GENERAL ELECTRIC CUSTOMER CHURN PREDICTION

ADAM M. LANG

DAT-690

SOUTHERN NEW HAMPSHIRE UNIVERSITY

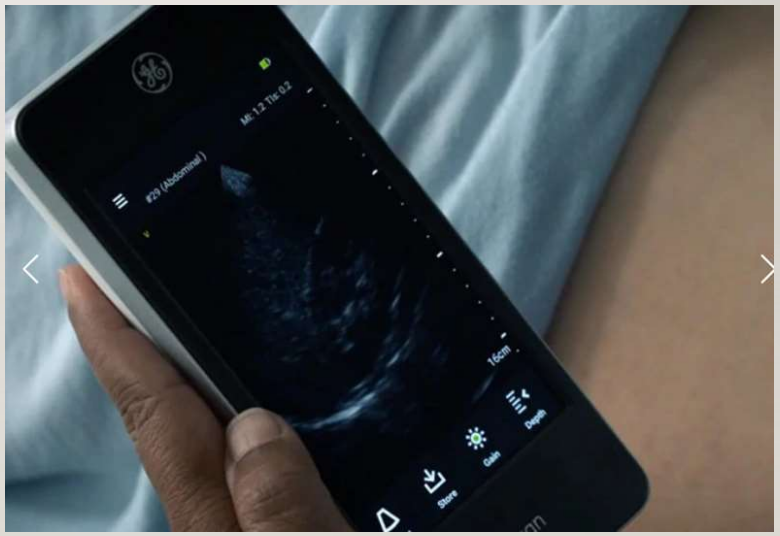
OVERVIEW

- Introduction
- Pilot Model Evaluation
- Plan Modification
- Plan Implementation and Results
- Conclusions and Implications
- Recommendations, Future Directions

INTRODUCTION

Background

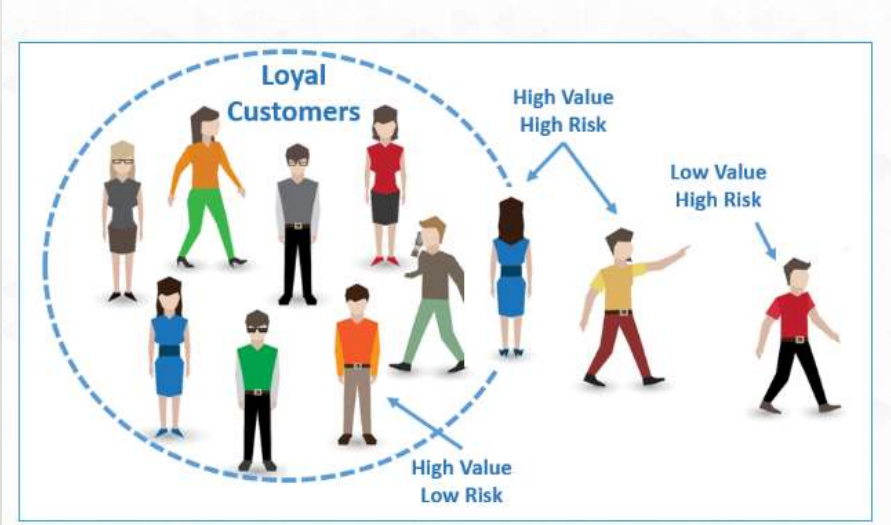
- General Electric artificial intelligence smart phone applications used by healthcare professionals (GE, 2016)
- $6.5 billion market by 2021 (Miliard, 2018)
- GE customers using resold mobile phones to access applications



GE Healthcare, 2020

PROBLEM

- Churn: Why? How many? What factors?
- 20-40% of customers will churn per year in telecommunications (Hashmi et al. 2013)
- Cost: 5-10 times more to add new customers than retain original customers
- Focus: reduce churn rates by 5% and increase profits by 85% (Ullah et al. 2019)



Churn Analysis Diagram

DunnSolutions, 2020

RESEARCH PLAN

PREDICTIVE MODELING (MACHINE LEARNING) – HIDDEN PATTERNS IN DATA

LOGISTIC REGRESSION PREDICTS CUSTOMER CHURN BY 94.8%
(MANDAK ET AL. 2019)

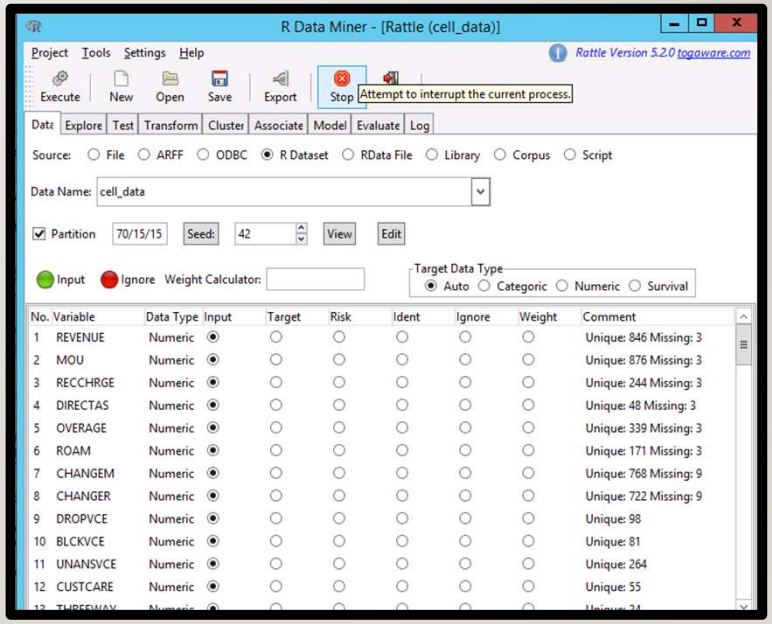
WHAT FACTORS (INDEPENDENT VARIABLES) ARE CONTRIBUTING TO CUSTOMER CHURN?

FACTORS + PERCENT CUSTOMERS CHURNING = RETENTION PLAN

PILOT EVALUATION

PILOT MODEL - OVERVIEW

- Raw Data Cleaning: missing values, factors
- Model: Logistic Regression (Stoltzfus, 2011)
- Variable Selection: multiple ANOVA tests (Kao et al. 2008)
- R-Rattle GUI Data Mining program



Rattle GUI Data Mining application

PILOT EVALUATION - SUCCESSES

```
Summary of the Logistic Regression model (built using glm):

Call:
glm(formula = TFC_CHURN ~ ., family = binomial(link = "logit"),
    data = crs$dataset[crs$train, c(crs$input, crs$target)])

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.563  -0.845  -0.681   1.164   2.524

Coefficients:
                     Estimate  Std. Error z value Pr(>|z|)
(Intercept)       -0.2708891   0.4206850  -0.644  0.51962
DROPVCE            0.0180423   0.0178778   1.009  0.31288
UNANSVCE          -0.0036490   0.0044958  -0.812  0.41699
MOUREC             0.0003046   0.0011633   0.262  0.79347
EQPDAYS            0.0007930   0.0003959   2.003  0.04518 *
TFC_Column.45(0,1] -0.1183329   0.4212610  -0.281  0.77879
TFC_Column.45(1,2] -0.0365952   0.1902792  -0.192  0.84749
TFC_Column.45(2,3] -0.9078847   0.3746637  -2.423  0.01538 *
TFC_REFURB(0,1]    0.6479820   0.2584047   2.508  0.01215 *
TFC_WEBCAP(0,1]   -0.6156180   0.2889541  -2.131  0.03313 *
TFC_RETCALLS(0,1]  1.1419654   0.4381403   2.606  0.00915 **
TFC_RETCALLS(1,2] 15.7945909 535.4113304   0.029  0.97647
IMO_MOU           -0.0007292   0.0004582  -1.592  0.11147
IMO_RECCHRGE      -0.0014507   0.0048749  -0.298  0.76602
IMO_DIRECTAS      -0.0802463   0.0679256  -1.181  0.23745
IMO_CHANGEM       -0.0010294   0.0004195  -2.454  0.01414 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 853.51  on 699  degrees of freedom
Residual deviance: 786.23  on 684  degrees of freedom
AIC: 818.23

Number of Fisher Scoring iterations: 12

Log likelihood: -393.113 (16 df)
Null/Residual deviance difference: 67.283 (15 df)
Chi-square p-value: 0.00000001
Pseudo R-Square (optimistic): 0.30544032
```

Multiple variables leading to churn identified

Good statistical outcomes

PILOT EVALUATION - SUCCESSES

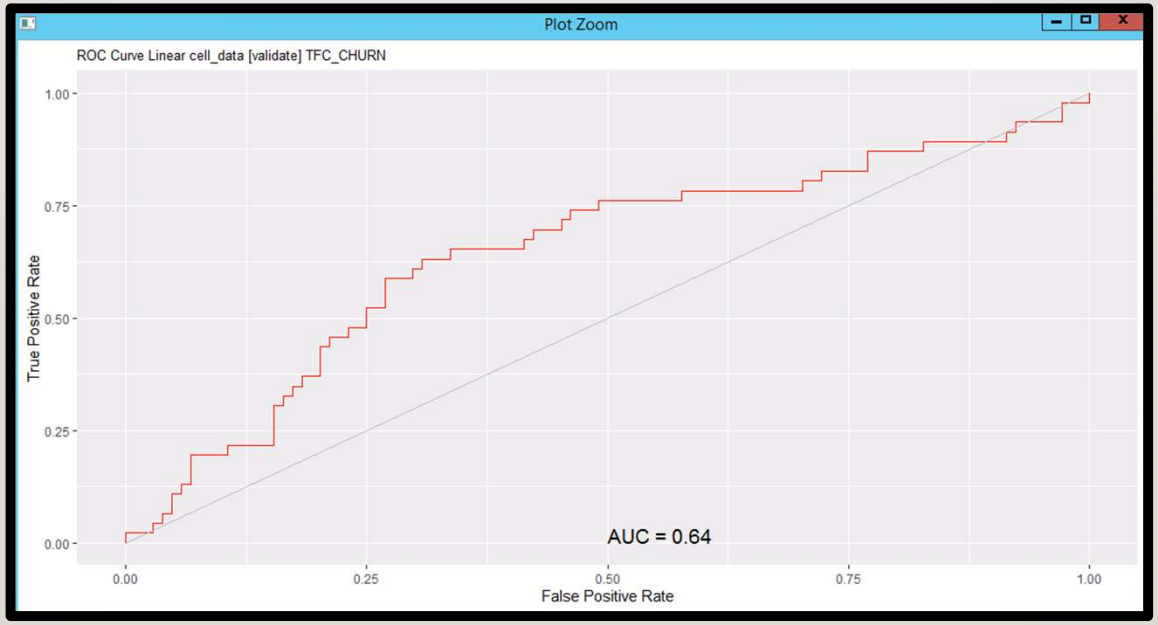
```
> vif(model)
                    GVIF Df GVIF^(1/(2*Df))
DROPVCE         2.102853  1        1.450122
UNANSVCE        1.983446  1        1.408349
MOUREC          3.315728  1        1.820914
EQPDAYS         1.321955  1        1.149763
TFC_Column.45   1.056795  3        1.009249
TFC_REFURB      1.053740  1        1.026518
TFC_WEBCAP      1.175240  1        1.084085
TFC_RETCALLS    1.036169  2        1.008922
IMO_MOU         5.179998  1        2.275961
IMO_RECCHRGE    1.449590  1        1.203989
IMO_DIRECTAS    1.186449  1        1.089242
IMO_CHANGEM     1.055300  1        1.027278
>
```

No collinearity

Variable Name	Variable Meaning	Level of Significance
EQP DAYS	Number of days with equipment	.05 level
TFC_COLUMN 45 (2,3)	PRIZM CODE. 2 = Urban, 3= Town	.05 level
TFC_REFURB	Handset refurbished	.05 level
TFC_WEBCAP	Handset web capable	.05 level
TFC_RETCALLS (0,1)	Number of calls previously made to retention team	.01 level
IMO_CHANGEM	% Change in minutes use	.05 level

Multiple variables: Calls made to Retention team!

PILOT EVALUATION – FAILURES

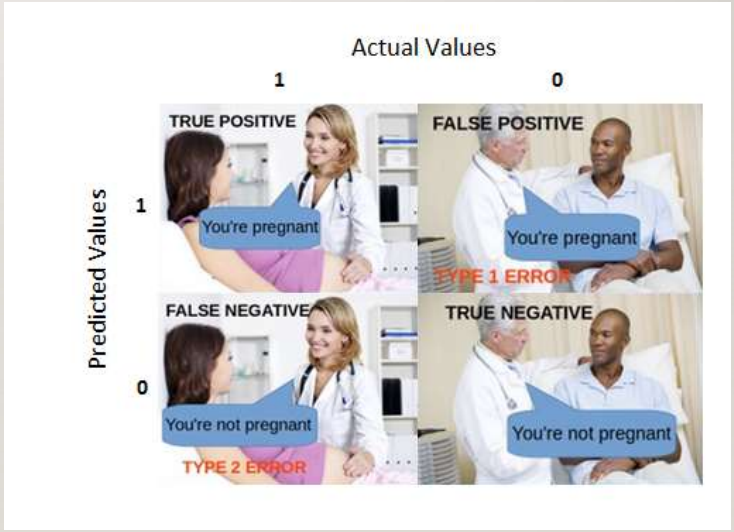


Poor Predictive capacity

PILOT EVALUATION - FAILURES

```
Error matrix for the Linear model on cell_data [validate] (counts):

        Predicted
Actual    [0,0] (0,1] Error
  [0,0]      99     5   4.8
  (0,1]      41     5  89.1

Error matrix for the Linear model on cell_data [validate] (proportions):

        Predicted
Actual    [0,0] (0,1] Error
  [0,0]    66.0   3.3   4.8
  (0,1]    27.3   3.3  89.1

Overall error: 30.7%, Averaged class error: 46.95%

Rattle timestamp: 2020-03-13 00:06:17 adam.lang_snhu
```

Confusion Matrix: Poor prediction



Narkhede, 2018

PILOT EVALUATION – RETURN ON INVESTMENT

- Remember this Focus: reduce churn rates by 5% and increase profits by 85% (Ullah et al. 2019)
- AUC curve → 64% of churners predicted
- Is the model predicting correctly? Do we see a real return on the investment?
- Feasibility of this prediction?

- Variables of interest = focus ROI!

Variable Name	Variable Meaning	Level of Significance
EQP DAYS	Number of days with equipment	.05 level
TFC_COLUMN 45 (2,3)	PRIZM CODE. 2 = Urban, 3= Town	.05 level
TFC_REFURB	Handset refurbished	.05 level
TFC_WEBCAP	Handset web capable	.05 level
TFC_RETCALLS (0,1)	Number of calls previously made to retention team	.01 level
IMO_CHANGEM	% Change in minutes use	.05 level

Current Pilot Model variables of interest. Are there others we are missing?

PILOT EVALUATION – AREAS OF CONCERN

DATA INTEGRITY/DATA QUALITY

NUMBER OF MODELS USED – BEST MODEL?

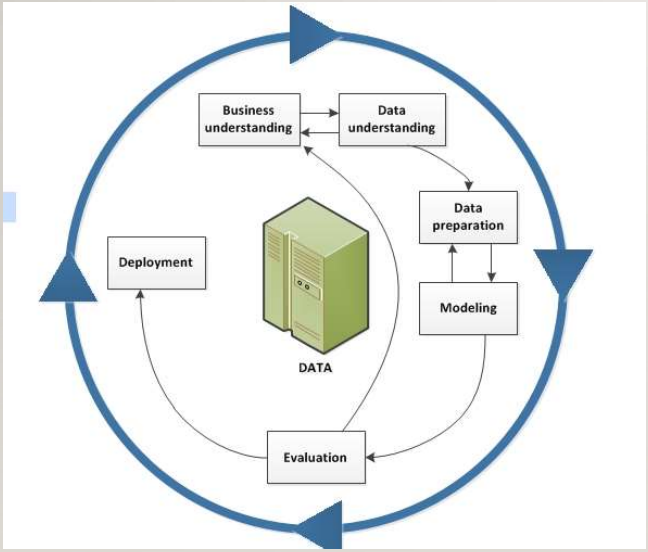
PROGRAMMING TECHNOLOGY

VARIABLES OF INTEREST – DID WE SELECT THE CORRECT ONES?

PLAN MODIFICATION

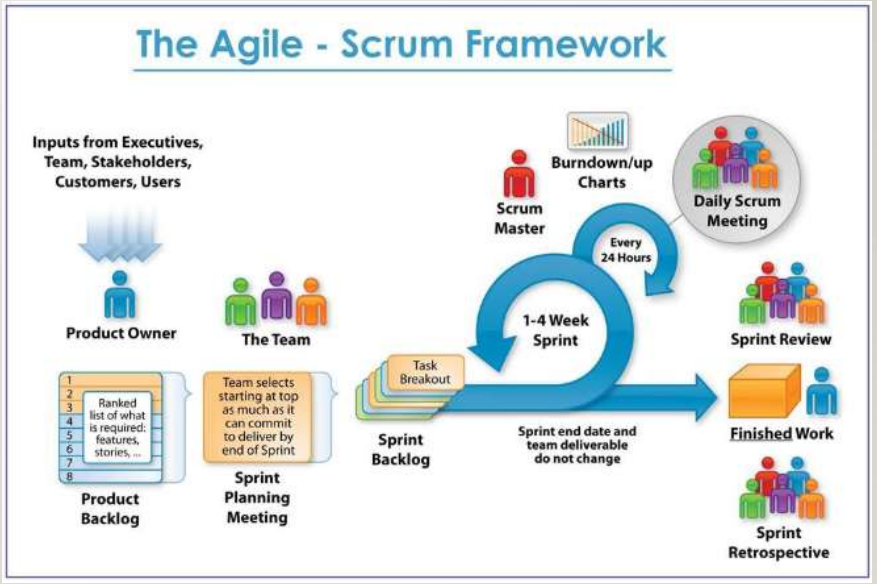
PLAN MODIFICATION – PLAN & PROCESS – CROSS-FUNCTIONAL TEAMWORK

- CRISP-DM process for data mining



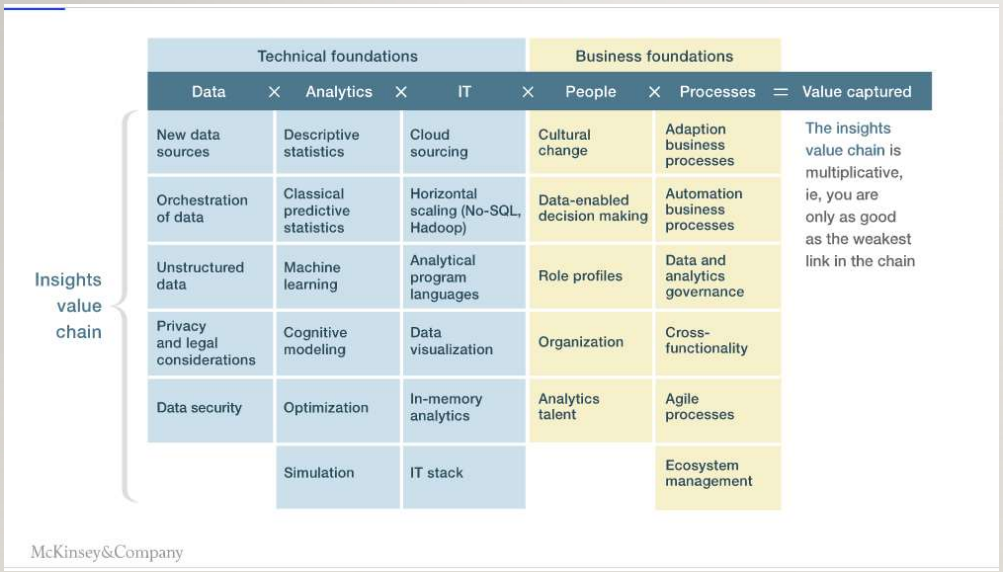
IBM, 2020

- Project Management Timeline: AGILE w/ SCRUM



Agile w/ Scrum: Streamlined Project Management. (Lara, 2018)

