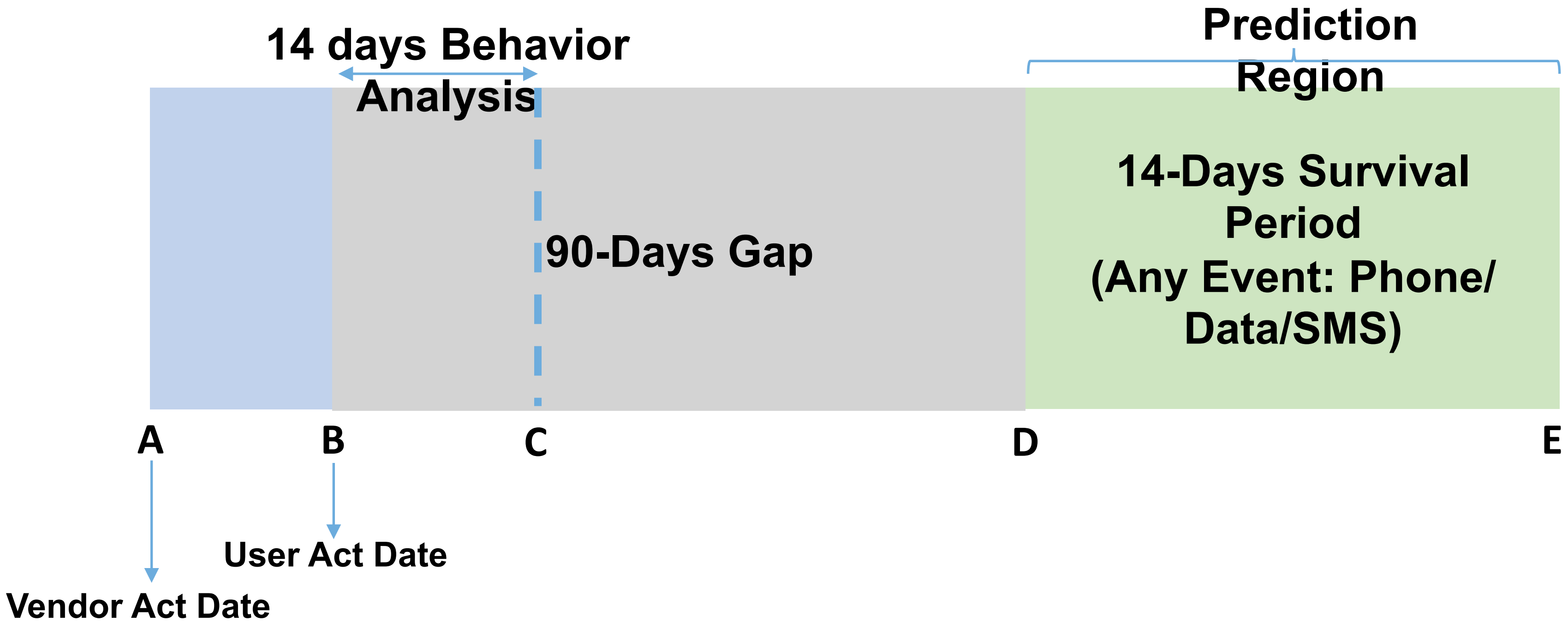
Predicting Survival Status of Pre-Paid SIM Card Users in Telecommunication Industry

