California Lutheran University

# Final Project

ECON-562: Advanced Analytics

Andrew Frohner
5-18-2025

**Predictive Modeling: Bank Data**

# Introduction:

An effective direct marketing campaign will help your business increase their revenues. It is not clear exactly how effective the current campaign is at generating revenue, but insight into what factors drive our sales are especially useful for developing new marketing strategies.

Our analysis uses account holder data from your firm to develop models for answering the following questions:

- What drives customers to buy new products?
- What drives the total sales?
- What contributes to the number of products the customer buys?

The analysis will include full description of the data set, data preparation steps, feature engineering, modeling output, evaluation, and performance.

## 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 custome | 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 |

The total data set contained 26 variables (23 predictors and 3 Target variables). The data was either categorical or numerical. Most of the observations were technically “incomplete.” Meaning that an observation contained a missing value in at least 1 of the 26 variables in our data set. This issue was confined to 4 of the features in the data (shown on next page).

| 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 |

The target variable for $ amount of new sales had over 80% missing values. Rather than remove these observations entirely, I chose to impute them for completion. Inspecting the data closer, these observations coincided with observations where the count of products purchased (CNT_tgt) was equal to zero. Missing values of sales Interval were imputed with 0 given the information that the customer had purchased zero products. Count of products purchased was also imputed. Initially imputed as the mean of CNT_tgt when the observation is male and their home value is less than $100K, but this resulted in a non-whole number of .244 – I chose to round down to 0 for this individual observation.

Customer age used a two-stage imputation method. The first stage used linear regression to predict a customer's age (formula below):

$$Demog_{Age} = \beta_0 + \beta_1 HomeValue_i + \beta_2 Income_i + \beta_3 PercentRetired_i + \beta_4 HomeOwner_i + \beta_5 rfm12_i + \beta_6 rfm1_i + \varepsilon_t$$

This successfully imputed 266,841 values, 20 observations were unsuccessfully imputed as *negative* values. A second step was necessary make these data points valid. The 20 negative observations were imputed as the 25% quantile value for customer age (51 for this data set).

Avg Sales past 3 years from Direct Promo (rfm3) was also imputed using linear regression:

$$rfm3 = \beta_0 + \beta_1 rfm1_i + \beta_2 rfm2_i + \beta_3 rfm4_i + \beta_2 rfm5_i + \beta_6 rfm7_i + \beta_7 rfm8_i + \varepsilon_t$$

A second step was required for negative observations. The negative values were imputed as the 5% quantile value for rfm3.

Incomplete observations were not the only issue in this data set. Erroneous values (data points that appear unrealistic or mistaken) also were present.

INT_tgt contained 55 observations over $100K. Of this subset, 11 observations were $500K. The other predictors (demographics and sales) had low variability when INT_tgt = $500K. This was a signal that these observations were erroneous. Erring on the side of preservation, these 11 observations were imputed as:

$$when\ INT_{tgt} = \$500K\ impute\ as: Mean\left(INT_{tgt}\right)\ when\ Income\ between\ \$60K\ and\ \$70K$$

The other 44 observations were INT_tgt was over $100K appeared to be real observations.

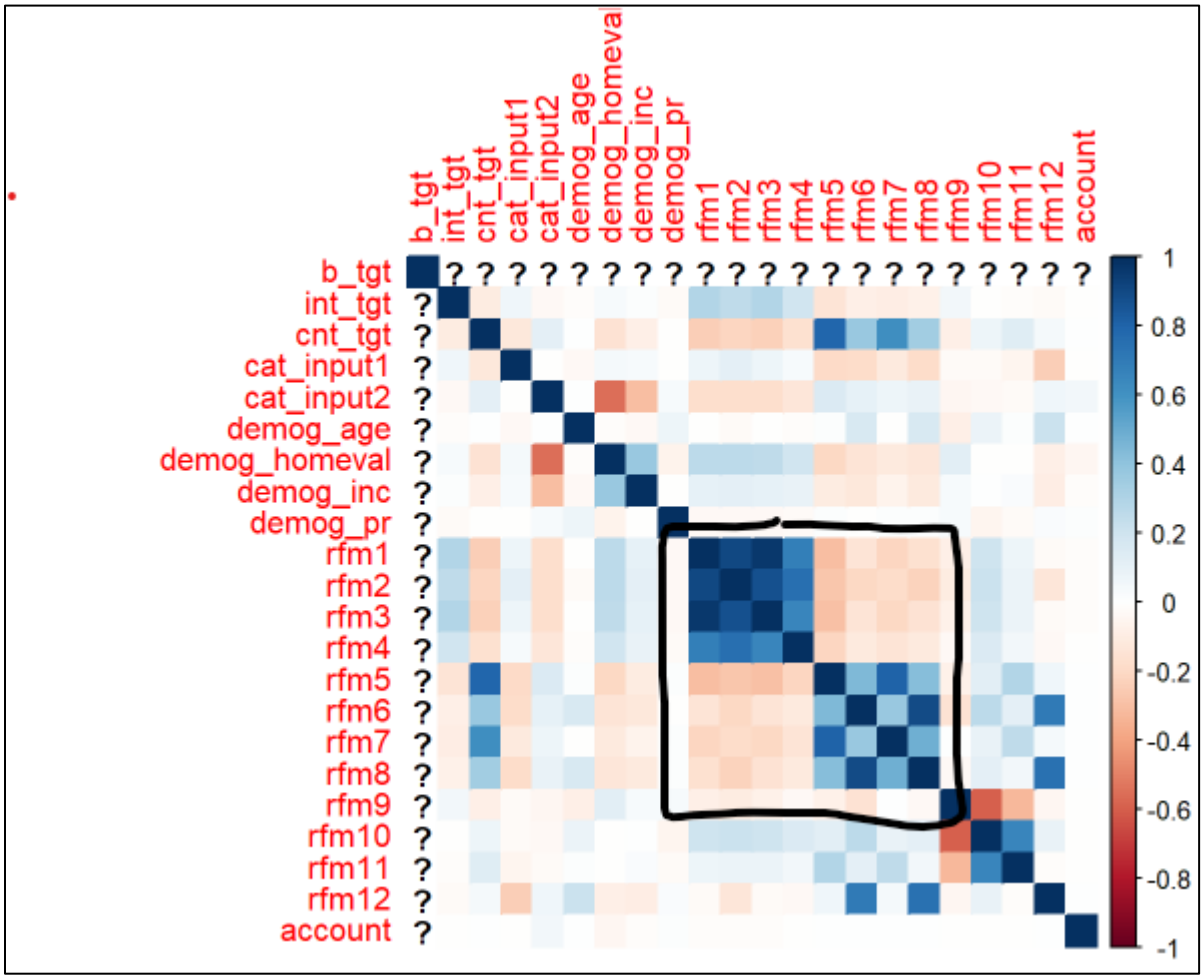
The investigation into erroneous values was conducted for other variables in the data set, and imputed data, rather than discarded it. Full detail of impacted variables in shown in the table on the next page.