PLAN MODIFICATION – STAKEHOLDER COLLABORATION



	Technical foundations			Business foundations		
	Data ×	Analytics ×	IT ×	People ×	Processes =	Value captured
Insights value chain	New data sources	Descriptive statistics	Cloud sourcing	Cultural change	Adaption business processes	The insights value chain is multiplicative, ie, you are only as good as the weakest link in the chain
	Orchestration of data	Classical predictive statistics	Horizontal scaling (No-SQL, Hadoop)	Data-enabled decision making	Automation business processes	
	Unstructured data	Machine learning	Analytical program languages	Role profiles	Data and analytics governance	
	Privacy and legal considerations	Cognitive modeling	Data visualization	Organization	Cross-functionality	
	Data security	Optimization	In-memory analytics	Analytics talent	Agile processes	
		Simulation	IT stack		Ecosystem management	

McKinsey&Company

Insights Value Chain (Hurtgen et al. 2018)

- 'Insights Value Chain' – Mckinsey & Company
- Cross functional collaboration
- Identify Stakeholders

1. Analytics team
2. IT department
3. Healthcare AI technology group
4. Business department
5. C-Suite executives

PLAN MODIFICATION – DATA PROTECTION

Data Protection/Security

- **Physical security: employee education, lock it up!**
- Firewalls
- Encryption
- Masking
- Tokenization
- Machine Learning: SVM, Clustering (Cheng et al. 2017)

✓ Anomaly detection

Data Ethics

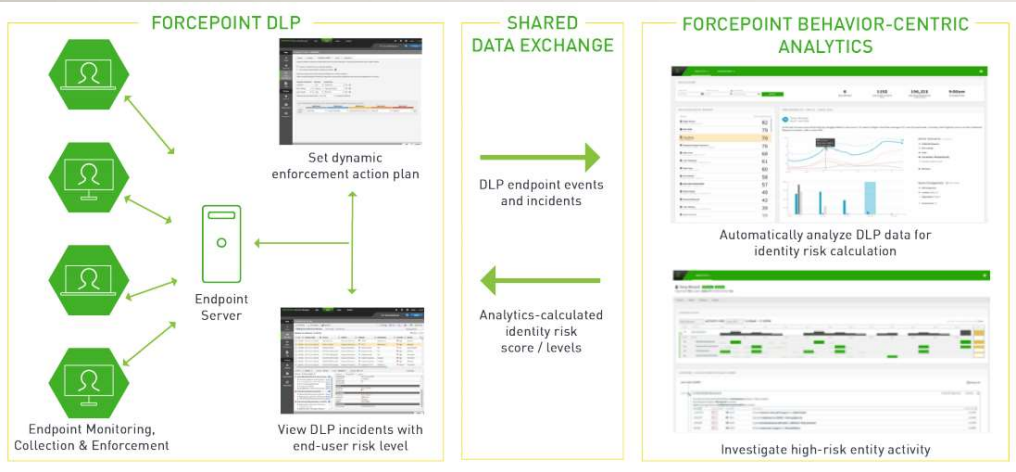
- **Informed consent from all customers**
- Version Control – Git
- **GLBA, CCPA, GDPR laws of data privacy – we do not own the data!** (McDaniel, 2019)

✓ Disclose to all customers our intentions

✓ Provide an opt out option to all customers

✓ Customers tell us what we can do with their data

DATA PROTECTION/SECURITY - TECHNOLOGY



Forcepoint, 2020

- Forcepoint Dynamic Data Protection

1. Behavior Centric Analytics: track each employee
2. Analytics driven forensics: normal vs. abnormal file/data movement
3. Policy Enforcement: individualized for each user and data file

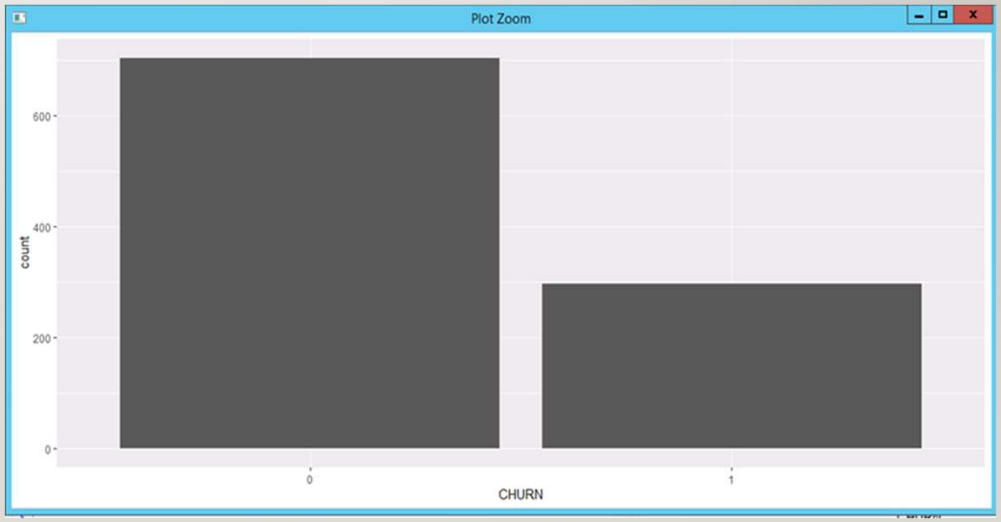
PLAN MODIFICATION – DATA QUALITY

- ***Imbalance in target variable CHURN***

✓ **Non-churners are favored by our model**

✓ **We will have to re-sample the target variable**

✓ **We will use the SMOTE technique (Chawla, 2002)**



DATA QUALITY ISSUES – MISSING VALUES

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

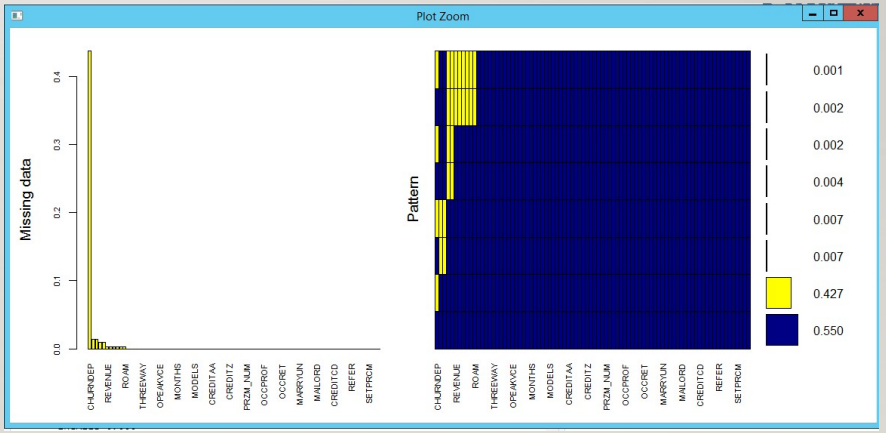
Source

Console Terminal Jobs

```
> #Aggregation Plot of missing values
> mice_plot <- aggr(cell_data, col=c('navyblue','yellow'),
+                   numbers=TRUE, sortVars=TRUE,
+                   labels=names(cell_data), cex.axis=.7,
+                   gap=3, ylab=c("Missing data","Pattern"))

 Variables sorted by number of missings:
      Variable Count
      CHURNDEP 0.437
          AGE1 0.014
          AGE2 0.014
       CHANGEM 0.009
       CHANGER 0.009
       REVENUE 0.003
           MOU 0.003
      RECCHRGE 0.003
      DIRECTAS 0.003
       OVERAGE 0.003
          ROAM 0.003
       DROPVCE 0.000
       BLCKVCE 0.000
      UNANSVCE 0.000
      CUSTCARE 0.000
      THREEWAY 0.000
        MOUREC 0.000
      OUTCALLS 0.000
       INCALLS 0.000
       PEAKVCE 0.000
```

RStudio table print of missing values

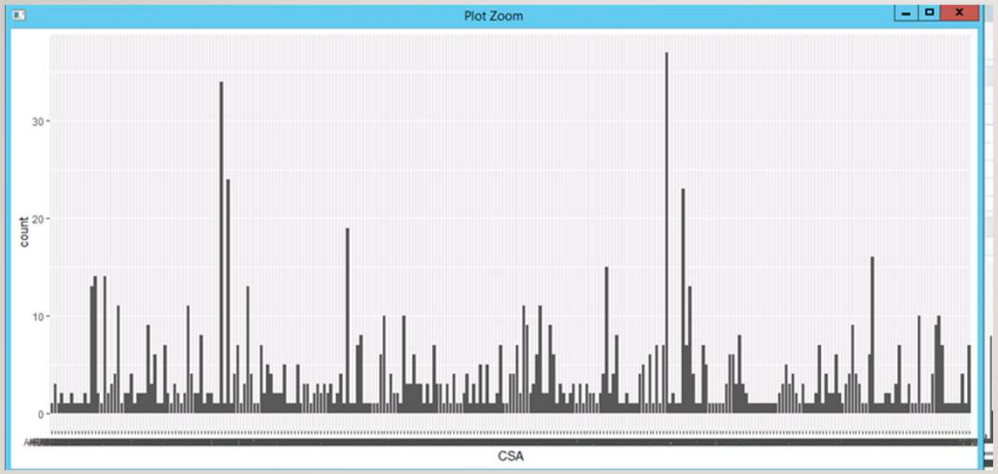
- 501 missing values
- "Missing Completely at Random" (Kang, 2013)



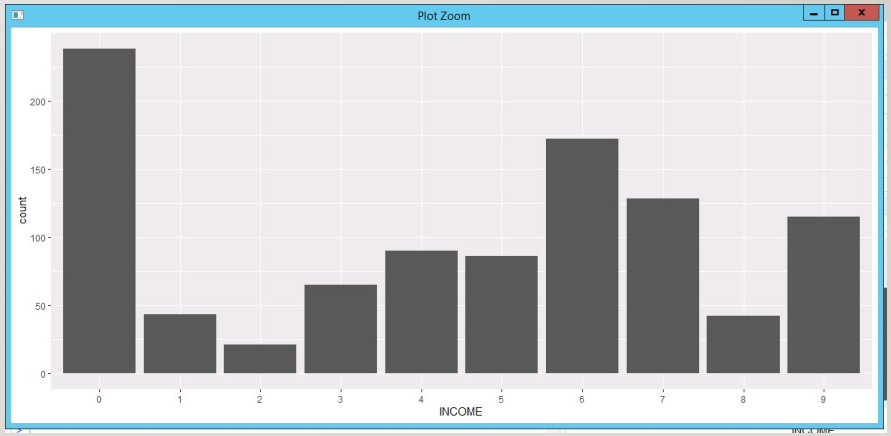
Aggregation Plot of missing values in data set

DATA QUALITY ISSUES – FACTORS WITH MULTIPLE LEVELS

- CSA is a factor with 277 levels

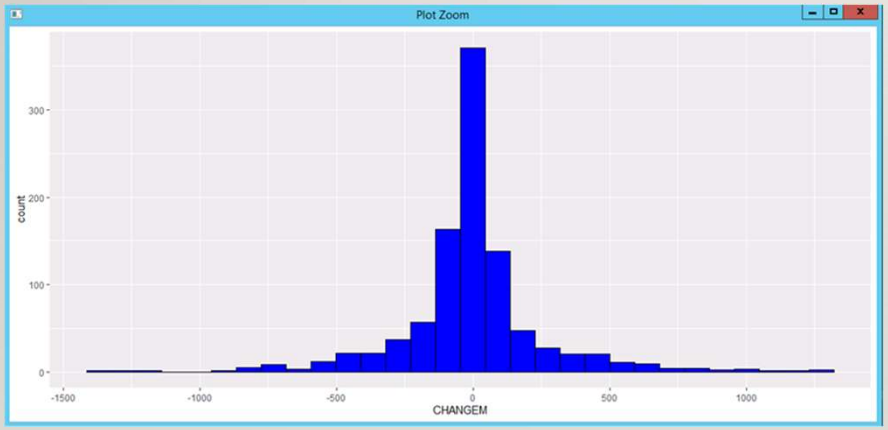


- INCOME factor variable has multiple levels and uneven distribution

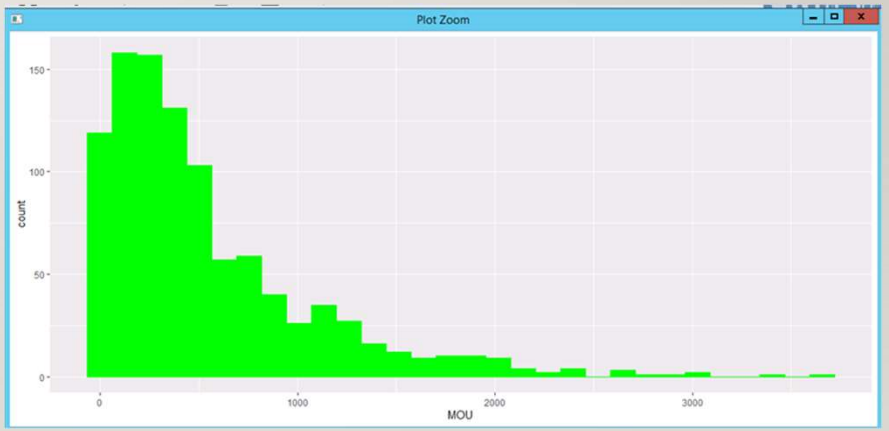


DATA QUALITY ISSUES – ABNORMAL DISTRIBUTIONS

- CHANGEM numeric normal distribution

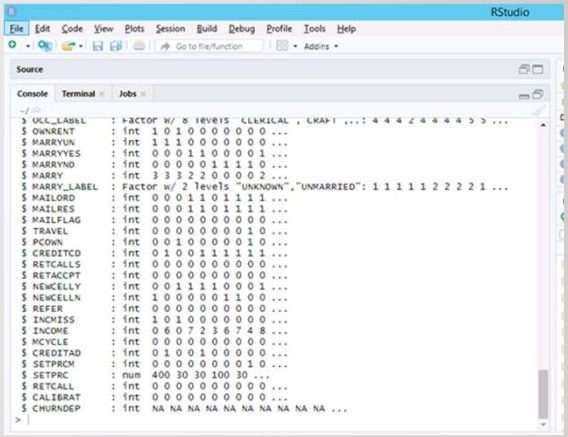


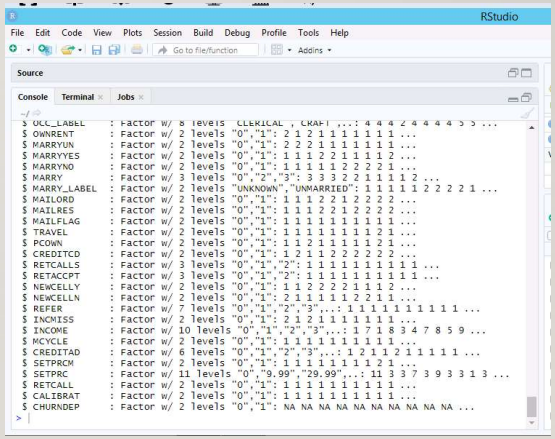
- MOU numeric variable is right skewed



DATA QUALITY ISSUES - OTHER

- **Integers that should be Factors**
- Improper column labeling
- 50% categorical, 50% continuous variables





PLAN MODIFICATION – PREDICTIVE MODELING

- Pilot model: Logistic regression
- Issues with Logistic Regression (Stoltzfus, 2011):

1. Multi-collinearity
2. Variable selection
3. Number of variables (10 events per 1)
4. Overfitting the model

Plan: Run multiple types of logistic regression, pick the best model

- Random Forest Model

1. Robust with multi-variate binary data (Courronne et al. 2018)
2. Random sample "bagging" technique
3. Variable importance, Gini Index to select variables for logistic regression

Plan: Run multiple Random Forest Models and compare to logistic regression models

PLAN MODIFICATION – VARIABLE SELECTION

- Multiple ANOVA's were used for Pilot
- ANOVA has 3 requirements to work (Kao et al. 2008):

1. Outcome variable normally distributed in each group
2. Variance in each group is the same
3. Observations not correlated with one another

- **Data set violates all 3!!!**

- Other options for variable selection:

1. **Penalized Regression (Desboulets, 2018)**

✓ **LASSO**

✓ **RIDGE**

✓ **ELASTIC NET**

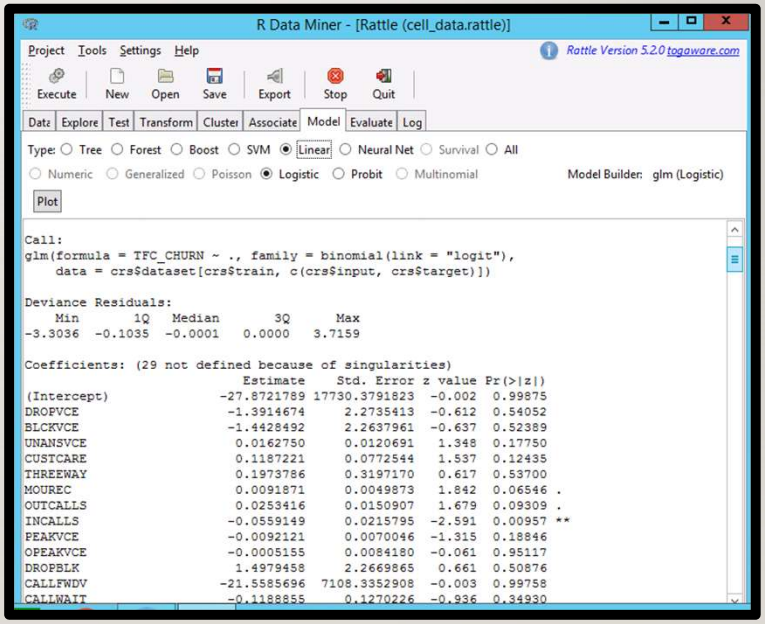
2 . Stepwise Regression (Desboulets, 2018)

3 . Random Forest (Couronne et al. 2018)

PLAN MODIFICATION – PROGRAMMING TECHNIQUES

- RATTLE-GUI used for Pilot
- RStudio hard coding will be used instead
- Reasoning:

1. Oversampling (SMOTE) technique
2. GLMNET function (lasso regression)
3. Stepwise regression
4. Cross validation comparison of models
5. More visualization techniques