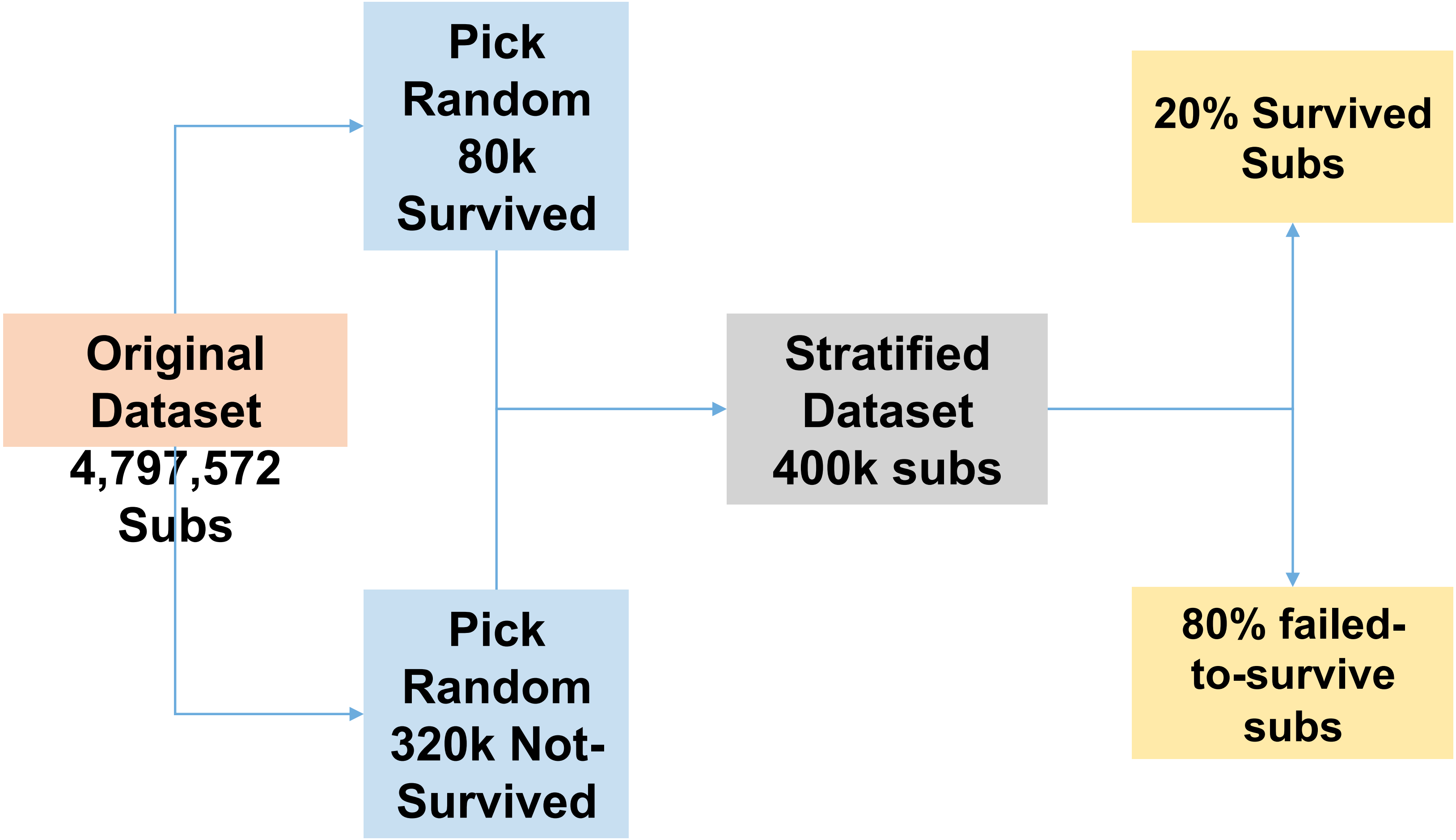
By: Vincent Adhi Handara

Introduction

Objectives: To predict survival status for all PREPAID subscribers **after** 90 days gap by analyzing behavior usage during **inital** 14 days after **user real activation date**