| 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 |

Given the small quantity of observations impacted, mean imputation did not affect the distribution of these predictors.

Additionally – an important exclusion was made here. All observations where the customer's age was under 21 were excluded from this data set (total of 12,066 observations). In general, younger customers are sold products that banks produce (equity lines of credit, auto loans). Most of this subset (10,627 observations) were under the age of 18.

## Exploration Analysis




The Correlation matrix of our complete predictor set *suggest* there may be some room for dimension reduction. As Average Sales variables (rfm1- rfm3) are positively correlated with each other, and so are Count of Purchase variables (rfm5-rfm8). I am not suggesting that we reduce dimensions so far as to only include one of each, rather perhaps *exclude* one of each. The Sales and Count variables measure short term and long term consumer behavior, and there is naturally some correlation between the two, but they are separate patterns, and both should be considered for prediction.

Another finding during exploration was class imbalance of account holders who did or did not buy new products (B_tgt). I considered synthetic sampling the minority class for the B_tgt model, but decided to forego maneuver and realization that this imbalance could seriously alter my coefficients estimates. Particularly for this model, where there is a cost to marketing towards customers who potentially do not buy the products. 60% increase is too risky, and a 60% decrease (via under sampling majority class) leaves out too much.

| 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% |

| 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 | - |

Outliers were defined as: any value 2x higher/lower than the 75% and the 25% quartiles respectively. The Table detailing outlier count indicates that the predictor set tends to skew more right than left.

I chose to forego addressing the complete set of outliers in favor of singling out observations that appeared erroneous, removing data, and employing modeling techniques that can handle outlier data.

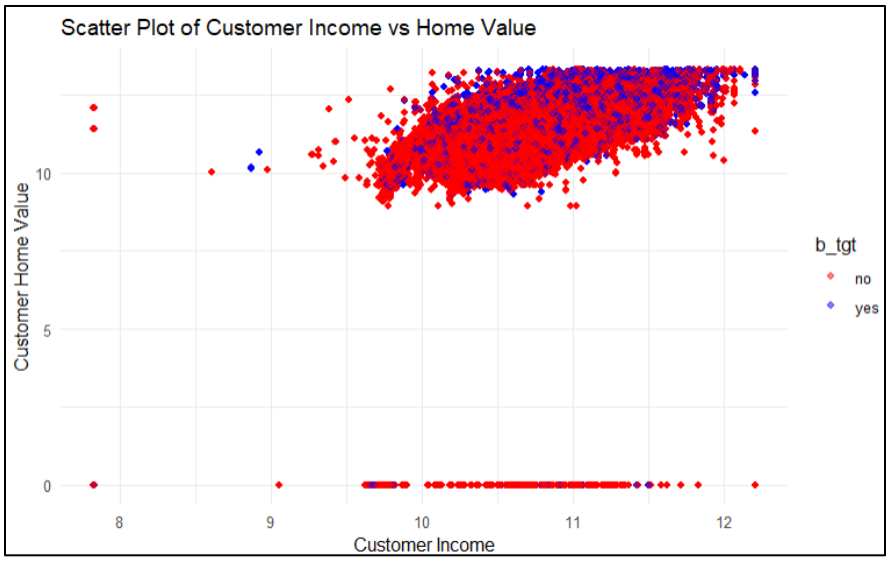
| Variable | Skewness | Transformation Applied | Skewness post Transformation |
|---|---|---|---|
| int_tgt | 4.84 | Log | 1.57 |
| cnt_tgt | 2.40 | | 2.40 |
| demog_age | -0.12 | Log | -0.77 |
| demog_homeval | 2.46 | Log | -5.99 |
| demog_inc | 0.23 | Log | -1.21 |
| ~~demog_pr~~ | ~~-0.15~~ | - | - |
| 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 |

The distribution of the variables was addressed through log transformations. At this time in the data processing stage, the data set had been free of negative values, and erroneous values. Some of the data was more skewed than others, but Log transformation was applied to all numerical predictors for consistency (with some exclusions).

CNT_tgt was not logged. While it is technically numeric, it is representative of a count. Log transformation would distort the

interpretation of this target variable and potentially mis specify the model. Percent Retired in area (Demog_pr) was not transformed to preserve interpretation.

## Feature Engineering




The Scatter plot shows inseparability between classes when plotting income vs home value. This amount of overlap between classes is not unexpected for a data set this size, but it is enough to rule out some algorithms for prediction. Most notably KNN, LDA, and QDA.

