ECON-561: Final Project

Predictive Modeling for Non-Profit organizations

Shaun Levenson | Francisco Mendoza | Andrew Frohner

John Garcia

February 18th, 2025

California Lutheran University

Master of Science in Quantitative Economics

Business Understanding

1. Your organization does not take a profit

- Business Success: depends on direct mail campaigns to solicit donations, yet the current response rate is only 10%, leading to inefficiencies and financial losses.
- Cost Structure & Risk: Outreach is $2 per recipient -- while the average donation from each responder is $14.50. Without precise targeting, many mailings result in a net loss of $0.55 per recipient.
- Strategic Goal: Implement predictive modeling to optimize direct marketing and improve donor targeting, lowering costs and increasing net donation revenue.

2. Marketing Campaign current status

- Current Response Rate: 10% of mail recipients donate, signifying that 90% of mailings do not generate revenue.
- Revenue vs. Cost Imbalance: Due to high marketing costs, sending mail to all potential donors is not cost-effective.
- Data-Driven Opportunity: By analyzing past donor behavior, we can identify high-probability donors and focus marketing efforts on them, enhancing ROI.

California Lutheran UNIVERSITY

Data Understanding

8,009 Observations in Total

49.75%
Training

Data Understanding – what’s in the box?

8,009 Observations in Total

49.75%
Training

25.19%
Validation

Data Understanding – what's in the box?

8,009 Observations in Total

49.75% Training

25.19% Validation

25.06% Test

Data Understanding – what’s in the box?

6,002 Obs used for Modeling

49.75%
Training

25.19%
Validation

25.06%
Test

Attribute Categories

- Demographic
- Financial
- Donation history

Data Understanding – Target Variables

DONR

Binary Outcome

- Nominal
- 1 IF person donates
- 0 IF person does not donate

Purposed for *Classification*

California Lutheran UNIVERSITY

Data Understanding – Target Variables

DONR

Binary Outcome

- Nominal
- 1 IF person donates
- 0 IF person does not donate

Purposed for *Classification*

DAMT

Donation Amount

- Continuous
- Dollar Amount

Purposed for *Prediction*

Data Understanding – General Notes

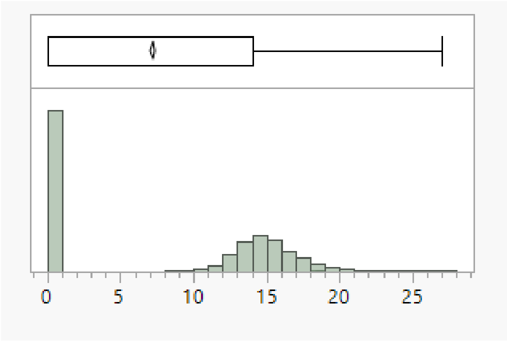
Our data lends insight into *historical engagement* and *prior response behavior*

Data Understanding – General Notes

Our data lends insight into *historical engagement* and *prior response behavior*

1. DAMT could be influenced by outliers

Summary Statistics	
Mean	7.21
Std Dev	7.36
Std Err Mean	0.10
Upper 95% Mean	7.40
Lower 95% Mean	7.02
N	6002.00
Skewness	0.12
N Missing	0.00
Median	0.00

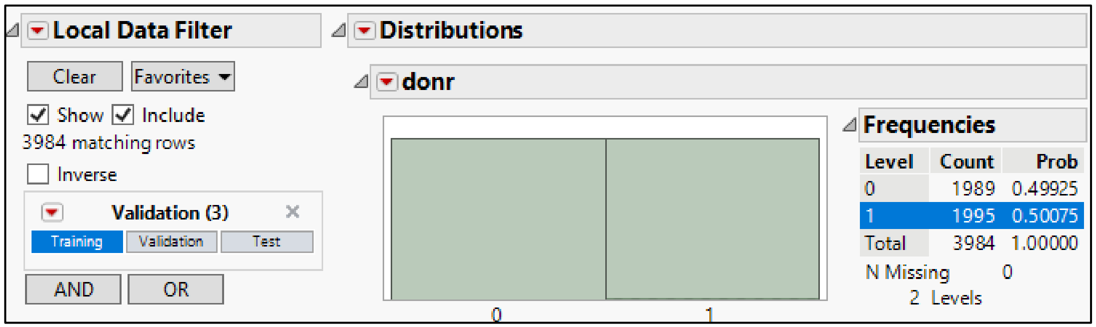


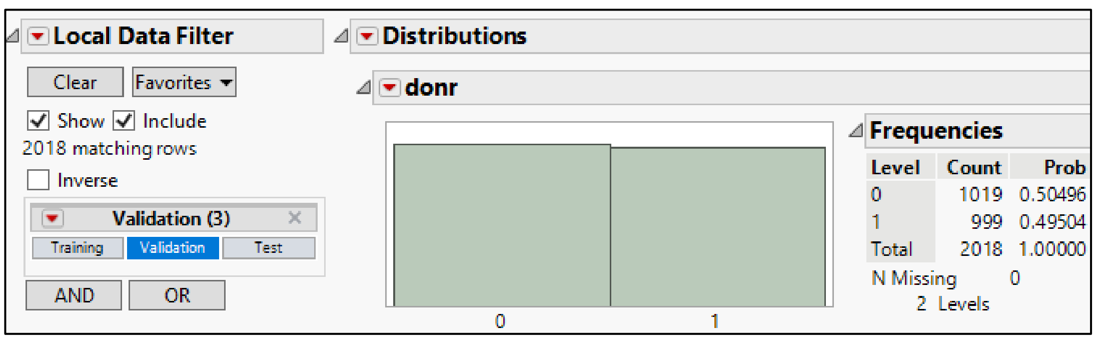
- Median < Mean
- Right skewed Distribution

Data Understanding – General Notes

Our data provides favorable test environment for a classification problem

2. DONR class in balanced in Training & Validation data





- Wei & Dunbrack (2013) found that balanced Training/Validation results in higher TPR, TNR, and AUC metrics.

Data Understanding

1. Data set allows us to answer…

- **Predict Donor Likelihood:** We can estimate which individuals are most likely to donate by analyzing past donation behavior and demographic factors.
- **Estimate Donation Amounts:** Using regression modeling, we can predict the expected donation amount from identified donors.
- **Segment Donor Profiles:** The dataset allows us to classify donors based on income, past giving history, and demographic features, enhancing targeted marketing strategies.
- **Optimize Marketing Efforts:** The organization can reduce marketing costs and maximize net revenue by identifying high-probability donors.

1. Data set has limits….

- **Imbalanced Test Data:** While the training and validation sets are balanced, the test set reflects real-world donor response rates (~10% donors, 90% non-donors), necessitating adjustments in classification modeling.
- **Limited Feature Scope:** The dataset includes demographic and donation history but lacks behavioral data such as engagement with past campaigns, online interactions, or social media activity.
- **Right-Skewed Donation Amounts:** Most donations are small, with a few high-value outliers, necessitating careful handling using transformations or robust regression techniques.
- **Potential Data Bias:** Features like income and wealth indicators may introduce biases, which could unintentionally exclude potential donors.

Modeling Objectives

Donor Classification

1. Identify predictive power: Use features that hold predictive power for Donor & Non-Donor class
2. Discriminate Ethically: avoid prediction using features that could put us at risk biasing in/out a certain demographic or population segment
3. Develop sustainable marketing Campaign: Share analytics that will generate a positive payoff under the current cost structure.

Donation Amount prediction

1. Maximize predictive capability

Modeling Process – Feature selection using Theory

1. Asked the question: Does this predictor hold a relationship with response feature?

- Aided in dimension reduction
- Kept circular logic or confusing logic out of the model
- Provided foundation for an initial prediction/classification

2. If yes – what direction do you <u>expect</u> its impact to be?

- Propels feature engineering
- Helps uncover non-linearities/interactions
- Provides the analyst with a sense *who* is donating and *how much* they generally donate

Modeling Process DONR– Feature engineering using filter

Tail	0.2
Quintile	3

Column	10% Quantile	90% Quantile	Low Threshold	High Threshold	Number of Outliers
ter1	0	1	-3	4	0
ter2	0	1	-3	4	0
ter3	0	1	-3	4	0
ter4	0	1	-3	4	0
ownd	0	1	-3	4	0
kids	0	4	-12	16	0
inc	2	6	-10	18	0
sex	0	1	-3	4	0
wlth	3	9	-15	27	0
hv	107	274	-394	775	0
incmed	20	74	-142	236	3
incavg	32	86	-130	248	4
low	1	33	-95	129	0
npro	21	102	-222	345	0
gifdol	49	206	-422	677	12
gifl	6	42	-102	150	76
gifr	5	31	-73	109	7
mdon	13	24	-20	57	0
lag	3	11	-21	35	0
gifa	5.07	20.01	-39.75	64.83	1

- Initial data contained Outliers

Modeling Process DONR–Standardization

Variable	skewness	transformation Performed	# of outliers 3x above 80% quintile range	Variable Created
~~wlth~~	~~-1.35~~	~~Col Standardize~~	~~0~~	~~std_wlth~~
hv	1.539	Col Standardize	0	std_hv
incmed	2.05	Col Standardize	3	std_incmed
incavg	1.9377	Col Standardize	4	std_incavg
low	1.36		0	
gifdol	6.54	Col Standardize	12	std_gifdol
gifl	7.81	Col Standardize	76	std_gifl
gifr	2.62	Col Standardize	7	std_gifr
mdon	1.1	Col Standardize	0	std_mdon
lag	2.42	Col Standardize	0	std_lag
gifa	1.78	Col Standardize	1	std_gifa

