ECON-562: Final Project

Data Analytics for Direct Marketing

Andrew Frohner

John Garcia

May 18th, 2025

California Lutheran University

Master of Science in Quantitative Economics

Executive Summary

1. **Project Overview**
   - Business Understanding
   - Key Objectives

2. **Data Understanding**
   - Data Overview
   - Data Quality
   - Key Findings

3. **Data Preparation**
   - Data Cleaning / Feature Engineering

4. **Modeling**
   - Selection / Tuning Approach
   - Metrics / Evaluation

5. **Insights / Deployment Strategy**

Business Understanding

Marketing Campaign

Purposed for:

1. Sending a message
2. Spreading awareness of our Financial Products

Business Understanding

Marketing Campaign

Purposed for:

1. Sending a message
2. Spreading awareness of our Financial Products

Revenue Sources

1. Credit
2. Car Loans
3. Short/Medium term Credit instruments

Business Understanding

Marketing Campaign

Purposed for:

1. Sending a message
2. Spreading awareness of our Financial Products

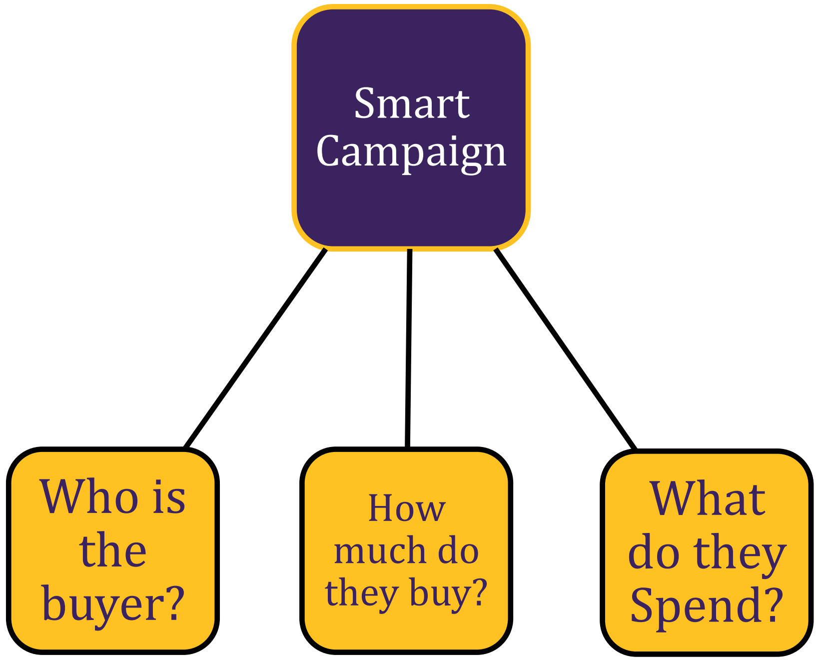
Revenue Sources

1. Credit
2. Car Loans
3. Short/Medium term Credit instruments

Objective

Use Data to spread awareness in a way that increases our Revenue

Business Understanding – Targets



Success requires answers to these questions

Data Understanding

Data Summary	
Numerical Features	19
Categorical Features	7
Number of Observations	1,060,038
Number of Incomplete observations	848,529
Target Set	**Type**
Customer bought a new product	Binary
Dollar amount of the product purchased	Numerical
Count of Products purchased by customer	Numerical
Predictor Set	**Type**
Account/Customer Characteristics	Numerical
Average Sales Measurements (in $)	Numerical
Average Sales Measurements (in count)	Numerical
Sales attributed to Promotions	Numerical
Count of Sales attributed to Promotions	Numerical
Time between purchases	Numerical
Customer loyalty	Categorical
Demographic information	Numerical/Categorical

- Large Dataset
- "Messy" Dataset

Data Understanding

Data Summary	
Numerical Features	19
Categorical Features	7
Number of Observations	1,060,038
Number of Incomplete observations	848,529
Target Set	**Type**
Customer bought a new product	Binary
Dollar amount of the product purchased	Numerical
Count of Products purchased by customer	Numerical
Predictor Set	**Type**
Account/Customer Characteristics	Numerical
Average Sales Measurements (in $)	Numerical
Average Sales Measurements (in count)	Numerical
Sales attributed to Promotions	Numerical
Count of Sales attributed to Promotions	Numerical
Time between purchases	Numerical
Customer loyalty	Categorical
Demographic information	Numerical/Categorical

- Multiple patterns in the in data.

Data Understanding

Data Summary	
Numerical Features	19
Categorical Features	7
Number of Observations	1,060,038
Number of Incomplete observations	848,529
Target Set	**Type**
Customer bought a new product	Binary
Dollar amount of the product purchased	Numerical
Count of Products purchased by customer	Numerical
Predictor Set	**Type**
Account/Customer Characteristics	Numerical
Average Sales Measurements (in $)	Numerical
Average Sales Measurements (in count)	Numerical
Sales attributed to Promotions	Numerical
Count of Sales attributed to Promotions	Numerical
Time between purchases	Numerical
Customer loyalty	Categorical
Demographic information	Numerical/Categorical

- 3 questions, but only 2 problems.

Data Understanding – Missing Data

Missing Values			
Code	**Description**	**Variable**	**# of missing**
int_tgt	$ amt of Product purchased	Target	848,529
cnt_tgt	Count of Products purchased	Target	1
rfm3	Avg Sales past 3 Years from Dir Promo	Predictor	225,786
demog_age	Customer Age	Predictor	266,861

Use 'em or lose 'em?

