Ruffalo Noel Levitz Project

Team 1: Jisoo Lee, Jordan Kloewer, Ran An, Xuan Zhang

Today’s Agenda

Introduction

Data Gathering and Preparation

Modeling

Conclusion

Introduction

1. Assumptions
2. Challenges
3. Original Dataset

Assumptions

- No external factors
- Correct data
- Use original dataset
- Most important target variable

Challenges

Other Challenges

- Group members in different sections

Data Challenges

- Ignoring columns
- Missing values
- Incomplete rows

Modeling Challenges

- Too much time prepping data
- Unknown modeling techniques

Original Dataset

Name: Pledge
Number/Letter: 39, AM
Description: Pledge flag
Format of Column: Single digit (1 or blank)

Column Type in R: Integer
Number of Missing/NAs in R: 244787

Questions About Column: What data is currently used that would cause someone to have a "pledge flag?"

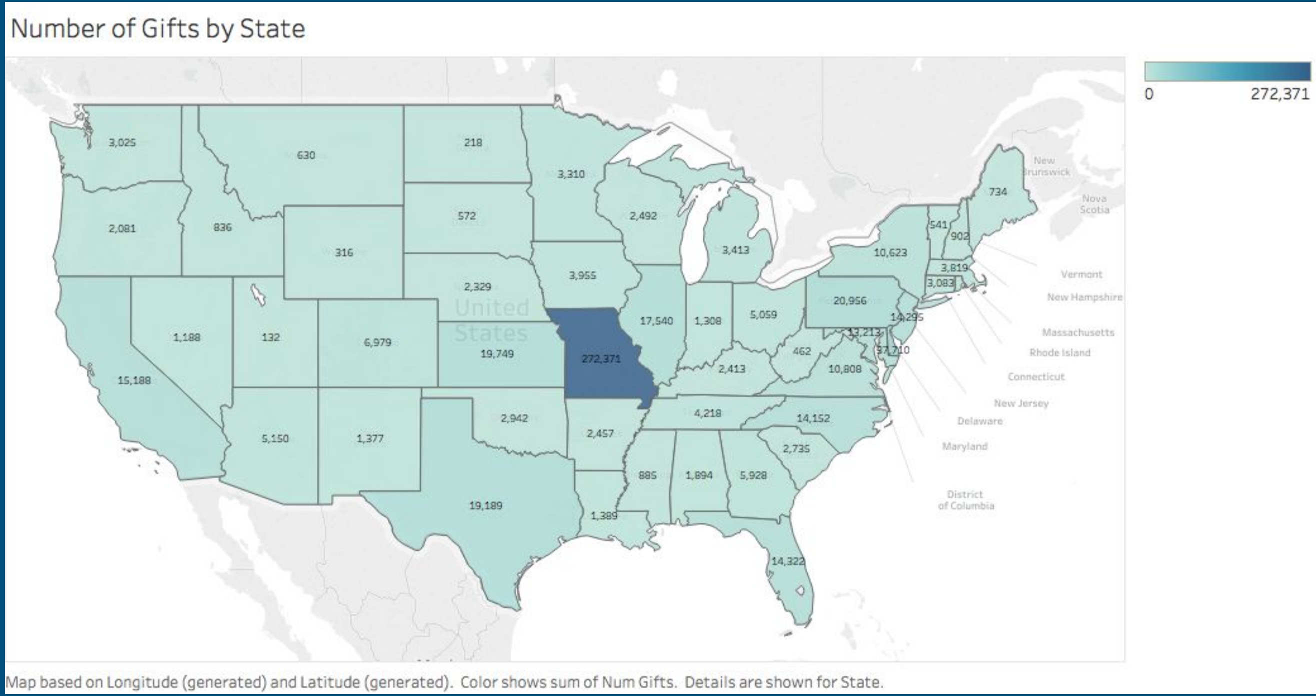
Any person with data populated in PledgeAmt (Column BI) and PledgeTot (Column BJ) (excluding where the pledge amounts/totals = 0) should be flagged as pledges. These fields are also numeric versions of the FY17_PLEDGE_AMOUNT column which was character in the original data for some reason

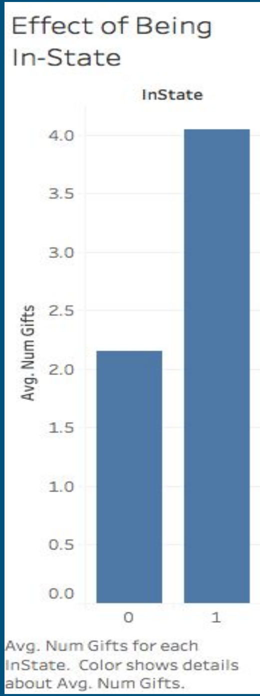
Any Similar Columns/Do Number Match Up?: Pldg

ACTION: DELETE - the column Pldg is the exact same but that one is formatted as a binary variable (0,1) with no blank lines

- 15,000,000+ cells
- Data Dictionary
  - Column descriptions
  - Number/text format
  - Column type in R
  - Missing values
  - Related columns
  - Questions
- Data Visualizations

Original Dataset





Data Gathering and Preparation

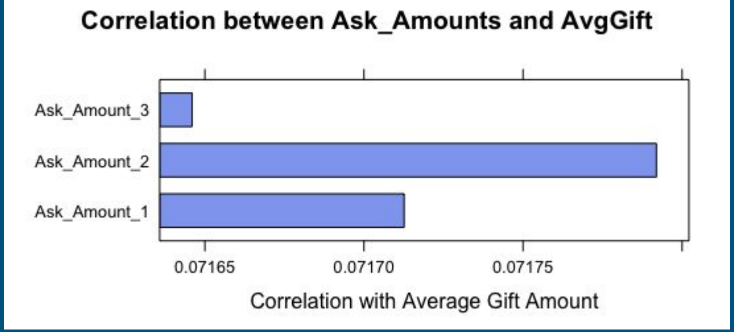
1. Data Cleaning
2. Data Selection
3. Dataset Used For Modeling