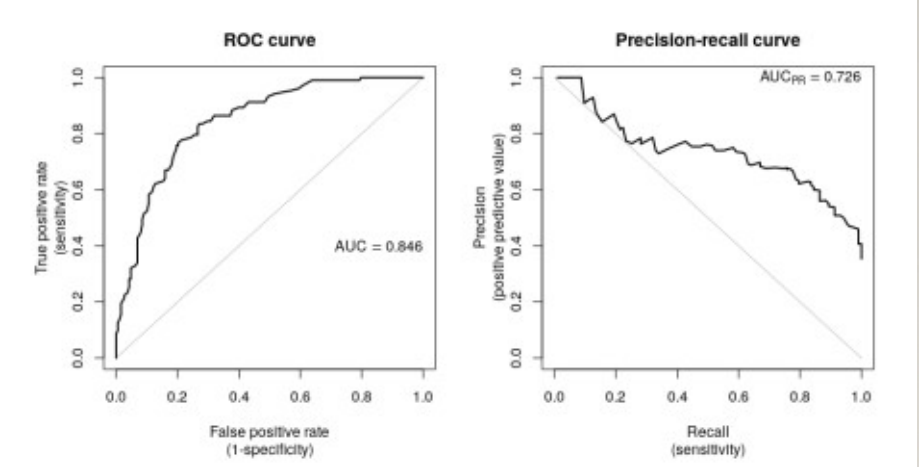


Example of RATTLE-GUI

PLAN MODIFICATION – COMPARING MODELS

- 10 folds cross validation (Krastajic et al. 2014)
- 'caret' and 'mlbench' packages (Brownlee, 2019)
- Resampling and trainControl
- Compare

1. Kappa values (Delgado et al. 2019)
2. Accuracy
3. AIC value (regression models only)
4. AUC
5. Confusion Matrices
6. Precision Recall Curves, f1 scores (Davis et al. 2006)



ROC Curve vs. Precision-Recall Curve (Barbosa, 2020)

PLAN IMPLEMENTATION & RESULTS

- Steps taken
- Methods and techniques used
- Results obtained

STEP 1: DATA CLEANING

- Missing Values
- Integer → Factor Conversion
- "Column 45" → PRZM_NUM
- Multi-level factor variables

STEP 2: CREATE TRAIN/TEST SPLIT

```
> #Create train/test split
> set.seed(42)
> #use caTools function to split, SplitRatio for 70%:30% splitting
> data1= sample.split(cell_data,SplitRatio = 0.7)
>
> #subsetting into Train data
> train =subset(cell_data,data1==TRUE)
>
> #subsetting into Test data
> test =subset(cell_data,data1==FALSE)
>
```

OVERVIEW OF MODELING PLAN FOR IMPLEMENTATION

- Run GLM pilot model
- Rebalance target variable
- Run GLM pilot with rebalanced data
- Run Random Forest #1 with normal data set
- Run Random Forest with rebalanced data
- Run GLM model with variables selected from Random Forest #1
- Run GLM model with variables selected from Random Forest #2
- Run LASSO penalized regression model

*Stepwise regression will be performed after each GLM model (4 stepwise regressions)

STEP 3: LOGISTIC REGRESSION PILOT MODEL (MODEL_GLM)

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.4572541  0.4237457  -1.079  0.28055
DROPVCE      0.0211056  0.0161138   1.310  0.19027
UNANSVCE     0.0024907  0.0037120   0.671  0.50223
MOUREC       0.0009828  0.0011008   0.893  0.37197
EQPDAYS      0.0006644  0.0003859   1.721  0.08517 .
PRZM_NUM1    0.1430509  0.4114868   0.348  0.72811
PRZM_NUM2    0.1574282  0.1909875   0.824  0.40978
PRZM_NUM3   -0.5940198  0.3476007  -1.709  0.08747 .
REFURB1      0.7120691  0.2502119   2.846  0.00443 **
WEBCAP1     -0.4707634  0.2989137  -1.575  0.11528
MOU         -0.0008347  0.0004398  -1.898  0.05771 .
RECCHRGE    -0.0053486  0.0047172  -1.134  0.25685
DIRECTAS     0.0034553  0.0573303   0.060  0.95194
CHANGEM     -0.0005096  0.0003819  -1.334  0.18213
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 821.79  on 676  degrees of freedom
Residual deviance: 785.47  on 663  degrees of freedom
AIC: 813.47

Number of Fisher Scoring iterations: 4
```

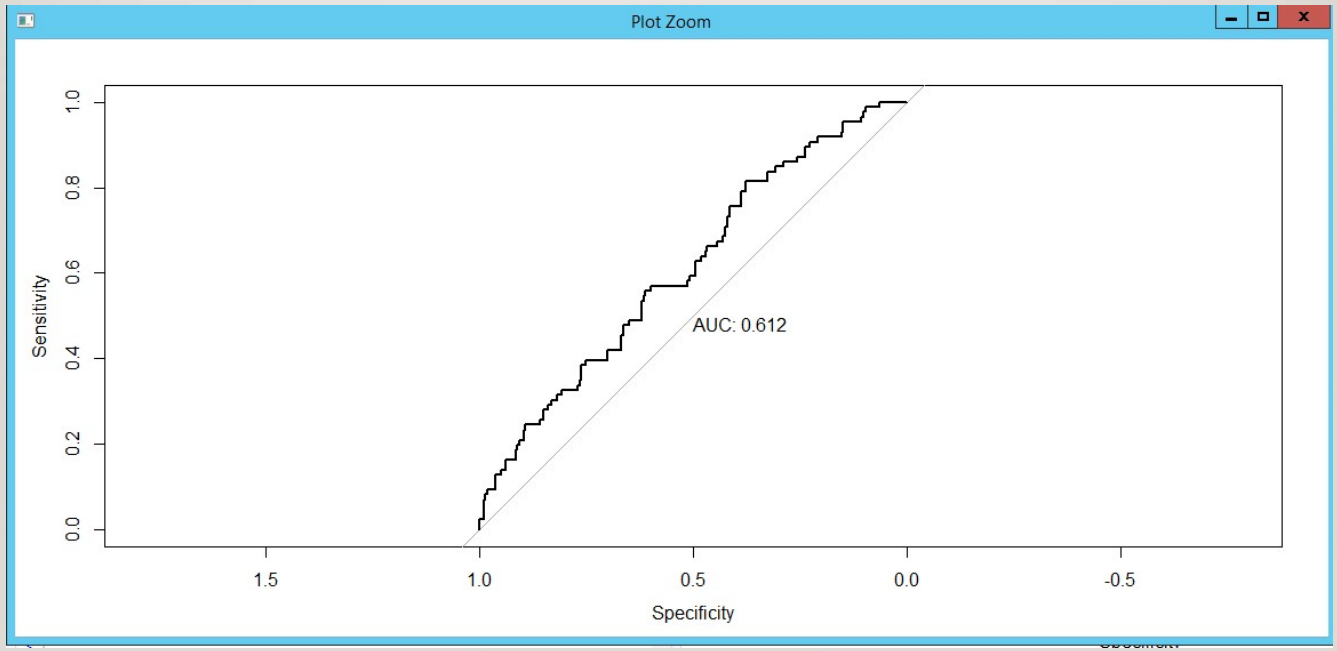
- AIC: 813.47
- Precision: 0.5
- Recall: 0.09302
- Accuracy: 0.71
- F1 score: 0.1568
- AUC: 0.61
- Confusion Matrix

	FALSE	TRUE
0	206	8
1	78	8

MODEL_GLM PILOT MODEL: COLLINEAR?

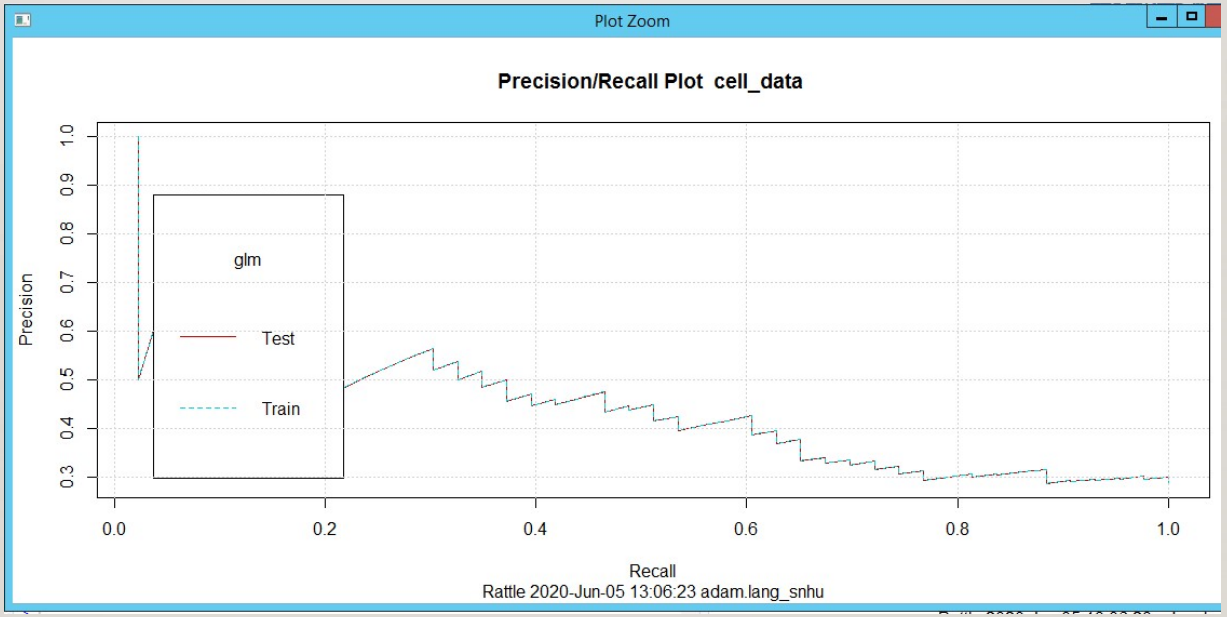
```
> vif(model_glm)
              GVIF Df GVIF^(1/(2*Df))
DROPVCE   2.284511  1        1.511460
UNANSVCE  2.181313  1        1.476927
MOUREC    4.023284  1        2.005813
EQPDAYS   1.349590  1        1.161719
PRZM_NUM  1.037234  3        1.006112
REFURB    1.035701  1        1.017694
WEBCAP    1.191401  1        1.091513
MOU       6.275647  1        2.505124
RECCHRGE  1.507447  1        1.227781
DIRECTAS  1.239559  1        1.113355
CHANGEM   1.022466  1        1.011171
>
```

MODEL_GLM: AUC-ROC CURVE



MODEL_GLM: PRECISION-RECALL CURVE

Low recall: poor ratio of positive instances correctly detected by the GLM model

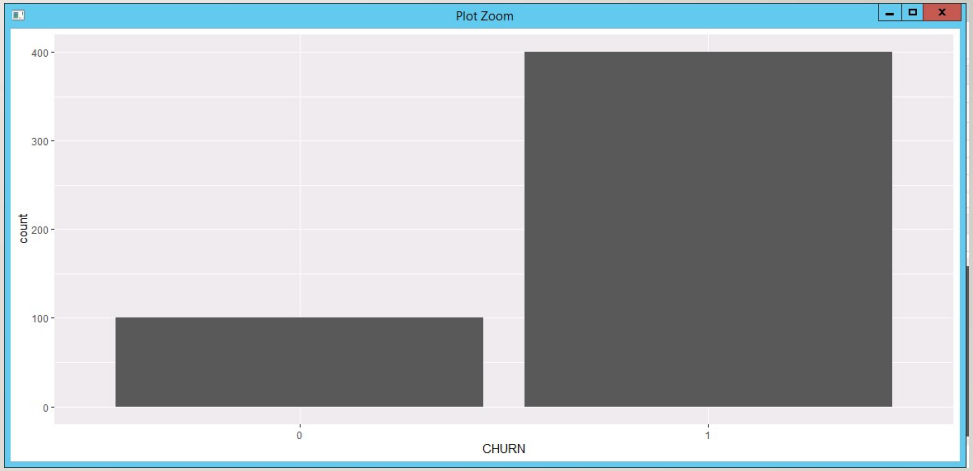


STEP 4: REBALANCE CHURN TARGET VARIABLE

```
192
193
194
195 ##Resample imbalanced target variable CHURN using SMOTE Technique##
196 ## Let's check the count of unique value in the target variable
197 as.data.frame(table(train$CHURN))
198 ## Smote : Synthetic Minority Oversampling Technique To Handle Class Imbalance In Binary Classification
199 balanced.data <- SMOTE(CHURN ~., train, perc.over = 100, k = 5, perc.under = 50)
200
```

SMOTE coding technique

- Perc.over is used to oversample the minority class
- Perc.under is used to undersample majority class



Newly balanced target variable

- Churners (1) now are oversampled

STEP 5: LOGISTIC REGRESSION PILOT WITH BALANCED DATA (MODEL_STEP2)

- Model and Stepwise Regression with pilot variables

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5305   0.3010   0.5501   0.6665   1.6471

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.7231123  0.3231293   2.238 0.025231 *
DROPVCE      0.0356285  0.0217914   1.635 0.102053
MOUREC       0.0031969  0.0015078   2.120 0.033983 *
EQPDAYS      0.0021009  0.0006320   3.324 0.000887 ***
PRZM_NUM1    0.9622185  0.7801053   1.233 0.217409
PRZM_NUM2    0.4583780  0.2654908   1.727 0.084252 .
PRZM_NUM3   -0.6523978  0.4600324  -1.418 0.156145
RETCALLS1    1.1430821  0.5485914   2.084 0.037190 *
MOU         -0.0018297  0.0004931  -3.710 0.000207 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 500.40  on 499  degrees of freedom
Residual deviance: 450.48  on 491  degrees of freedom
AIC: 468.48

Number of Fisher Scoring iterations: 5
```

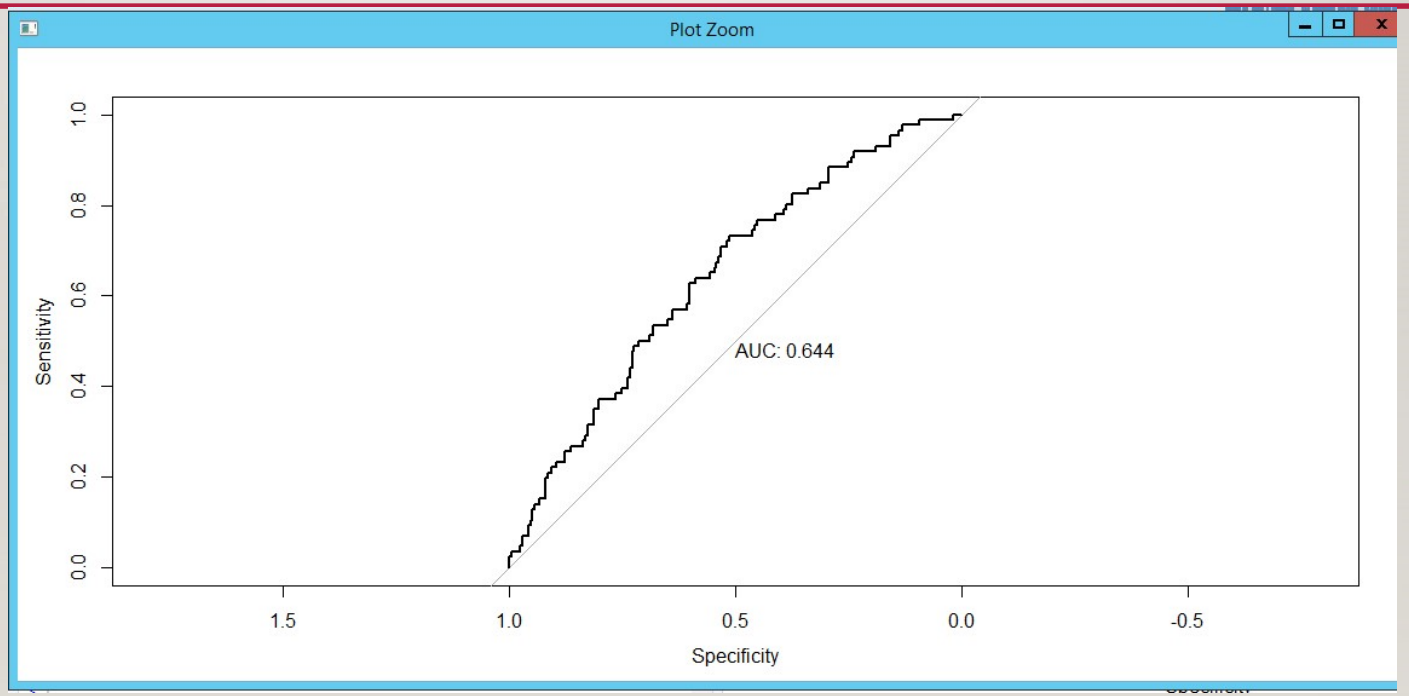
- AIC: 468.48
- Precision: 0.5
- Recall: 0.09302
- Accuracy: 0.71
- F1 score: 0.1568
- **AUC: 0.64**
- Confusion Matrix:

	Predicted		
Actual	0	1	Error
0	0	20	100
1	0	80	0

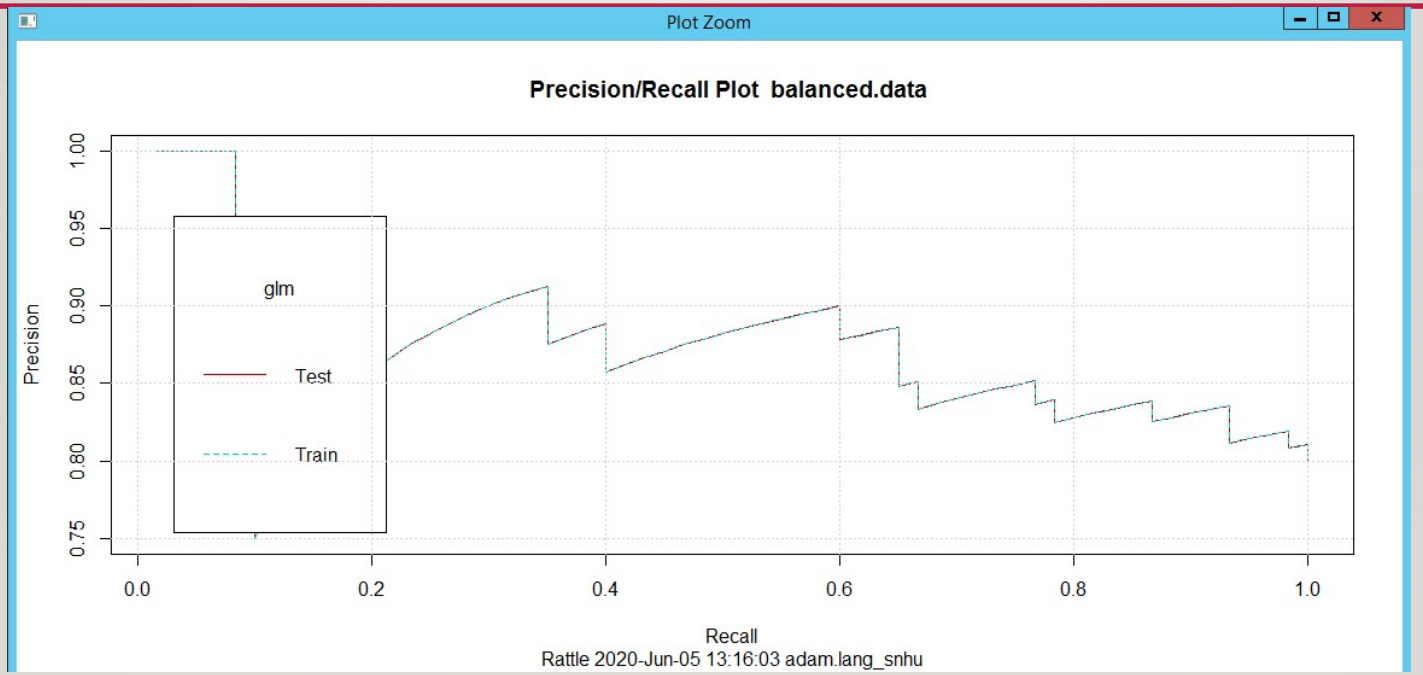
LOGISTIC REGRESSION PILOT WITH BALANCED DATA (MODEL_STEP2): COLLINEAR?

```
> vif(model_step2)
              GVIF Df GVIF^(1/(2*Df))
DROPVCE   2.355236  1         1.534678
MOUREC    5.097302  1         2.257720
EQPDAYS   1.136209  1         1.065931
PRZM_NUM  1.062623  3         1.010175
RETCALLS  1.018070  1         1.008995
MOU       5.855702  1         2.419856
```