Stratified Sampling Method



Original Dataset 4,797,572 Subs

Pick Random 80k Survived

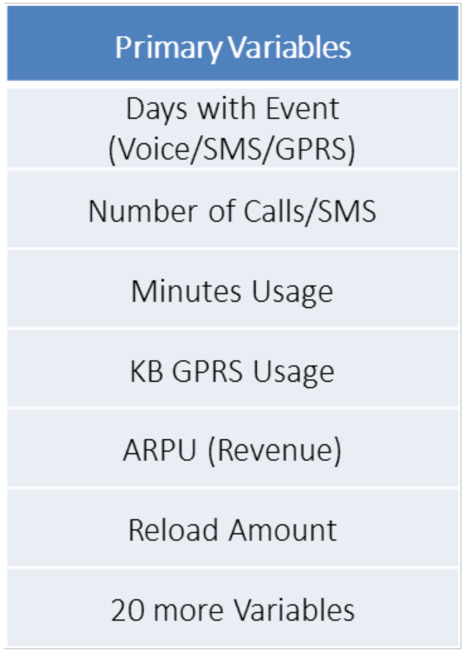
Pick Random 320k Not-Survived

Stratified Dataset 400k subs

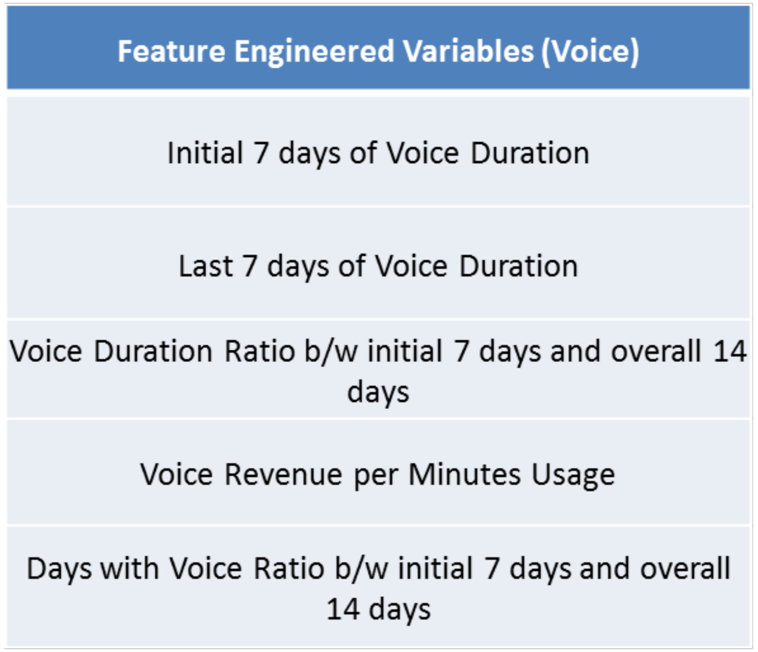
20% Survived Subs

80% failed-to-survive subs

Feature Engineering

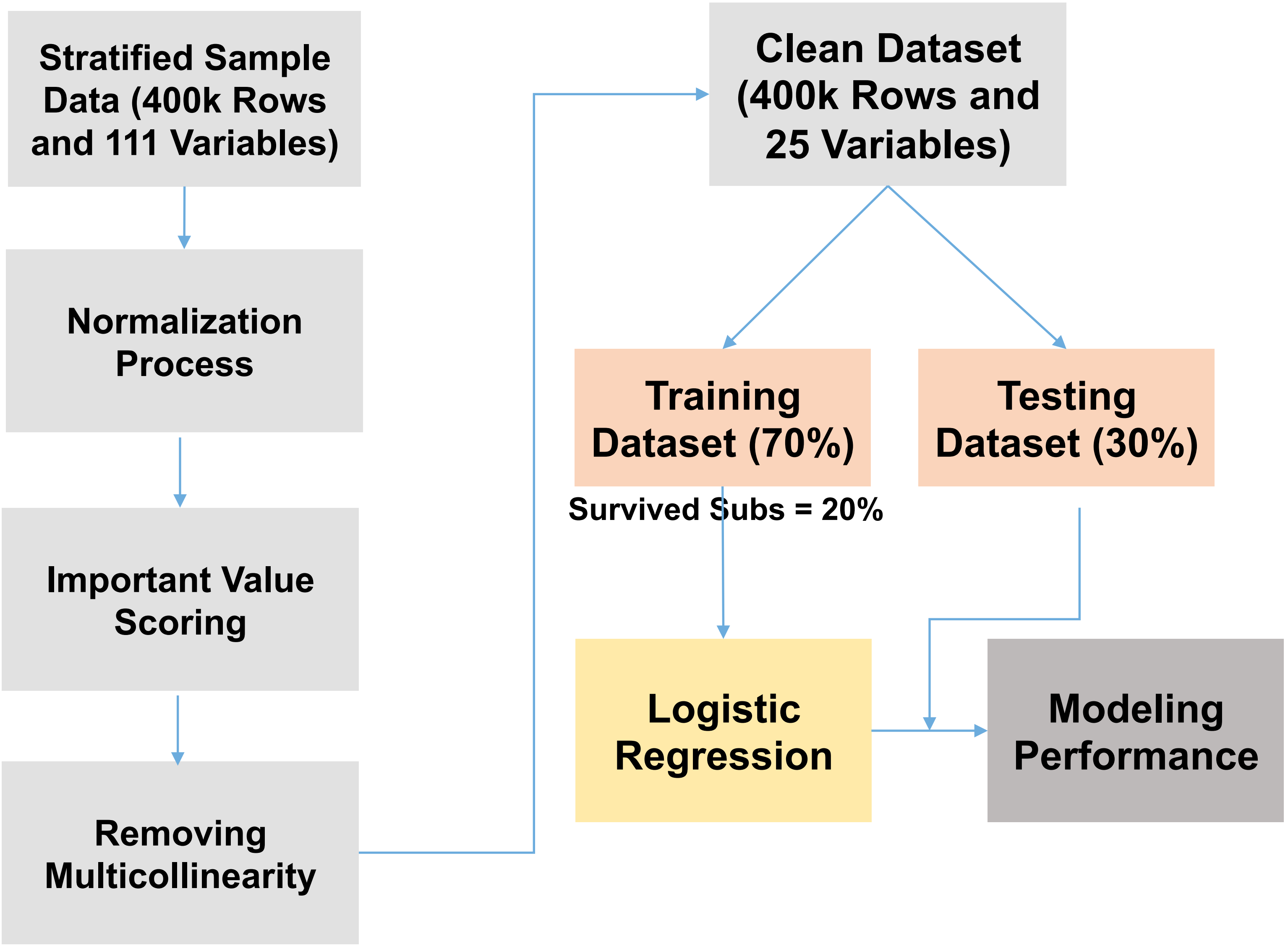


Primary Variables
Days with Event (Voice/SMS/GPRS)
Number of Calls/SMS
Minutes Usage
KB GPRS Usage
ARPU (Revenue)
Reload Amount
20 more Variables



Feature Engineered Variables (Voice)