Data Cleaning

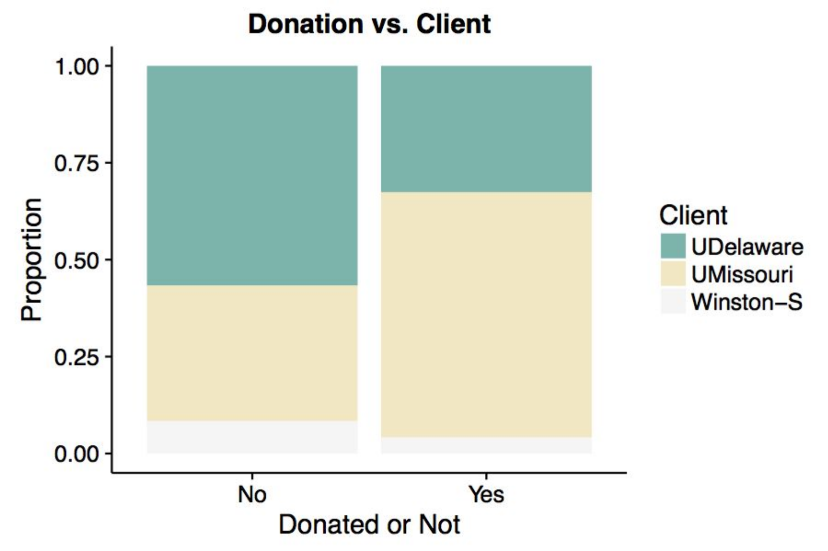
- After examination,
  - Removed errors and outliers: negative talk time and age below 3
  - Transformed some columns to improve interpretability
    - STATE: change to CRegion to reduce the number of levels
    - RecentGradYr, LastGradYr: change to YearSinceRecentGrad to have numeric values
  - Deleted 32 irrelevant and duplicate columns such as PHONE_RESEARCH, PAYMENT_TYPE
- Changed missing values to 0, except of AGE
  - Normally missing values of AGE are replaced to mean/median of AGE, but to avoid misleading impact of 53,161 NAs of AGE on models, separated dataset into:
    - 1) Dataset without NAs of AGE
    - 2) Dataset with replaced NAs of AGE to median

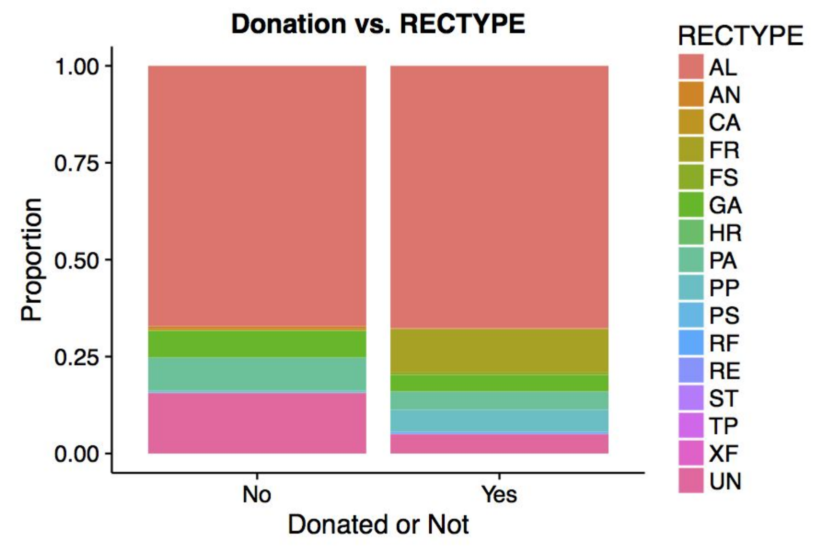
Data Selection

- For models to predict whether donate or not,
  - Transformed NumGifts into Donor01 to have binary values
  - Further removed 11 columns like FSTGFTA, FSTGFTD, LSTGFTD
  - Extracted 4 highly correlated columns:
    - AGE - YearSinceGrad & YearSinceRecentGrad
    - Ask_Amount_2 - Ask_Amount_1 & Ask_Amount_3

