons	-0.0105	IC7H	-0.0239
PercMale	0.0172	IC2F	0.0046
PercWhite	0.0023	IC5F	0.0005
PercBlack	-0.0270	LFAM	0.0246
PercAsianInd	-0.0013	PercManage	-0.0234
PercChin	-0.0271	PercCler	-0.0050
PercPhil	-0.0084	PercFarm	0.0258
PercKor	0.0160	PercTran	0.0084
PercViet	-0.0031	PercLab	-0.0040
PercHaw	0.0023	EMP2	0.0023
AgeC3	0.0066	EMP6	-0.0136
AgeC6	0.0091	EMP7	-0.0219
ChildC1	0.0007	EMP8	0.0019
ChildC2	0.0157	EMP10	-0.0075
PercElder	0.0511	EMP11	-0.0070
PercSepDiv	0.0117	EMP13	-0.0183
PercSingle	-0.0297	MedEdu	-0.0291
AvgHomeVal	0.1019	Edu1	-0.0140
PercStateBr	-0.0418	Edu3	-0.0382
LangOth	-0.0178	Edu5	0.0275
		Edu6	0.0014

Market recommendation

- **Personal financial status plays an important role of donation.**

  (Income, Settlement type, Social Status)

- **Focus on middle class within each settlement type.**

- **Pay more attention on frequency and amount of donation of individual.**

- **Individual information is more important than neighborhood census.**

  (Education, Profession, Working load, Household ,Income, Marriage, …etc.)

Future Improvement

Feature Engineering

- e.g. convert zip code into continuous variable via leave-one-out encoding

Ensemble Lasso & Random Forest & Gradient Boosting, etc.

Q & A