LOGISTIC PILOT WITH BALANCED DATA (MODEL_STEP2): AUC-ROC CURVE

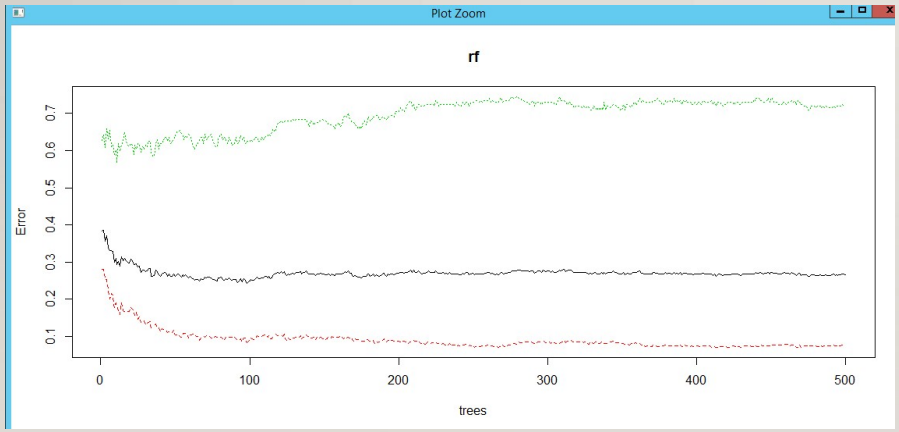


LOGISTIC PILOT WITH BALANCED DATA (MODEL_STEP2): PRECISION-RECALL CURVE

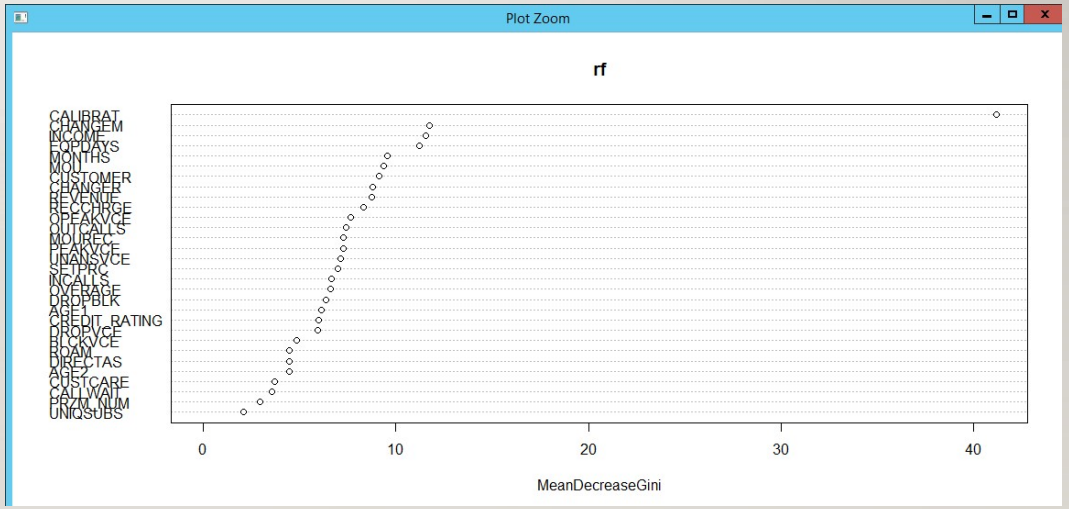


STEP 6: RANDOM FOREST MODEL #1

500 trees, 9 variables split at each node
OOB estimate of error rate: 26.59%
Class error for 0: 0.0754
Class error for 1: 0.720000



OOB error plot



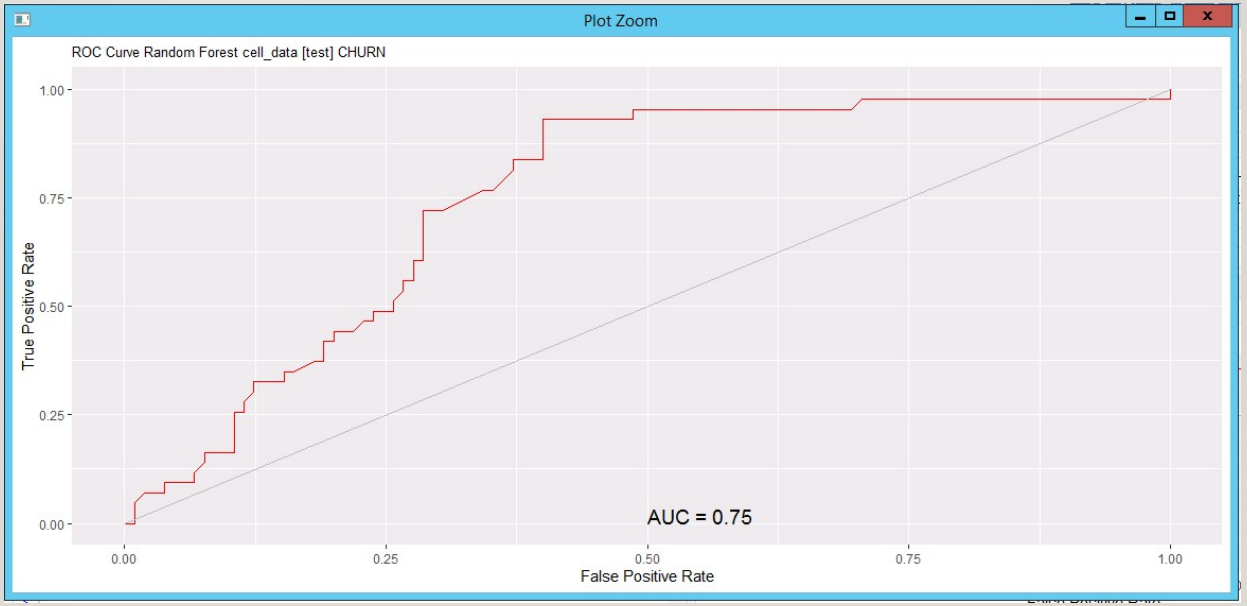
Variable Importance Plot

RANDOM FOREST MODEL #1

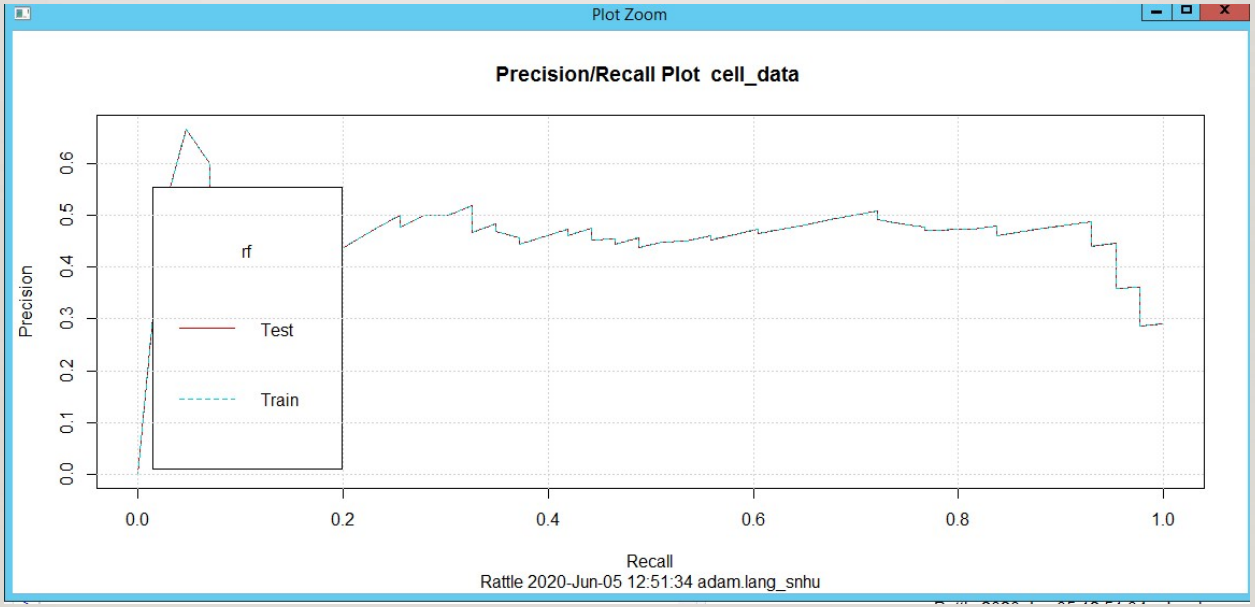
- Precision: 0.58
- Recall: 0.64
- F1 score: 0.61
- AUC: 0.75

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 198  62
         1  16  24

               Accuracy : 0.74
                 95% CI : (0.6865, 0.7887)
    No Information Rate : 0.7133
    P-Value [Acc > NIR] : 0.1693

                  Kappa : 0.2432

 Mcnemar's Test P-Value : 0.0000003483

            Sensitivity : 0.2791
            Specificity : 0.9252
         Pos Pred Value : 0.6000
         Neg Pred Value : 0.7615
             Prevalence : 0.2867
         Detection Rate : 0.0800
   Detection Prevalence : 0.1333
      Balanced Accuracy : 0.6022

       'Positive' Class : 1
```

RANDOM FOREST #1: AUC-ROC CURVE

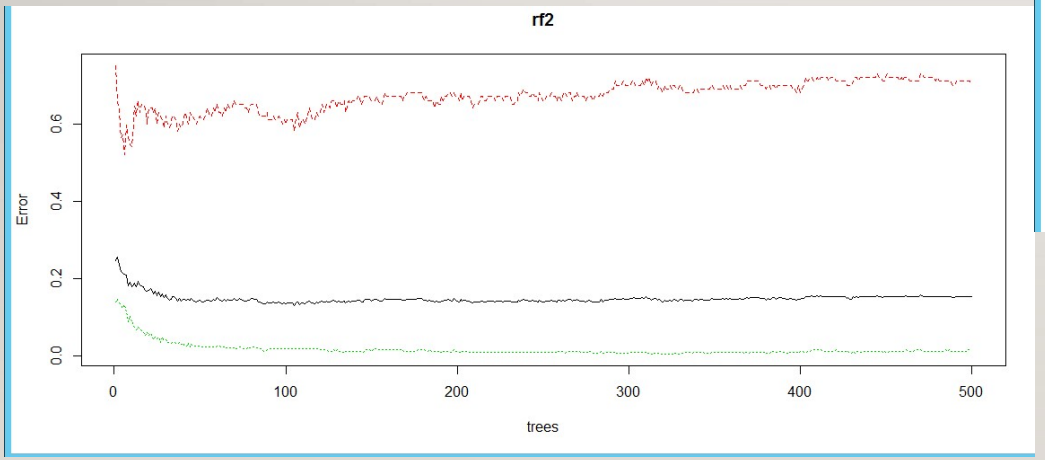


RANDOM FOREST #1: PRECISION-RECALL CURVE

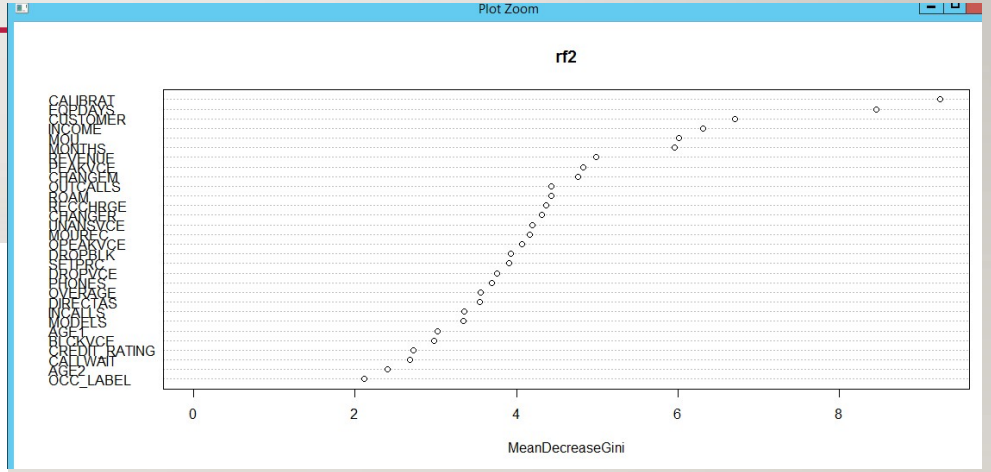


STEP 7: RANDOM FOREST #2: BALANCED DATA

500 trees, 9 variables split at each node
OOB estimate of error rate: 13%
Class error for 0: 0.6200
Class error for 1: 0.0075



OOB error plot.



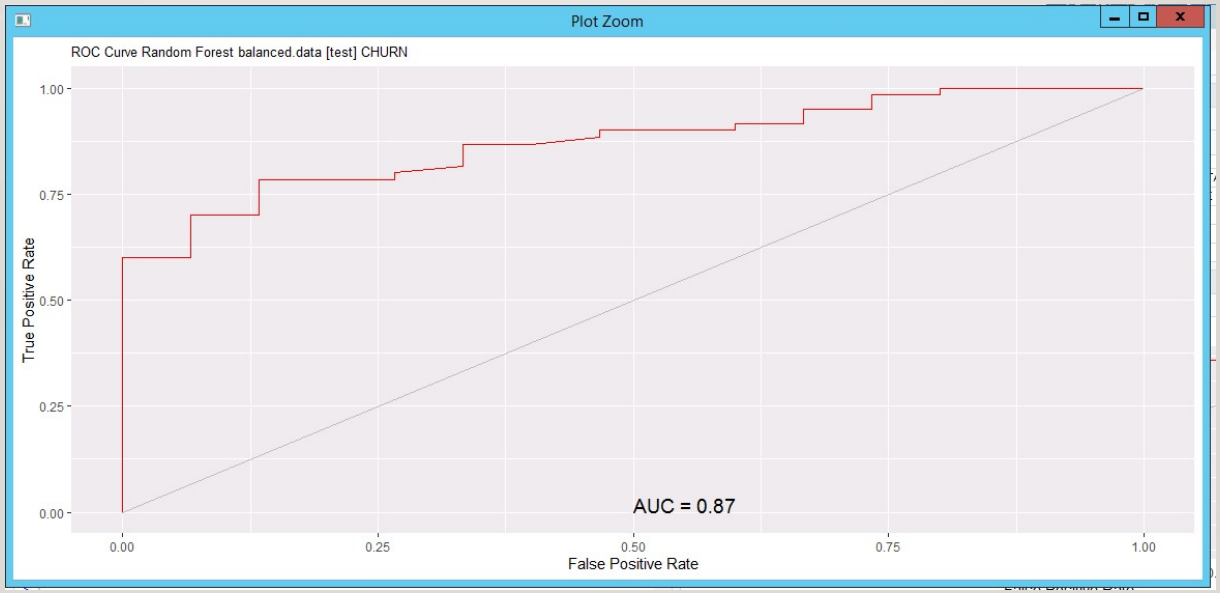
Variable Importance Plot

RANDOM FOREST #2

- Precision: 0.335
- Recall: 1.0
- F1 score: 0.50
- AUC: 0.87

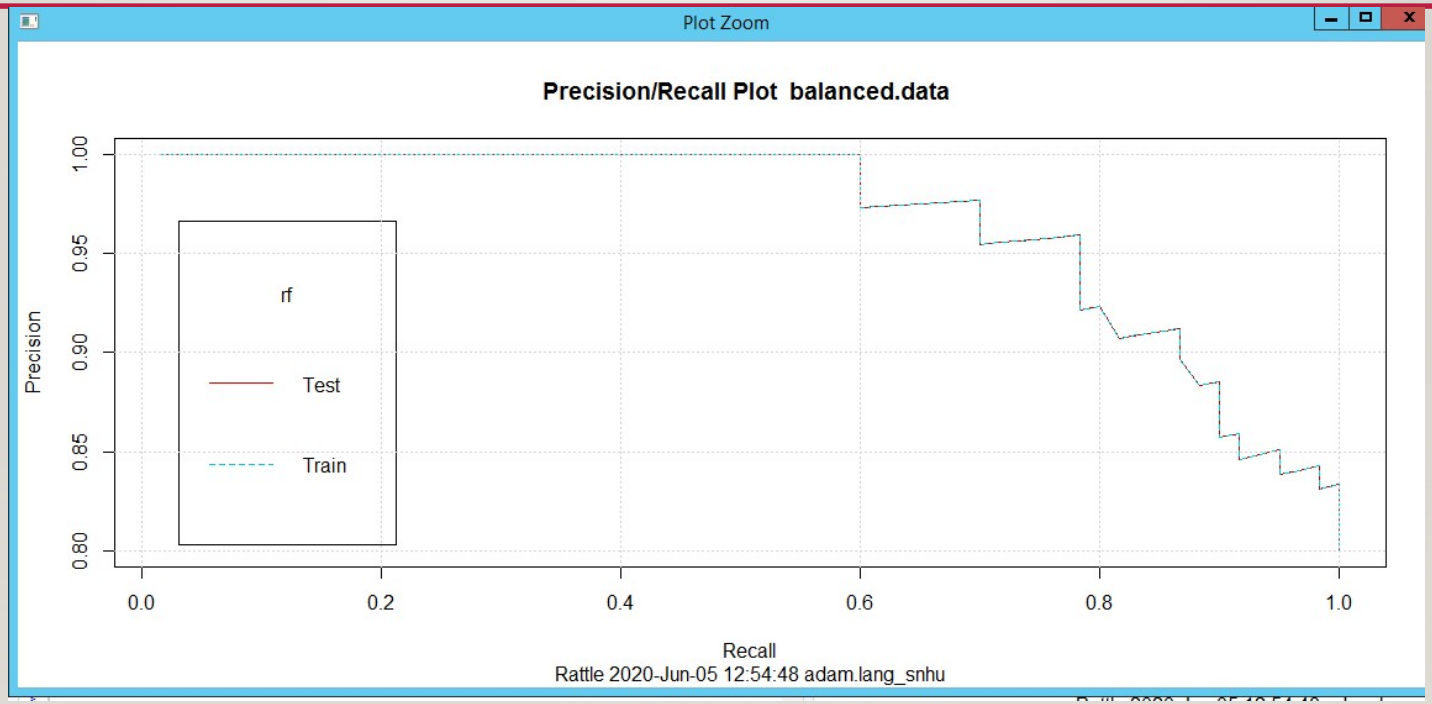
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  36   2
         1 178  84

               Accuracy : 0.4
                 95% CI : (0.3441, 0.4579)
    No Information Rate : 0.7133
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0899

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9767
            Specificity : 0.1682
         Pos Pred Value : 0.3206
         Neg Pred Value : 0.9474
             Prevalence : 0.2867
         Detection Rate : 0.2800
   Detection Prevalence : 0.8733
      Balanced Accuracy : 0.5725

       'Positive' Class : 1
```