Initial 7 days of Voice Duration
Last 7 days of Voice Duration
Voice Duration Ratio b/w initial 7 days and overall 14 days
Voice Revenue per Minutes Usage
Days with Voice Ratio b/w initial 7 days and overall 14 days

Predictive Modeling Flowchart



Stratified Sample Data (400k Rows and 111 Variables)

Normalization Process

Important Value Scoring

Removing Multicollinearity

Clean Dataset (400k Rows and 25 Variables)

Training Dataset (70%)

Survived Subs = 20%

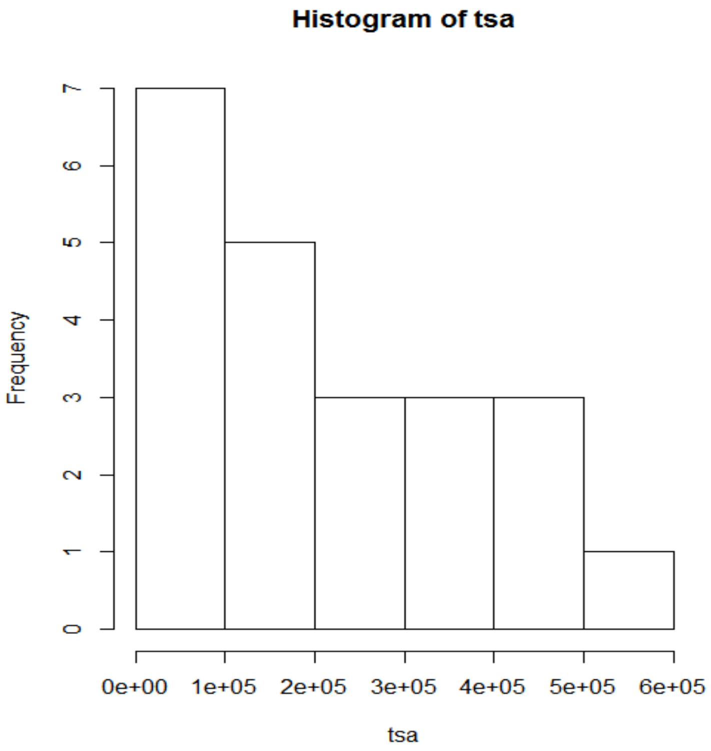
Testing Dataset (30%)