Data Understanding – Missing Data

Missing Values			
Code	**Description**	**Variable**	**# of missing**
int_tgt	$ amt of Product purchased	Target	848,529
cnt_tgt	Count of Products purchased	Target	1
rfm3	Avg Sales past 3 Years from Dir Promo	Predictor	225,786
demog_age	Customer Age	Predictor	266,861

USE THEM!

Convenient

1. INT_TGT imputed as 0
2. CNT_TGT imputed as 0

Data Understanding

Missing Values			
Code	**Description**	**Variable**	**# of missing**
int_tgt	$ amt of Product purchased	Target	848,529
cnt_tgt	Count of Products purchased	Target	1
rfm3	Avg Sales past 3 Years from Dir Promo	Predictor	225,786
demog_age	Customer Age	Predictor	266,861

USE THEM!

Complicated

1. Rfm3 imputed with *Linear Regression*
2. Demog_age imputed with *Linear Regression*

Data Understanding

Missing Values			
Code	**Description**	**Variable**	**# of missing**
int_tgt	$ amt of Product purchased	Target	848,529
cnt_tgt	Count of Products purchased	Target	1
rfm3	Avg Sales past 3 Years from Dir Promo	Predictor	225,786
demog_age	Customer Age	Predictor	266,861

USE THEM!

Complicated

1. Rfm3 imputed with *Linear Regression*
2. Demog_age imputed with *Linear Regression*

Side Effect

1. Generated 22 negative Observations
2. Generated 20 negative Observations

Now use quantile imputation

Data Understanding – Erroneous data

Code	Desc	Erroneous Observation	Imputation Method	Impacted Observations
CNT_tgt	Count of Products purchased	CNT_tgt = 6	Mean of CNT_tgt when Income is between $30K and $35K	11
rfm4	Last Product Purchase Amt	rfm4 > 8000	Impute as 0	11
rfm2	Avg Lifetime Sales	rfm2 > 500	Mean of rfm2 when rfm2 < 500	11
rfm3	Avg 3yr Sales from Dir Promo	rfm3 > 3000	Mean of rfm2 when rfm3 < 3000	11
rfm5	Count of Products purchased last 3Yrs	rfm5 = 18	When int_tgt = 5 impute as Mean of rfm5 when Int_tgt = 5	3
rfm5	Count of Products purchased last 3Yrs	rfm5 = 18	When int_tgt = 0 impute as Mean of rfm5 when Int_tgt = 0	2
rfm6	Count of Products purchased Lifetime	rfm6 >100	when Int_tgt < $20K impute as Mean of rfm6	11
rfm8	Count of Products purchased from Dir Promo	rfm8 = 46	Impute as 0	5
rfm9	Months Since Last Purchase	Age < 21	impute rfm9 as 0 when demog_age < 21	11

Count of observations when Customer age less than 21	12,066

* Observations were removed

Spot an observation and question its validity in the data set

California Lutheran UNIVERSITY

Data Understanding – Key Findings

1. Skewed Distributions

Variable	Skewness	Transformation Applied	Skewness post Transformation
int_tgt	4.84	Log	1.57
cnt_tgt	2.40		2.40
demog_age	-0.12	Log	-0.77
demog_homeval	2.46	Log	-5.99
demog_inc	0.23	Log	-1.21
~~demog_pr~~	~~-0.15~~	-	-
rfm1	103.15	Log	-1.11
rfm2	8.54	Log	0.36
rfm3	40.24	Log	0.12
rfm4	88.75	Log	-0.48
rfm5	1.23	Log	-0.22
rfm6	1.89	Log	-0.21
rfm7	1.23	Log	-0.03
rfm8	1.43	Log	-0.14
rfm9	-0.60	Log	-2.52
rfm10	2.86	Log	0.72
rfm11	0.32	Log	-1.35
rfm12	0.31	Log	-0.38
~~account~~	~~0.00~~	~~Log~~	-
demog_inc2	1.31	Log	-0.10
demog_inc2_sq	3.98	Log	-0.10
rfm6_sq	7.88	Log	-0.20
prospect_ho	8.84	Log	8.84
rfm2_inc2	9.81	Log	0.13

Data Understanding – Key Findings

2. Class Imbalance (B_TGT)

Training Data Set	Obs	Percentage
Buy a new product = YES	125285	20%
Buy a new product = NO	503655	80%
Total	**628940**	**100%**

Has implications on Model performance – decide now

Data Understanding – Key Findings

2. Class Imbalance (B_TGT)

Training Data Set	Obs	Percentage
Buy a new product = YES	125285	20%
Buy a new product = NO	503655	80%
Total	**628940**	**100%**

SMOTE Data Set	Obs	Percentage	Percent Increase in Obs
Buy a new product = YES	503655	50%	**302%**
Buy a new product = NO	503655	50%	0%
Total	**1007310**	**100%**	60%

Synthetic Observations?