RANDOM FOREST #2: AUC-ROC CURVE



RANDOM FOREST #2: PRECISION-RECALL CURVE

High recall: improved ratio of positive instances correctly detected by the Random Forest



STEP 8: LOGISTIC REGRESSION WITH BALANCED DATA, RANDOM FOREST #1 VARIABLE SELECTION (MODEL_STEP3)

Stepwise Regression results are below

Console | Terminal | Jobs

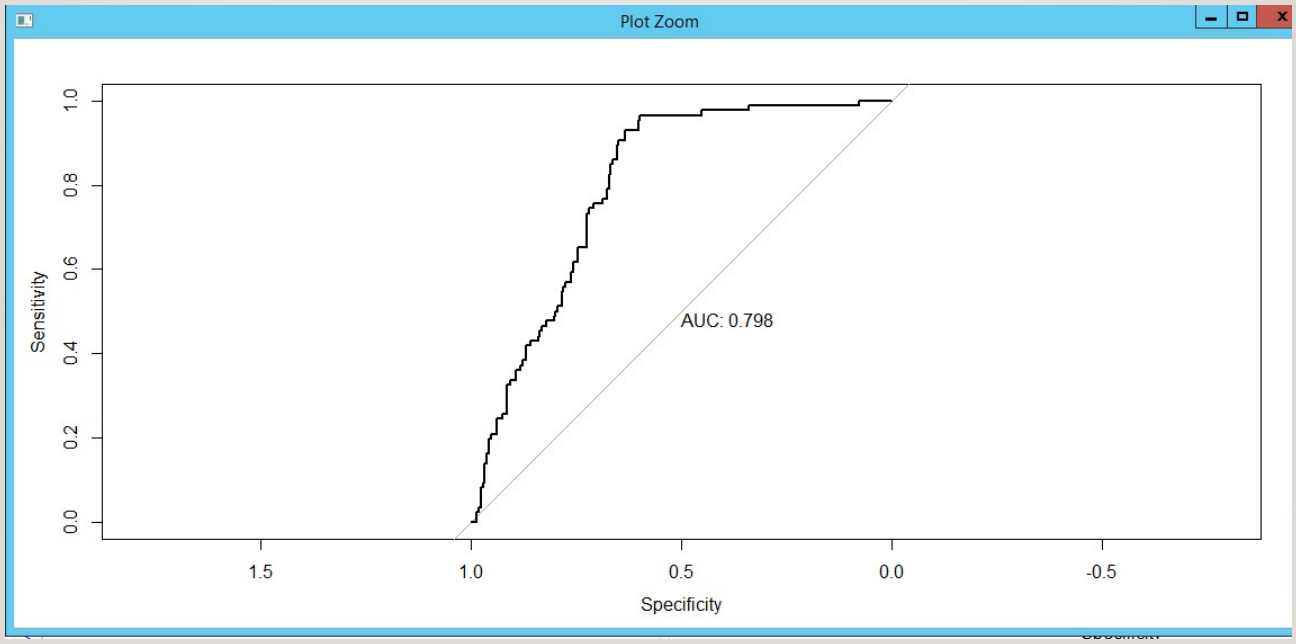
~/

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7604  -0.9255  -0.2266   0.9832   2.8942

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.3492851  0.4638610  -7.220 5.18e-13 ***
CALIBRAT1    3.6222736  0.3750297   9.659  < 2e-16 ***
EQPDAYS      0.0005957  0.0004137   1.440   0.1499
MOU         -0.0004252  0.0002936  -1.448   0.1475
RECCHRGE    -0.0150139  0.0060165  -2.495   0.0126 *
REVENUE      0.0073321  0.0036983   1.983   0.0474 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 821.79  on 676  degrees of freedom
Residual deviance: 585.03  on 671  degrees of freedom
AIC: 597.03

Number of Fisher Scoring iterations: 6

> vif(model_step3)
CALIBRAT  EQPDAYS      MOU RECCHRGE  REVENUE
1.005184 1.138090 2.082521 1.694487 2.290819
>
```

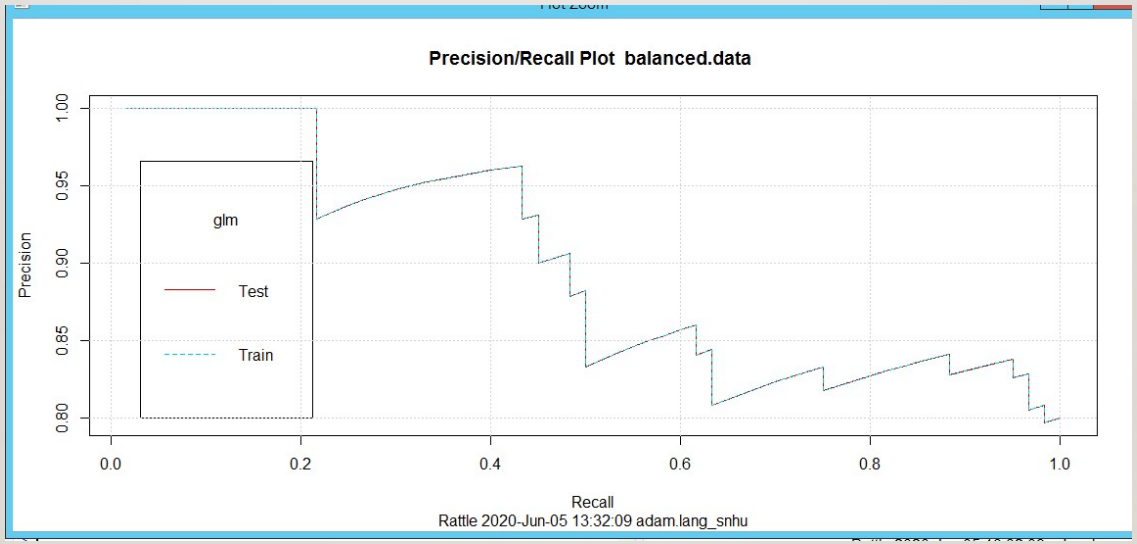
- AIC: 597.03
- Precision: 0.54
- Recall: 0.62
- Accuracy: 0.74
- F1 score: 0.58
- AUC: 0.79
- Confusion Matrix:

	Predicted		
Actual	0	1	Error
0	1.3	18.7	93.3
1	2.7	77.3	3.3

LOGISTIC REGRESSION WITH BALANCED DATA, RANDOM FOREST #1 VARIABLE SELECTION (MODEL_STEP3): AUC-ROC CURVE



LOGISTIC REGRESSION WITH BALANCED DATA, RANDOM FOREST #1 VARIABLE SELECTION (MODEL_STEP3): PRECISION-RECALL CURVE



STEP 9: LOGISTIC REGRESSION WITH BALANCED DATA AND RANDOM FOREST #2 VARIABLES (MODEL_GLM4)

Source

Console | Terminal | Jobs

```
                 Estimate   Std. Error z value Pr(>|z|)
(Intercept)    25.001424491  8.164372861   3.062 0.002197 **
CALIBRAT1       2.484652280  0.322412751   7.706 1.29e-14 ***
MONTHS         -0.083863507  0.024936207  -3.363 0.000771 ***
CUSTOMER       -0.000021030  0.000007504  -2.803 0.005070 **
INCOME1        -0.581732168  0.789278670  -0.737 0.461096
INCOME2        -0.279044238  1.236500067  -0.226 0.821456
INCOME3         0.191131442  0.602508141   0.317 0.751072
INCOME4        -1.890879693  0.564442965  -3.350 0.000808 ***
INCOME5         0.708155241  0.747279876   0.948 0.343311
INCOME6        -1.405597504  0.456427899  -3.080 0.002073 **
INCOME7        -1.447212864  0.467286847  -3.097 0.001955 **
INCOME8        -1.582106250  0.689556382  -2.294 0.021769 *
INCOME9        -1.174296935  0.588898117  -1.994 0.046146 *
EQPDAYS         0.001206118  0.000736916   1.637 0.101691
CREDIT_RATING2 -0.753189129  0.449382601  -1.676 0.093728 .
CREDIT_RATING3  0.956336568  0.526445302   1.817 0.069280 .
CREDIT_RATING4 -1.190245170  0.576976595  -2.063 0.039122 *
CREDIT_RATING5 -0.787570152  0.576627126  -1.366 0.171995
CREDIT_RATING6 -4.117826533  1.279114668  -3.219 0.001285 **
CREDIT_RATING7  1.480277618  1.249448187   1.185 0.236118
MOU            -0.001728954  0.000618721  -2.794 0.005200 **
RECCHRGE       -0.001806889  0.008256432  -0.219 0.826770
OPEAKVCE        0.002144731  0.002843875   0.754 0.450754
CHANGEM         0.000296204  0.000681358   0.435 0.663761
DROPBLK        -0.016447473  0.019530646  -0.842 0.399711
AGE2           -0.008326309  0.006761528  -1.231 0.218164
PEAKVCE         0.005377514  0.002605597   2.064 0.039034 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 500.40  on 499  degrees of freedom
Residual deviance: 351.16  on 473  degrees of freedom
AIC: 405.16

Number of Fisher Scoring iterations: 6
```

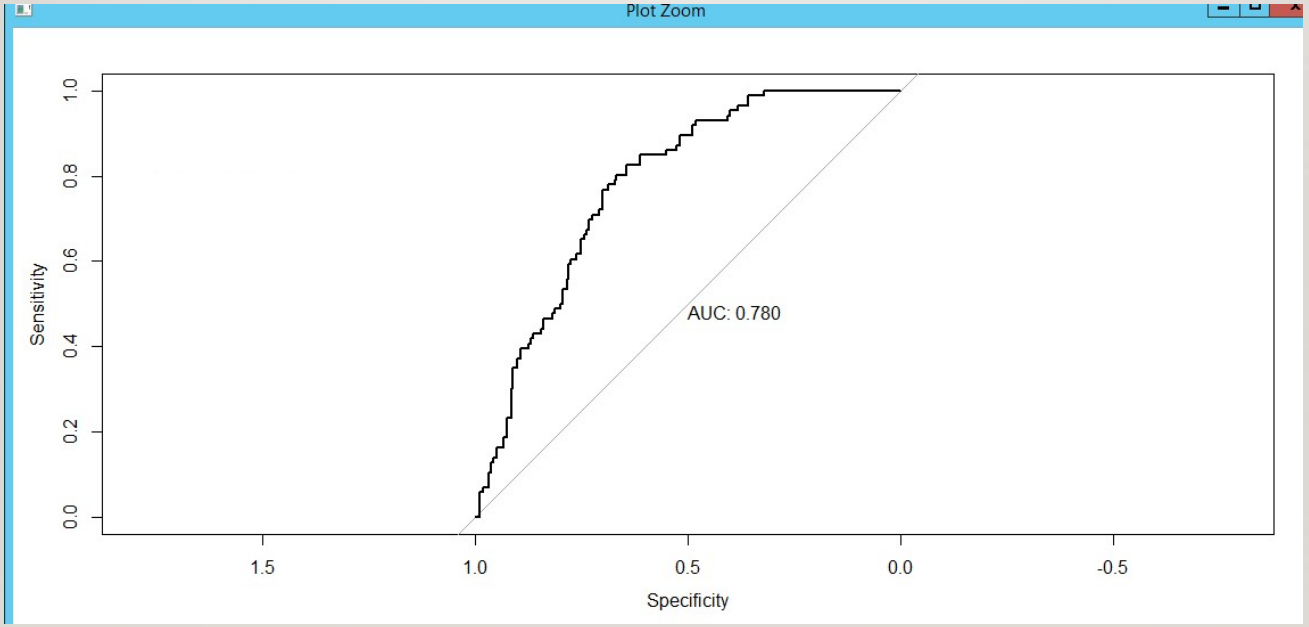
- AIC: 405.16
- Precision: 0.36
- Recall: 1.0
- Accuracy: 0.49
- F1 score: 0.53
- AUC: 0.78
- Confusion Matrix:

Actual	Predicted 0	1	Error
0	5.3	14.7	73.3
1	2.7	77.3	3.3

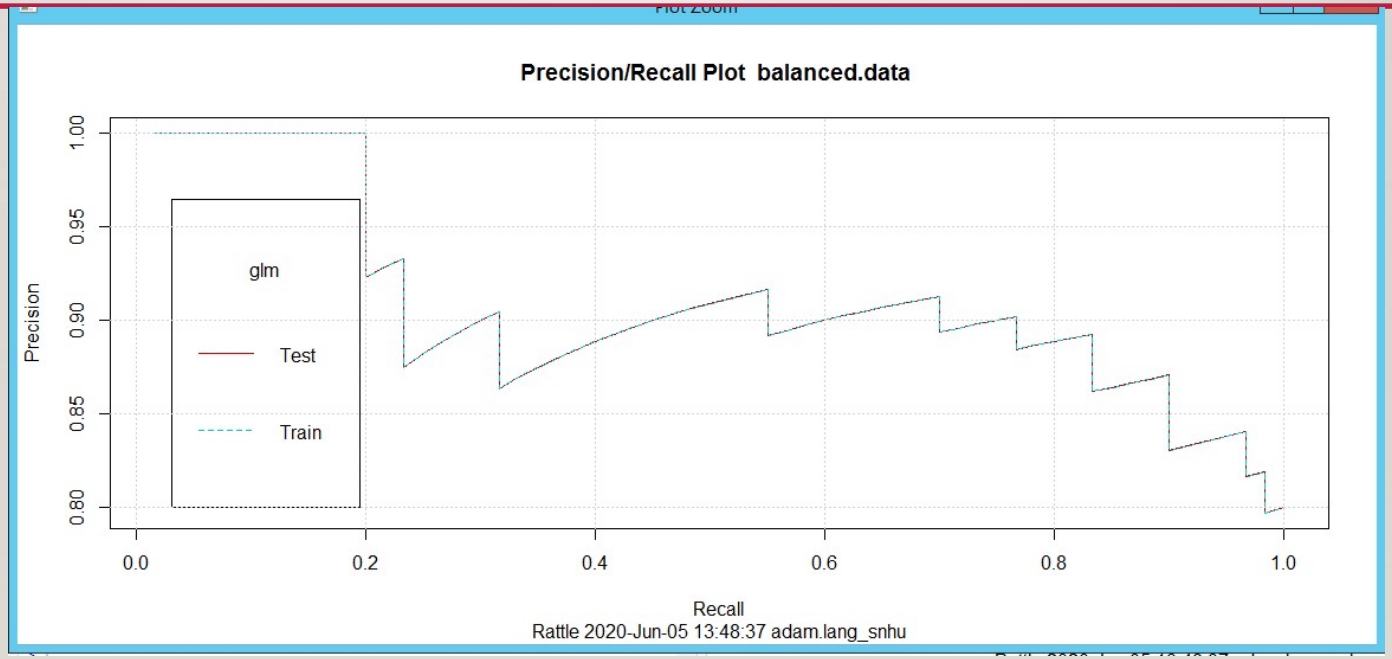
LOGISTIC REGRESSION WITH BALANCED DATA AND RANDOM FOREST #2 VARIABLES (MODEL_GLM4): COLLINEARITY?

```
> vif(model_glm4)
                    GVIF Df GVIF^(1/(2*Df))
CALIBRAT        1.405892  1        1.185703
MONTHS          2.796378  1        1.672238
CUSTOMER        2.068351  1        1.438176
INCOME          2.783430  9        1.058520
EQPDAYS         1.566203  1        1.251480
CREDIT_RATING   2.552363  6        1.081215
MOU             4.154080  1        2.038156
RECCHRGE        1.554540  1        1.246812
OPEAKVCE        2.650199  1        1.627943
CHANGEM         1.186855  1        1.089429
DROPBLK         2.248168  1        1.499389
AGE2            1.358398  1        1.165503
PEAKVCE         2.983455  1        1.727268
>
```

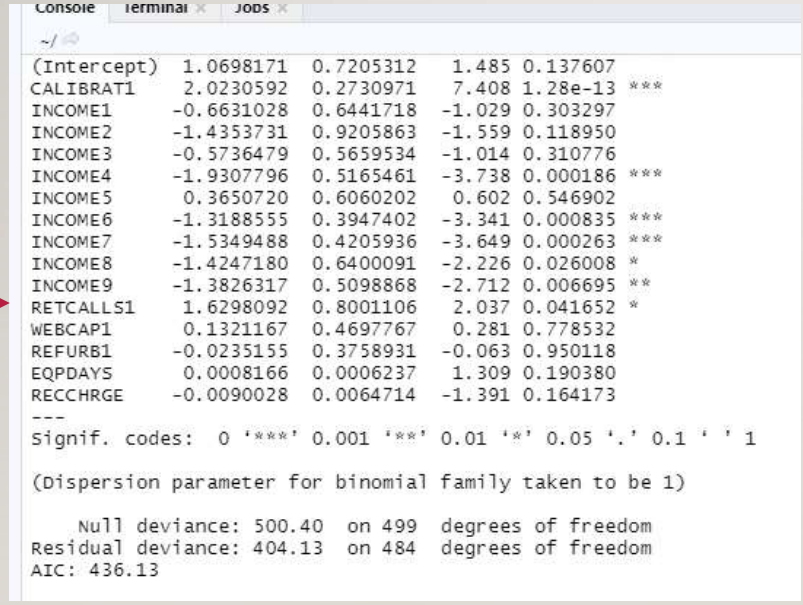
LOGISTIC REGRESSION WITH BALANCED DATA AND RANDOM FOREST #2 VARIABLES (MODEL_GLM4): AUC-ROC CURVE



LOGISTIC REGRESSION WITH BALANCED DATA AND RANDOM FOREST #2 VARIABLES (MODEL_GLM4): PRECISION-RECALL CURVE



STEP 10: LASSO REGRESSION FOR VARIABLE SELECTION (LASSO1)