A KNN model may struggle even if we choose a large K due to the overlap being dense throughout the relationship of these two predictors. It would also be computationally intense with 20 predictors at over 600K training observations.

LDA and QDA would struggle with the large predictor set as well. Additionally, those models would not be optimal for the log transformations applied to the data. All the predictors are on the same scale, but they do not carry the same mean-variance structure that these methods assume.

I am holding favor for GLM, Random Forest, Neural Network, and Gradient Boosted algorithms based on their ability to handle non-linearity, more relaxed assumptions regarding predictor distributions, robustness to outliers, and interpretability.

To address the non-linearity that I suspect is present in the relationships under examination. Certain features were created. The engineered features incorporate non-linearity in the areas of Sales, Product Purchase counts, and demographics. Only four features were created out of parsimony.

| 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 |

**Feature Selection: B_TGT**

| 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 | |

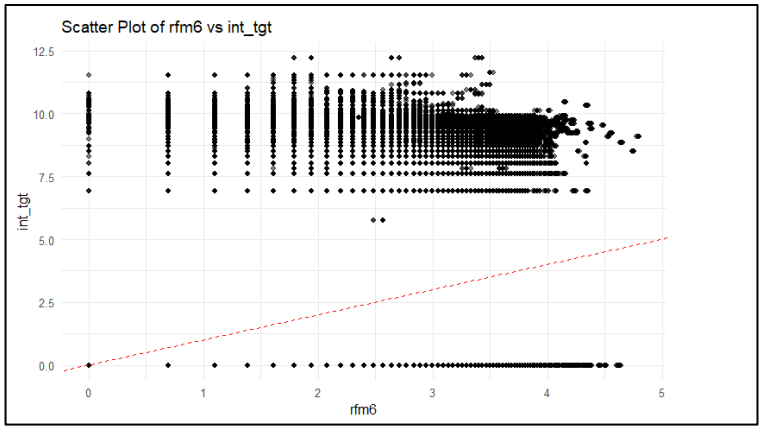
The B_TGT model included twenty predictors and excluded four. The exclusions are additional dimensions that already occupy space in the model. The interactions and squared terms are included to better classify the linearly inseparable data.

## **Feature Selection:** INT_TGT & CNT_TGT

The predictor sets for the prediction models incorporate RFM1 variable (Avg Sales past 3 years). I chose to include these in the prediction models because they incorporate a different pattern in customer behavior that I think is more useful when the outcome is non-binary. They likely exhibit some multi-collinearity but including both can help us characterize customers which customers are buying more products and more/less expensive products. RFM6^2 was included in the INT_tgt regression because it appeared to hold non-linear relationship with the target. Table breakouts are shown on the next page for the two models.

| 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) |




As the number of lifetime purchases increases (rfm6) we see the clustering of Int_tgt change in density. Suggestive that Int_tgt does not purely increase as rfm6 increases.

The feature selection process did not provide as much room as I initially thought was possible after initial EDA.

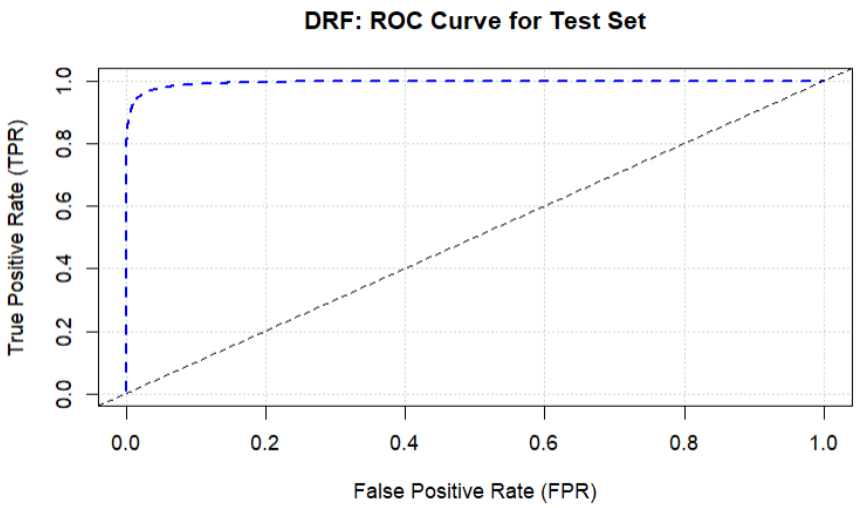
**Model Results: B_TGT** (Green Denotes 1st ranked Candidate / Yellow denotes 2nd Ranked Candidate)

**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 Performance:** Candidate DRF

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




The B_TGT prediction was made most accurately by the Random Forest model. The relative importance of predictor set for this model was led by:

RFM2 Average Sales Lifetime
RFM3 Average Sales Past Three Years Dir Promo Resp
RFM4 Last Product Purchase Amount
DEMOG_HOMEVAL Home Value
RFM9 Months Since Last Purchase

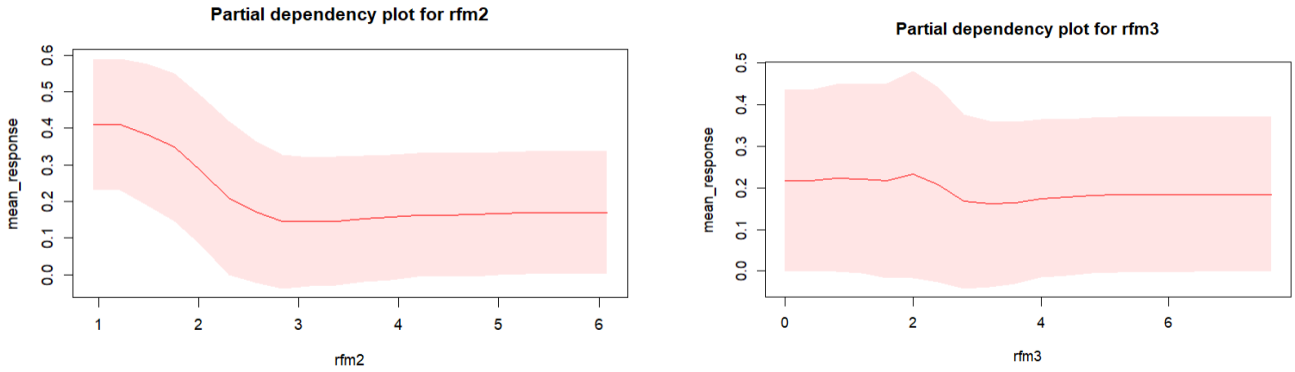
The probability threshold used for classifying a customer as a buyer of a new product was 37%.