Has implications on Model performance – decide now

Data Understanding – Key Findings

STOP

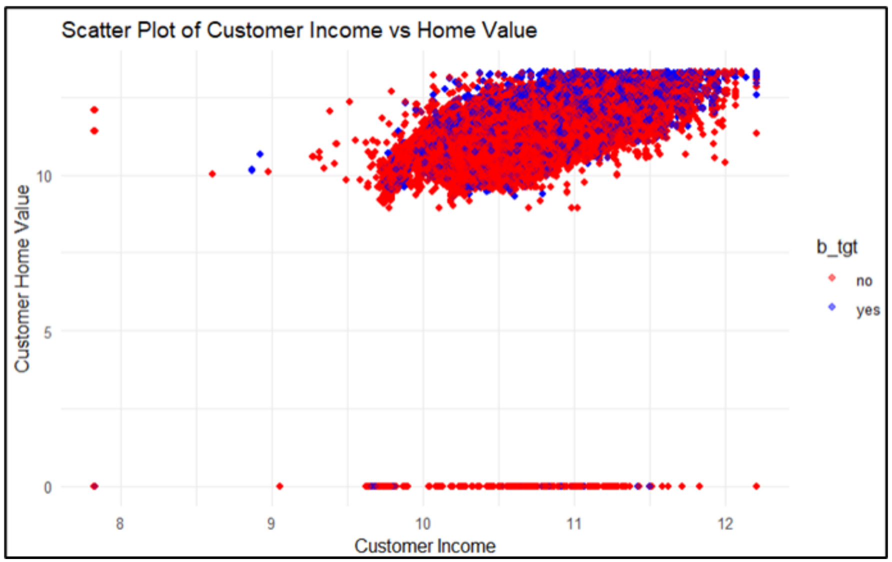
3. Outliers / Non-Linearities / Distributions

Variable	High Outliers	Low Outliers
int_tgt	207,593	-
cnt_tgt	211,509	-
demog_age	-	5,329
demog_homeval	73,306	-
demog_inc	8,470	-
demog_pr	7,858	31,245
rfm1	22,992	-
rfm2	29,359	-
rfm3	24,256	-
rfm4	30,067	-
rfm5	4,702	-
rfm6	21,343	-
rfm7	58,447	-
rfm8	21,247	-
rfm9	4	32,026
rfm10	75,991	21,319
rfm11	16,726	23,987
rfm12	66	-

Winsorize ?

Data Understanding – Key Findings

3. Outliers / Non-Linearities / Distributions

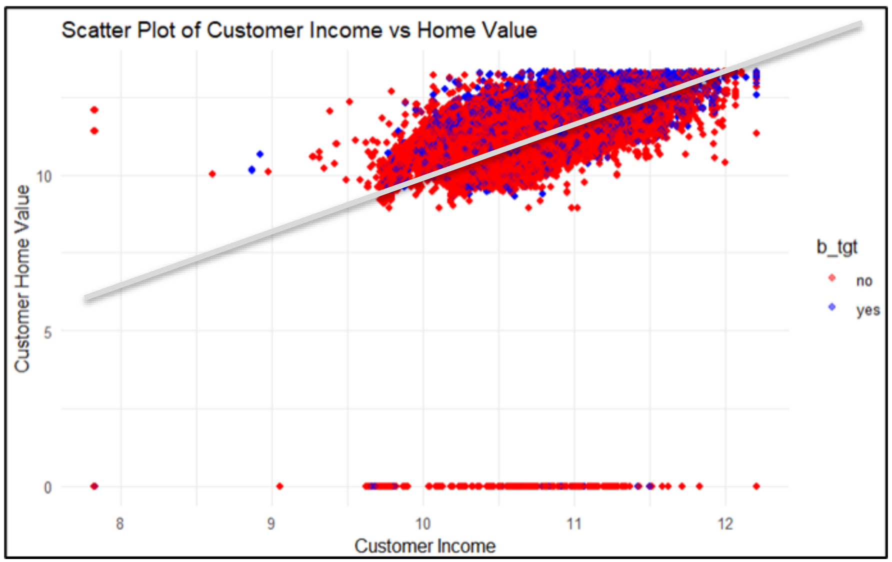


Variable	High Outliers	Low Outliers
int_tgt	207,593	-
cnt_tgt	211,509	-
demog_age	-	5,329
demog_homeval	73,306	-
demog_inc	8,470	-
demog_pr	7,858	31,245
rfm1	22,992	-
rfm2	29,359	-
rfm3	24,256	-
rfm4	30,067	-
rfm5	4,702	-
rfm6	21,343	-
rfm7	58,447	-
rfm8	21,247	-
rfm9	4	32,026
rfm10	75,991	21,319
rfm11	16,726	23,987
rfm12	66	-

Can I separate the classes with a straight line?