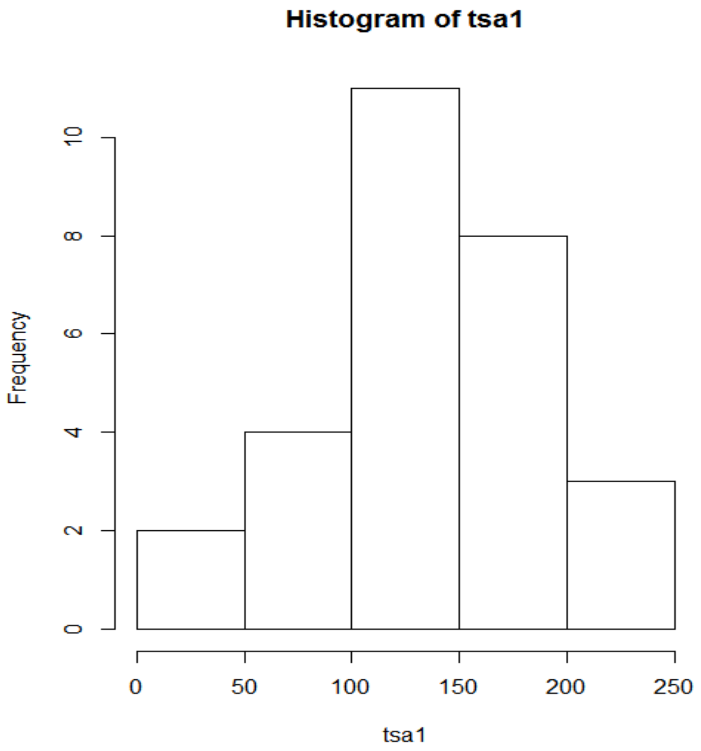
Logistic Regression

Modeling Performance

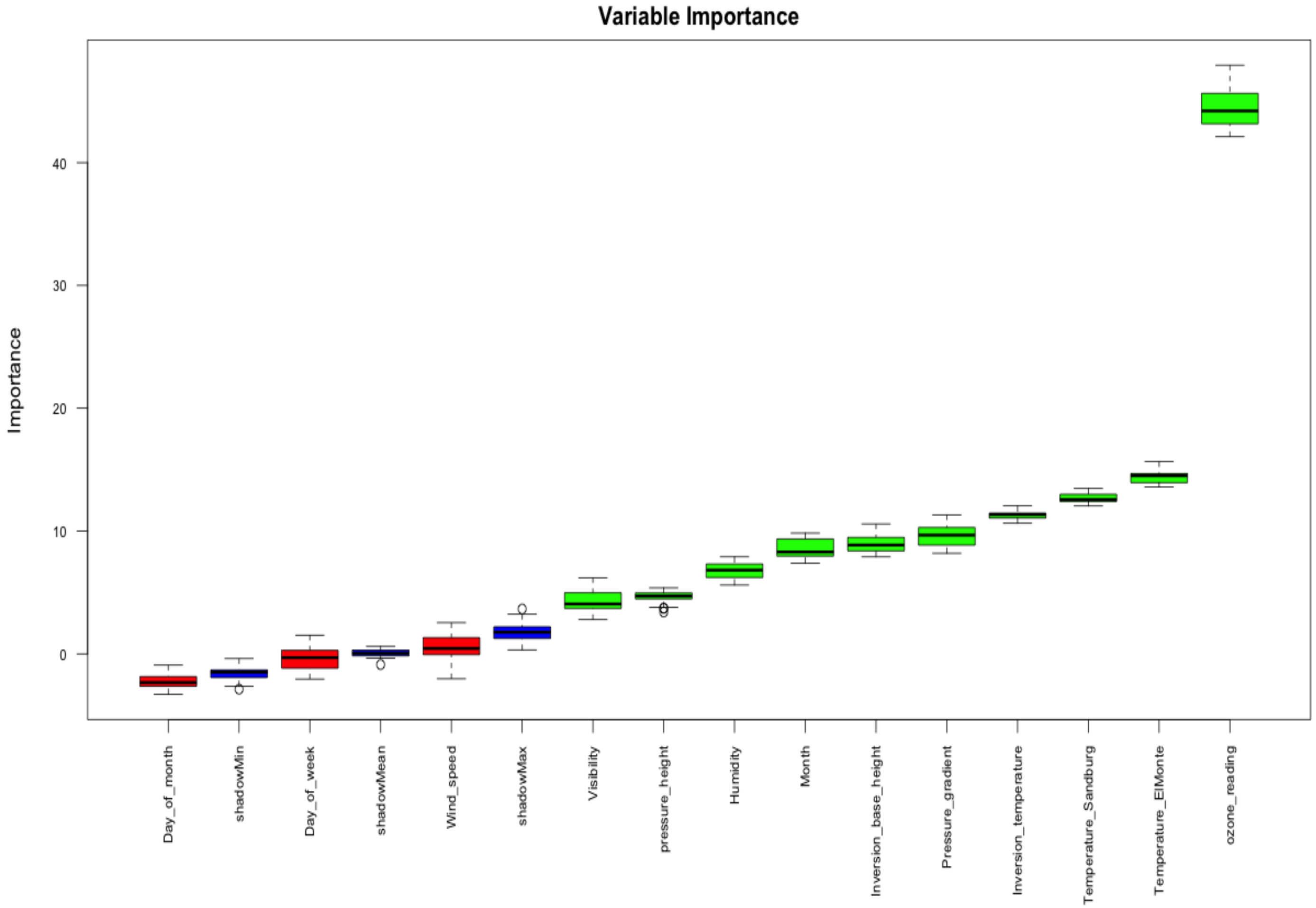
Normalization through BoxCox Transformation