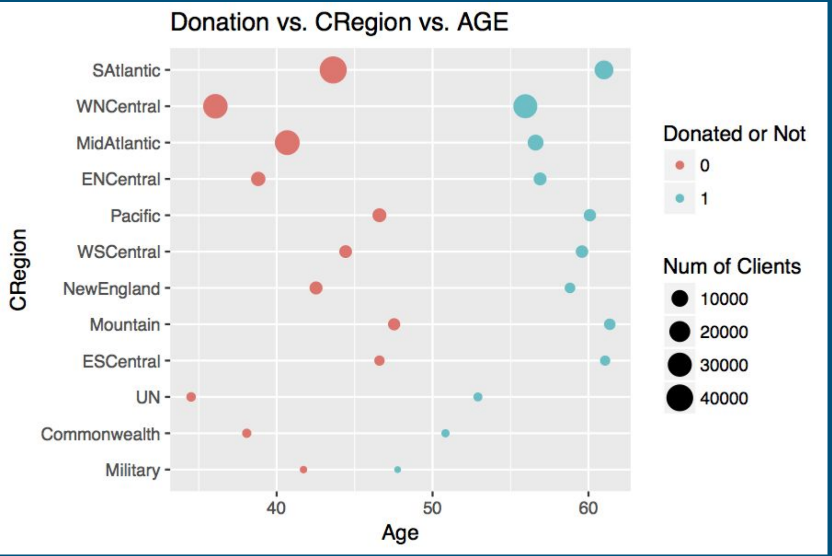


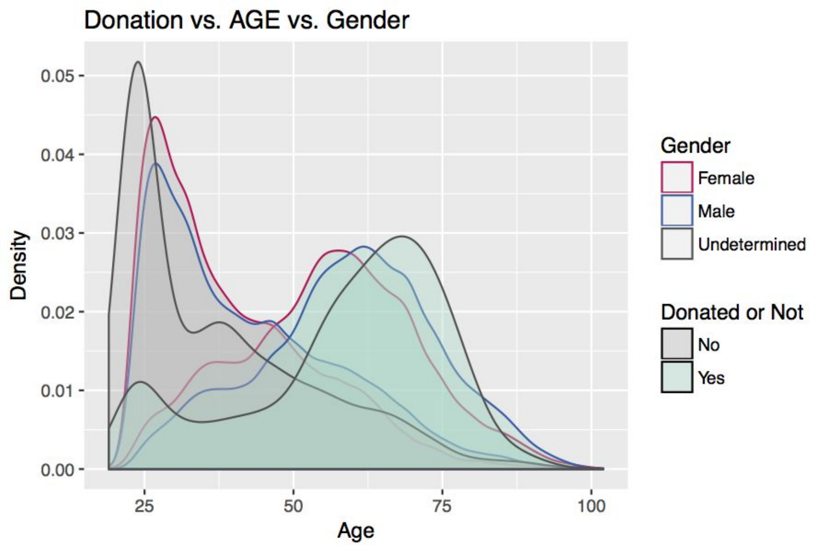
Data Selection



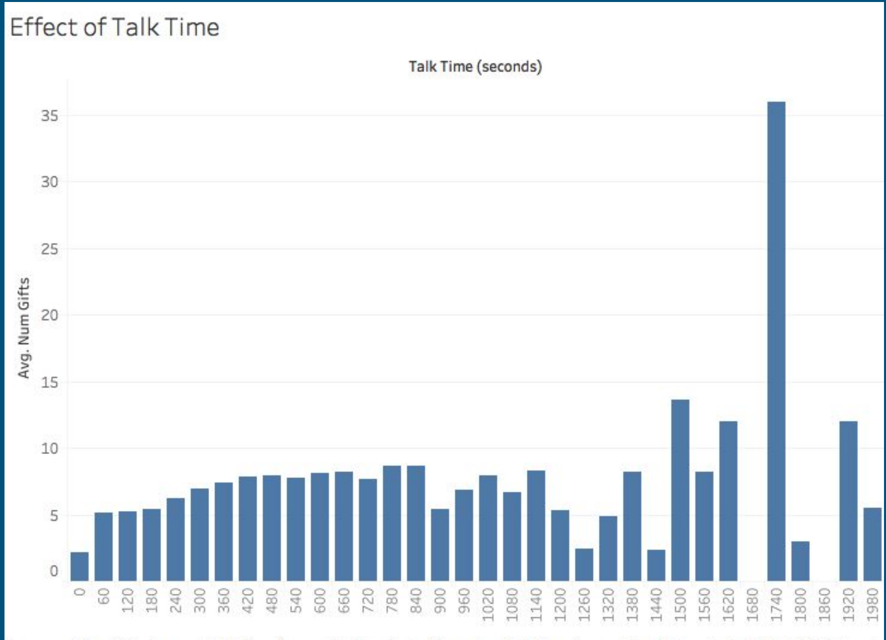


Data Selection

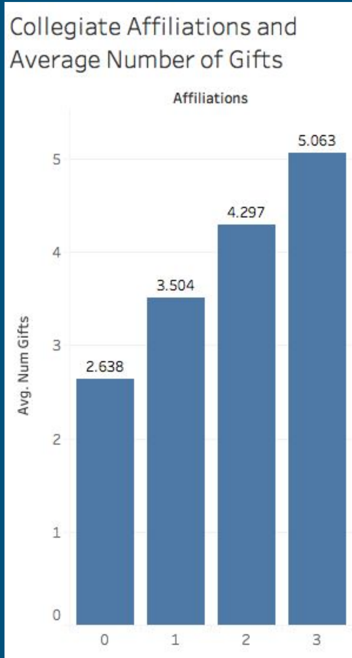




Data Selection



Average of Num Gifts for each Talk Time (seconds). The view is filtered on Talk Time (seconds), which excludes Null, -60, 2220 and 3120.



Average of Num Gifts for each Affiliations.

Data Selection

- Left 14 columns with 252,725 rows
  - Predictors: SCHOOL, AGE, RECTYPE, GENDER, NumDegrees, SuccCont, AffilCount, client, Ask_Amount_2, InState, TotAttempts, TALK_TIME, and CRegion
  - Target: Donor01

Dataset Used for Modeling

- Had multiple datasets that we could use:
    - Depending on AGE:
        - trainset_rmAGE - Dataset without NAs of AGE
        - trainset_medAGE - Dataset with replaced NAs of AGE to median of AGE (45)
    - Depending on client, (no Winston-S due to too many incomplete rows):
        - trainset_UD - Dataset for UDelaware
        - trainset_UM - Dataset for UMissouri
    - Depending on balancing methods that we used*:
        - trainset_under - Dataset with undersampling target variable
        - trainset_both - Dataset with over and undersampling target variable

*undersampling: reduce the number of majority class in target variable
*oversampling: increase the number of minority class in target variable