Data Understanding – Key Findings

3. Outliers / Non-Linearities / Distributions

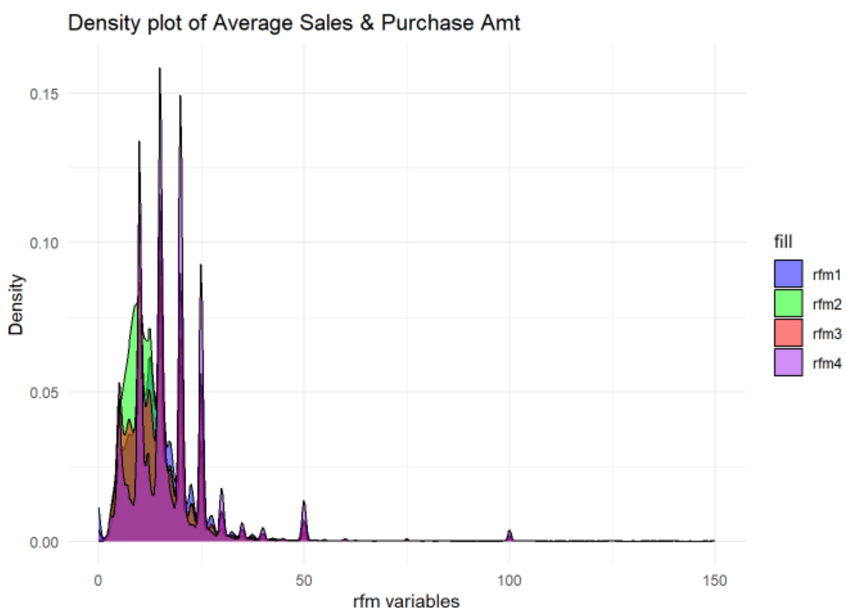


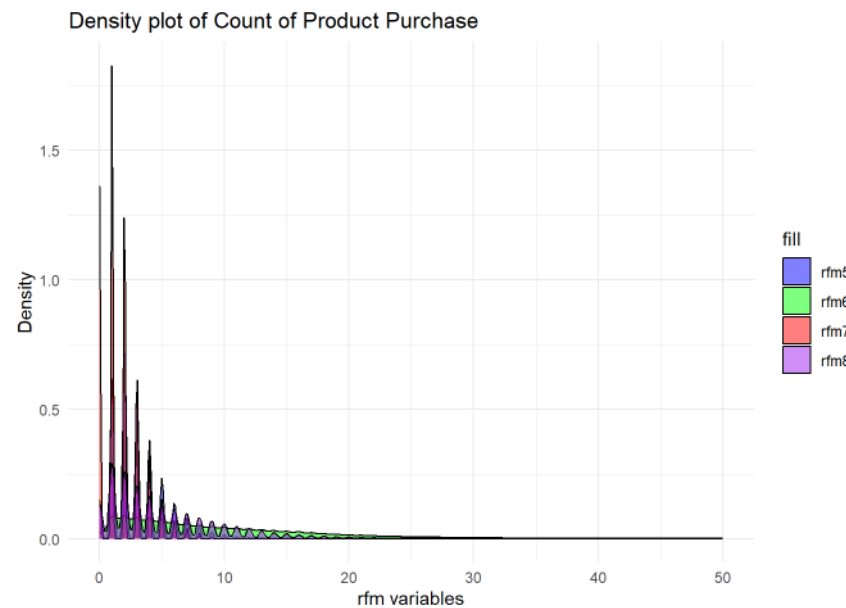
Variable	High Outliers	Low Outliers
int_tgt	207,593	-
cnt_tgt	211,509	-
demog_age	-	5,329
demog_homeval	73,306	-
demog_inc	8,470	-
demog_pr	7,858	31,245
rfm1	22,992	-
rfm2	29,359	-
rfm3	24,256	-
rfm4	30,067	-
rfm5	4,702	-
rfm6	21,343	-
rfm7	58,447	-
rfm8	21,247	-
rfm9	4	32,026
rfm10	75,991	21,319
rfm11	16,726	23,987
rfm12	66	-

I don't think so…

Data Understanding – Key Findings

3. Outliers / Non-Linearities / Distributions





Certain models struggle *more* with distributions like this

Data Preparation – Feature Engineering

Variable Interaction	Code
Count of Purchases Lifetime	rfm6^2
Lifetime Sales * Income	rfm2*demog_inc2
Income Squared	demog_Inc^2
Income * Home value	demog_Inc_Homeval

Non-linearities captured across demographics and Sales measures

California Lutheran
UNIVERSITY

Data Preparation – Feature Selection

Classification (B_TGT)

- 20 predictors (16 numeric , 4 categorical)

Total Sales (INT_TGT)

- 19 predictors (18 numeric , 1 categorical)

Number of Products (CNT_TGT)

- 21 predictors (17 numeric , 4 categorical)

Non-linearities captured across demographics and Sales measures

Modeling – ML selection

Classification (B_TGT) – 4 Fits

- Random Forest , Logisitic (Regularized) , Neural Net , Gradient Boosted Model

Total Sales (INT_TGT) - 5 Fits

- Random Forest , Logisitic (Regularized) , Neural Net , Generalized Additive Model (GAM) , Ensemble Model

Number of Products (CNT_TGT) - 4 Fits

- Random Forest , Logisitic (Regularized) , Neural Net , Generalized Additive Model (GAM)

Modeling – ML selection

Classification (B_TGT) – 4 Fits

- Random Forest , Logistic (Regularized) , Neural Net , Gradient Boosted Model

Total Sales (INT_TGT) - 5 Fits

- Random Forest , Logistic (Regularized) , , Neural Net , Generalized Additive Model (GAM) , Ensemble Model

Number of Products (CNT_TGT) - 4 Fits

- Random Forest , Logistic (Regularized) , Neural Net , Generalized Additive Model (GAM)

Capable of handling outliers/Skewness

✓ Random Forest
✓ Neural Net

Capture Non-Linearity well

✓ Random Forest
✓ Neural Net
✓ Gradient Boosted Method

Can Resist favoring Majority Class

✓ Neural Net

Offer Concrete & Clear Interpretation

✓ Logisitic Regression
✓ Generalized Additative Model

Why these?