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln(y_i) & \text{if } \lambda = 0, \end{cases}$$


Histogram of tsa

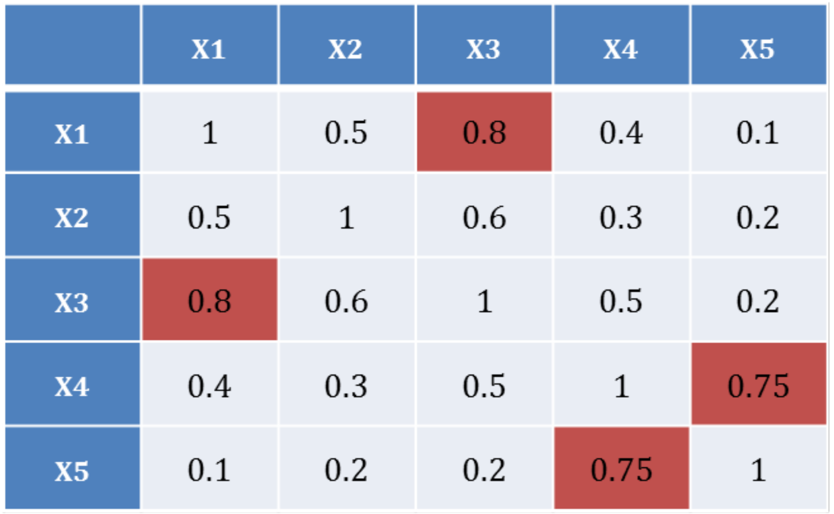

Histogram of tsa1

Finding Important Variables (Boruta Package in R)



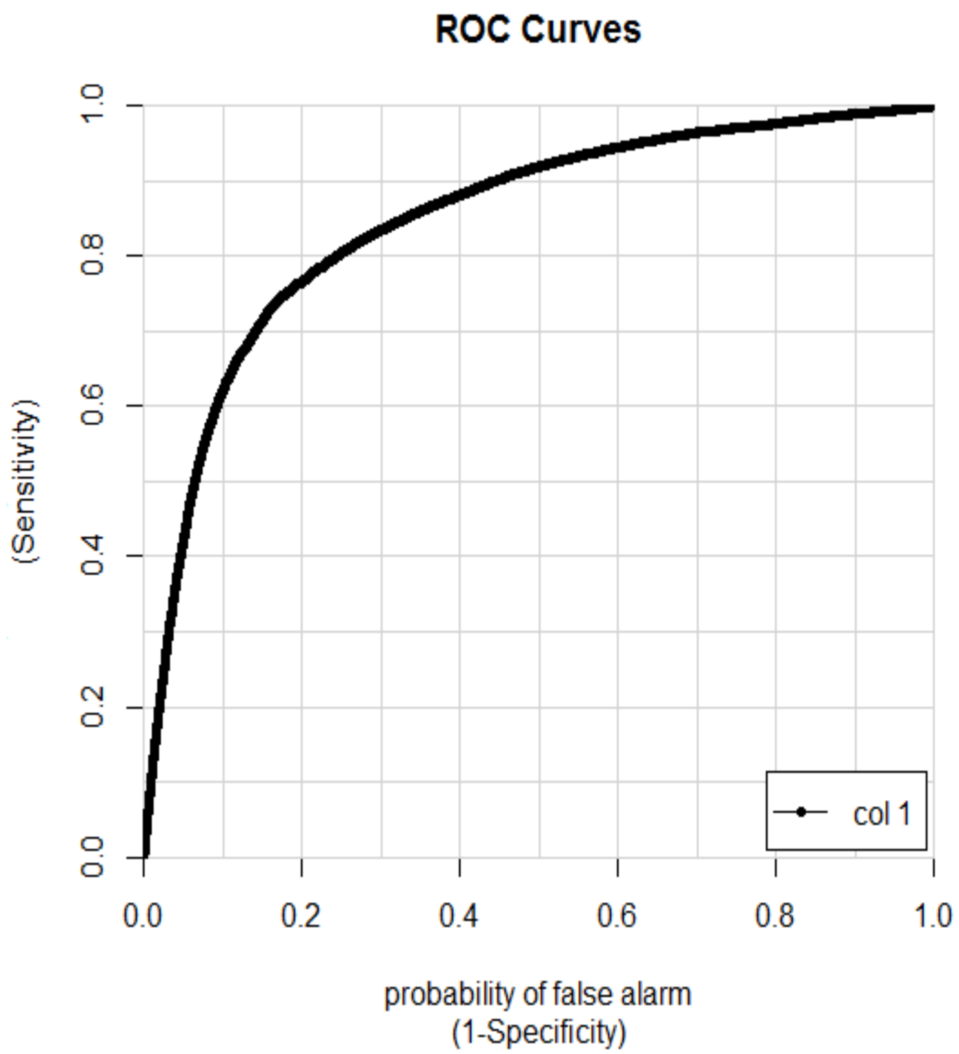
Removing Multicollinearity (Correlation > 0.7)