The algorithm used for this model tuned this threshold to maximize the balance of sensitivity and precision. This model's top 2 priorities were:

1. Correctly assigning "buyer" to true buyers.
2. Minimizing prediction of "buyer" when the customer is not buying.

The cost structure of your firm is not explicit, but this model is assuming that there is a greater cost to missing customers who buy products than marketing to customers who are not buying products.

Though a DRF is not directly interpretable, we can gain some insight into how the response variable reacts to a change in the predictor through analysis of Partial Dependence Plots.




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.

Model Takeaways/Notes:

- This model exhibits strength in predicting both classes (Sensitivity and Specificity metrics in 90% range).
- Gini Coefficient and ROC Curve indicate that this model discriminates far more reliably than pure chance.
- The model maintained its accuracy on the Test data set.

**Model Results: INT_TGT** (Green Denotes 1st ranked Candidate / Yellow denotes 2nd Ranked Candidate)

**Validation Data**

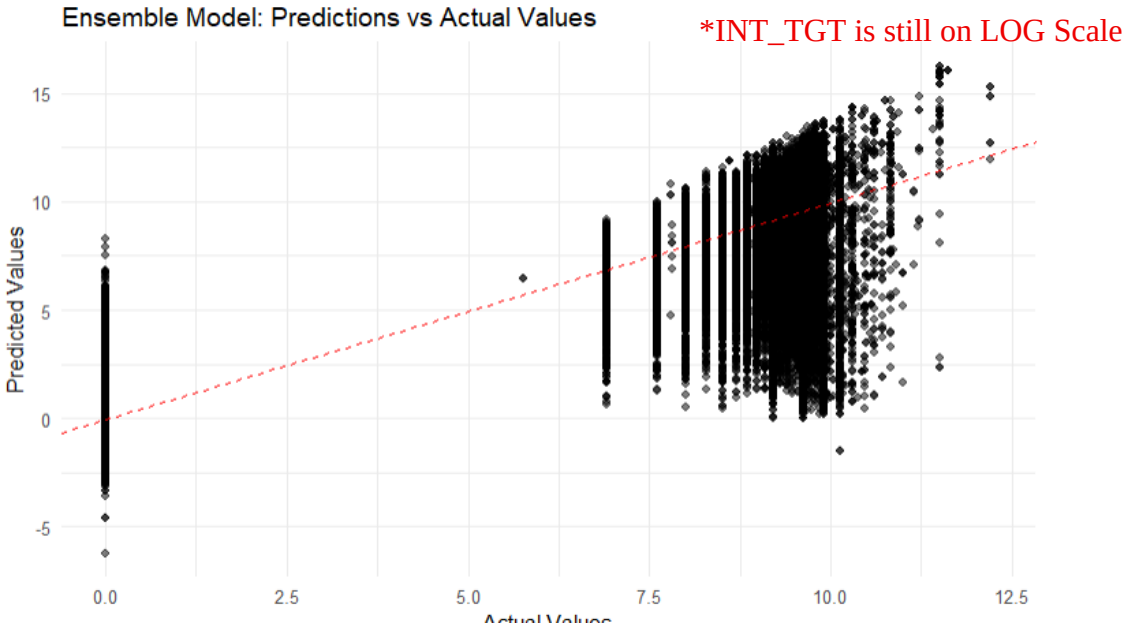
| Model Metrics | DRF | Neural Net | GLM | GAM | Ensemble Method |
|---|---|---|---|---|---|
| MSE | 2.115 | 9.055 | 10.211 | 10.170 | 1.520 |
| RMSE | 1.454 | 3.009 | 3.196 | 3.189 | 1.233 |
| MAE | 0.804 | 1.953 | 2.370 | 2.418 | 0.640 |
| Mean Resid Deviance | 2.115 | 9.055 | 10.211 | 10.170 | 1.520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