- Ordinal features were skewed but not standardized

- Variables had moderate skewness

Modeling Process DONR– Feature engineering

Wlth	Count N	1	proportion	Weight Assigned
0	152	1	2.532%	39.48684211
1	130	1	2.166%	46.16923077
2	149	1	2.483%	40.28187919
3	215	1	3.582%	27.91627907
4	319	1	5.315%	18.81504702
5	303	1	5.048%	19.80858086
6	415	1	6.914%	14.4626506
7	379	1	6.315%	15.83641161
8	2314	1	38.554%	2.59377701
9	1626	1	27.091%	3.691266913
Total	**6002**		**100%**	

- Sample was skewed towards the Wealthier groups

Weighting scheme:

$$\frac{n(\,wlth\ (1))}{N} = proportion$$

$$\frac{1}{proportion} = Weight$$

$$\boxed{weight\ * Wlth = Adj\_wlth}$$

Modeling Process DONR– Feature engineering

	Corresponds to
Wlth (ordinal ranking	**Adj_Wlth**
0	0.000
1	46.169
2	80.537
3	83.749
4	75.260
5	99.043
6	86.776
7	112.337
8	20.750
9	33.221

Before

- Wlth purely measure of Inc relative to average income in the state

New Ordinal Ranking	
Wlth (former)	**Adj_Wlth (new)**
0	0.000
8	20.750
9	33.221
1	46.169
4	75.260
2	80.564
3	83.749
6	86.776
5	99.043
7	112.337

After

- Wlth measured relative to Income AND sample representation

Modeling Process DONR– Feature selection

Elastic Net

	Estimate	Std Error	Siginficant	Odds Increase	% Probability they will donate
Intercept	-3.84459	1.055613		0.021395	2%
ter1[0-1]	-1.72161	0.181053	*	0.178778	15%
ter2[0-1]	-3.29141	0.182388	*	0.037201	4%
ter3[0-1]	0.094151	0.217178		1.098726	52%
ter4[0-1]	0.050378	0.20797		1.051669	51%
ownd[0-1]	-4.62713	0.239486	*	0.009783	1%
kids	-3.06992	0.133699	*	0.046425	4%
inc[2-1]	1.759407	0.349403	*	5.808993	85%
inc[3-2]	1.643453	0.246723	*	5.172999	84%
inc[4-3]	1.567659	0.198592	*	4.795407	83%
inc[5-4]	-1.84156	0.172634	*	0.158570	14%
inc[6-5]	-1.54342	0.272159	*	0.213650	18%
inc[7-6]	-1.47603	0.415197		0.228544	19%
Adj_Wlth[20.750216-0]	5.606004	0.691016	*	272.055040	100%
Adj_Wlth[33.221403-20.750216]	-0.15365	0.135227		0.857576	46%
Adj_Wlth[46.16923-33.221403]	-4.39029	0.466317	*	0.012397	1%
Adj_Wlth[75.2602-46.16923]	2.806979	0.524057	*	16.559820	94%
Adj_Wlth[80.56376-75.2602]	-1.04224	0.481665		0.352663	26%
Adj_Wlth[83.74884-80.56376]	-0.68187	0.530634		0.505669	34%
Adj_Wlth[86.7759-83.74884]	3.520804	0.42231	*	33.811585	97%
Adj_Wlth[99.0429-86.7759]	-1.91535	0.35898	*	0.147290	13%
Adj_Wlth[112.33691-99.0429]	1.702912	0.367136	*	5.489913	85%
Inc^2	-0.03148	0.030374		0.969008	49%
sex	-0.10392	0.115793		0.901294	47%
Kids_Squared	0.374043	0.030181	*	1.453600	59%
npro	-0.001	0.0114		0.998997	50%
nor_incmed	0.720405	0.076425	*	2.055265	67%
Root_npro	0.300079	0.168651		1.349965	57%

- Penalized Regression used as final test of predictor set

Modeling Process – Feature Selection Results

DONR Predictors

- 14 Total features
- 3 Non-Linear terms
- 3 Indicators
- 2 Ordinal rankings
- 6 Numerical measures

Ter1, Ter2, Ter3, Ter4	Nominal
Kids	Continuous
Inc	Ordinal
$Kids^2$	Continuous
Sex	Nominal
Adj_Wlth	Ordinal
Npro	Continuous
Inc_avg^2	Continuous
Inc_med	Continuous
Ownd	Nominal
$\sqrt{npro}$	Continuous

Modeling Process – Feature Selection Results

DONR Predictors

- 14 Total features
- 3 Non-Linear terms
- 3 Indicators
- 2 Ordinal rankings
- 6 Numerical measures

Notable Exclusions

- Gift Variables: gifdol | gifr | Mdon | Lag | gifa
- low
- Hv

Ter1, Ter2, Ter3, Ter4	Nominal
Kids	Continuous
Inc	Ordinal
$Kids^2$	Continuous
Sex	Nominal
Adj_Wlth	Ordinal
Npro	Continuous
Inc_avg^2	Continuous
Inc_med	Continuous
Ownd	Nominal
$\sqrt{npro}$	Continuous

Gift Variables assume DONR = 1

Modeling Process – Feature Selection Results

DONR Predictors

- 14 Total features
- 3 Non-Linear terms
- 3 Indicators
- 2 Ordinal rankings
- 6 Numerical measures

Ter1, Ter2, Ter3, Ter4	Nominal
Kids	Continuous
Inc	Ordinal
$Kids^2$	Continuous
Sex	Nominal
Adj_Wlth	Ordinal
Npro	Continuous
Inc_avg^2	Continuous
Inc_med	Continuous
Ownd	Nominal
$\sqrt{npro}$	Continuous

Notable Exclusions

- Gift Variables: gifdol | gifr | Mdon | Lag | gifa
- low
- Hv

Multivariate

Correlations

	hv	incmed	incavg	inc
hv	1.0000	0.7229	0.8400	0.0060
incmed	0.7229	1.0000	0.8685	0.0072
incavg	0.8400	0.8685	1.0000	0.0101
inc	0.0060	0.0072	0.0101	1.0000

HV was removed to reduce dimensionality

Modeling Process – Feature Selection Results

DONR Predictors

- 14 Total features
- 3 Non-Linear terms
- 3 Indicators
- 2 Ordinal rankings
- 6 Numerical measures

Risk

- Ownd – could be associated with socioeconomic factors that disproportionately favor wealthier classes

Ter1, Ter2, Ter3, Ter4	Nominal
Kids	Continuous
Inc	Ordinal
$Kids^2$	Continuous
Sex	Nominal
Adj_Wlth	Ordinal
Npro	Continuous
Inc_avg^2	Continuous
Inc_med	Continuous
$\boxed{\text{Ownd}}$	$\boxed{\text{Nominal}}$
$\sqrt{npro}$	Continuous

Kept in model – because it is effective for modeling class partitioning and just because you own a home does not mean you must be wealthy

Modeling Process DAMT– Feature engineering

- Squared terms were included to account for potential nonlinearity present in $Kids$ and $Wealth$
- If the variable had a skew of greater than 3, the term was logged x+1 to account for individuals with 0s

Variable	skewness	transformation Performed	Variable Created
Kids	0.39	^2	Kids^2
Wlth		^2	Wlth^2
gifdol	7	Log(gifdol +1)	log(gifdol)
gifl	8	Log(gifl +1)	log(gifl)
gifr	3	Log(gifr +1)	log(gifr)

Modeling Process DAMT– Feature engineering

Interactions – 8 total

Inc*Kids
- Higher income families with more children may donate less due to financial strain

Wlth*kids
- Wealthier households with more children may still donate at a high level

Inc*Sex
- If gender influences donation behavior at different income levels, this interaction will capture it

Gifa*Inc
- High income donors with high past gift amounts may donate even more

Gifr*Wlth
- High wealth donors with a high recent gift amount may be repeat high donors

Mdon*Lag
- If a donor took a long gap between their first and second donation, but donated recently, they may be reactivated donors

Mdon*Gifr
- Donors who recently gave a large amount may be high value repeat donors
- Donors who gave a large amount a long time ago may be at risk of discontinuing donations

Ter1-4*Wlth
- Regional Wealth Effects

Interaction terms support our model when the effect of 1 predictor varies by another

Modeling Process – Feature Selection Results

DAMT Predictors

- 25 Total features
- 1 Non-Linear terms
- 3 Indicators
- 2 Ordinal rankings
- 6 Numerical measures

- Ter1-4
- Ownd
- Kids
- Kids^2
- Inc with binary for specific income ranges
- Sex
- Wlth with binary for specific wealth ranges
- Hv
- Incmed
- Incavg
- Low
- Npro
- Log(gifdol)
- Log(gifr)
- Mdon
- Lag
- gifa

Interactions Terms

- Inc*kids
- Inc*sex
- Wlth*kids
- Gifa*inc
- Gifr*wlth
- Mdon*lag
- Mdon*gifr
- Ter1-4*wlth