```
Console   Terminal   Jobs
~/
(Intercept)  1.0698171  0.7205312   1.485 0.137607
CALIBRAT1    2.0230592  0.2730971   7.408 1.28e-13 ***
INCOME1     -0.6631028  0.6441718  -1.029 0.303297
INCOME2     -1.4353731  0.9205863  -1.559 0.118950
INCOME3     -0.5736479  0.5659534  -1.014 0.310776
INCOME4     -1.9307796  0.5165461  -3.738 0.000186 ***
INCOME5      0.3650720  0.6060202   0.602 0.546902
INCOME6     -1.3188555  0.3947402  -3.341 0.000835 ***
INCOME7     -1.5349488  0.4205936  -3.649 0.000263 ***
INCOME8     -1.4247180  0.6400091  -2.226 0.026008 *
INCOME9     -1.3826317  0.5098868  -2.712 0.006695 **
RETCALLS1    1.6298092  0.8001106   2.037 0.041652 *
WEBCAP1      0.1321167  0.4697767   0.281 0.778532
REFURB1     -0.0235155  0.3758931  -0.063 0.950118
EQPDAYS      0.0008166  0.0006237   1.309 0.190380
RECCHRGE    -0.0090028  0.0064714  -1.391 0.164173
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 500.40  on 499  degrees of freedom
Residual deviance: 404.13  on 484  degrees of freedom
AIC: 436.13
```

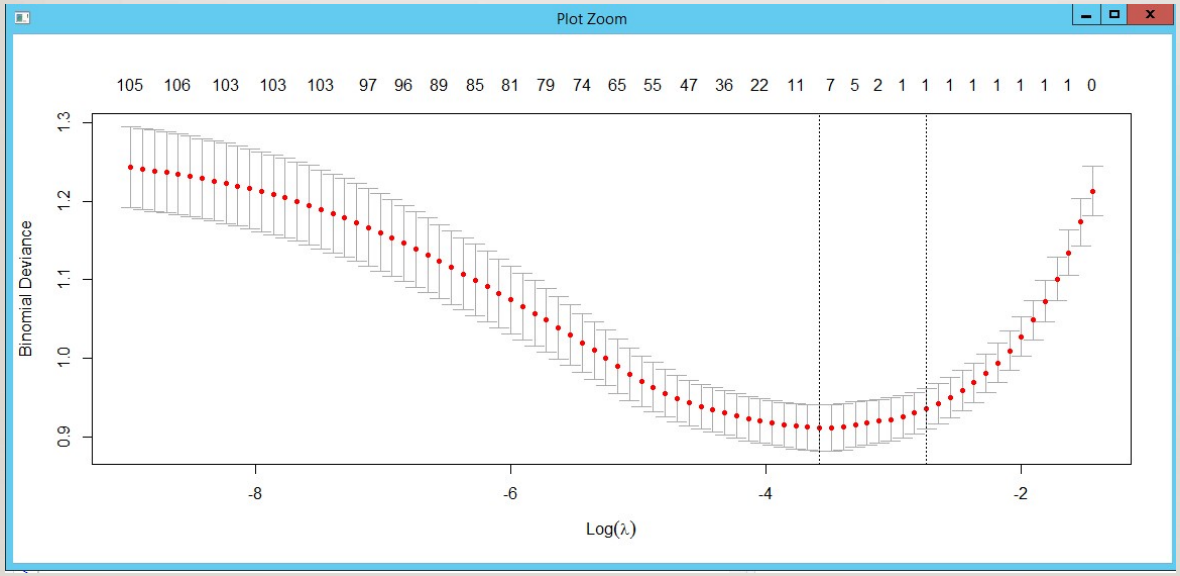
- AIC: 436.13
- Precision: 0.37
- Recall: 0.97
- Accuracy: 0.51
- F1 score: 0.536
- AUC: 0.75
- Confusion Matrix:

Actual	Predicted 0	1	Error
0	5.3	14.7	73.3
1	4.0	76.0	5.0

LASSO REGRESSION (LASSO1): COLLINEARITY?

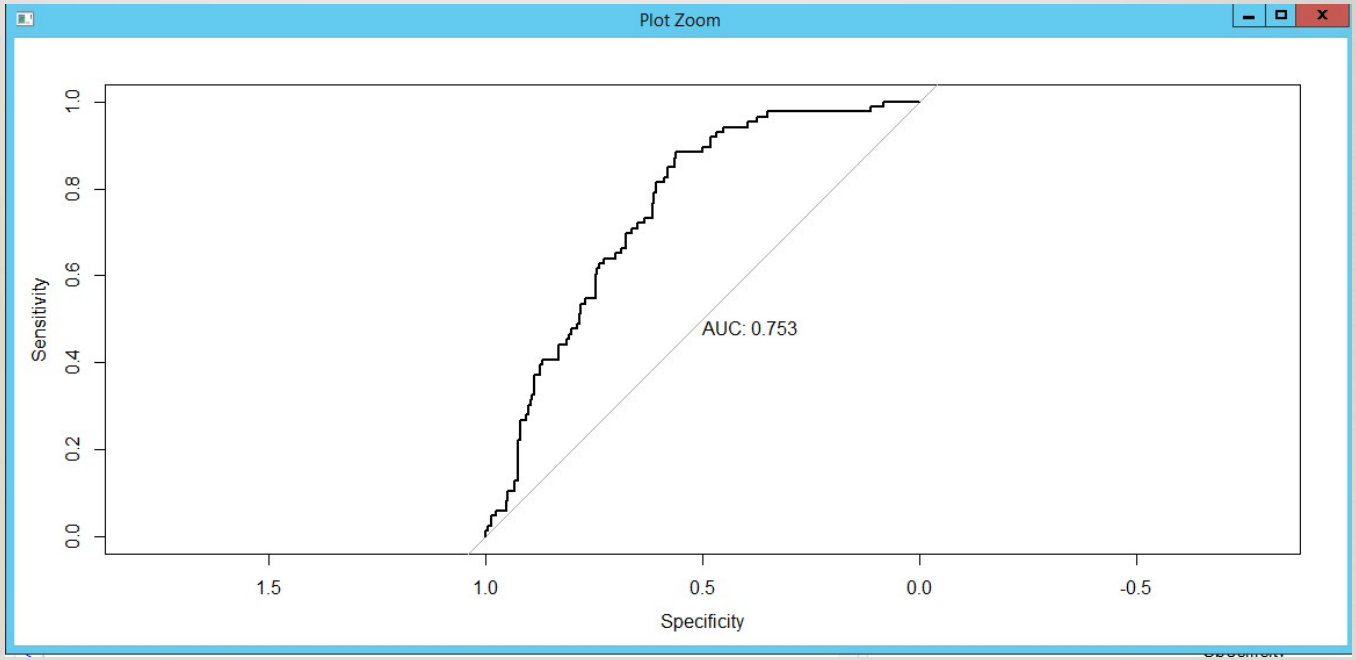
```
> #Check collinearity of Lasso model
> vif(lasso1)
             GVIF Df GVIF^(1/(2*Df))
CALIBRAT 1.244389  1        1.115522
INCOME   1.428140  9        1.019996
RETCALLS 1.167977  1        1.080730
WEBCAP   1.077022  1        1.037797
REFURB   1.060416  1        1.029765
EQPDAYS  1.190152  1        1.090941
RECCHRGE 1.088158  1        1.043148
>
```

LASSO MODEL (LASSO1): LOG OF COEFFICIENTS



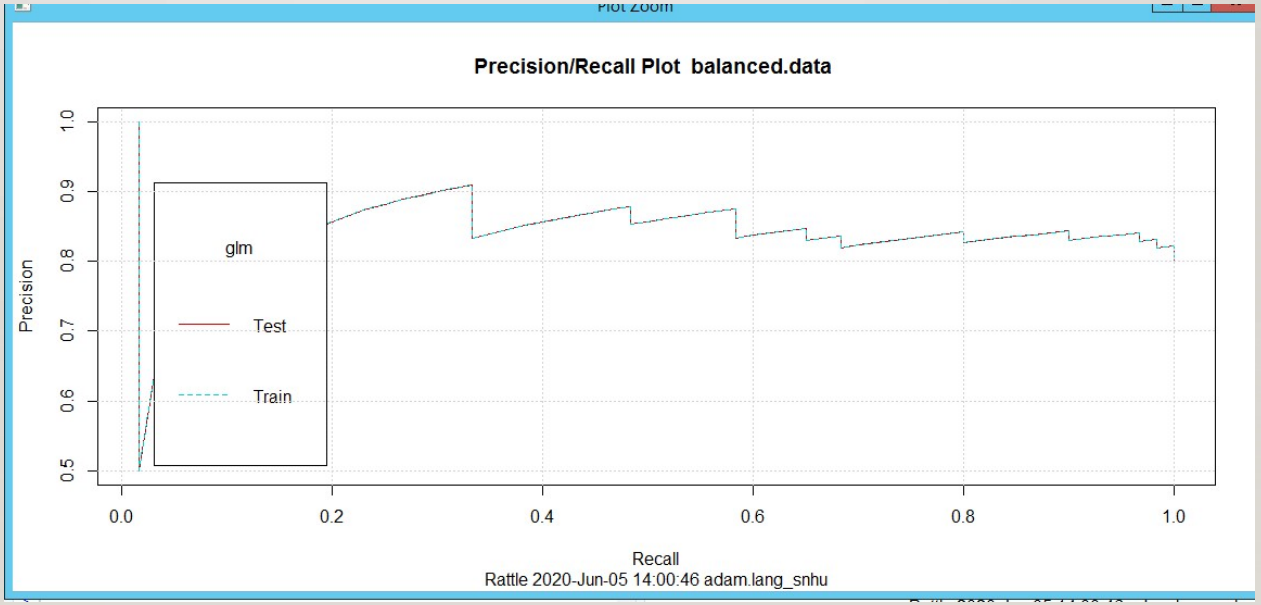
As coefficients are forced to zero, the model improves at predicting coefficients for variable selection

LASSO MODEL (LASSO1): AUC-ROC CURVE



LASSO MODEL (LASSO1): PRECISION-RECALL CURVE

Curve is similar to Random Forest #1



GOODNESS OF FIT TESTS: LOGISTIC REGRESSION

- Likelihood Ratio (lower = better)

Lasso: 0.0000001821

Model Step3: 0.471

Model Step 2: 0.9102

Model Step: 0.2851

- Pseudo R2 (McFadden) (higher = better)

Lasso: 0.19

Model Step4: 0.28809

Model 4: 0.2982

Model Step3: 0.28809

Model 3: 0.30601

Model Step 2: 0.05568

Model 2: 0.0635

Model Step: 0.0323

Model pilot: 0.044

GOODNESS OF FIT: LOGISTIC REGRESSION

- BIC values (lower = better)

```
             df      BIC
model_glm    14 876.7220
model_glm2   15 561.7965
model_glm3   21 707.1800
model_glm4   27 518.9592
lasso1       16 503.5646
```

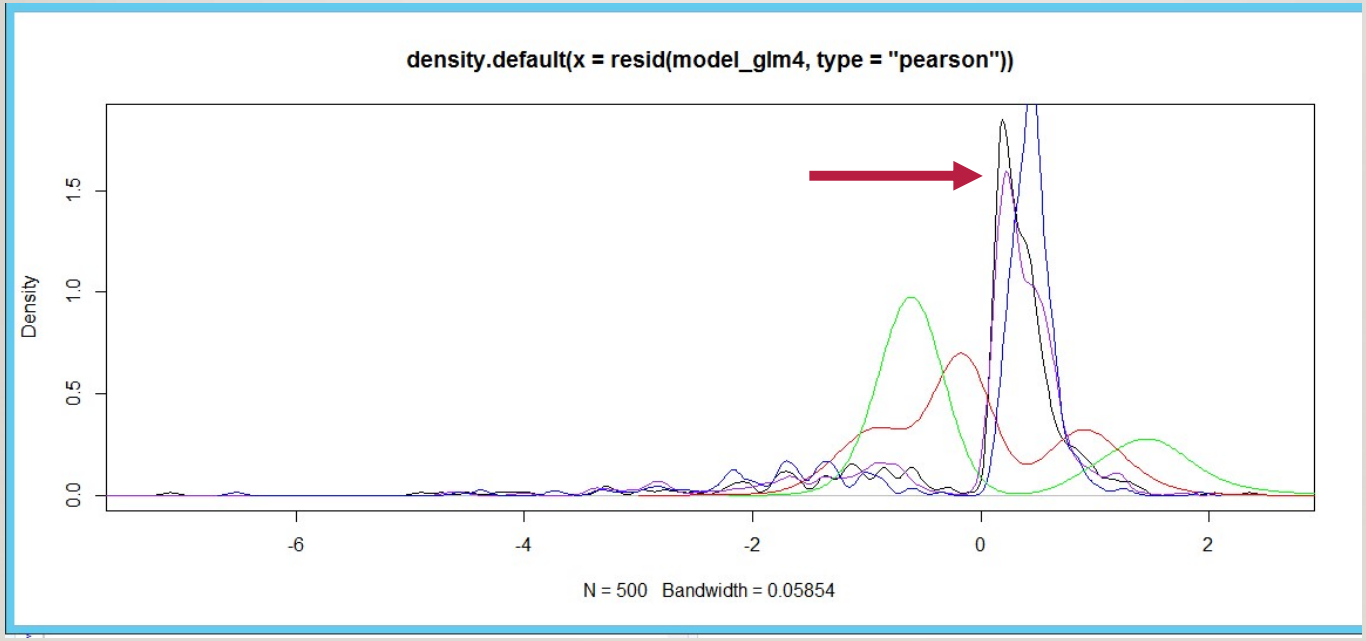
```
> BIC(model_step,model_st
             df      BIC
model_step    6 834.3021
model_step2   6 509.8834
model_step3   6 624.1394
model_step4   6 624.1394
```

- AIC values (lower = better)