**Test Data Performance:** Candidate Ensemble

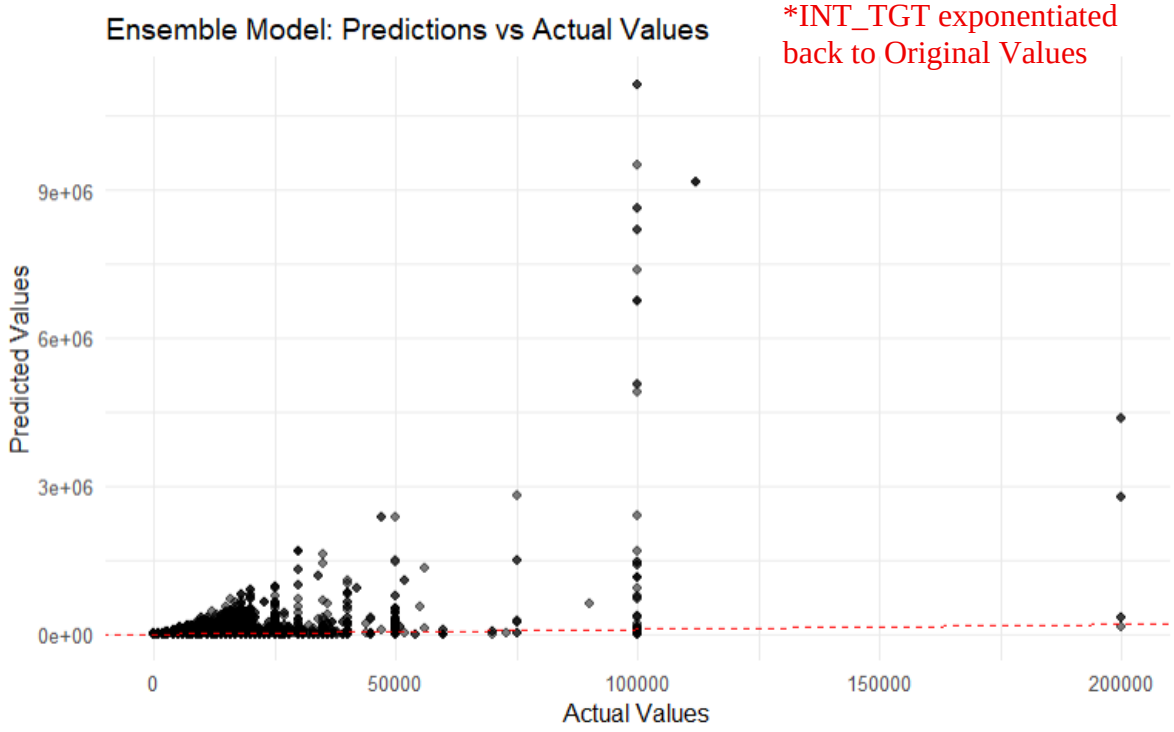
| | |
|---|---|
| MSE | 1.521 |
| RMSE | 1.235 |
| MAE | 0.642 |
| RMSLE | NaN |
| Mean Residual Deviance | 1.525 |
| Rsquared | 0.884 |

An additional model was fit to improve performance of prediction. The validation metrics above do not show strong performance for prediction of Total Sales. The ensemble model was fit using the top 5 performing models on the validation data from the automation algorithm. It did not consider the GAM as it was trained outside the automation loop and performed worse than the top 5 models.

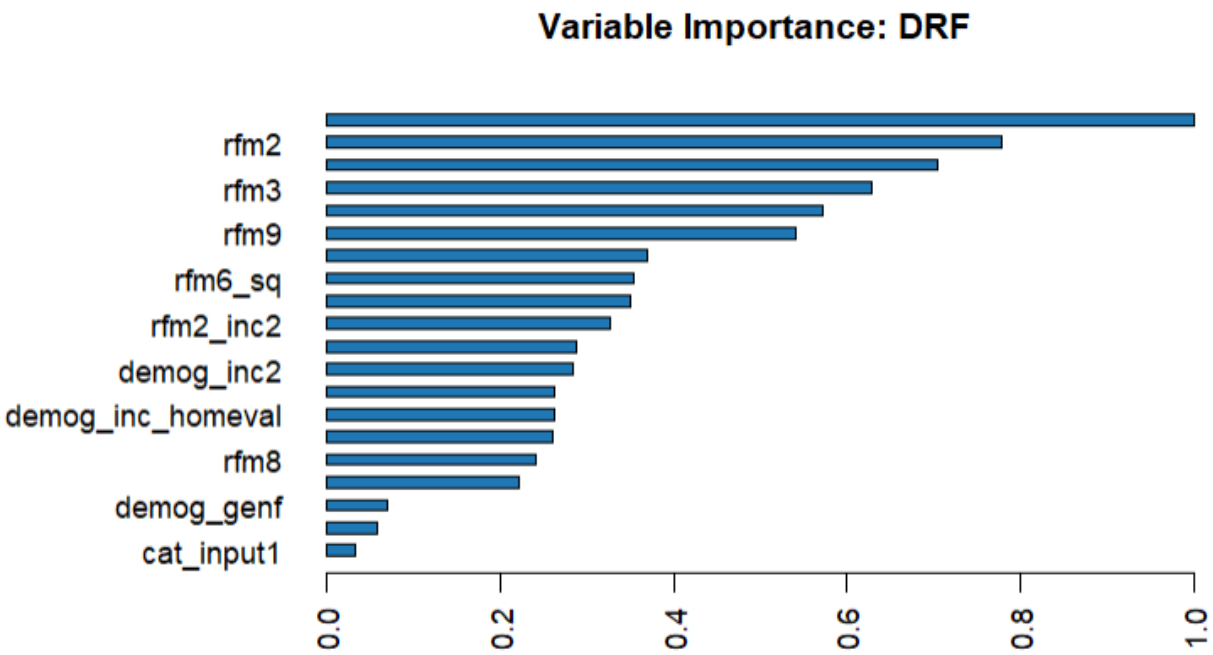
The Ensemble model includes 2 Random Forests, 2 Neural Networks, and 1 GLM. The actual vs predicteds are shown in their logged form (as they were trained, validated, and tested on), and in exponentiated form (to return INT_tgt back to its original value for interpretation purposes).



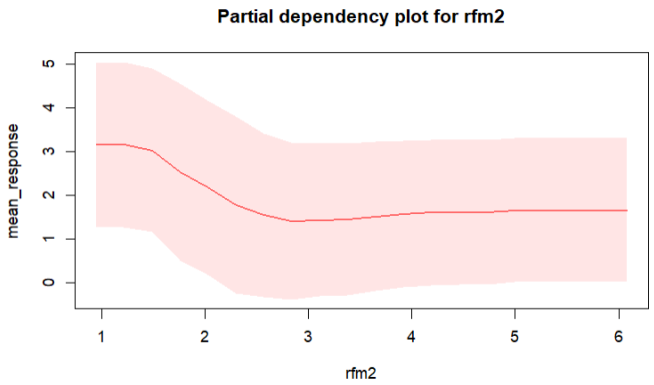
This pattern seems ok. The model is predicting smaller value better, and underpredicts larger values. However, the interpretation flips when we return the Total sales back to their original values. Taken off the log scale we see the ensemble model starts to drastically overshoot the sales value for customers starting around the $23K range.