Descendingly Sorted by
Important Value Score



	X1	X2	X3	X4	X5
X1	1	0.5	0.8	0.4	0.1
X2	0.5	1	0.6	0.3	0.2
X3	0.8	0.6	1	0.5	0.2
X4	0.4	0.3	0.5	1	0.75
X5	0.1	0.2	0.2	0.75	1

Logistic Regression Performance (ROC Curves and Confusion Matrix) on Testing Dataset



		No	Yes
Predicted Survival Status	No	90449	12832
	Yes	5523	11196

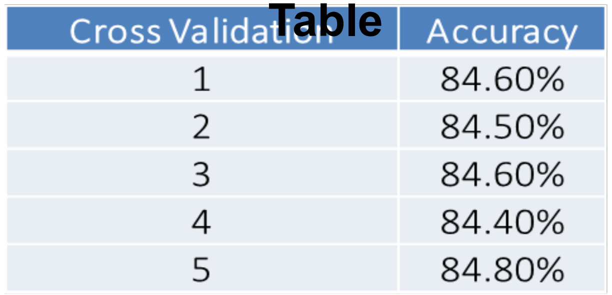
Accuracy	84.7%

Testing Subs = 120000

5 Fold Cross Validation

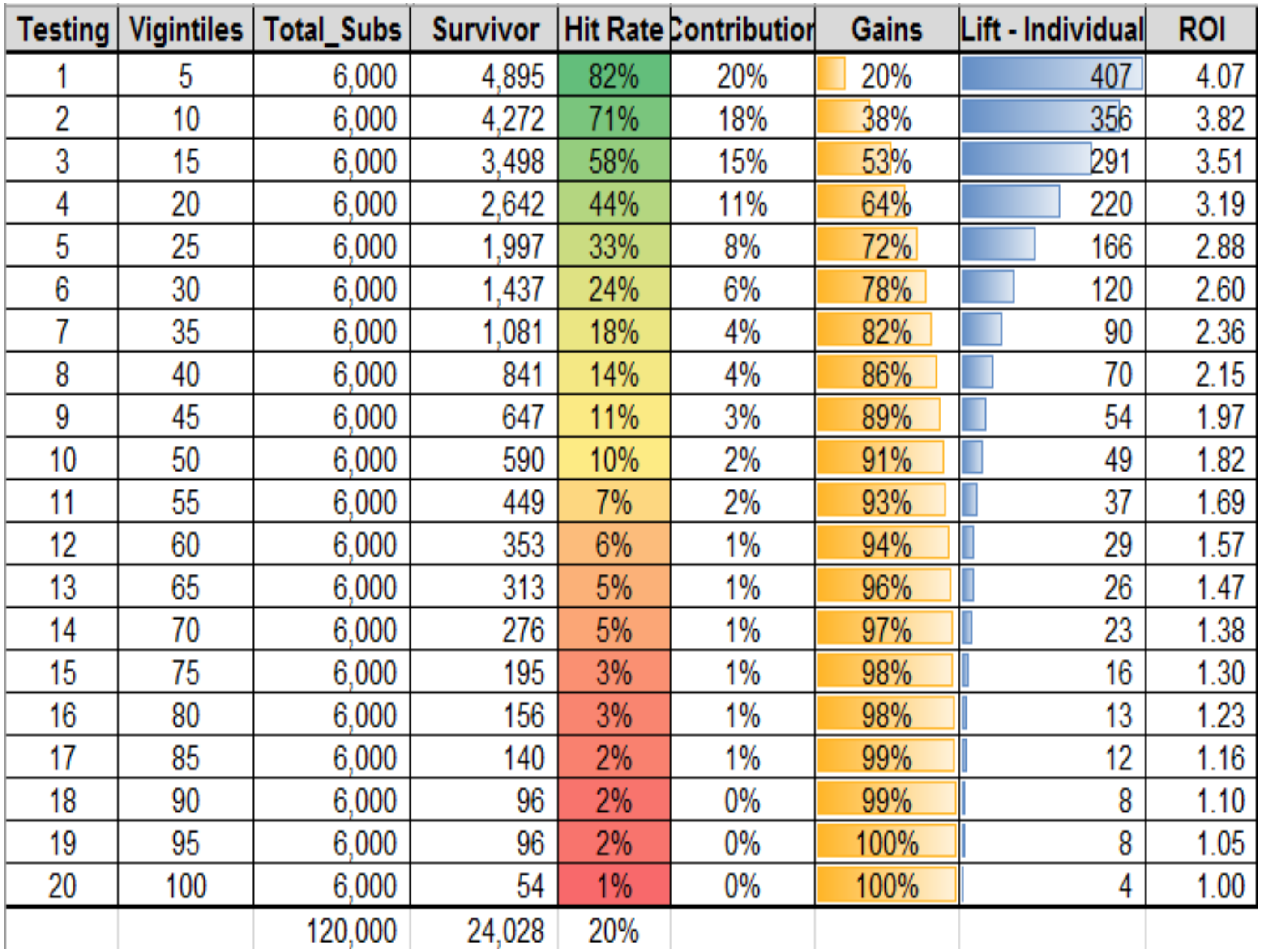
Iteration 1	Test	Train	Train	Train	Train
Iteration 2	Train	Test	Train	Train	Train
Iteration 3	Train	Train	Test	Train	Train
Iteration 4	Train	Train	Train	Test	Train
Iteration 5	Train	Train	Train	Train	Test

5 Fold Cross Validation Table



Cross Validation	Accuracy
1	84.60%
2	84.50%
3	84.60%
4	84.40%
5	84.80%

Gain Table for Testing Dataset - Logistic Regression