```
> AIC(model_glm,model_glm2,mo
             df      AIC
model_glm    14 813.4746
model_glm2   15 498.5773
model_glm3   21 612.3089
model_glm4   27 405.1647
lasso1       16 436.1309
```

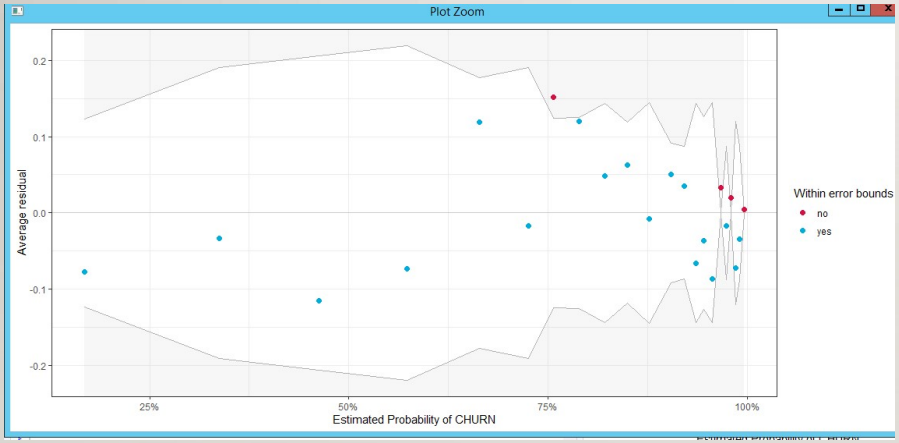
```
> AIC(model_step,model_ste
             df      AIC
model_step    6 807.1961
model_step2   6 484.5957
model_step3   6 597.0334
model_step4   6 597.0334
```

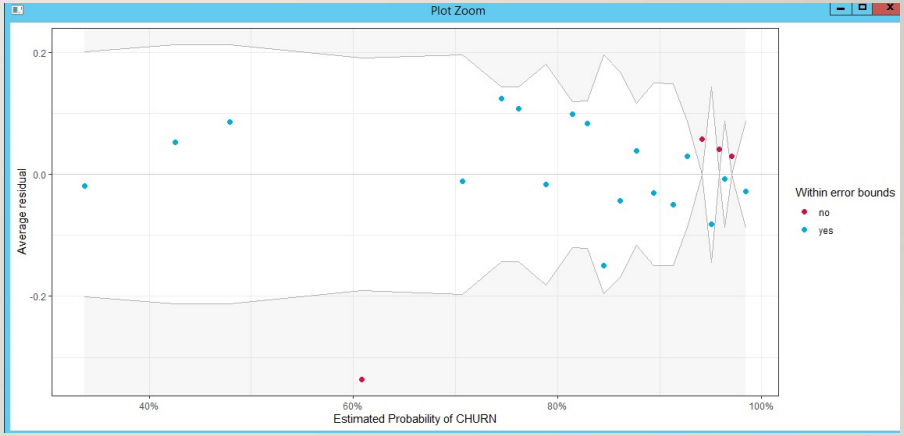
GOODNESS OF FIT GLM MODELS: PEARSON RESIDUALS PLOT



Glm4 = black
Lasso = purple
Glm3 = red
Glm2 = blue
Glm= green

GOODNESS OF FIT GLM MODELS: BINNED RESIDUALS

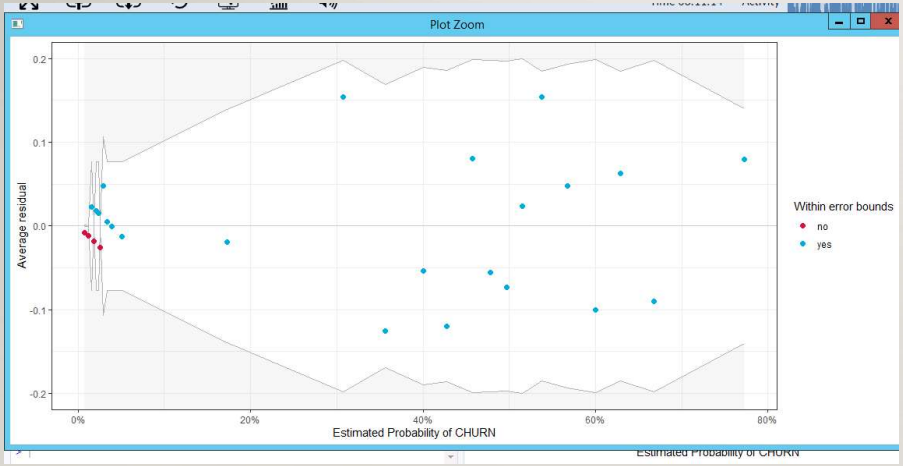


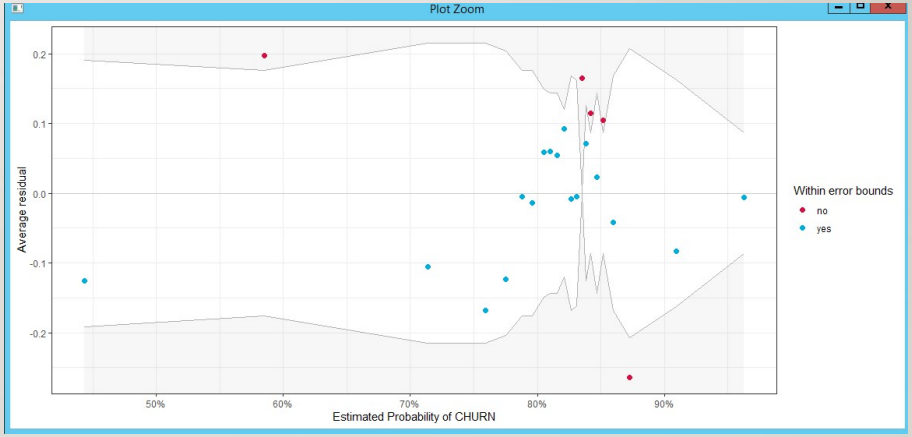


- Model_glm4 has 82% of residuals in error bounds

- Lasso model has 82% of residuals in error bounds

GOODNESS OF FIT GLM MODELS: BINNED RESIDUALS

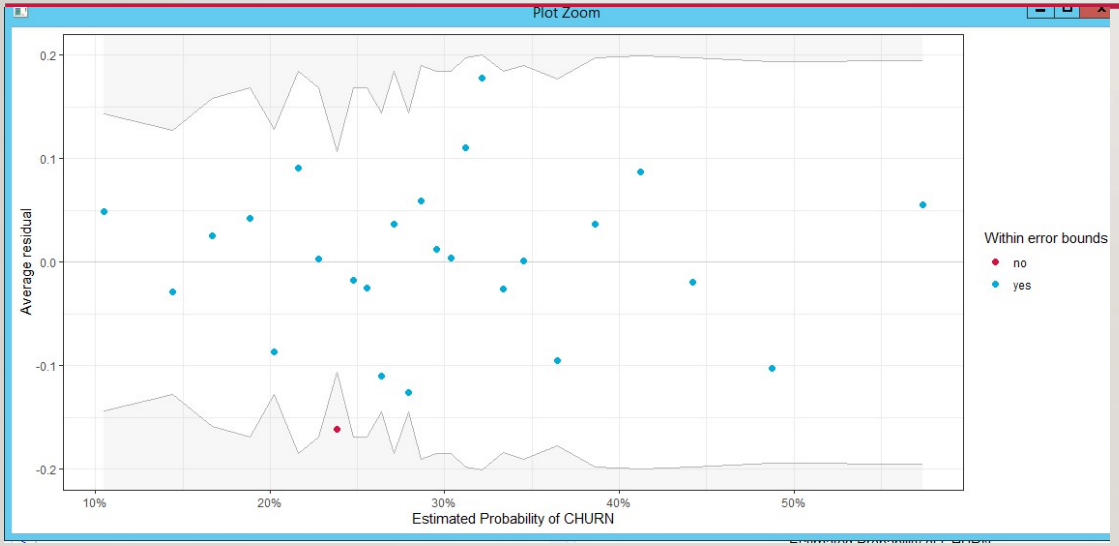




- Model_glm3 has 85% within error bounds

- Model_glm2 has 77% of residuals within bounds

GLM MODELS: BINNED RESIDUALS



Ironically the original pilot model has 96% of the residuals within the error bounds

WHAT IS THE BEST REGRESSION MODEL?

MODEL_GLM4

- AUC: 0.78
- AIC: 405.1
- Conf Matrix:
- Precision: 0.36
- Recall: 1.0

	Predicted		
Actual	0	1	Error
0	5.3	14.7	73.3
1	2.7	77.3	3.3

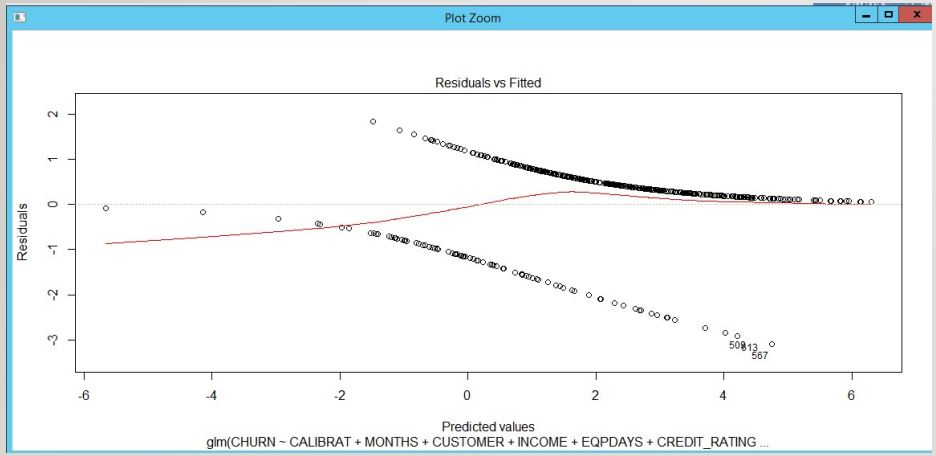
Next Best GLM Models?

- LASSO
- MODEL_STEP2 (stepwise regression)

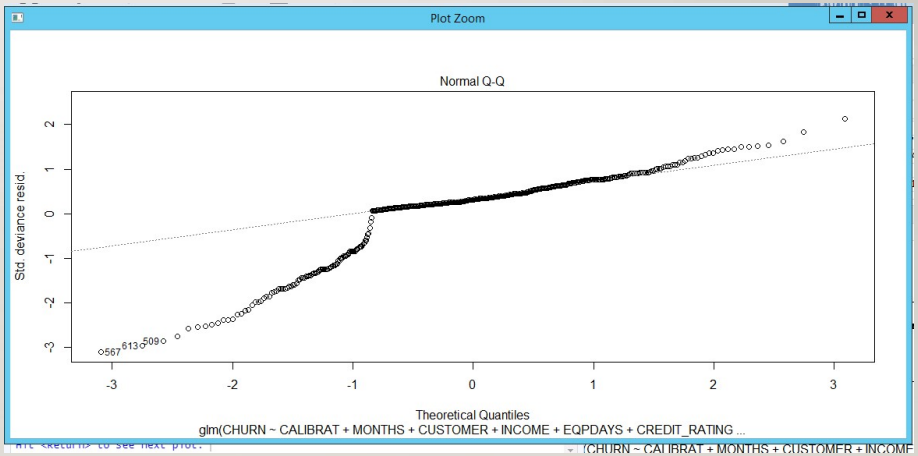
Variable selection techniques

- All 3 are represented here
- These techniques worked!

FINAL PLOTS OF BEST MODEL: MODEL_GLM4

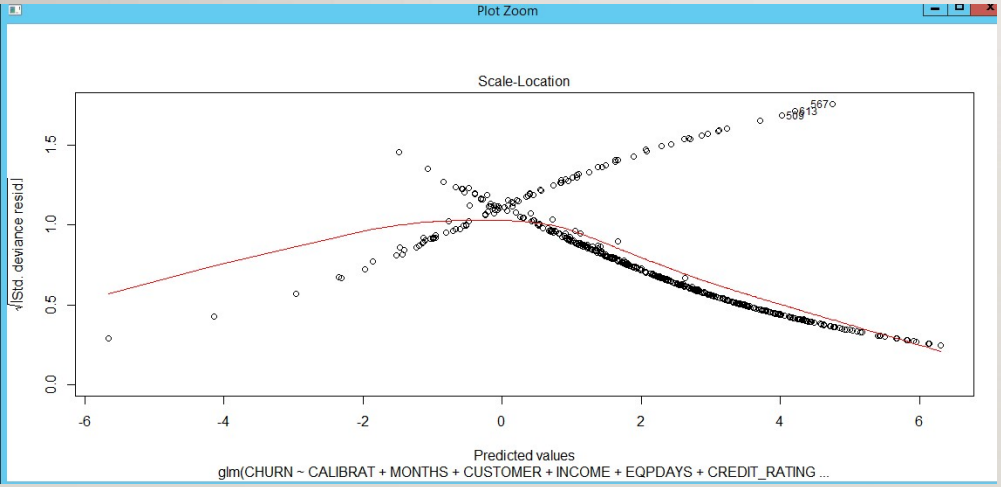


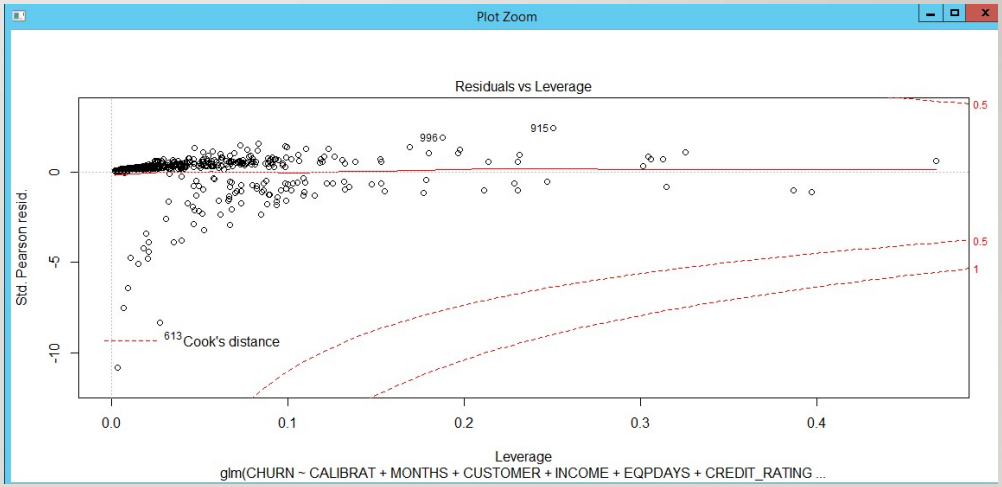
Residuals vs. Fitted Plot



Q-Q Plot: CHURN variable relatively normally distributed

FINAL PLOTS OF BEST MODEL: MODEL_GLM4

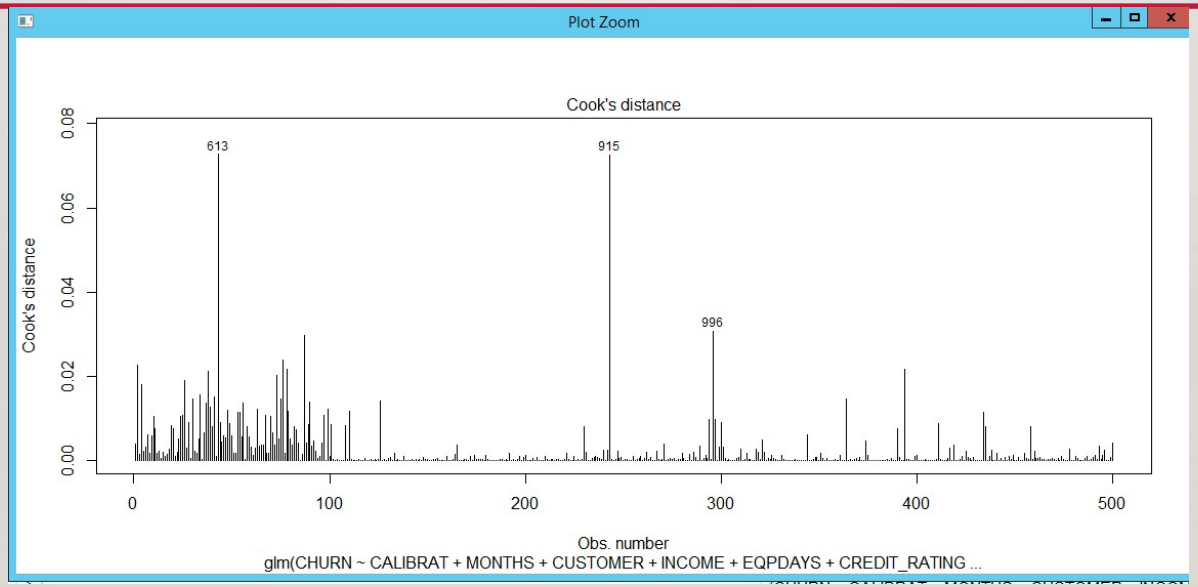




Scale Location Plot: uniform variance, "homoscedastic" predictor values

Residuals vs. Leverage Plot: influential values!

COOK'S DISTANCE PLOT: MODEL_GLM4



Cooks Distance: Model outliers? 3 influential observations are numbered

COOK'S DISTANCE: TOP 3 INFLUENTIAL VALUES (MODEL_GLM4)

```
# A tibble: 3 x 23
  .rownames CHURN CALIBRAT MONTHS CUSTOMER INCOME EQPDAYS CREDIT_RATING   MOU
  <chr>     <fct> <fct>      <dbl>    <dbl> <fct>    <dbl> <fct>          <dbl>
1 613       0     1             12  1062597 8          367 7              478
2 915       1     1             11  1066289 6          346 6              221.
3 996       1     1             41  1098468 6          179 5              406
```

- Strong influence on the fitted values of the regression model
- We start to see some trends with the churners in INCOME, CREDIT RATING

MOST INFLUENTIAL OBSERVATION (MODEL_GLM4)

```
~/
> model.data %>%
+   filter(abs(.std.resid) > 3)
# A tibble: 1 x 23
  .rownames CHURN CALIBRAT MONTHS CUSTOMER INCOME EQPDAYS CREDIT_RATING   MOU
  <chr>     <fct> <fct>     <dbl>    <dbl> <fct>    <dbl> <fct>          <dbl>
1 567       0     1            13  1049453 0          396 1               52.2
```

Strongest influence: profile of a non-churning customer?

HOW DO WE COMPARE MODELS?

MODEL_GLM4

- AIC: 405.16
- Precision: 0.36
- Recall: 1.0
- Accuracy: 0.49
- F1 score: 0.53
- AUC: 0.78
- Confusion Matrix:

Actual	Predicted 0	1	Error
0	5.3	14.7	73.3
1	2.7	77.3	3.3

Random Forest model #2

- Precision: 0.335
- Recall: 1.0
- F1 score: 0.50
- AUC: 0.87

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  36   2
         1 178  84

               Accuracy : 0.4
                 95% CI : (0.3441, 0.4579)
    No Information Rate : 0.7133
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0899

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9767
            Specificity : 0.1682
         Pos Pred Value : 0.3206
         Neg Pred Value : 0.9474
             Prevalence : 0.2867
         Detection Rate : 0.2800
   Detection Prevalence : 0.8733
      Balanced Accuracy : 0.5725

       'Positive' Class : 1
```

MODEL COMPARISON: CROSS VALIDATION

- Library 'caret' (Brownlee, 2019)
- Library 'mlbench'
- 'Repeatedcv' = 10 folds cross validation (Krstajic et al. 2014)
  - ✓ To overcome the "Bias-variance trade-off"

3 folds cross validation example (Yiu, 2020)

CROSS VALIDATION STEPS

```
#First set up trainControl for repeated 10 folds cross validation
set.seed(123)
train.control <- trainControl(method = "repeatedcv",
                              number = 10, repeats = 3)
# Train the models
```

Step 1: Set up train control for 10 folds cross validation

```
                              number = 10, repeats = 3)
# Train the models
##1. model_glm4 will be trained/cross validated first as this was the best GLM model
glm4 <- train(CHURN~ CALIBRAT + MONTHS + CUSTOMER + INCOME +
                EQPDAYS + CREDIT_RATING + MOU + RECCHRGE +
                OPEAKVCE + CHANGEM + DROPBLK + AGE2 + PEAKVCE
              , data = balanced.data, method = "glm",
              trControl = train.control)
# Summarize the results for this model
print(glm4)
```

Step 2: Train each model on the train control

```
##Now Perform resampling using caret package to compare results of all models
results <- resamples(list(R1=rfor1, R2=rfor2,log4=glm4,log2=glm2,lass=lasso))
#Print summary of all the models together
#This will print Accuracy and Kappa values of the cross validation process
#with summary statistics.
summary(results)
```

Step 3: Resample all the models together

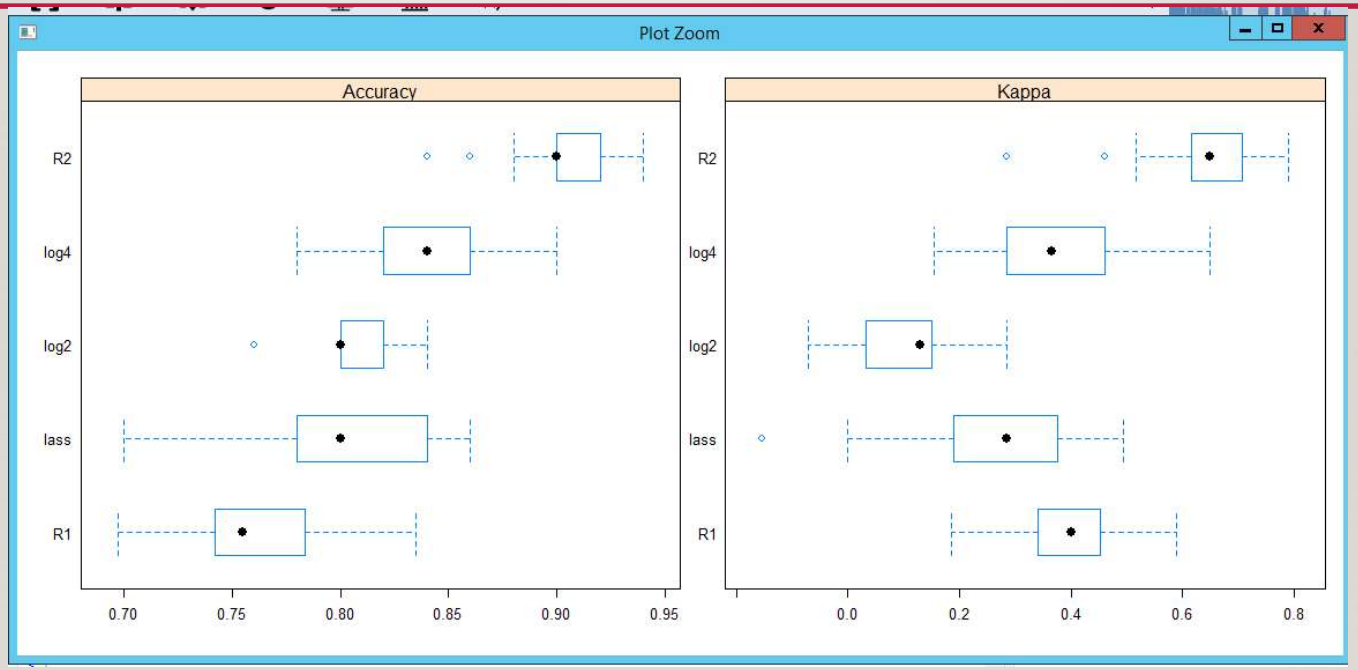
```
##Now we can print comparison plots of the cross validation of all models
# box and whisker plots to compare all cross validated models (resamples)
scales <- list(x=list(relation="free"), y=list(relation="free"))
bwplot(results, scales=scales)

# density plots of accuracy of cross validated models (resamples)
scales <- list(x=list(relation="free"), y=list(relation="free"))
densityplot(results, scales=scales, pch = "|")
```

Step 4: Create Comparison Plots

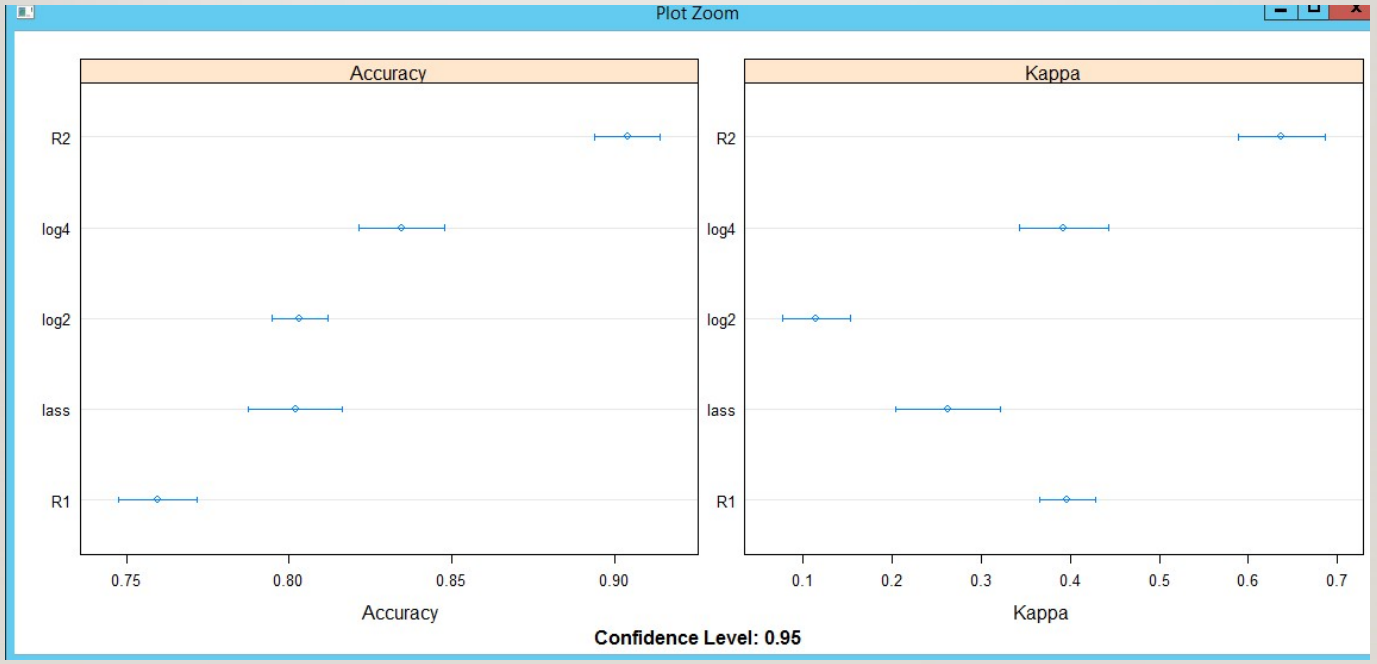
MODEL COMPARISON: CROSS VALIDATION

```
Call:
summary.resamples(object = results)

Models: R1, R2, log4, log2, lass
Number of resamples: 30

Accuracy
          Min.  1st Qu.   Median      Mean   3rd Qu.      Max. NA's
R1   0.6969697 0.742268 0.755102 0.7594916 0.7815064 0.8350515    0
R2   0.8400000 0.900000 0.900000 0.9040000 0.9200000 0.9400000    0
log4 0.7800000 0.820000 0.840000 0.8346667 0.8550000 0.9000000    0
log2 0.7600000 0.800000 0.800000 0.8033333 0.8200000 0.8400000    0
lass 0.7000000 0.780000 0.800000 0.8020000 0.8350000 0.8600000    0

Kappa
             Min.    1st Qu.    Median      Mean   3rd Qu.      Max. NA's
R1    0.18630137 0.34256511 0.4014436 0.3968336 0.4510882 0.5896351    0
R2    0.28571429 0.61538462 0.6478873 0.6375138 0.7058824 0.7887324    0
log4  0.15384615 0.29536680 0.3661972 0.3926706 0.4610187 0.6478873    0
log2 -0.07142857 0.03225806 0.1290431 0.1146242 0.1509434 0.2857143    0
lass -0.15384615 0.19027899 0.2857143 0.2630720 0.3727993 0.4943820    0

>
```