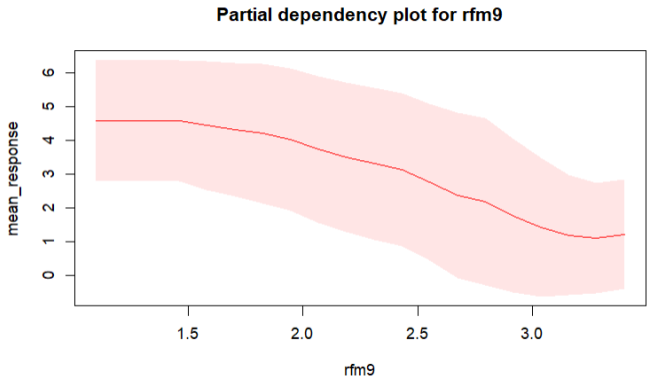



For some cases the model is over 9X in its prediction (Predicting sales to be $900K when they're actually $100K).




The Stacked ensemble does not naturally have variable importance, so I am showing the metrics from the top performing Random Forest (DRF). Many of the features play a role in prediction for Total Sales. The engineered features were among the top five most important predictors in the model. A more detailed list of interpretable coefficients from the top performing GLM model and the stacked ensemble are available in the appendix. Brief interpretation is given by PDP (rfm2 & rfm9):



Like the prediction of new customer, increases in Avg Lifetime Sales (rfm2) appear to lower the value of total sales generated by a customer. And interestingly, the time since the last purchase has the same directional effect (with a higher magnitude).

Model Takeaways/Notes:

- This model has predictive power for all levels of Sales
- This model is unreliable at predicting higher value customers
- Deploy this model strategically, this model is missing a pattern in the data that hurts its predictive power.
- Direct the marketing campaign towards customers who have lower lifetime sales, and avoid sending to customers who have not purchased in a while.