Modeling – ML Tuning

H2o AI

- Offers comprehensive model training with extensive hyperparameter tuning capability
- Easy train/validation/test split
- Quick model comparisons

Modeling – Classification (B_TGT) results

Threshold used *0.377* *0.377* *0.377* *0.377*

Validation Data

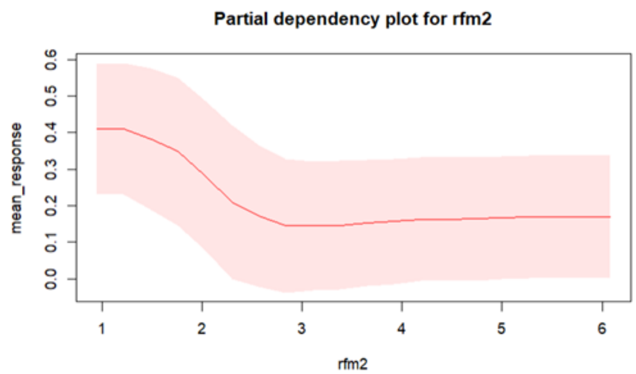
Classification Metrics	DRF	GLM	Neural Net	GBM
Sensitivity	0.932	0.665	0.670	0.928
Specificity	0.988	0.860	0.869	0.989
Precision	0.952	0.538	0.557	0.954
Accuracy	0.977	0.821	0.829	0.977
Recall	0.932	0.665	0.670	0.928
F1	0.942	0.595	0.608	0.941
Model Metrics	**DRF**	**GLM**	**Neural Net**	**GBM**
MSE	0.028	0.112	0.107	0.026
RMSE	0.168	0.335	0.328	0.162
LogLoss	0.120	0.364	0.346	0.113
AUC	0.995	0.852	0.864	0.994
Gini	0.990	0.704	0.728	0.988
Rsquare	0.823	0.293	0.823	0.834
Lambda		0.00001		
AIC		152500.20		

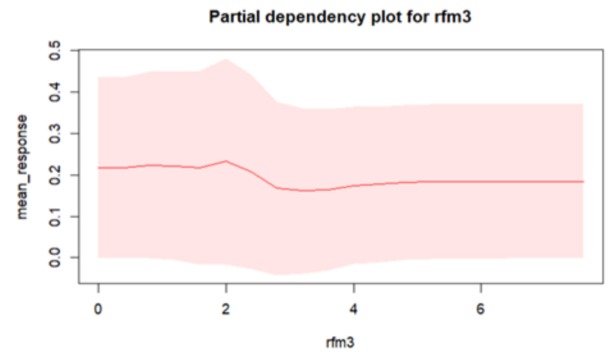
Test Data

Sensitivity	0.932
Specificity	0.989
Precision	0.954
Recall	0.932
F1 Score	0.943
Accuracy	0.978
AUC	0.995