Modeling

1. Hypotheses
2. Modeling Techniques
3. Results
4. Evaluation Criteria

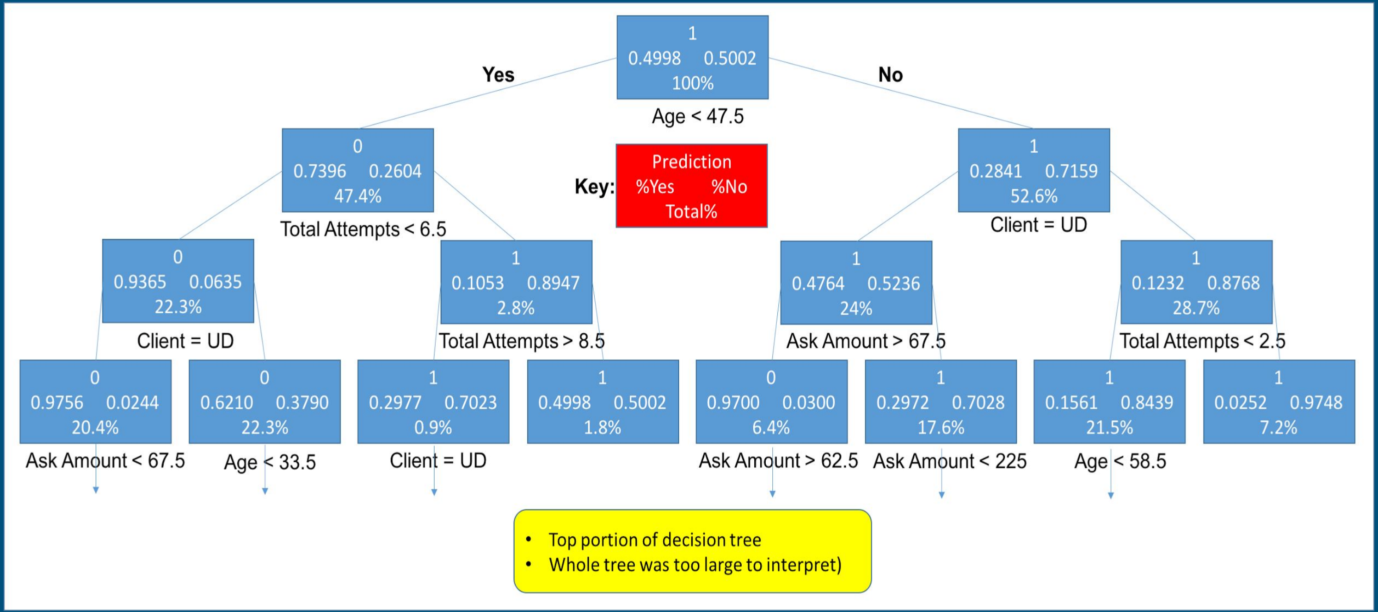
Hyphotheses

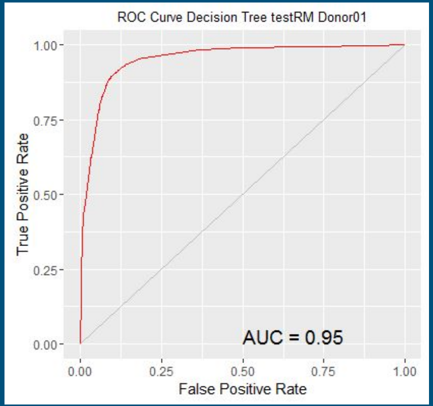
- Based on the findings:
  - Chance to donate
    - People living in certain areas or in-state tend to donate more than those living in other areas or out-of-state
    - Older individuals are more likely to donate than younger individuals
    - People with more call attempts tend to donate more than those with less call attempts
    - People with longer talk time have a higher chance of donating
    - Males have a higher chance to donate

Modeling Techniques Explored

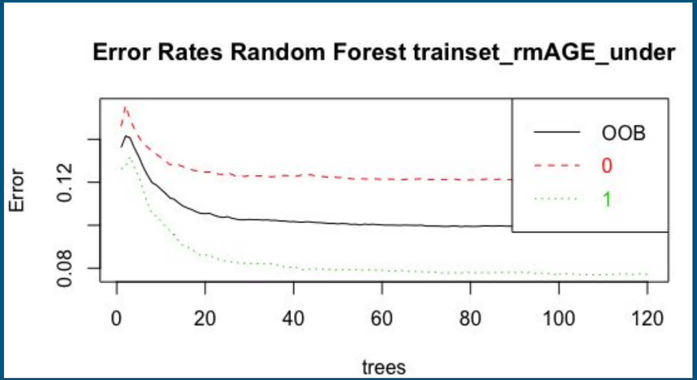
- Decision Tree:
  - Used R-packages to help pick optimum parameters for min split, min bucket, max depth and complexity values
- Random Forest: advanced decision tree
  - Number of trees: the larger the better, but the longer to compute, results stops getting significantly better beyond a critical number
  - Number of variables: the lower the greater reduction of variance, but also the greater increase in bias
- Support Vector Machine (SVM):
  - Kernel: Radial Basis (rbfdot)
  - Options: The parameter has been set from 0-1.
- Used Rattle and R-packages to build three models