**Model Results:** **CNT_TGT** (Green Denotes 1st ranked Candidate / Yellow denotes 2nd Ranked Candidate)

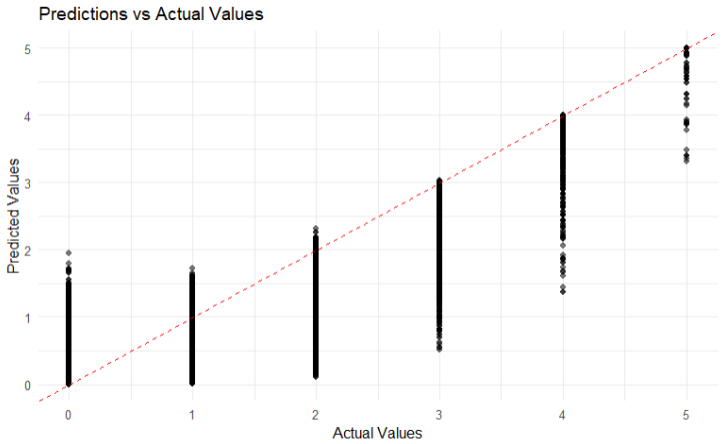
**Validation Data**

| 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 | |

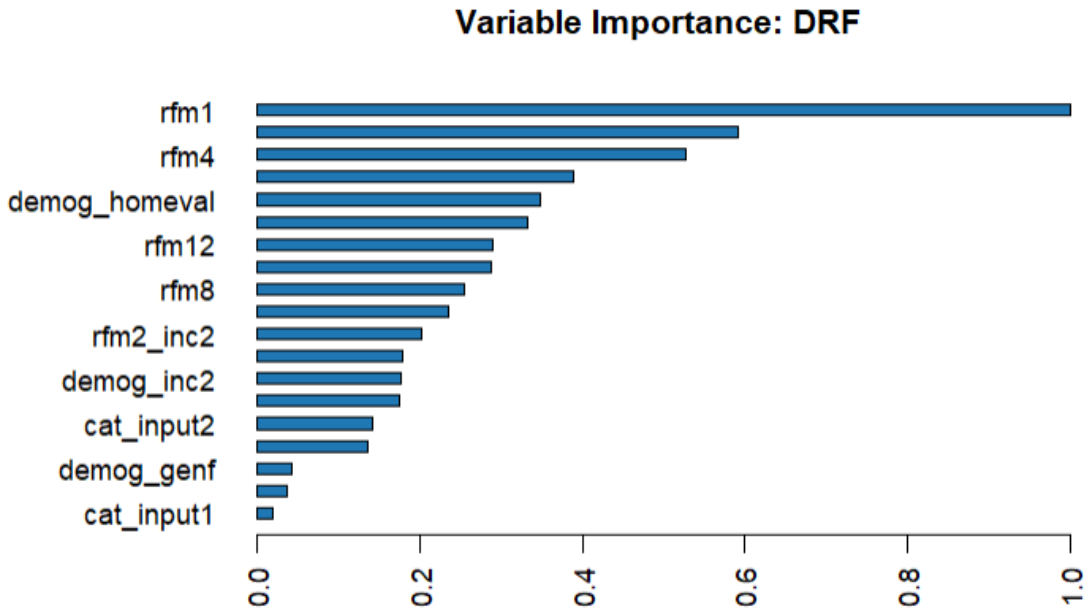
---

**Test Data Performance:** Candidate DRF

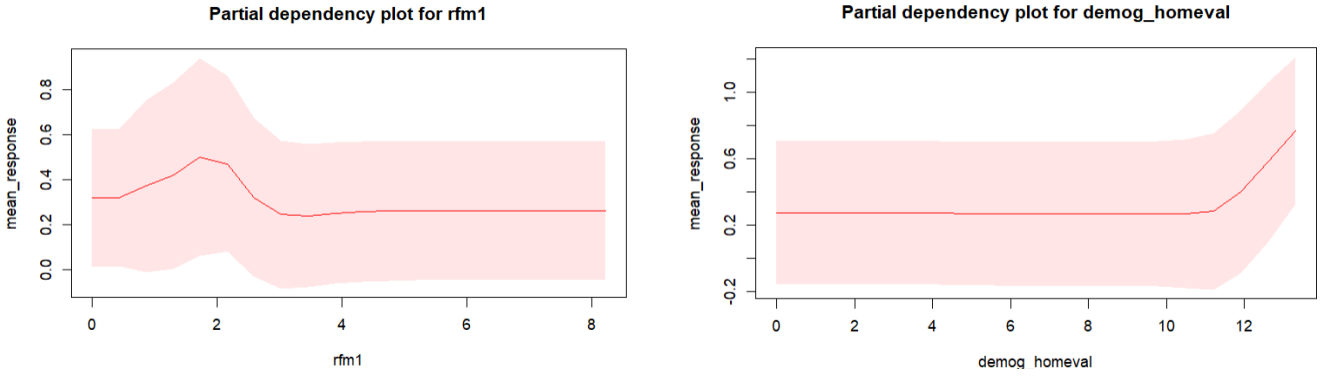
| | |
|---|---|
| MSE | 0.058 |
| RMSE | 0.241 |
| MAE | 0.126 |
| RMSLE | 0.141 |
| Mean Residual Deviance | 0.058 |



The CNT_TGT prediction was made most accurately by the Random Forest model. The relative importance of predictor set for this model was led by:




Average 3Yr Sales (rfm1) and Home value (demog_homeval) examined for further understanding below:




The Partial plots reveal interesting dynamics regarding how the count of products purchased reacts to Sales and homeval, Avg Sales appear to, ultimately, have no impact on response at the low/high ends. But in the lower/middle tier there is a “sweet spot” where This variable is meaningful to increasing purchase count.

The count of products does not react much at all for many of the customers regardless of their home is cheap or expensive. There is just one exception, customers living in the most expensive homes appear to move the count of purchases up almost 100% in magnitude (from .3 to .6).

Modeling Takeaways/Notes:

- RMSE indicates predictions are off my about .24 units of the actual count of products a customer purchases.
- All algorithms predicted within 1 unit of the actual count
- The DRF model most strongly predicts customers who purchase 1 product
- The DRF model consistently underestimates customers who order 2 or more products.
- The average of recent sales (rfm1), last product purchase amount, and homeownership are the most useful indicators for predicting the count of products purchased.

**Summary & Conclusion:**

The account holder data was best modeled by the Random Forest algorithm. most effective at handling data that contained nonlinearities, outliers, and skewed distributions.

When working with data such as this, it is important to remember to spend time thinking about ways make the data more palatable for machine learning. Imputation (where necessary), transformations, and feature engineering on the front end can help deliver insights that are not obvious upon first look.

When trying to entice a customer towards buying a product it is helpful to keep in mind higher average sales indicate a lower likelihood of acquiring a new customer, and a cheaper one as well.

When trying to characterize customers who may be interested in multiple products, target the ones might have a high value home.

Models for predicting customers and the count of products can be reliably deployed for future marketing strategies. The Model for predicting sales amount (and potentially profitability for short/medium term business finance objectives), should be deployed with caution. And I would advise against using it for profitability measurement at this moment in time. It is overly optimistic in its predictions.

## References

GitHub. (2025). GitHub Copilot: Your AI pair programmer. Retrieved from https://github.com/features/copilot

California Lutheran University. (2025, May 15th ). School of Management Introduces AI Tutors to Students. Retrieved from Advanced Analytics AI Tutor: ECON-562-1/MBA-580-01 - 25/SP/A

# Appendix

## R Markdown file for Modeling and EDA

ECON_562_FINALPROJ_mod_AF_vFinal.R

ECON_562_FINALPROJ_AF_V4.html

## INT_TGT model coefficients

GLM Model Coefficients

| variable | Coefficient |
|---|---|
| rfm2 | 17.48711583 |
| rfm2_inc2 | -17.30835907 |
| demog_inc2 | 12.84937141 |
| Intercept | -4.691321633 |
| demog_inc2_sq | 2.892043665 |
| rfm9 | -2.546512296 |
| rfm6 | -1.31556689 |
| rfm6_sq | 1.019364677 |
| rfm12 | -0.992174883 |
| rfm3 | -0.777493912 |
| rfm1 | 0.313137596 |
| demog_homeval | 0.258746187 |
| rfm8 | 0.248902968 |
| rfm4 | -0.203650242 |
| rfm10 | -0.045508421 |
| demog_inc_homeval | -0.008159624 |
| demog_age | 0.006146897 |
| demog_ho.yes | -0.001213708 |
| demog_ho.no | 0.000680984 |
| demog_pr | 8.07881447802002e-45 |

Stacked Ensemble Coefficients

| variable | coefficient |
|---|---|
| DRF_1_AutoML_6_20250518_35511 | 0.909682065 |
| XRT_1_AutoML_6_20250518_35511 | 0.510607542 |
| DeepLearning_1_AutoML_6_20250518_35511 | -0.341121201 |
| DeepLearning_grid_1_AutoML_6_20250518_35511_model_1 | -0.043050921 |
| Intercept | -0.033395814 |
| GLM_1_AutoML_6_20250518_35511 | -0.029069098 |