Random Forest Selected for Test Performance

Modeling – Classification (B_TGT) results

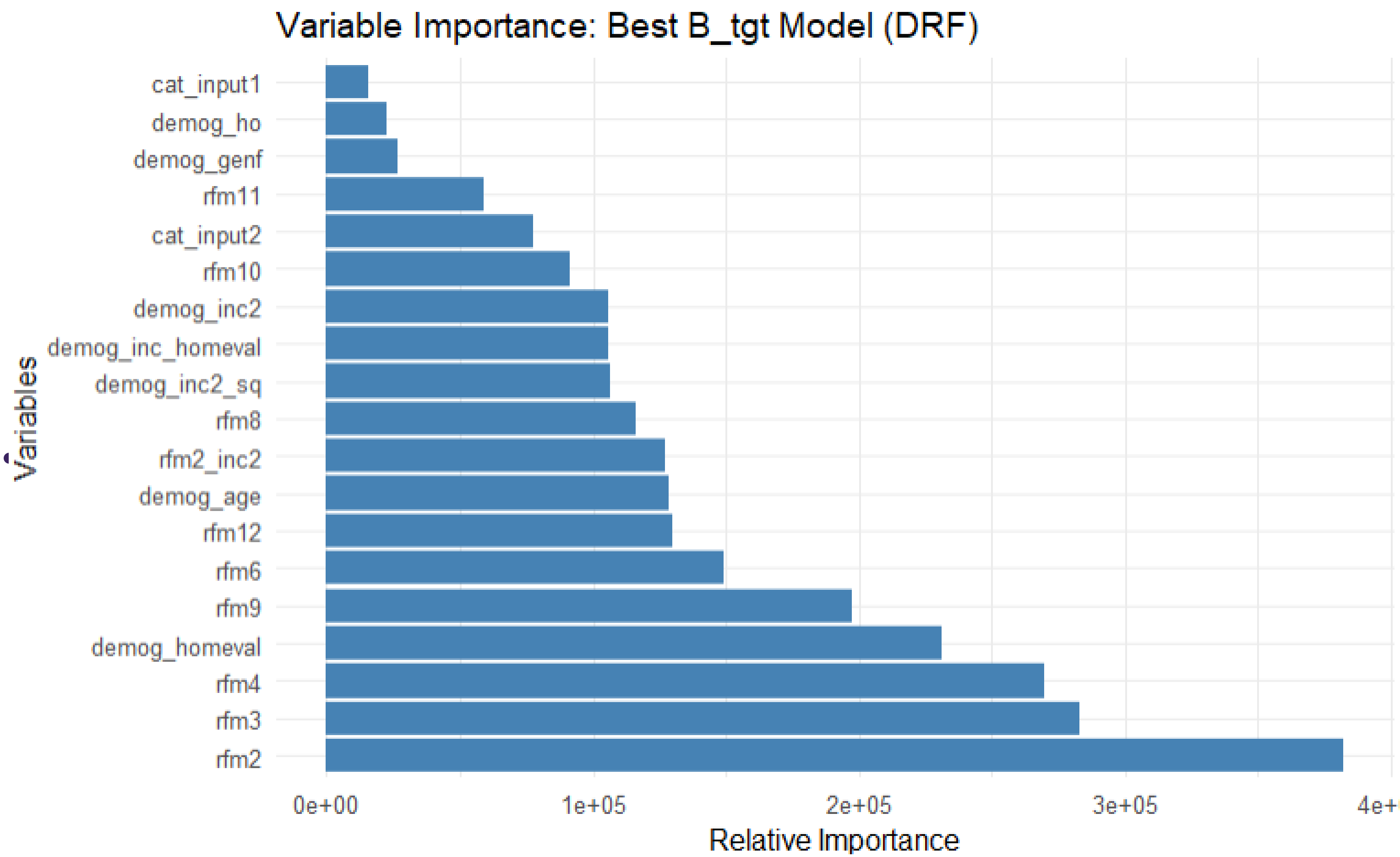




- When the Average Lifetime Sales and the Average 3Yr Sales from Dir Promos are increasing (or higher)… the customer is less likely to purchase a product.

PDP offers *indirect* method of Model interpretation

Modeling – Classification (B_TGT) results



Average Sales and Count of Sales most useful for training splits

Modeling – Prediction (INT_TGT) results

Validation Data

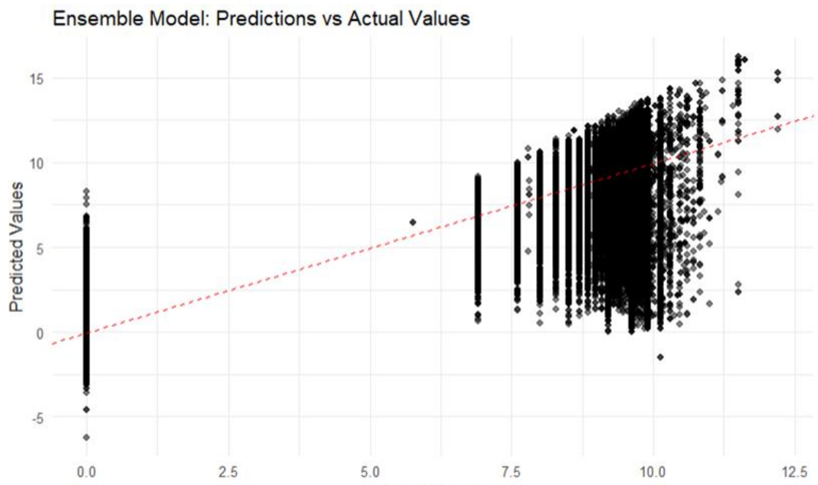
Model Metrics	DRF	Neural Net	GLM	GAM	Ensemble Method
MSE	2.115	9.055	10.211	10.170	1.520
RMSE	1.454	3.009	3.196	3.189	1.233
MAE	0.804	1.953	2.370	2.418	0.640
Mean Resid Deviance	2.115	9.055	10.211	10.170	1.520
Rsquare	0.838	0.302	0.220	0.223	0.884
AIC			1080662	1079853	
Lambda (Ridge Regression)			0.0001197	0.010	

MSE	1.521
RMSE	1.235
MAE	0.642
RMSLE	NaN
Mean Residual Deviance	1.525
Rsquared	0.884

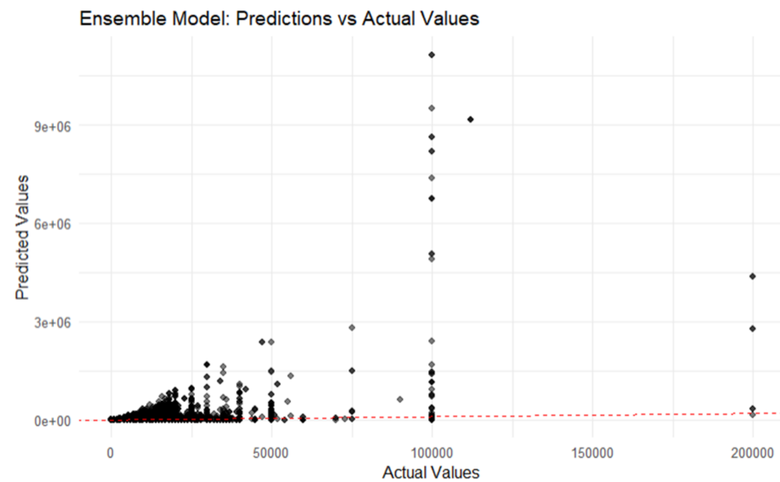
Ensemble Method used for Test Performance