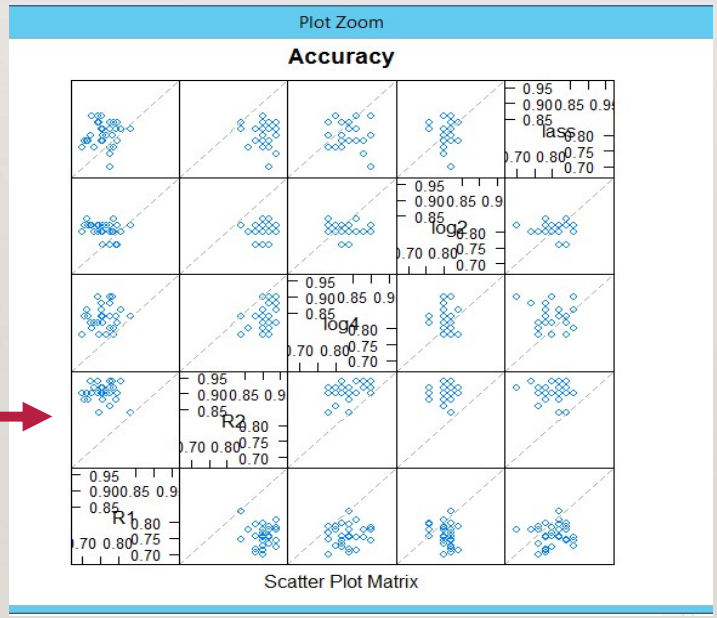
MODEL COMPARISON: BOX PLOTS



MODEL COMPARISON: DOT PLOT



MODEL COMPARISON: SCATTER PLOT MATRIX



MODEL COMPARISON: P VALUE COMPARISON

```
Call:
summary.diff.resamples(object = diffs)

p-value adjustment: bonferroni
Upper diagonal: estimates of the difference
Lower diagonal: p-value for H0: difference = 0

Accuracy
     R1          R2          log4       log2       lass
R1               -0.144508 -0.075175 -0.043842 -0.042508
R2   < 2.2e-16              0.069333  0.100667  0.102000
log4 4.092e-09 3.736e-10              0.031333  0.032667
log2 9.450e-05 3.832e-14 0.0067462              0.001333
lass 0.0004052 3.715e-11 0.0345836 1.0000000

Kappa
     R1          R2          log4       log2       lass
R1               -0.240680  0.004163  0.282209  0.133762
R2   9.478e-08              0.244843  0.522890  0.374442
log4 1.000000  1.272e-08              0.278046  0.129599
log2 2.018e-10 1.628e-15 4.068e-08              -0.148448
lass 0.002026  2.686e-09 0.028264   0.001341
```

Random Forest #2 has the lowest P values of all the models

CONCLUSIONS AND IMPLICATIONS

CONCLUSIONS AND IMPLICATIONS

Best Models (Top 3)

- Random Forest #2 with balanced data
- Model_GLM4(log4) with balanced data and variable selection using Random Forest #2
- Lasso

Model Comparison

- Random Forest vs. Logistic Regression
- Others to consider?

Data Integrity

- Cleaning strategy improved modeling
- Did omitting variables skew the models?
- Big improvement from pilot!

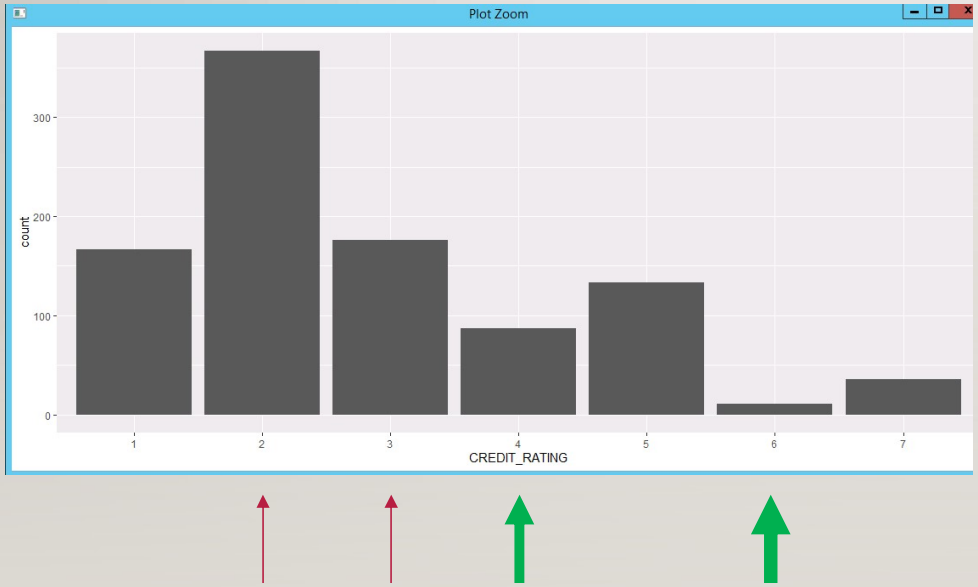
Balancing the data

- Target variable frequency had strong influence on predictive capacity of models
- Oversampling improved outcome

CONCLUSIONS AND IMPLICATIONS: VARIABLES OF IMPORTANCE

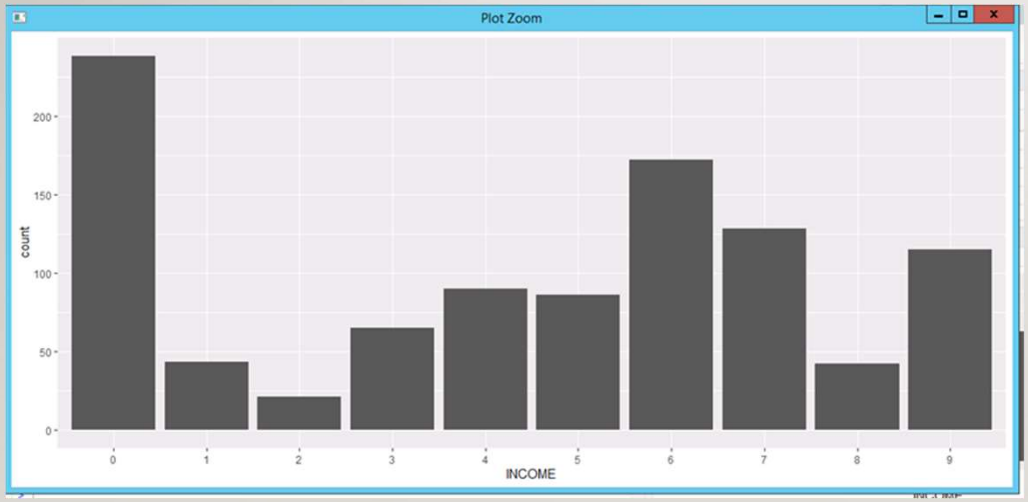
Variable	Meaning	Significance level
CALIBRAT1	Calibration sample	0 level
MONTHS	# of months in service	0 level
CUSTOMER	Customer ID #	.001 level
INCOME: 4,6,7,8,9	Income level	.01, .001 levels
CREDIT RATING: 2,3,4,6	Credit rating levels	.05, .01 levels
MOU	Mean monthly mins of use	.001 level
PEAKVCE	Mean # of in and out peak voice calls	.01 level
RETCALLS1	# calls previously made to retention team	.01 level

CREDIT RATING: A CLOSER LOOK



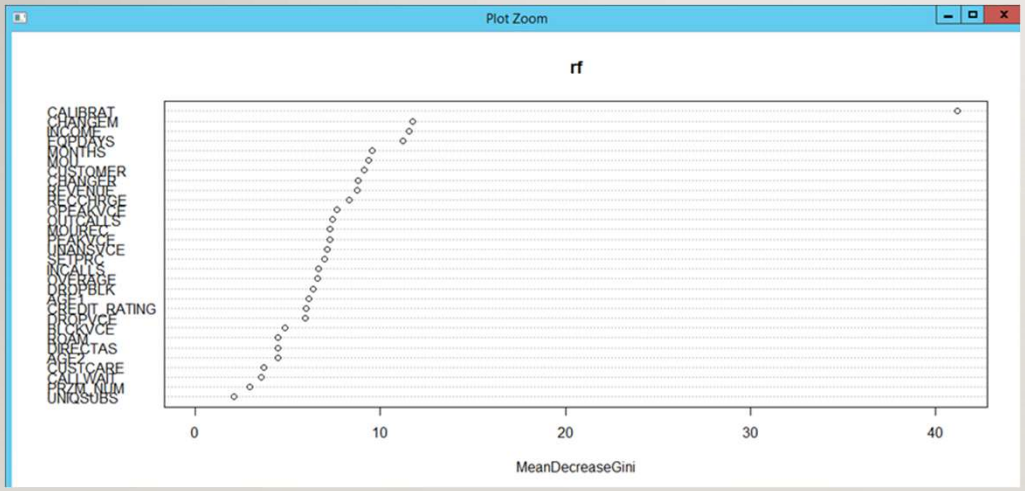
- 1-7 levels corresponds to each rating
- 2 = 'A' = Highest credit rating: .05 level
- 3 = 'B' = Good credit rating: .05 level
- 4 = 'C' = Medium credit rating: .01 level
- 6 = 'GY' = Very low credit rating: .001 level

INCOME LEVELS: DO YOU SEE A TREND?



- 0-7 levels, 0 means 'missing'
- Income level 4: 0 level
- Income level 6: .001 level
- Income level 7: .001 level
- Income level 8: .01 level
- Income level 9: .01 level

FUTURE VARIABLES TO STILL CONSIDER: PHONE USAGE



Random Forest #2 Variables of Importance

- Number of days with current equipment
- % change in minutes of use
- % change in revenues
- Mean # of in and out off peak voice calls
- Mean total recurring charge
- Mean # outbound voice calls
- Mean unrounded monthly minutes of use voice calls

CONCLUSIONS: ORGANIZATIONAL IMPACT AND FUTURE DIRECTIONS

- Create Customer Profiles

1. Income levels
2. Credit Ratings
3. Socio-economic data?

- Create Profiles of Phone data

1. Calls made to retention team
2. Change in use of the phone in any way
3. Make customer profiles of each type of user?

- We did not see any correlations with the WEBCAP variable

1. **How do we detect healthcare phone application usage?**
2. Future research in correlation

CUSTOMER PROFILES: DEEPER LOOK

```
# A tibble: 3 x 23
  .rownames CHURN CALIBRAT MONTHS CUSTOMER INCOME EQPDAYS CREDIT_RATING   MOU
  <chr>     <fct> <fct>      <dbl>    <dbl> <fct>    <dbl> <fct>         <dbl>
1 613       0     1             12  1062597 8          367 7               478
2 915       1     1             11  1066289 6          346 6               221.
3 996       1     1             41  1098468 6          179 5               406
```

- Is this the standard profile of each class of churner or are these just outliers?
- Creating customer profiles should be a future direction of our data analytics initiative

COST, BENEFIT ANALYSIS OF CONCLUSIONS

- Remember from the beginning of this presentation:

✓ 20-40% of customers will churn per year in telecommunications (Hashmi et al. 2013)

✓ Cost: 5-10 times more to add new customers than retain original customers

✓ **Focus: reduce churn rates by 5% and increase profits by 85% (Ullah et al. 2019)**

- Our results:

➢ We identified between 77-84% of churners based on our confusion matrices

➢ If we can now retain 5% of those we can increase our profits by 85%?!

➢ Income and Credit Rating are influential in retaining customers – focus to increase our profits?

➢ Phone usage is influential in retaining customers – but how many are using the mobile phone apps?

➢ Return on investment: create customer profiles!

RECOMMENDATIONS

RECOMMENDATIONS

- Data Science and Technology

✓ Python vs. RStudio

✓ Machine Learning and Deep Learning

✓ Different modeling techniques

✓ Tableau – maintain user interactive dashboards

- Project applications and future use

✓ **Web deployment of machine learning model for end business users at GE**

✓ **Future use:**

  ✓ **Customer Profiles,**

  ✓ **Phone and Healthcare app usage**

  ✓ **GE IoT sensor technology (Winig, 2016)**

RECOMMENDATIONS

- Future alterations to be made

✓ More raw data: mobile phones, customers, apps, sensors

✓ Retrain the models on new data (training data decays!)

✓ Data privacy: cloud vs. data lake vs. enterprise database

✓ Team focused approach to data analytics: different teams for different projects?

✓ Data integrity: different management of missing values, multi-level factor variables?

✓ Feature Engineering: CREDIT RATING, INCOME levels need to be numerical not categorical for correlation analysis

REFERENCES

- Atmathew (2015) Evaluating Logistic Regression Models. Retrieved from: https://www.r-bloggers.com/evaluating-logistic-regression-models/
- Barbosa, AM (2020) Area under the precision-recall curve. Retrieved from: https://www.r-bloggers.com/area-under-the-precision-recall-curve/
- Brownlee, J. (2019) Compare The Performance of Machine Learning Algorithms in R. Retrieved from: https://machinelearningmastery.com/compare-the-performance-of-machine-learning-algorithms-in-r/
- Chawla et al. (2002) SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research. Vol. 16:321-357.
- Cheng et al. (2017) Enterprise data breach: causes, challenges, prevention, and future directions. WIREs Data Mining Knowl Discov 2017, 7:e1211. doi: 10.1002/widm.1211
- Couronne et al. (2018) Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics. Vol. 19(270):1-14.
- Davis et al. (2006) The Relationship Between Precision-Recall and ROC Curves. ICML '06: Proceedings of the 23rd international conference on Machine Learning. Pgs: 233-240.
- Delgado et al. (2019) Why Cohen's Kappa should be avoided as performance measure in classification. PLOS ONE. Vol. 14(9): e0222916. https://doi.org/10.1371/journal.pone.0222916
- Desboulets, D. (2018) A Review on Variable Selection in Regression Analysis. Econometrics. Vol. 6(45):1-27.

REFERENCES

- Forcepoint (2020) Dynamic Data Protection. Retrieved from: https://www.forcepoint.com/solutions/need/dynamic-data-protection
- GE (2016) Big Data, Analytics & Artificial Intelligence. Healthcare Whitepaper.
- GE Healthcare (2020) Vscan Extend Auto Optimize App. Retrieved from: https://apps.gehealthcare.com/app-products/vscan-auto-optimize
- Guru99 (2020) GLM in R: Generalized Linear Model with Example. Retrieved from: https://www.guru99.com/r-generalized-linear-model.html
- Hagerman, I. (2017) Residuals Plots Part 1 – Residuals vs. Fitted Plot. Retrieved from: https://medium.com/data-distilled/residual-plots-part-1-residuals-vs-fitted-plot-f069849616b1
- Hagerman, I. (2017) Residuals Plots Part 2 – Normal QQ Plots. Retrieved from: https://medium.com/data-distilled/residual-plots-part-2-normal-qq-plots-c220ee9ed9fc
- Hagerman, I. (2017) Residuals Plots Part 3 – Scale-Location Plot. Retrieved from: https://medium.com/data-distilled/residual-plots-part-3-scale-location-plot-113e469b99c
- Hagerman, I. (2017) Residuals Plots Part 4 – Residuals vs. Leverage Plot. Retrieved from: https://medium.com/data-distilled/residual-plots-part-4-residuals-vs-leverage-plot-14aeed009ef7

REFERENCES

- Hashmi et al. (2013). Customer Churn Prediction in Telecommunication: A Decade Review and Classification. International Journal of Computer Science Issues. 10(5): p.271–282.
- Hurtgen et al. (2018) Achieving business impact with data. Retrieved from: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/achieving-business-impact-with-data
- Horvath, B. (2019) Generalized Linear Models: Residuals and Diagnostics. Retrieved from: https://rpubs.com/benhorvath/glm_diagnostics
- IBM (2020) CRISP-DM Help Overview. Retrieved from: https://www.ibm.com/support/knowledgecenter/SS3RA7_sub/modeler_crispdm_ddita/clementine/crisp_help/crisp_overview.html

REFERENCES

- Kao et al. (2008) Analysis of Variance: Is There a Difference in Means and What Does It Mean? J Surg Res. Vol. 144(1):158-170.
- Kang, H. (2013) The prevention and handling of the missing data. Korean J Anesthesiol. Vol. 64(5): 402-406.
- Krstajic et al. (2014) Cross-validation pitfalls when selecting and assessing regression and classification models. Journal of Chemoinformatics. Vol. 6(10):1-15.
- Lara, J. (2018) Agile Scrum Process in a Nutshell
- Mandak et al. (2019). Use of Logistic Regression for Understanding and Prediction of Customer Churn in Telecommunications. STATISTIKA. Vol. 99(2): 129-141.
- McDaniel, S. (2019) Data Privacy: Safeguarding Trusted Data. Retrieved from: https://www.talend.com/resources/data-privacy/
- Miliard, M. (2018) GE launches new Edison platform with AI apps. Retrieved from: https://www.healthcareitnews.com/news/ge-launches-new-edison-platform-ai-apps
- Narkhede, S. (2018) Understanding Confusion Matrix Retrieved from: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

REFERENCES

- Sliger, M. (2011). Agile project management with Scrum. Paper presented at PMI® Global Congress 2011—North America, Dallas, TX. Newtown Square, PA: Project Management Institute.
- Stoltzfus, J. (2011) Logistic Regression: A Brief Primer. ACADEMIC EMERGENCY MEDICINE. Vol. 18(10): 1099-1104.
- Ullah et al. (2019) A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector. Vol. 7:134-149.
- Winig, L. (2016) GE'S BIG BET ON DATA AND ANALYTICS. Retrieved from: https://sloanreview.mit.edu/case-study/ge-big-bet-on-data-and-analytics/
- Yanagihara et al. (2012) Bias-corrected AIC for selecting variables in multinomial logistic regression models. Linear Algebra and its Applications. Vol. 436: 4329-4341.
- Yiu, T. (2020) Understanding Cross Validation. Retrieved from: https://towardsdatascience.com/understanding-cross-validation-419dbd47e9bd
- Zhang et al. (2016) Residuals and regression diagnostics: focusing on logistic regression. Ann Transl Med. Vol. 4(10):1-8.