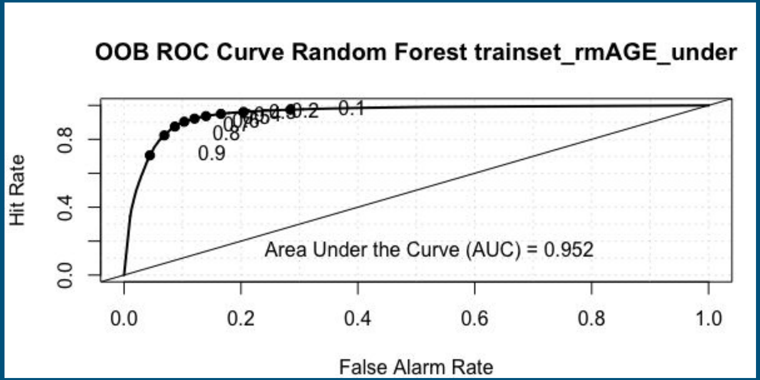
Results -Decision Tree





Results -Random Forest

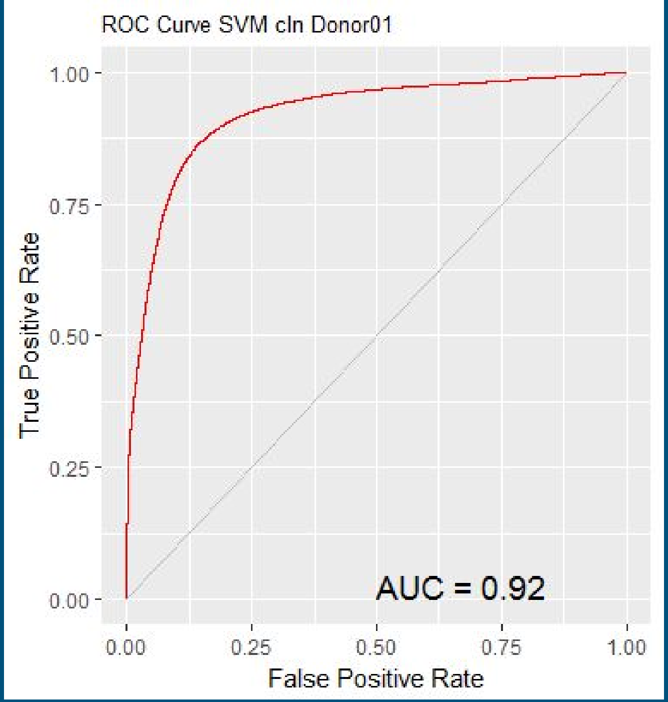




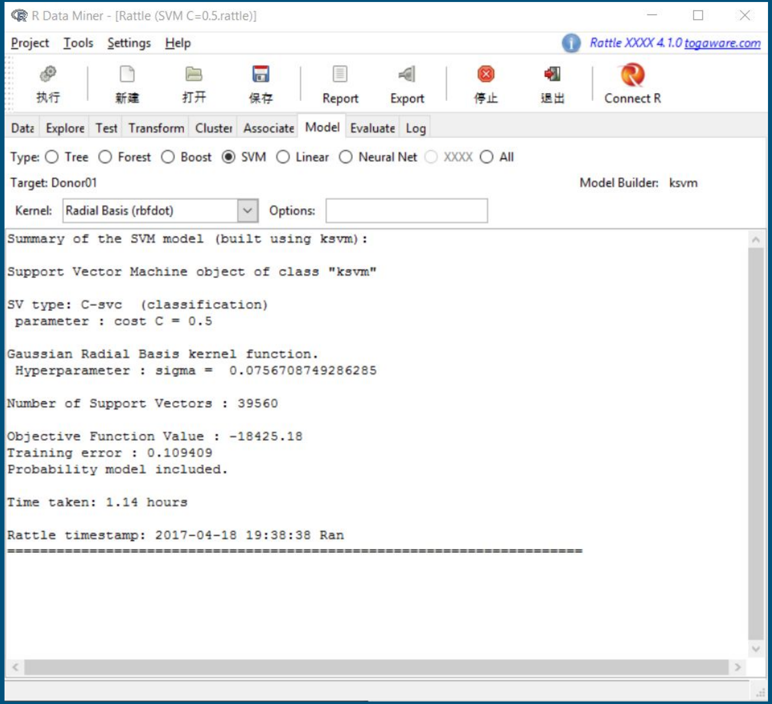

```
------------------------------------------------------------------
Tree 1 Rule 34 Node 2796 Decision 1

1: TALK_TIME <= 41.5
2: CRegion IN ("Military", "Pacific", "ESCentral", "WSCentral", "Mountain", "WNCentral", "ENCentral"
3: SCHOOL IN ("Agriculture and Natural Resources", "Applied Science", "Continuing Education", "Educa
4: client IN ("UDelaware")
5: AffilCount <= 0.5
6: NumDegrees > 0.5
7: SuccCont IN ("0")
8: TotAttempts > 2.5
9: Ask_Amount_2 > 67.5
10: TotAttempts <= 7
11: TotAttempts <= 4.5
12: GENDER IN ("M", "U")
13: CRegion IN ("Military", "Pacific", "ESCentral", "Mountain", "NewEngland", "Commonwealth", "SAtla
------------------------------------------------------------------
```

Results -SVM



ROC Curve SVM cln Donor01
1.00
0.75
0.50
0.25
0.00
True Positive Rate
AUC = 0.92
0.00
0.25
0.50
0.75
1.00
False Positive Rate



R Data Miner - [Rattle (SVM C=0.5.rattle)]
Project Tools Settings Help
Rattle XXXX 4.1.0 togaware.com
执行 新建 打开 保存 Report Export 停止 退出 Connect R
Data Explore Test Transform Cluster Associate Model Evaluate Log
Type: Tree Forest Boost SVM Linear Neural Net XXXX All
Target: Donor01
Model Builder: ksvm
Kernel: Radial Basis (rbfdot)
Options:
Summary of the SVM model (built using ksvm):
Support Vector Machine object of class "ksvm"
SV type: C-svc (classification)
parameter : cost C = 0.5
Gaussian Radial Basis kernel function.
Hyperparameter : sigma = 0.0756708749286285
Number of Support Vectors : 39560
Objective Function Value : -18425.18
Training error : 0.109409
Probability model included.
Time taken: 1.14 hours
Rattle timestamp: 2017-04-18 19:38:38 Ran