Modeling – Donor Classification

Selected Models

1. Logistic Regression
2. Bootstrap Forest
3. Neural Network
4. K-Nearest Neighbors

Discussion

- Modeling Methodology
- Results
- Potential Profitability

DONR Classification – Logistic Regression

Term	Estimate	Std Error	Siginficant	Odds Increasing?	Change in Odds they will dontate
Adj_Wlth[20.750216-0]	5.714	0.704	*	303.186	100%
Adj_Wlth[86.7759-83.74884]	3.638	0.451	*	38.024	97%
Adj_Wlth[75.2602-46.16923]	3.127	0.764	*	22.801	96%
inc[2-1]	1.789	0.383	*	5.982	86%
Adj_Wlth[112.33691-99.0429]	1.743	0.362	*	5.714	85%
inc[3-2]	1.668	0.252	*	5.300	84%
inc[4-3]	1.592	0.197	*	4.912	83%
nor_incmed	0.735	0.075	*	2.086	68%
Kids_Squared	0.387	0.032	*	1.473	60%
ter1[0]	-0.875	0.094	*	0.417	29%
inc[6-5]	-1.564	0.290	*	0.209	17%
ter2[0]	-1.668	0.095	*	0.189	16%
inc[5-4]	-1.869	0.174	*	0.154	13%
Adj_Wlth[99.0429-86.7759]	-1.974	0.349	*	0.139	12%
ownd[0]	-2.351	0.148	*	0.095	9%
kids	-3.142	0.146	*	0.043	4%
Adj_Wlth[46.16923-33.221403]	-4.705	0.729	*	0.009	1%
Intercept	-8.901	1.039	*	0.000	0%

More impactful to the Odds that someone will Donate

DONR Classification – Logistic Regression

Term	Estimate	Std Error	Siginficant	Odds Increasing?	Change in Odds they will dontate
Adj_Wlth[20.750216-0]	5.714	0.704	*	303.186	100%
Adj_Wlth[86.7759-83.74884]	3.638	0.451	*	38.024	97%
Adj_Wlth[75.2602-46.16923]	3.127	0.764	*	22.801	96%
inc[2-1]	1.789	0.383	*	5.982	86%
Adj_Wlth[112.33691-99.0429]	1.743	0.362	*	5.714	85%
inc[3-2]	1.668	0.252	*	5.300	84%
inc[4-3]	1.592	0.197	*	4.912	83%
nor_incmed	0.735	0.075	*	2.086	68%
Kids_Squared	0.387	0.032	*	1.473	60%
ter1[0]	-0.875	0.094	*	0.417	29%
inc[6-5]	-1.564	0.290	*	0.209	17%
ter2[0]	-1.668	0.095	*	0.189	16%
inc[5-4]	-1.869	0.174	*	0.154	13%
Adj_Wlth[99.0429-86.7759]	-1.974	0.349	*	0.139	12%
ownd[0]	-2.351	0.148	*	0.095	9%
kids	-3.142	0.146	*	0.043	4%
Adj_Wlth[46.16923-33.221403]	-4.705	0.729	*	0.009	1%
Intercept	-8.901	1.039	*	0.000	0%

Less impactful to the Odds that someone will Donate

DONR Classification – Logistic Regression

Term	Estimate	Std Error	Siginficant	Odds Increasing?	Change in Odds they will dontate
Adj_Wlth[20.750216-0]	5.714	0.704	*	303.186	100%
Adj_Wlth[86.7759-83.74884]	3.638	0.451	*	38.024	97%
Adj_Wlth[75.2602-46.16923]	3.127	0.764	*	22.801	96%
inc[2-1]	1.789	0.383	*	5.982	86%
Adj_Wlth[112.33691-99.0429]	1.743	0.362	*	5.714	85%
inc[3-2]	1.668	0.252	*	5.300	84%
inc[4-3]	1.592	0.197	*	4.912	83%
nor_incmed	0.735	0.075	*	2.086	68%
Kids_Squared	0.387	0.032	*	1.473	60%
ter1[0]	-0.875	0.094	*	0.417	29%
inc[6-5]	-1.564	0.290	*	0.209	17%
ter2[0]	-1.668	0.095	*	0.189	16%
inc[5-4]	-1.869	0.174	*	0.154	13%
Adj_Wlth[99.0429-86.7759]	-1.974	0.349	*	0.139	12%
ownd[0]	-2.351	0.148	*	0.095	9%
kids	-3.142	0.146	*	0.043	4%
Adj_Wlth[46.16923-33.221403]	-4.705	0.729	*	0.009	1%
Intercept	-8.901	1.039	*	0.000	0%

Less impactful to the Odds that someone will Donate

- This is nice – we can easily SEE which predictors can impact the odds of someone donating

DONR Classification – Logistic Regression Profitability

Cost to Mail	$2.00
Profit from Donation	$12.50

Profit Matrix sepcified Classification Threshold .1071

True Positive Rate	100%
True Positives	996
False Negatives	3
Positive Class	999
P-Threshold	0.1071
Expected True Positives	996
Profit	$12,450

False Positive rate	35%
False Positives	354
True Negatives	665
Negative Class	1019
P-Threshold	0.1071
Expected False Positives	354
Profit	($4,425)

False Negative	0.30%
False Negatives	3
True Positives	996
Positive Class	999
P-Threshold	0.1071
Expected True Positives	3
Profit	($38)

Profit Comparison	
Analytics Model	
Profit from correctly Identifying Donors	$12,450
Loss avoided from Mis-Identifying Donors	($4,425)
Loss avoided from Not soliciting Donors	($38)
Profit	$7,988
Current Campaign	
Profit from Donors	$2,523
Cost from Solicitation	$4,036
Profit	($1,514)
Profit/Loss over current Campaign >	**$9,501**

P threshold = .1071

Method	TP	FN	FP	TN	Sensitivity	Specificity	Precision	Accuracy	F1	MCC	Profit
Fit Nominal Logistic	996	3	354	665	0.997	0.6526	0.7378	0.8231	0.848	0.6902	2.417

Specified Profit Matrix		
Decision		
Actual	1	0
1	12.5	-12.5
0	-2	1

Actual	**Decision Count**	
donr	1	0
1	996	3
0	354	665

DONR Classification – Logistic Regression Profitability

Cost to Mail	$2.00
Profit from Donation	$12.50

Profit Matrix sepcified Classification Threshold .1071

True Positive Rate	100%
True Positives	996
False Negatives	3
Positive Class	999
P-Threshold	0.1071
Expected True Positives	996
Profit	$12,450

False Positive rate	35%
False Positives	354
True Negatives	665
Negative Class	1019
P-Threshold	0.1071
Expected False Positives	354
Profit	($4,425)

False Negative	0.30%
False Negatives	3
True Positives	996
Positive Class	999
P-Threshold	0.1071
Expected True Positives	3
Profit	($38)

Profit Comparison	
Analytics Model	
Profit from correctly Identifying Donors	$12,450
Loss avoided from Mis-Identifying Donors	($4,425)
Loss avoided from Not soliciting Donors	($38)
Profit	$7,988
Current Campaign	
Profit from Donors	$2,523
Cost from Solicitation	$4,036
Profit	($1,514)
Profit/Loss over current Campaign >	**$9,501**

Notes:

- Average Probability for Donor Class **54%**
- Average Probability for Non-Donor Class **46%**

Potential Drawbacks:

- Model struggles to predict non-Donors
- Specified Profit matrix allows for easier entry into Donor class – not optimal for Profit

California Lutheran UNIVERSITY

DONR Classification –Logistic Regression Profitability

Cost to Mail	$2.00
Profit from Donation	$12.50

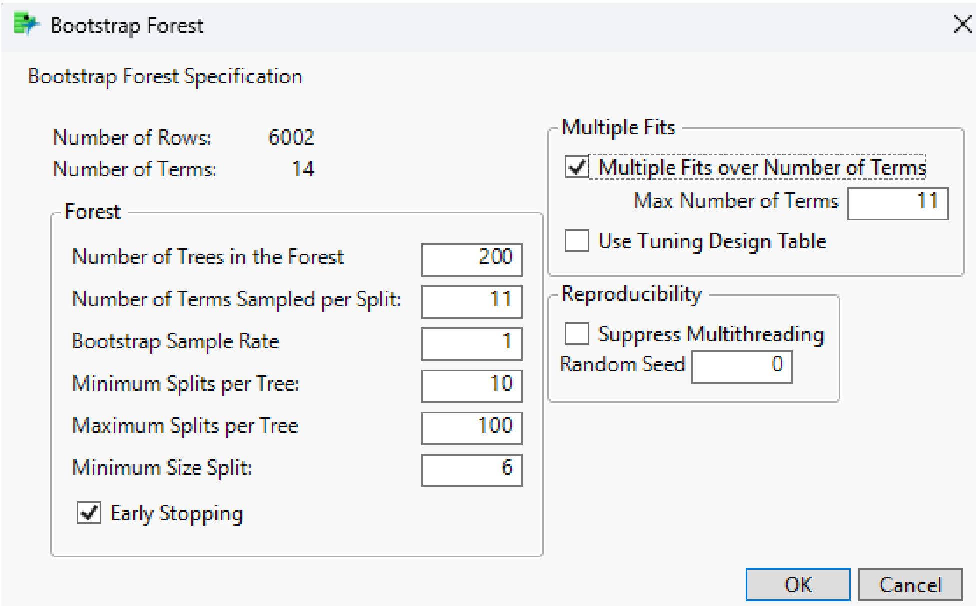
Optimal Classification Threshold .4777

True Positive Rate	92%	**False Positive rate**	14%	**False Negative**	7.51%
True Positives	924	False Positives	141	False Negatives	75
False Negatives	75	True Negatives	878	True Positives	924
Positive Class	999	Negative Class	1019	Positive Class	999
P-Threshold	0.4777	P-Threshold	0.4777	P-Threshold	0.4777
Expected True Positives	924	Expected False Positives	141	Expected True Positives	75
Profit	$11,550	Profit	($1,763)	Profit	($938)