Modeling – Prediction (INT_TGT) results

Response on Log Scale



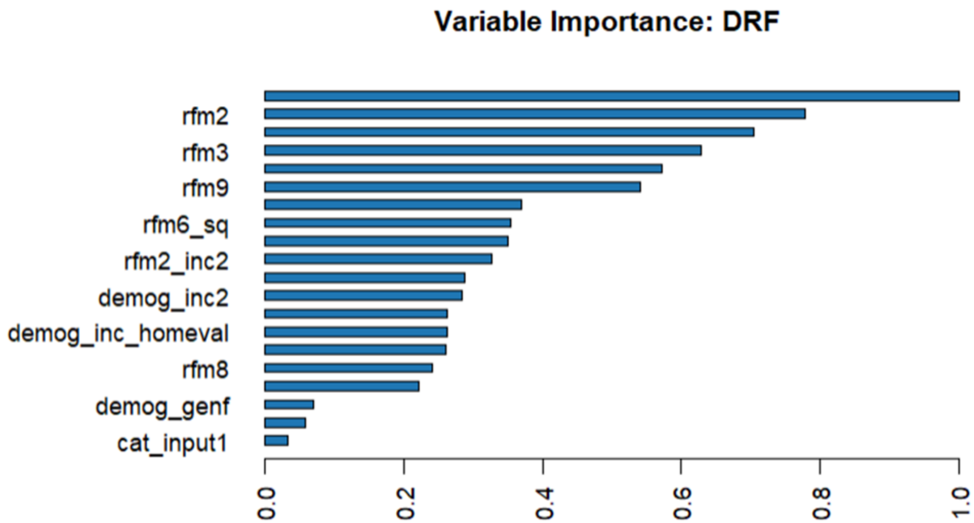
Response on Units Scale

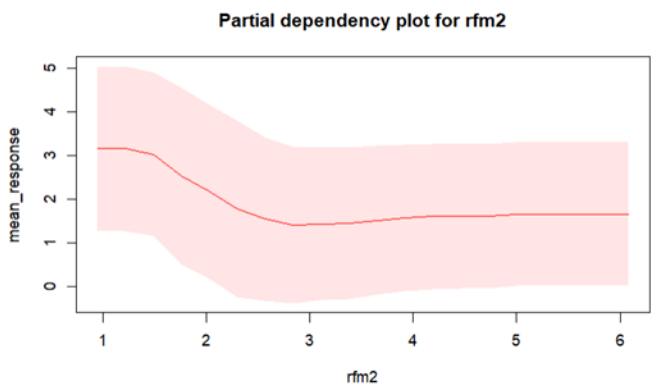


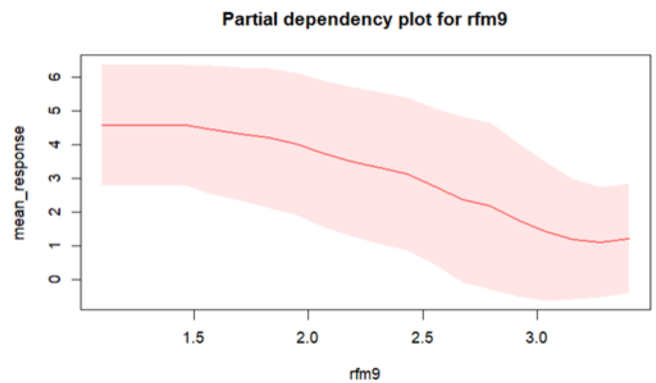
Interpretation of model changes

California Lutheran University

Modeling – Prediction (INT_TGT) results







Avg Sales hold similar relationship with New Customer & Total Sales

Modeling – Prediction (CNT_TGT) results

Validation Data

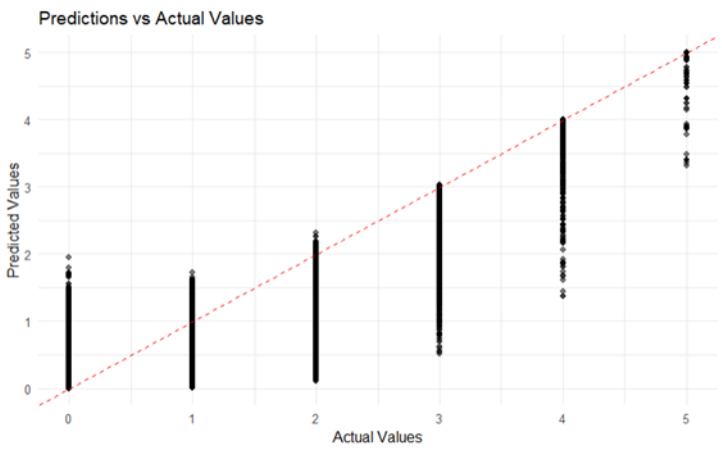
Model Metrics	DRF	Neural Net	GLM	GAM
MSE	0.058	0.301	0.350	0.352
RMSE	0.242	0.548	0.592	0.593
MAE	0.126	0.328	0.404	0.404
Mean Resid Deviance	0.058	0.301	0.350	0.352
Rsquare	0.879	0.378	0.276	0.272
AIC			374436	375805
Lambda (Ridge Regression)			0.00002767	

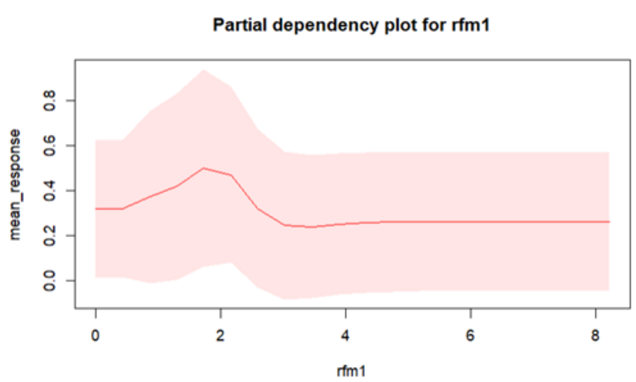
MSE	0.058
RMSE	0.241
MAE	0.126
RMSLE	0.141
Mean Residual Deviance	0.058