Evaluation Criteria - Overall Dataset

Dataset	rmAGE	medAGE	rmAGE_under	rmAGE_both
Tree	**15, 5, 15, 0.0001**	**21, 7, 10, 0.0001**	**21, 7, 10, 0.0001**	**21, 7, 10, 0.0001**
FPR	0.0816	0.0707	0.1077	0.1071
FNR	0.1311	0.1639	0.1089	0.1061
F	0.8695	0.8457	0.8464	0.8462
Forest	**120, 3**	**120, 3**	**30, 3**	**30, 3**
FPR	0.0827	0.0693	0.1190	0.1118
FNR	0.1239	0.1402	0.0812	0.0882
F	0.8623	0.8605	0.8575	0.8594
SVM	**-**	**0.5**	**-**	**-**
FPR	-	0.0984		
FNR	-	0.2034		
F	-	0.8444		

FPR: false positive rate (predicted 0 was really 1)
FNR: false negative rate (predicted 1 was really 0)
F score: overall accuracy for binary target

Evaluation Criteria - Client Specific

Dataset	UDelaware (UD)	UMissouri (UM)	UM_under	UM_both
Tree	15, 5, 5, 0.0001	15, 5, 15, 0.0001	90, 30, 10, 0.0001	90, 30, 5, 0.0001
FPR	0.0110	0.5014	0.4866	0.7421
FNR	0.1556	0.1420	0.8543	0.6128
F	0.8943	0.5674	0.1279	0.2487
Forest	60, 3	90, 3	30, 3	30, 3
FPR	0.0115	0.2879	0.8137	0.8239
FNR	0.0969	0.1133	0.8401	0.8127
F	0.9268	0.7045	0.1251	0.1251
SVM	0.5	0.5	-	-
FPR	0.0186	0.2199		
FNR	0.2232	0.1447		
F	0.9628	0.5745		

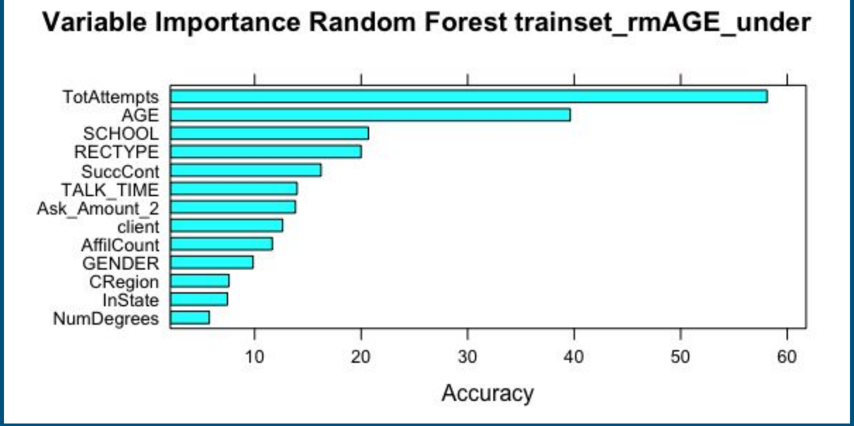
FPR: false positive rate (predicted 0 was really 1)
FNR: false negative rate (predicted 1 was really 0)
F score: overall accuracy for binary target

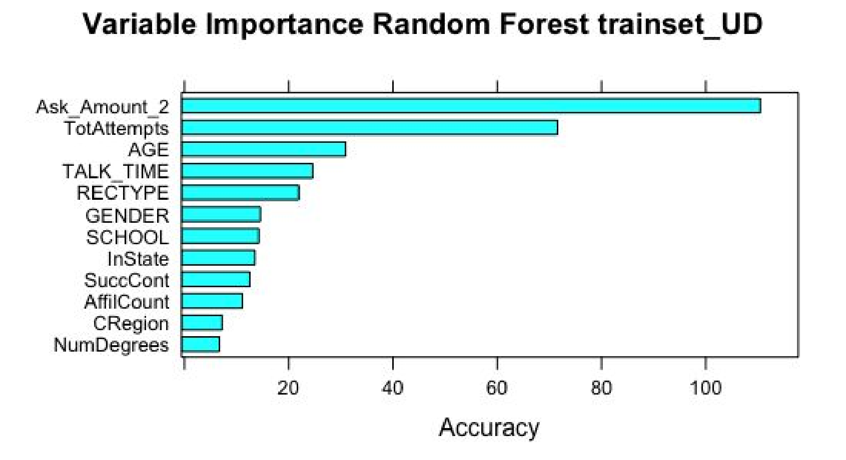
Conclusion

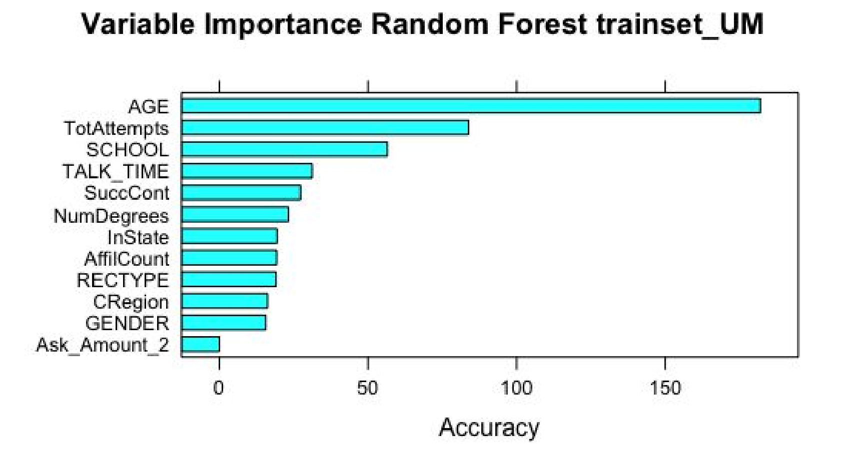
1. Final Best Model
2. Insights
3. Recommendations
4. Final Takeaways