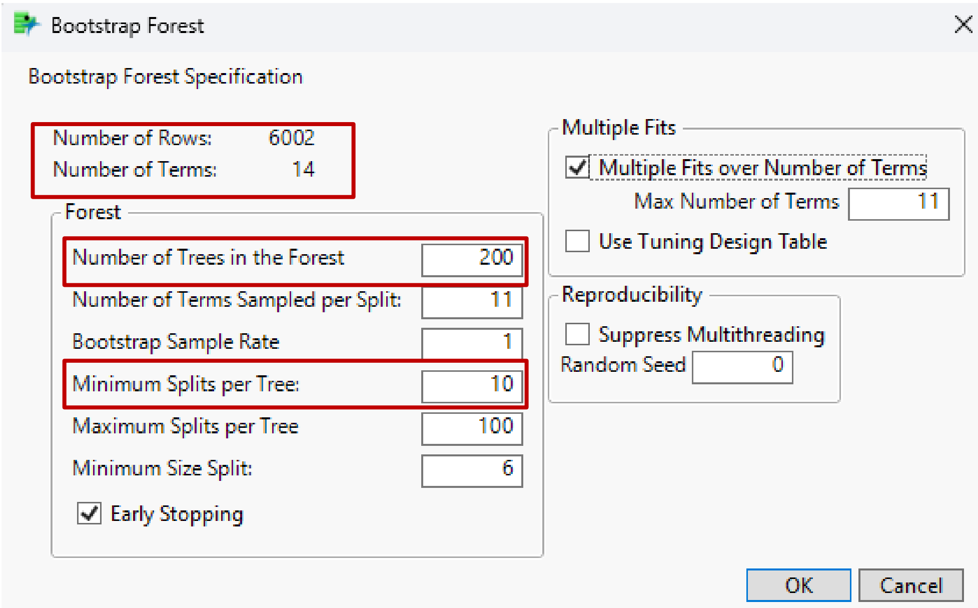
Profit Comparison	
Analytics Model	
Profit from correctly Identifying Donors	$11,550
Loss avoided from Mis-Identifying Donors	($1,763)
Loss avoided from Not soliciting Donors	($938)
Profit	$8,850
Current Campaign	
Profit from Donors	$2,523
Cost from Solicitation	$4,036
Profit	($1,514)
Profit/Loss over current Campaign >	**$10,364**

Defining Stricter entry criteria via Predicted Probability generates higher profit.

DONR Classification – Bootstrap Forest



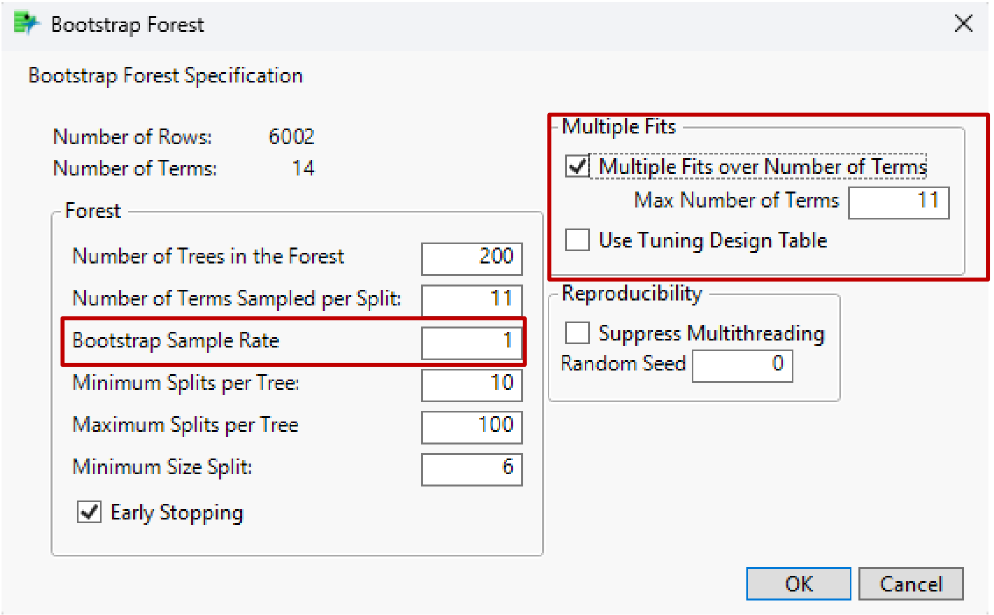
DONR Classification – Bootstrap Forest



Model Structure

- Accommodative of large data set and large predictor set

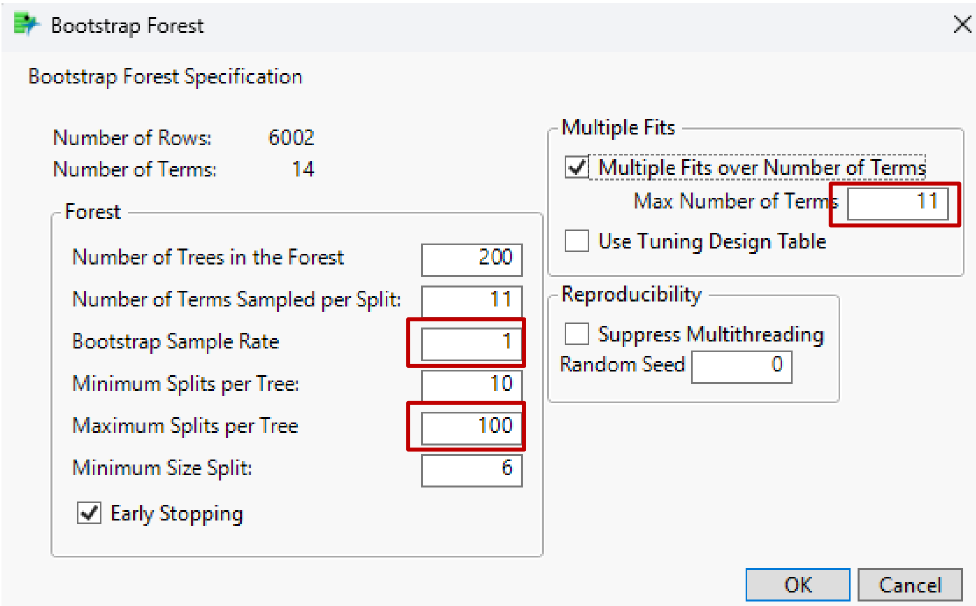
DONR Classification – Bootstrap Forest



Model Structure

- Accommodative of large data set and large predictor set
- Introduces Variation into the data set

DONR Classification – Bootstrap Forest

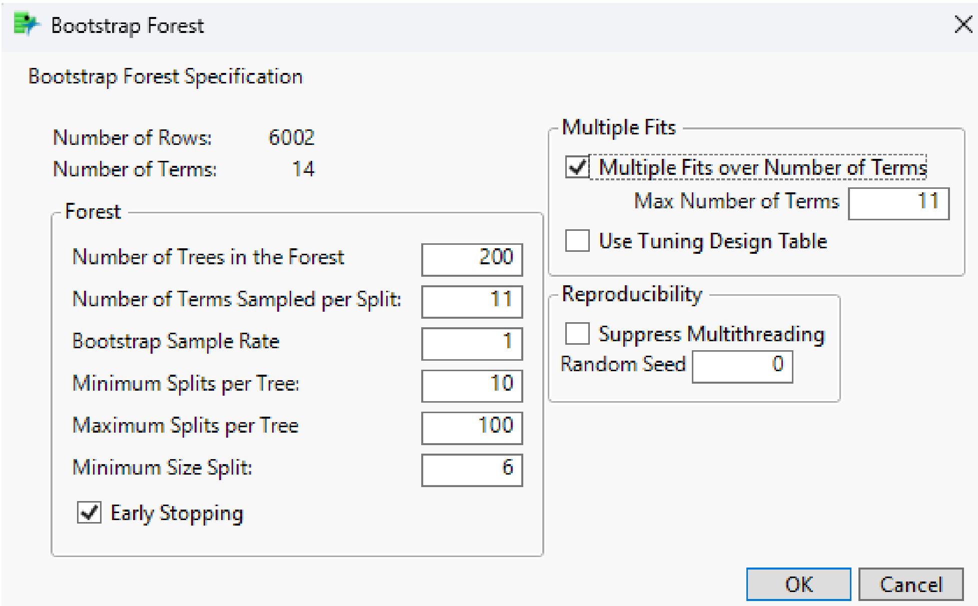


erm	Number of Splits	G^2		Portion
kids	930	461.875	+++++++++++++++++	0.1981
Kids_Squared	878	415.529	+++++++++++++++	0.1782
inc	3062	406.3673	++++++++++++++	0.1743
ter2	684	267.4954	++++++++++	0.1147
ownd	601	249.4581	+++++++++	0.107
nor_incmed	2032	145.4467	+++++	0.0624
ter1	971	74.16847	++	0.0318
Root_npro	1185	71.89991	++	0.0308
Adj_Wlth	944	71.74417	++	0.0308
npro	1168	66.5508	++	0.0285
Inc^2	1317	58.32026	++	0.025
sex	805	20.22205		0.0087
ter3	340	12.14274		0.0052
ter4	329	10.72247		0.0046

Model Structure

- Accommodative of large data set and large predictor set
- Introduces Variation into the data set
- Conscious of Overfitting Risk and aware of column contributions

DONR Classification – Bootstrap Forest



Model Structure

- Accommodative of large data set and large predictor set
- Introduces Variation into the data set
- Conscious of Overfitting Risk and aware of column contributions

Model Results

		FN	FP	TN	Sensitivity	Specificity	Precision	Accuracy	F1	MCC	Profit
Bootstrap I	907	92	171	848	0.9079	0.8322	0.8414	0.8697	0.8734	0.7418	5.2991