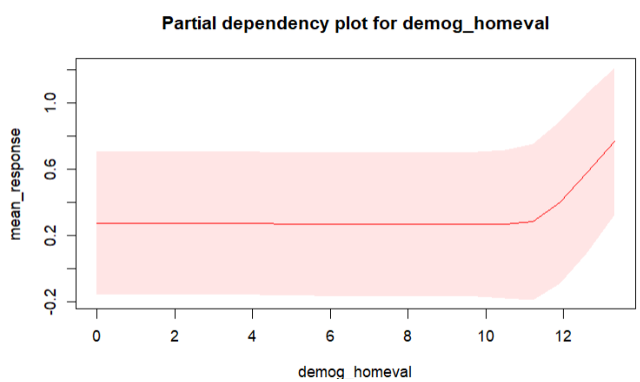
Random Forest used for Test Performance

California Lutheran UNIVERSITY

Modeling – Prediction (CNT_TGT) results







Takeaways - Summary

1. **All Models Generalize well**
   - Performance from Training ⇨ Validation ⇨ Test is consistent
2. **Average Sales and Number of Products sold are at the center of a smarter Marketing strategy**
   - VIPs across models show similar predictors near the top
3. **Model benefitted from Data preprocessing**
   - High R^2 metrics across classification and prediction confirm our predictor set is explaining variability in the response

Deployment Strategies

1. **Classification Model**
   - Can be deployed to immediately divide your account holder did into prospective customers/non-customers .
   - Can help business estimate (non-statistically) where sales stand for non-deposits product line.
2. **Total Sales Model**
   - Use in tandem with Classification model to develop financial forecast of sales if we successfully prospective customers
   - Focus Sales estimation for customers who generate lower sales – model is not reliable to enough to generate accurate Sales forecasts across entire customer base
3. **Count of Products Model**
   - Use in tandem with Classification model to develop tailored product offers to segments of Account holder based in relevant target range for Avg. Sales level specific and Home value.

Slide intentionally left blank

Appendix

Full predictor Set

B_tgt Predictor set	
Included	*Excluded*
RFM2 Average Sales Lifetime	RFM1 Average Sales Past Three Years
RFM3 Average Sales Past Three Years Dir Promo Resp	RFM5 Count Purchased Past 3 Years
RFM4 Last Product Purchase Amount	RFM7 Count Purchased Past 3 Years Dir Promo Resp
RFM6 Count Purchased Lifetime	DEMOG_GENM Male Binary (yes/no)
RFM8 Count Purchased Lifetime Dir Promo Resp	
RFM9 Months Since Last Purchase	
RFM10 Count Total Promos Past Year	
RFM11 Count Direct Promos Past Year	
RFM12 Customer Tenure	
DEMOG_AGE Customer Age	
DEMOG_GENF Female Binary (yes/no)	
DEMOG_HO Homeowner Binary (yes/no)	
DEMOG_HOMEVAL Home Value	
DEMOG_INC2 Income	
DEMOG_PR Percentage retired in the area	
CAT_INPUT1 Account Activity Level	
CAT_INPUT2 Customer Value Level	
DEMOG_INC2_sq Income Squared	
rfm2_inc2 Sales * Income interaction	
DEMOG_INC_HomeVal Income Homevalue interaction	

Appendix

Full predictor Set

CNT_tgt Predictor set
Included
RFM1 Average Sales Past Three Years
RFM2 Average Sales Lifetime
RFM3 Average Sales Past Three Years Dir Promo Resp
RFM4 Last Product Purchase Amount
RFM6 Count Purchased Lifetime
RFM8 Count Purchased Lifetime Dir Promo Resp
RFM9 Months Since Last Purchase
RFM10 Count Total Promos Past Year
RFM11 Count Direct Promos Past Year
RFM12 Customer Tenure
DEMOG_AGE Customer Age
DEMOG_GENF Female Binary (yes/no)
DEMOG_HO Homeowner Binary (yes/no)
DEMOG_HOMEVAL Home Value
DEMOG_INC2 Income
DEMOG_PR Percentage retired in the area
CAT_INPUT1 Account Activity Level
CAT_INPUT2 Customer Value Level
DEMOG_INC2_sq Income Squared
rfm2_inc2 Sales * Income interaction
DEMOG_INC_HomeVal Income Homevalue interaction
Excluded
RFM5 Count Purchased Past 3 Years
RFM7 Count Purchased Past 3 Years Dir Promo Resp
DEMOG_GENM Male Binary (yes/no)

INT_tgt Predictor set
Included
RFM1 Average Sales Past Three Years
RFM2 Average Sales Lifetime
RFM3 Average Sales Past Three Years Dir Promo Resp
RFM4 Last Product Purchase Amount
RFM6 Count Purchased Lifetime
RFM6^2 Count Purchased Lifetime Squared
RFM8 Count Purchased Lifetime Dir Promo Resp
RFM9 Months Since Last Purchase
RFM10 Count Total Promos Past Year
RFM11 Count Direct Promos Past Year
RFM12 Customer Tenure
DEMOG_AGE Customer Age
DEMOG_HO Homeowner Binary (yes/no)
DEMOG_HOMEVAL Home Value
DEMOG_INC2 Income
DEMOG_PR Percentage retired in the area
DEMOG_INC2_sq Income Squared
rfm2_inc2 Sales * Income interaction
DEMOG_INC_HomeVal Income Homevalue interaction
Excluded
RFM5 Count Purchased Past 3 Years
RFM7 Count Purchased Past 3 Years Dir Promo Resp
DEMOG_GENM Male Binary (yes/no)
CAT_INPUT1 Account Activity Level
CAT_INPUT2 Customer Value Level
DEMOG_GENF Female Binary (yes/no)