RUFFALO NOEL LEVITZ

Final Best Model







Insights

Important Factors

- Age
- Total Attempts
- Talk Time

Insignificant Factors

- Number of degrees
- Client's location
- Gender

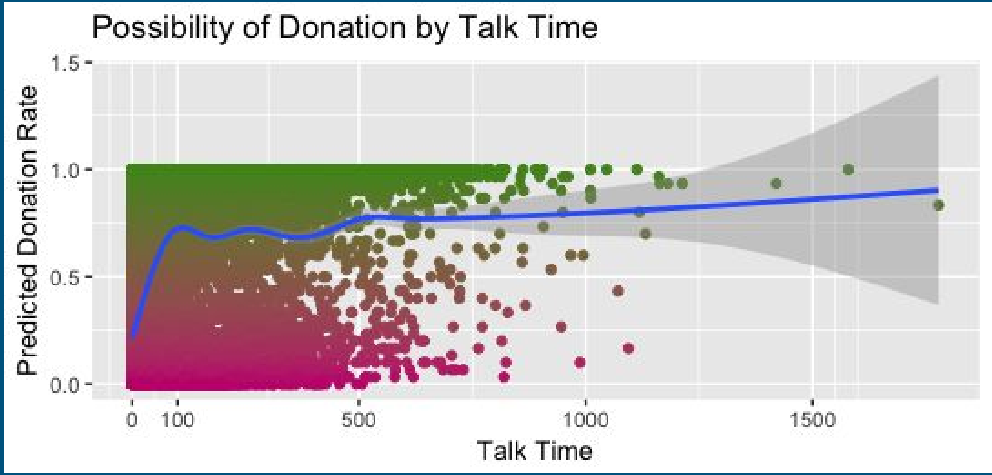
Strategic Recommendations

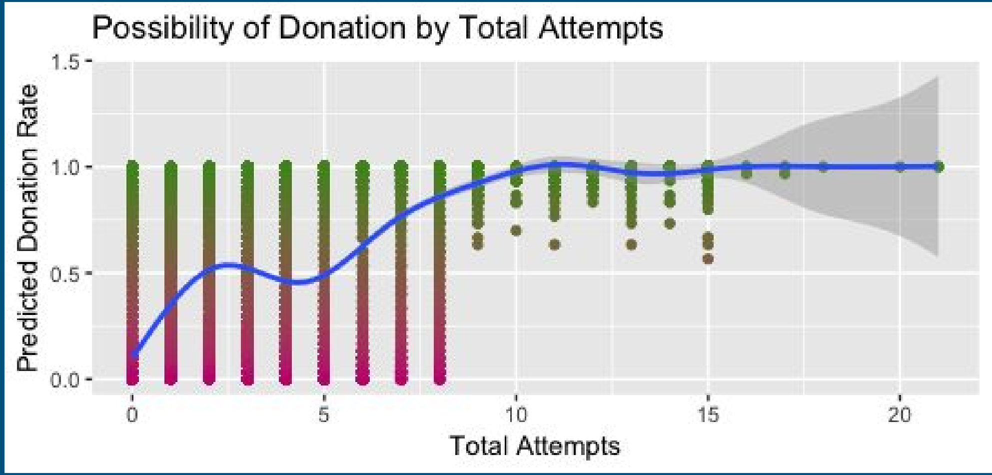
- Target people in the 55-75 age group

Tactical Recommendations

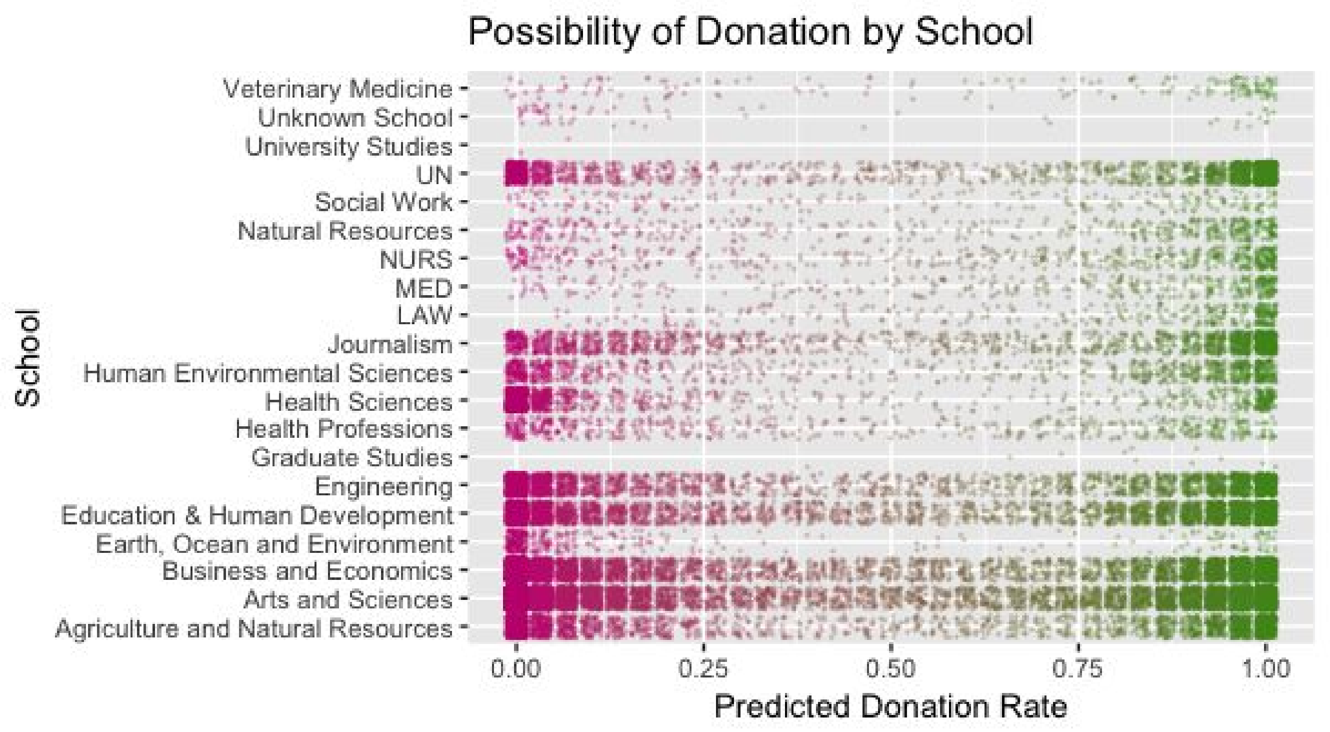
- Delaware clients: set optimal ask amount
- Missouri clients: focus on older people