DONR Classification – Bootstrap Forest Profitability

Cost to Mail	$2.00
Profit from Donation	$12.50

Optimal Classification Threshold .4745

True Positive Rate	99%	**False Positive rate**	33%	**False Negative**	1.40%	***Profit Comparison***	
True Positives	985	False Positives	333	False Negatives	14	*Analytics Model*	
False Negatives	14	True Negatives	686	True Positives	985	Profit from correctly Identifying Donors	$12,313
Positive Class	999	Negative Class	1019	Positive Class	999	Loss avoided from Mis-Identifying Donors	($4,163)
P-Threshold	0.2227	P-Threshold	0.2227	P-Threshold	0.2227	Loss avoided from Not soliciting Donors	($175)
Expected True Positives	985	Expected False Positives	333	Expected True Positives	14	**Profit**	$7,975
Profit	$12,313	Profit	($4,163)	Profit	($175)		
2018						*Current Campaign*	
						Profit from Donors	$2,523
						Cost from Solicitation	$4,036
						Profit	($1,514)
						Profit/Loss over current Campaign >	**$9,489**

Cost to Mail	$2.00
Profit from Donation	$12.50

Profit Matrix sepcified Classification Threshold .1071

True Positive Rate	99%	**False Positive rate**	45%	**False Negative**	0.50%	***Profit Comparison***	
True Positives	994	False Positives	455	False Negatives	5	*Analytics Model*	
False Negatives	6	True Negatives	564	True Positives	994	Profit from correctly Identifying Donors	$12,413
Positive Class	999	Negative Class	1019	Positive Class	999	Loss avoided from Mis-Identifying Donc	($5,688)
P-Threshold	0.1071	P-Threshold	0.1071	P-Threshold	0.1071	Loss avoided from Not soliciting Donor:	($63)
Expected True Positives	993.006	Expected False Positives	455	Expected True Positives	5	**Profit**	$6,663
Profit	$12,413	Profit	($5,688)	Profit	($63)		
						Current Campaign	
						Profit from Donors	$2,523
						Cost from Solicitation	$4,036
						Profit	($1,514)
						Profit/Loss over current Campaign >	**$8,176**

Takeaway

- Model is classifying all participants with predicted probability >47.4% as a Donor

California Lutheran UNIVERSITY

DONR Classification – Bootstrap Forest Profitability

Cost to Mail	$2.00
Profit from Donation	$12.50

Optimal Classification Threshold .4745

True Positive Rate	99%	False Positive rate	33%	False Negative	1.40%
True Positives	985	False Positives	333	False Negatives	14
False Negatives	14	True Negatives	686	True Positives	985
Positive Class	999	Negative Class	1019	Positive Class	999
P-Threshold	0.2227	P-Threshold	0.2227	P-Threshold	0.2227
Expected True Positives	985	Expected False Positives	333	Expected True Positives	14
Profit	$12,313	Profit	($4,163)	Profit	($175)

2018

Profit Comparison	
Analytics Model	
Profit from correctly Identifying Donors	$12,313
Loss avoided from Mis-Identifying Donors	($4,163)
Loss avoided from Not soliciting Donors	($175)
Profit	$7,975
Current Campaign	
Profit from Donors	$2,523
Cost from Solicitation	$4,036
Profit	($1,514)
Profit/Loss over current Campaign >	**$9,489**

Cost to Mail	$2.00
Profit from Donation	$12.50

Profit Matrix sepcified Classification Threshold .1071

True Positive Rate	99%	False Positive rate	45%	False Negative	0.50%
True Positives	994	False Positives	455	False Negatives	5
False Negatives	6	True Negatives	564	True Positives	994
Positive Class	999	Negative Class	1019	Positive Class	999
P-Threshold	0.1071	P-Threshold	0.1071	P-Threshold	0.1071
Expected True Positives	993.006	Expected False Positives	455	Expected True Positives	5
Profit	$12,413	Profit	($5,688)	Profit	($63)

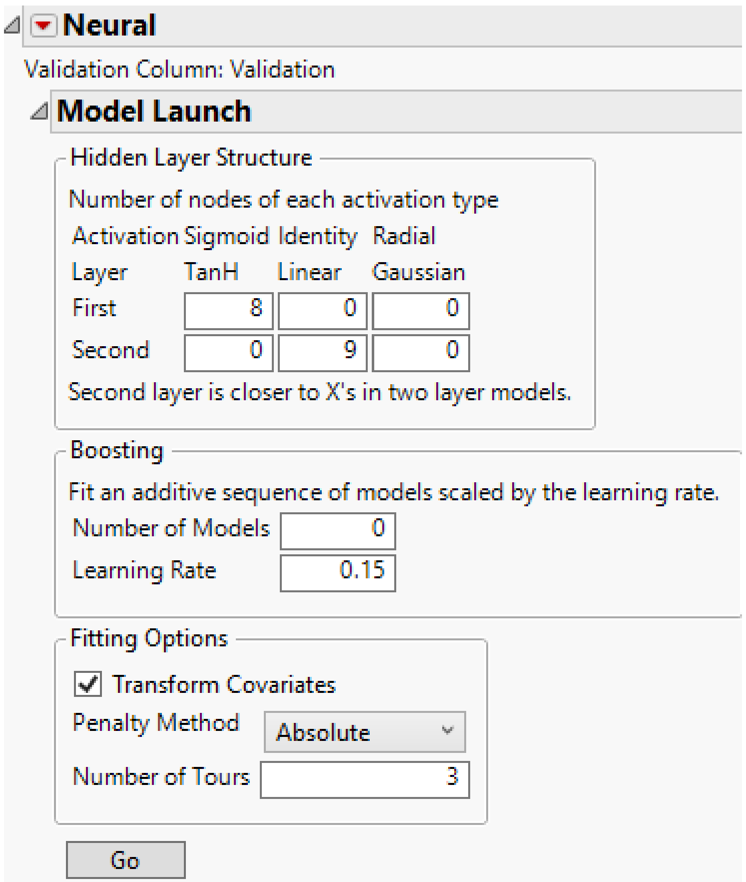
Profit Comparison	
Analytics Model	
Profit from correctly Identifying Donors	$12,413
Loss avoided from Mis-Identifying Donc	($5,688)
Loss avoided from Not soliciting Donor:	($63)
Profit	$6,663
Current Campaign	
Profit from Donors	$2,523
Cost from Solicitation	$4,036
Profit	($1,514)
Profit/Loss over current Campaign >	**$8,176**

Takeaway

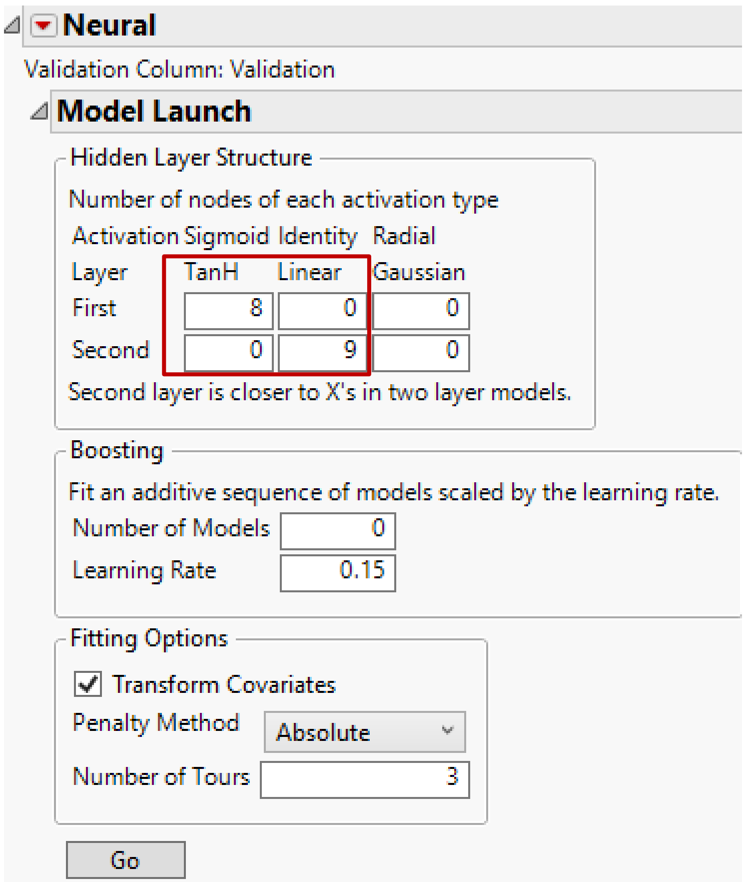
- Using model to solicit people with >47% chance of being a Donor will generate $9.5K more profit than the current campaign