## Full Data Set used for Training ,, Validation, Test and Splits

| | vars | n | mean | sd | median | min | max | range |
|---|---|---|---|---|---|---|---|---|
| b_tgt | 1 | 1047972 | NaN | *NA* | *NA* | Inf | -Inf | -Inf |
| int_tgt | 2 | 1047972 | 1.78E+00 | 3.63E+00 | 0.00E+00 | 0.00E+00 | 1.22E+01 | 1.22E+01 |
| cnt_tgt | 3 | 1047972 | 3.10E-01 | 7.00E-01 | 0.00E+00 | 0.00E+00 | 5.00E+00 | 5.00E+00 |
| cat_input1 | 4 | 1047972 | NaN | *NA* | *NA* | Inf | -Inf | -Inf |
| cat_input2 | 5 | 1047972 | NaN | *NA* | *NA* | Inf | -Inf | -Inf |
| demog_age | 6 | 1047972 | 4.07E+00 | 2.50E-01 | 4.08E+00 | 3.09E+00 | 4.50E+00 | 1.41E+00 |
| demog_ho | 7 | 1047972 | NaN | *NA* | *NA* | Inf | -Inf | -Inf |
| demog_homeval | 8 | 1047972 | 1.12E+01 | 1.38E+00 | 1.12E+01 | 0.00E+00 | 1.33E+01 | 1.33E+01 |
| demog_inc | 9 | 1047972 | 8.23E+00 | 4.63E+00 | 1.07E+01 | 0.00E+00 | 1.22E+01 | 1.22E+01 |
| demog_pr | 10 | 1047972 | 1.52E+39 | 1.71E+41 | 2.90E+13 | 0.00E+00 | 7.31E+43 | 7.31E+43 |
| rfm1 | 11 | 1047972 | 2.69E+00 | 5.70E-01 | 2.77E+00 | 0.00E+00 | 8.22E+00 | 8.22E+00 |
| rfm2 | 12 | 1047972 | 2.55E+00 | 4.70E-01 | 2.54E+00 | 9.50E-01 | 6.08E+00 | 5.13E+00 |
| rfm3 | 13 | 1047972 | 2.69E+00 | 5.20E-01 | 2.77E+00 | 0.00E+00 | 7.60E+00 | 7.60E+00 |
| rfm4 | 14 | 1047972 | 2.76E+00 | 5.40E-01 | 2.77E+00 | 0.00E+00 | 8.22E+00 | 8.22E+00 |
| rfm5 | 15 | 1047972 | 1.23E+00 | 5.20E-01 | 1.10E+00 | 0.00E+00 | 2.89E+00 | 2.89E+00 |
| rfm6 | 16 | 1047972 | 2.04E+00 | 8.10E-01 | 2.08E+00 | 0.00E+00 | 4.80E+00 | 4.80E+00 |
| rfm7 | 17 | 1047972 | 8.20E-01 | 5.70E-01 | 6.90E-01 | 0.00E+00 | 2.48E+00 | 2.48E+00 |
| rfm8 | 18 | 1047972 | 1.52E+00 | 7.60E-01 | 1.61E+00 | 0.00E+00 | 3.61E+00 | 3.61E+00 |
| rfm9 | 19 | 1047972 | 2.93E+00 | 2.60E-01 | 2.94E+00 | 1.10E+00 | 3.40E+00 | 2.30E+00 |
| rfm10 | 20 | 1047972 | 2.59E+00 | 2.70E-01 | 2.56E+00 | 0.00E+00 | 4.36E+00 | 4.36E+00 |
| rfm11 | 21 | 1047972 | 1.82E+00 | 2.30E-01 | 1.95E+00 | 0.00E+00 | 3.14E+00 | 3.14E+00 |
| rfm12 | 22 | 1047972 | 4.06E+00 | 6.30E-01 | 4.17E+00 | 0.00E+00 | 6.35E+00 | 6.35E+00 |
| demog_genf | 23 | 1047972 | NaN | *NA* | *NA* | Inf | -Inf | -Inf |
| demog_genm | 24 | 1047972 | NaN | *NA* | *NA* | Inf | -Inf | -Inf |
| account | 25 | 1047972 | 1.01E+08 | 3.06E+05 | 1.01E+08 | 1.00E+08 | 1.01E+08 | 1.06E+06 |
| dataset | 26 | 1047972 | NaN | *NA* | *NA* | Inf | -Inf | -Inf |
| int_tgt_bin | 27 | 1047972 | NaN | *NA* | *NA* | Inf | -Inf | -Inf |
| demog_inc2 | 28 | 1047972 | 1.07E+01 | 4.30E-01 | 1.07E+01 | 7.82E+00 | 1.22E+01 | 4.38E+00 |
| demog_inc2_sq | 29 | 1047972 | 2.13E+01 | 8.60E-01 | 2.14E+01 | 1.56E+01 | 2.44E+01 | 8.77E+00 |
| rfm6_sq | 30 | 1047972 | 3.80E+00 | 1.79E+00 | 3.91E+00 | 0.00E+00 | 9.58E+00 | 9.58E+00 |
| prospect_ho | 31 | 1047972 | 1.00E-02 | 1.10E-01 | 0.00E+00 | 0.00E+00 | 1.00E+00 | 1.00E+00 |
| rfm2_inc2 | 32 | 1047972 | 1.31E+01 | 6.90E-01 | 1.31E+01 | 9.78E+00 | 1.72E+01 | 7.43E+00 |
| demog_inc_homeval | 33 | 1047972 | 9.29E+01 | 5.32E+01 | 1.18E+02 | 0.00E+00 | 1.62E+02 | 1.62E+02 |