Insights

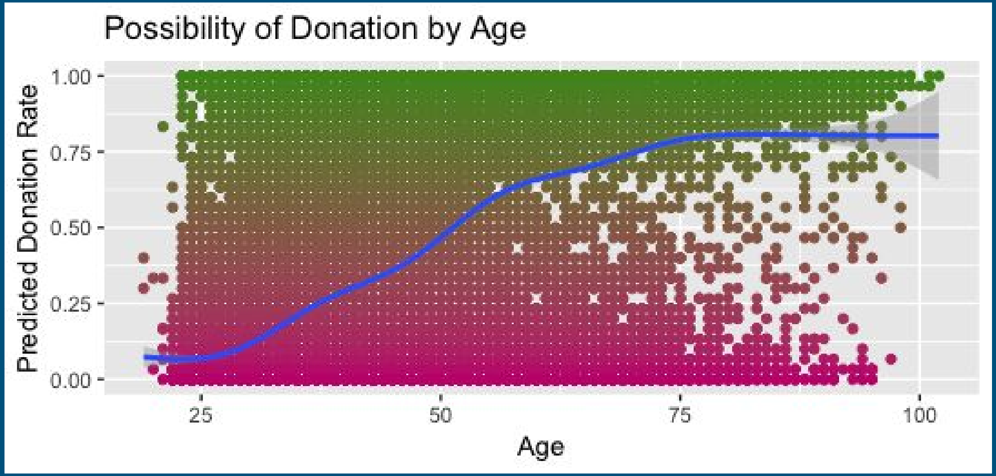

Possibility of Donation by Talk Time
Predicted Donation Rate
1.5
1.0
0.5
0.0
0
100
500
1000
1500
Talk Time

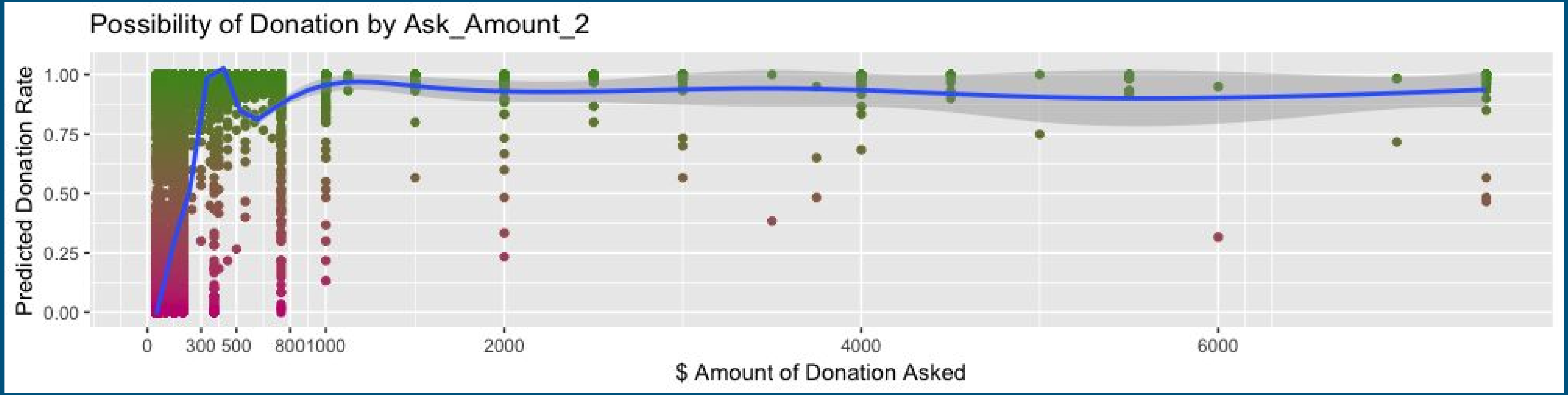

Possibility of Donation by Total Attempts
Predicted Donation Rate
1.5
1.0
0.5
0.0
0
5
10
15
20
Total Attempts

Insights


Possibility of Donation by School
School
Veterinary Medicine
Unknown School
University Studies
UN
Social Work
Natural Resources
NURS
MED
LAW
Journalism
Human Environmental Sciences
Health Sciences
Health Professions
Graduate Studies
Engineering
Education & Human Development
Earth, Ocean and Environment
Business and Economics
Arts and Sciences
Agriculture and Natural Resources
0.00
0.25
0.50
0.75
1.00
Predicted Donation Rate

Insights


Possibility of Donation by Age
Predicted Donation Rate
1.00
0.75
0.50
0.25
0.00
25
50
75
100
Age


Possibility of Donation by Ask_Amount_2
Predicted Donation Rate
1.00
0.75
0.50
0.25
0.00
0
300 500
8001000
2000
4000
6000
$ Amount of Donation Asked

Final Takeaways

THANK YOU!