California Lutheran UNIVERSITY

DONR Classification – Neural Net



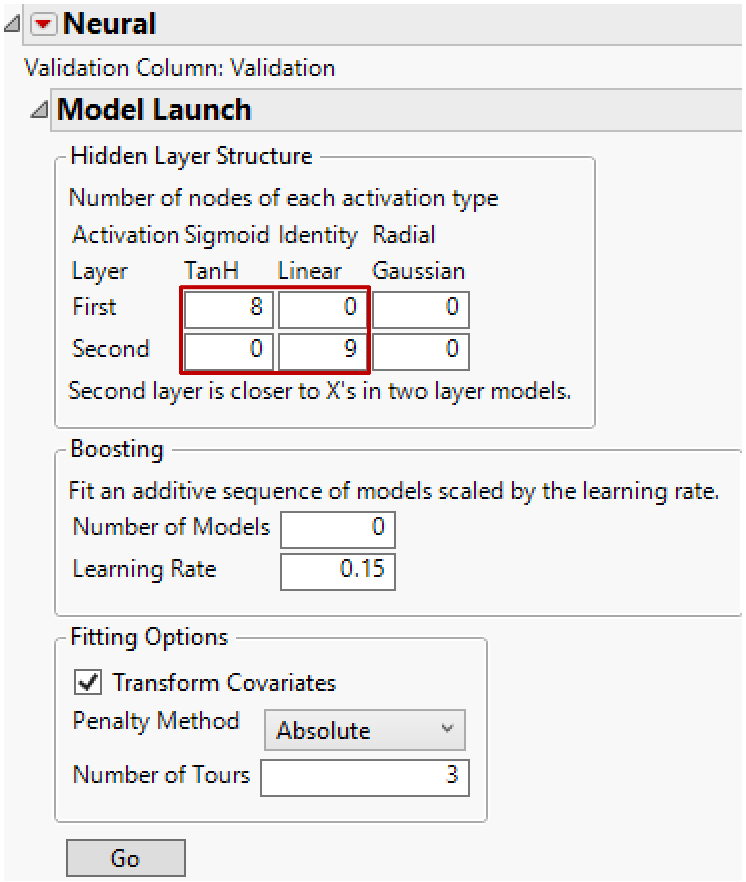
DONR Classification – Neural Net



Model Structure

- Dual activation functions
  - TanH bounds signal in 1st hidden layer
  - Linear sends “full” signal to output layer

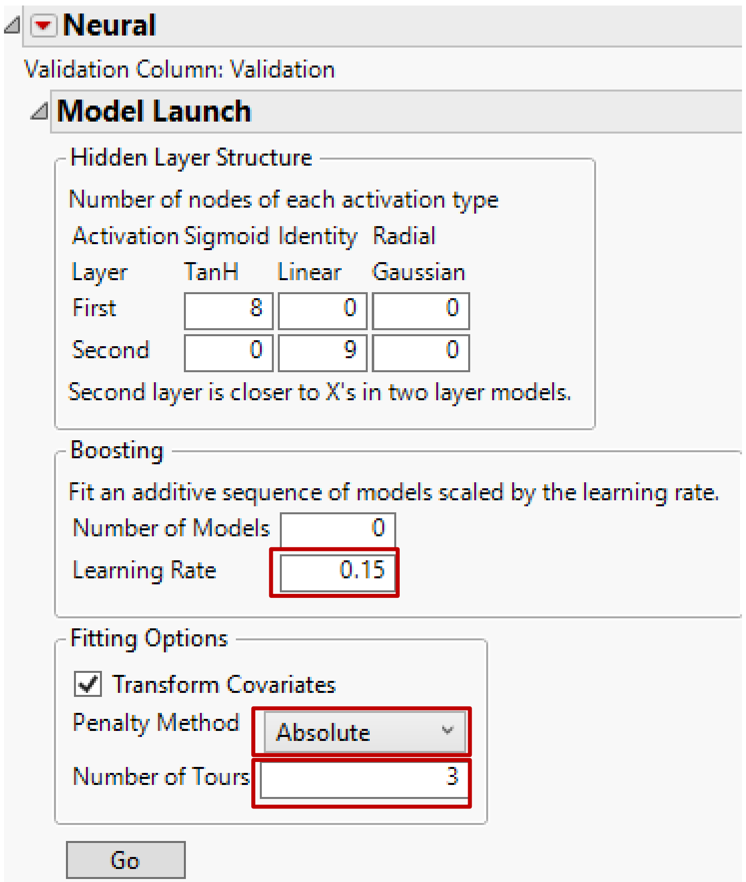
DONR Classification – Neural Net



Model Structure

- Node count chosen via modelers discretion.
  - Tuning Feature suggested Node counts of 8 and 9 respectively citing "explained variation" (R^2 metric)
  - 3 and 3 node count provided competitive results.

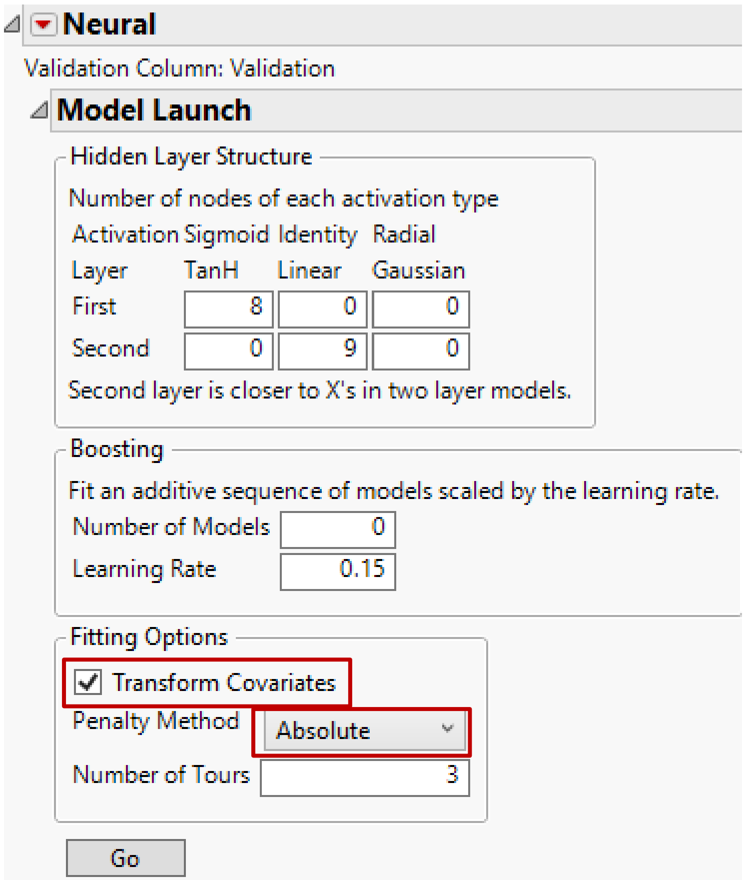
DONR Classification – Neural Net



Model Structure

- Mitigate Overfitting Risk
  - Increased learning rate because of possible collinearity of predictors
  - Penalization applied/chosen from profiler revealing that a few predictors dictate the model's prediction
  - Allowance of multiple (but not excessive epochs)

DONR Classification – Neural Net



Model Structure

- Transform covariates because our data was *standardized*
    - Model will be reading skewed data if not addressed here.

DONR Classification – Neural Net results

Training

Measures	Value
Generalized RSquare	0.802068
Entropy RSquare	0.663767
RASE	0.2685
Mean Abs Dev	0.148177
Misclassification Rate	0.101657
-LogLikelihood	928.5048
Sum Freq	3984

Validation

Measures	Value
Generalized RSquare	0.767227
Entropy RSquare	0.617958
RASE	0.282816
Mean Abs Dev	0.159152
Misclassification Rate	0.10555
-LogLikelihood	534.3518
Sum Freq	2018

Actual	Predicted Count	
donr	0	1
0	1758	231
1	174	1821

Actual	Predicted Count	
donr	0	1
0	884	135
1	78	921

Actual	Predicted Rate	
donr	0	1
0	0.884	0.116
1	0.087	0.913

Actual	Predicted Rate	
donr	0	1
0	0.868	0.132
1	0.078	0.922

Method	TP	FN	FP	TN	Sensitivity	Specificity	Precision	Accuracy	F1	MCC
Neural	915	84	133	886	0.9159	0.8695	0.8731	0.8925	0.894	0.7859

Classification

- Model appears to generalize well
- Model is **89%** accurate
- P-threshold is realistic: **58.6%**

DONR Classification – Neural Net Profitability

Cost to Mail	$2.00
Profit from Donation	$12.50

Optimal Classification Threshold .5861

True Positive Rate	90%	**False Positive rate**	11%	**False Negative**	9.61%
True Positives	903	False Positives	115	False Negatives	96
False Negatives	96	True Negatives	904	True Positives	903
Positive Class	999	Negative Class	1019	Positive Class	999
P-Threshold	0.2227	P-Threshold	0.2227	P-Threshold	0.2227
Expected True Positives	903	Expected False Positives	115	Expected True Positives	96
Profit	$11,288	Profit	($1,438)	Profit	($1,200)
2018					

Profit Comparison	
Analytics Model	
Profit from correctly Identifying Donors	$11,288
Loss avoided from Mis-Identifying Donors	($1,438)
Loss avoided from Not soliciting Donors	($1,200)
Profit	$8,650
Current Campaign	
Profit from Donors	$2,523
Cost from Solicitation	$4,036
Profit	($1,514)
Profit/Loss over current Campaign >	**$10,164**

Takeaway

- Using model to solicit people with >50% chance of being a Donor will generate $10K more profit than the current campaign

California Lutheran UNIVERSITY

DONR Classification – K-Nearest

Training					
K	Count	RSquare	Misclassification Rate	Misclassifications	
1	3984	27%	0.202	805	
2	3984	33%	0.220	878	
3	3984	36%	0.184	732	
4	3984	38%	0.188	747	
5	3984	40%	0.178	708	
6	3984	40%	0.178	709	
7	3984	41%	0.179	712	
8	3984	41%	0.179	712	
9	3984	41%	0.176	703	
10	3984	41%	0.175	697	*
11	3984	42%	0.177	706	
12	3984	42%	0.176	701	
13	3984	43%	0.180	717	
14	3984	43%	0.180	716	
15	3984	43%	0.181	721	

K	Count	RSquare	Misclassification Rate	Misclassifications	
1	2018	23%	0.219	442	
2	2018	32%	0.218	439	
3	2018	35%	0.184	371	
4	2018	37%	0.195	394	
5	2018	38%	0.177	358	
6	2018	39%	0.178	359	
7	2018	40%	0.167	336	
8	2018	41%	0.166	335	*
9	2018	40%	0.169	342	
10	2018	42%	0.173	350	
11	2018	42%	0.173	350	
12	2018	42%	0.177	357	
13	2018	41%	0.178	359	
14	2018	41%	0.186	376	
15	2018	41%	0.178	360	

Metrics	
Sensitivity	0.944
Specificity	0.726
Accuracy	0.834
Precision	0.772
F1	0.849
MCC	0.686

Take note of:

- Competitive classification capability for Donors
- Shortcomings
  - Adding new predictors will require a lot more data.
  - No insight into *which* predictor is more/less useful (non-parametric model)
  - May struggle outside this particular test environment.