Testing	Vigintiles	Total_Subs	Survivor	Hit Rate	Contribution	Gains	Lift - Individual	ROI
1	5	6,000	4,895	82%	20%	20%	407	4.07
2	10	6,000	4,272	71%	18%	38%	356	3.82
3	15	6,000	3,498	58%	15%	53%	291	3.51
4	20	6,000	2,642	44%	11%	64%	220	3.19
5	25	6,000	1,997	33%	8%	72%	166	2.88
6	30	6,000	1,437	24%	6%	78%	120	2.60
7	35	6,000	1,081	18%	4%	82%	90	2.36
8	40	6,000	841	14%	4%	86%	70	2.15
9	45	6,000	647	11%	3%	89%	54	1.97
10	50	6,000	590	10%	2%	91%	49	1.82
11	55	6,000	449	7%	2%	93%	37	1.69
12	60	6,000	353	6%	1%	94%	29	1.57
13	65	6,000	313	5%	1%	96%	26	1.47
14	70	6,000	276	5%	1%	97%	23	1.38
15	75	6,000	195	3%	1%	98%	16	1.30
16	80	6,000	156	3%	1%	98%	13	1.23
17	85	6,000	140	2%	1%	99%	12	1.16
18	90	6,000	96	2%	0%	99%	8	1.10
19	95	6,000	96	2%	0%	100%	8	1.05
20	100	6,000	54	1%	0%	100%	4	1.00
		120,000	24,028	20%				

Testing Subs = 120000