DONR Classification – K-Nearest

Cost to Mail	$2.00
Profit from Donation	$12.50

No P-Value Threshold Specified

True Positive Rate	94%
True Positives	943
False Negatives	56
Positive Class	999
P-Threshold	-
Expected True Positives	943
Profit	$11,788

False Positive rate	27%
False Positives	279
True Negatives	740
Negative Class	1019
P-Threshold	0.2227
Expected False Positives	279
Profit	($3,488)

False Negative	5.61%
False Negatives	56
True Positives	943
Positive Class	999
P-Threshold	0.2227
Expected True Positives	56
Profit	($700)

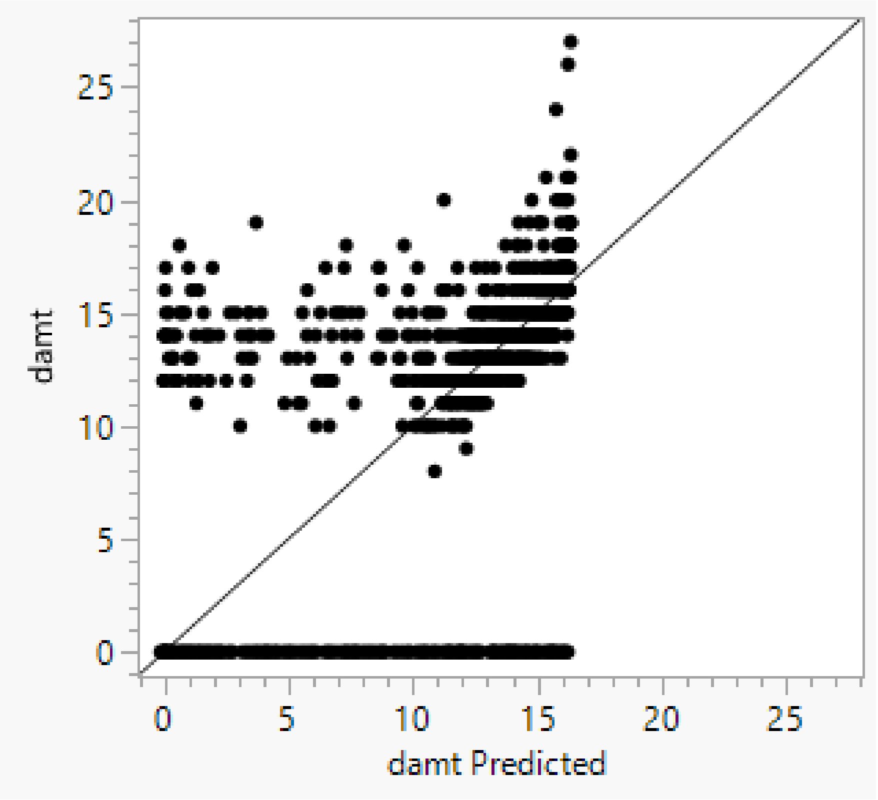
Profit Comparison	
Analytics Model	
Profit from correctly Identifying Donors	$11,788
Loss avoided from Mis-Identifying Donc	($3,488)
Loss avoided from Not soliciting Donor:	($700)
Profit	$7,600
Current Campaign	
Profit from Donors	$2,523
Cost from Solicitation	$4,036
Profit	($1,514)
Profit/Loss over current Campaign >	**$9,114**

Modeling – Donation Amount Prediction

1. Neural Network
2. XGBoost
3. Bootstrap Forest
4. Linear Regression

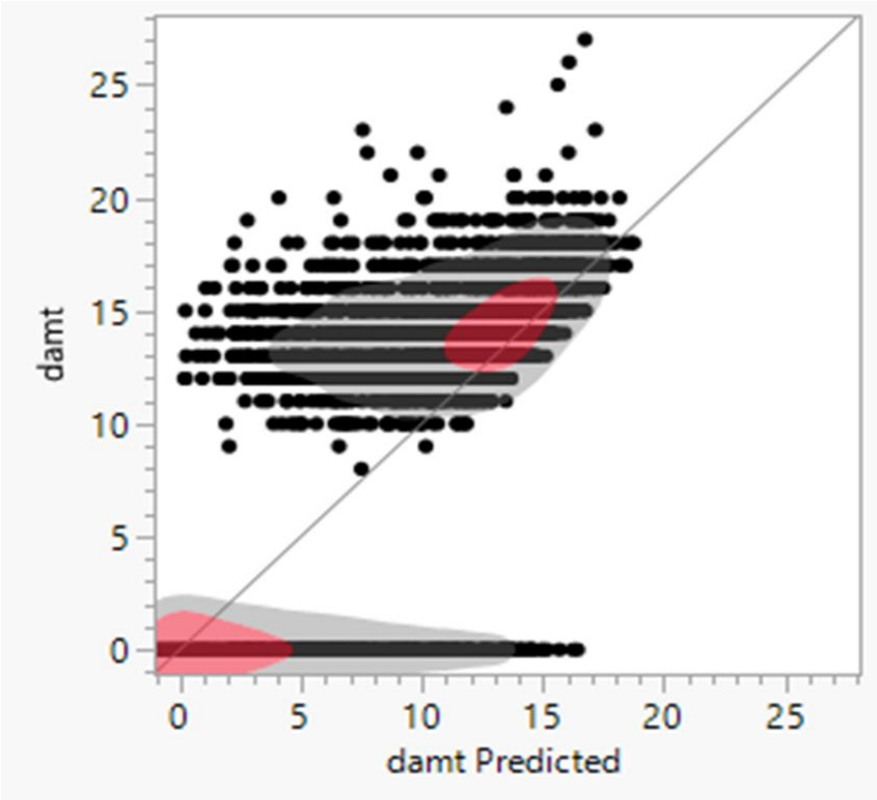
DAMT Prediction – Residuals vs Predicted

Neural Network



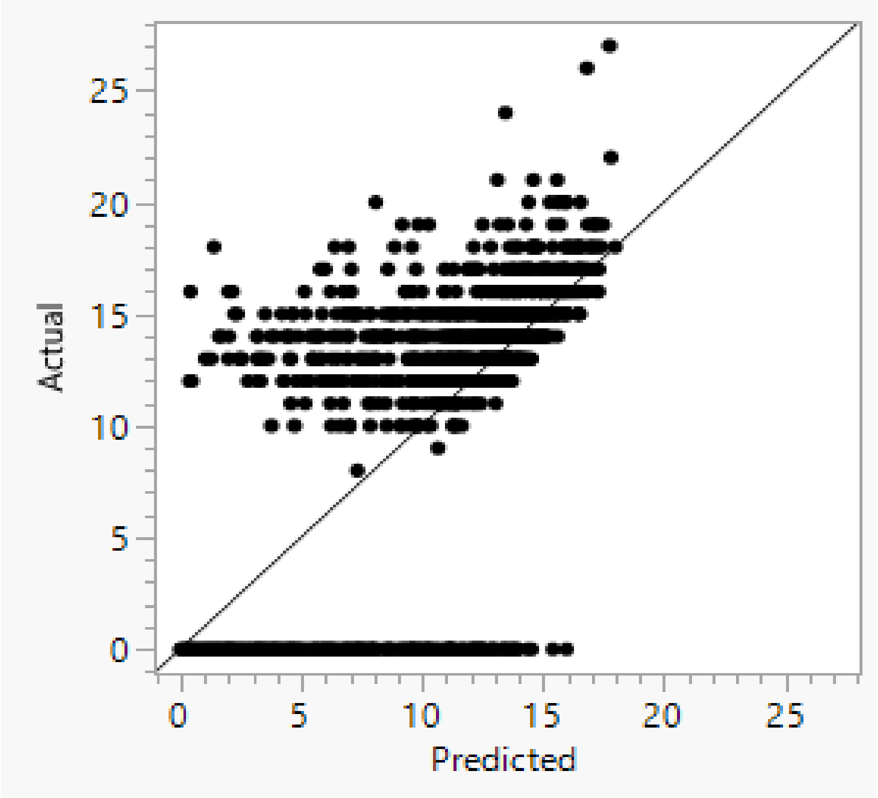
DAMT Prediction – Residuals vs Predicted

XGBoost



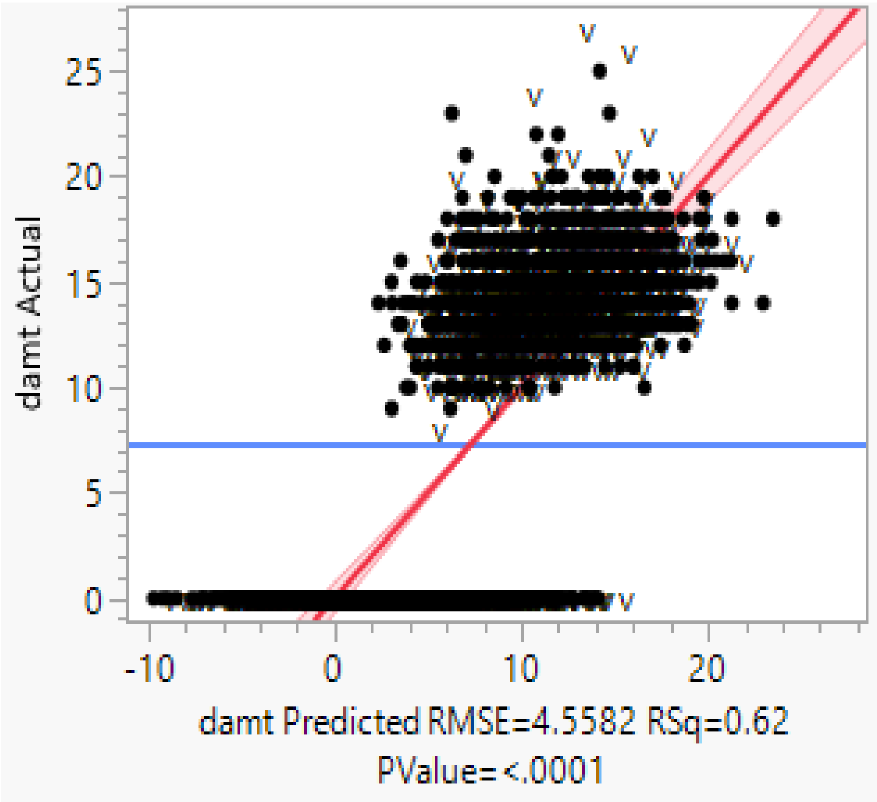
DAMT Prediction – Residuals vs Predicted

Bootstrap Forest



DAMT Prediction – Residuals vs Predicted

Linear Regression



DAMT Prediction – Model Comparison

Model Comparison

Neural Network

Measures	Value
RSquare	0.9274883
RASE	4.1016141
Mean Abs Dev	1.9141827
-LogLikelihood	4727.0397
SSE	33949.295
Sum Freq	2018

XGBoosted Tree

Measure	Value
RSquare	0.6816
Correlation	0.8268
RMSE	4.1533
MAE	2.7363

Bootstrap Forest

	RSquare	RASE	N
Training	0.858	2.7820214	3984
Validation	0.679	4.1530079	2018

Linear Regression

Source	RSquare	RASE	Freq
Training Set	0.6227	4.5318	3984
Validation Set	0.5881	4.7022	2018

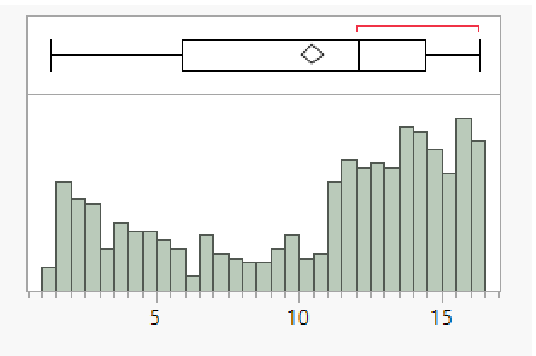
DAMT Prediction – Model Selection

Neural Network

- Neural network had the lowest RASE and highest R^2
- Returns an average predicted donation amount of $10.48
- Less than the average donation amount from the initial problem
- Likely due to the models struggling to predict who donates on its own
- May not be a problem for predicting lower than the average because:
  - Planning for more conservative predictions allows for better cash flow handling by ensuring liquidity
  - And allows for upside surprises in the case of exceeded expectations

Mean	10.482498
Std Dev	4.7530772
Std Err Mean	0.2026719
Upper 95% Mean	10.880605
Lower 95% Mean	10.084391
N	550
N Missing	2

100.0%	maximum	16.354828
99.5%		16.311546
97.5%		16.176484
90.0%		15.750464
75.0%	quartile	14.388443
50.0%	median	12.088546
25.0%	quartile	5.9137316
10.0%		2.6883805
2.5%		1.687637
0.5%		1.3809455
0.0%	minimum	1.3132042



Next Steps

Implement Campaign using Analytics

1. Donor classification: Distribute mail to those our Neural Network model says have at least a 58% probability of being a Donor
2. Donation Amount: Apply Neural Network model to predict donation amounts in the Donor class

Business Benefit: Minimizes wasted mailers & refines donation forecasts for allocating future budgets.

Next Steps

ROI & Campaign Benefits

- **Reduced Mailing Costs**
  - Exclude 50+% of individuals with low donation probability
- **Higher Accuracy**
  - 89.5% classification success → fewer misses on potential donors
  - RMSE ~4.1 for amounts → improved revenue forecasts
- **Revenue Stability**
  - Under-predicting top donors means upside if actuals exceed predictions

Next Steps

Further Analytics through Donor Segmentation

- **Purpose:** Combine donation likelihood (DONR) with predicted gift amounts (DAMT)
  - Exclude 50+% of individuals with low donation probability
- **Approach:**
  - Sort donors by territory, wealth, income, kids, etc.
  - Data-driven thresholds yield color-coded personas for marketing focus

Next Steps

Persona Blue

Profile

- High wealth rating (WLTH≥7), moderate income (INC=4–5), homeowner
- Neural Net DONR Probability ~70–80%

DAMT:

- \$20–\$25 predicted (far above the \$2 cost)

Business Case:

- Fewer mailers, significant average gift → robust ROI

Persona Yellow

Profile

- Possibly territory 2/3, moderate wealth/income, stable home
- DONR Probability ~60–70%

DAMT:

- \$5–\$10 each, but higher frequency of giving

Business Angle:

- Frequent smaller gifts → cumulative annual total is substantial

Persona Green

Profile

- Mid-level income (INC=3–4), territory 4, 1–2 kids
- DONR Probability ~65–75% (Neural Net)

DAMT:

- \$12–\$18 typical donation

Value:

- Reliable mid-tier donor base, easy net profit margin

California Lutheran UNIVERSITY

Next Steps - Deployment

1. Score Everyone

- Neural Net classification for DONR
- Retain only those > X% threshold

2. Estimate Gift (DAMT)

- Neural Net regression → budgeting & planning

3. Tailor Marketing

- Personalize mailers for Persona Blue, Green, or Yellow

4. Monitor & Retrain

- Keep track of actual donation rates/amounts; update models every cycle

Final Takeaways

Modeling Insight:

- Neural Net outperforms other classifiers for DONR (~89.5% accuracy).
- Neural Net also excels in DAMT (~0.927 R2, RMSE ~4.1).

Campaign improvement has Two-Step solution:

- Identify donors (DONR)
- Predict amounts (DAMT)

Build upon your Analytics

- Personas help refine marketing strategies and mailers.

Conclusion: From a –$0.55 baseline to a **positive** margin per piece, this data-driven approach significantly increases fundraising efficiency and ROI

References

- Garcia, John. “Week 5 Overview.” Foundation of Analytics. K Nearest Neighbors, 10 Dec. 2024, Westlake Village, Hub101.
- Garcia, John. “Week 8 Overview.” Foundation of Analytics. K Nearest Neighbors, 7 Jan. 2025, Westlake Village, Hub101.
- Garcia, John. “Week 7 Overview.” Foundation of Analytics. K Nearest Neighbors, 17 Dec. 2024, Westlake Village, Hub101.
- ML Explainer Chatbot. (2025). Response to “Why does my profit Matrix not give me the optimal P-threshold” prompt. Garcia, John, California Lutheran University, 10 Feb 2025.
- Wei, Qiong, and Roland L. Dunbrack. “The role of balanced training and testing data sets for binary classifiers in bioinformatics.” *PLoS ONE*, vol. 8, no. 7, 9 July 2013, https://doi.org/10.1371/journal.pone.0067863.
- SHMUELI, Galit, et al. “7,9,10,11.” Machine Learning for Business Analytics, Concepts, Techniques and Applications with JMP Pro, 2nd ed., vol. 1, Machine Learning for Business Analytics, Concepts, Techniques and Applications with JMP Pro, Hoboken, NJ, 2023, pp. 147–311.

DONR Classification – Neural Net without Ownd

Training	
Measures	**Value**
Generalized RSquare	0.741781
Entropy RSquare	0.58623
RASE	0.300013
Mean Abs Dev	0.184774
Misclassification Rate	0.126506
-LogLikelihood	1142.622
Sum Freq	3984

Validation	
Measures	**Value**
Generalized RSquare	0.698828
Entropy RSquare	0.535679
RASE	0.313796
Mean Abs Dev	0.196042
Misclassification Rate	0.1333
-LogLikelihood	649.433
Sum Freq	2018

Model Structure

- Model still generalized well, but loses 2% accuracy
- Probability threshold: **51.9%**

Actual	**Predicted Count**	
donr	**0**	**1**
0	1725	264
1	240	1755

Actual	**Predicted Count**	
donr	**0**	**1**
0	867	152
1	117	882

Actual	**Predicted Rate**	
donr	**0**	**1**
0	0.867	0.133
1	0.12	0.88

Actual	**Predicted Rate**	
donr	**0**	**1**
0	0.851	0.149
1	0.117	0.883

Method	TP	FN	FP	TN	Sensitivity	Specificity	Precision	Accuracy	F1	MCC
Neural	1744	251	251	1738	0.8742	0.8738	0.8742	$\boxed{0.874}$	0.8742	0.748