2024 Travelers University Modeling Competition

Maksuda Toma, Aarif Baksh   University of Nebraska Lincoln
Team: Roaming Residuals

2024-12-05

Business Problem

- **Problem**:CloverShield Insurance is facing high call center costs caused by inefficient resource allocation due to unpredictable policyholder call behavior.
- **Objective**: Reduce call center costs while maintaining operational efficiency.
- **Challenge**: Forecast the number of calls policyholders are likely to make.
- **Approach**: Develop a predictive model leveraging segmentation data.
- **Outcome**: Enable optimized resource allocation and improved cost management.

Data Overview

- The data was compiled by our Business Intelligence department at CloverShield.
- Training Set: 80,000 records
- Test Set: 20,000 records

Distribution of `call_counts`

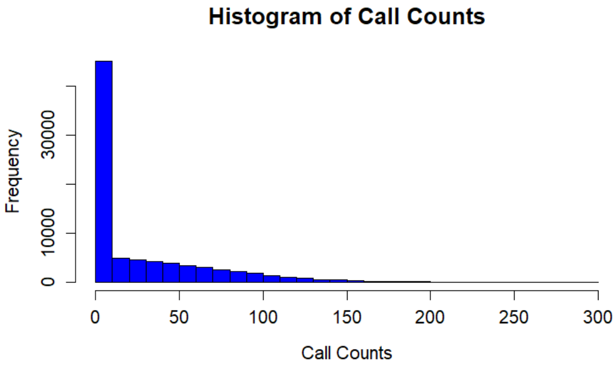


Figure 1: Fig-1: Call Counts Distribution

Observations:

1. This graph shows that call_count is rightly skewed.
2. About half (50.18%) of the customers did not make any calls.

Data Cleaning and Missing Value count

First, we prepares the data by cleaning and transforming it (e.g., converting characters to factors, marking missing values.)

Variable	Number of missing values
acq_method	16,066
newest_veh_age	58,015
pol_edeliv_ind	838
telematics_ind	58,015

Zero Values

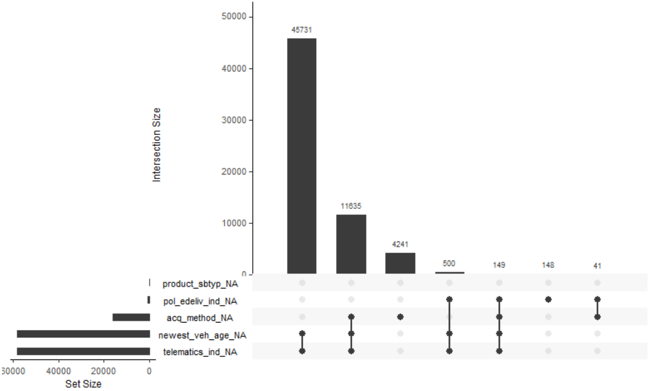
50.18% of the rows in the call_counts column are zeros, indicating that most customers made no calls. This is significant and might suggest using models like Zero-Inflated Poisson (ZIP) to handle the high frequency of zeros. The dataset contains both numeric and categorical variables, with some columns having significant missing values. - The target variable (call_counts) is heavily zero-inflated and skewed, which may require specialized modeling approaches. - Some numeric variables, like ann_prm_amt and home_lot_sq_footage, have wide ranges and outliers, suggesting that data transformation or scaling may be beneficial.

Missing Data Summary

Variable	Missing (%)
telematics_ind	72%
newest_veh_age	72%

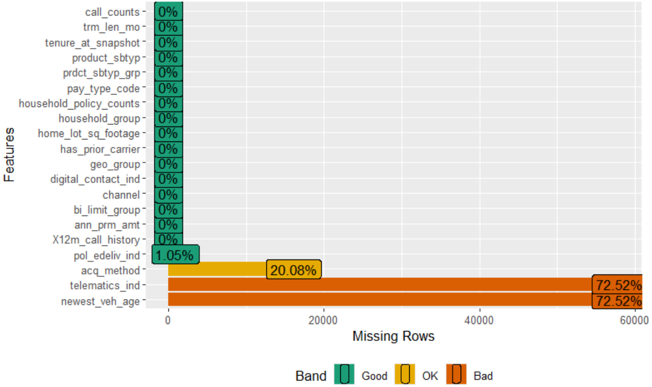
Missing Value display-1

The UpSet Plot visualizes missing data patterns across variables, with newest_veh_age and telematics_ind having the highest missingness. Most rows (∼45,731) have missing values only in newest_veh_age, while overlapping missingness across multiple variables is less common. This suggests prioritizing simple imputation for isolated missingness and predictive methods for overlapping patterns.



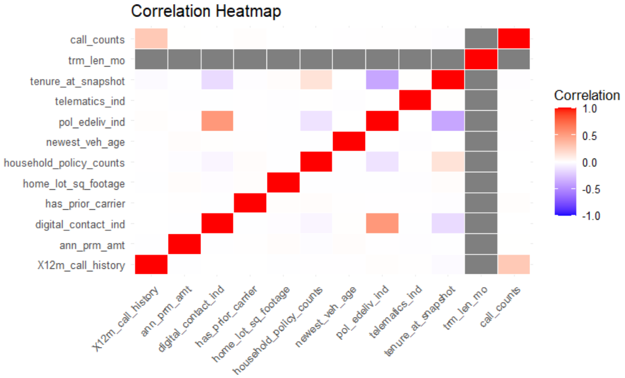
Missing Value display-2

This chart highlights missing data percentages across features. Most features have no missing values, but newest_veh_age and telematics_ind (72.52% missing) require advanced handling, while acq_method (20.08%) needs simpler imputation. Minimal effort is required for features like pol_edeliv_ind (1.05%).



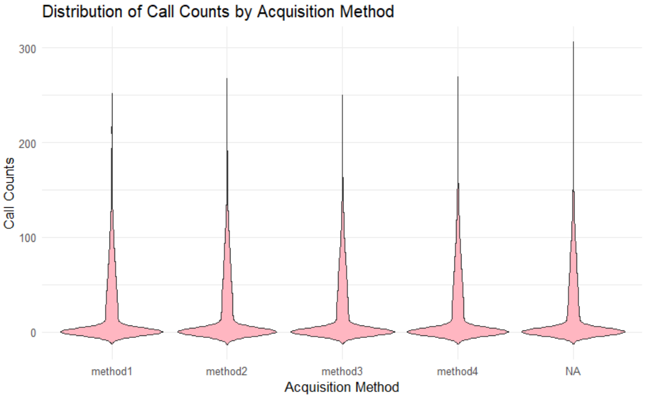
Correlation Matrix

The correlation heatmap identifies X12m_call_history as the strongest predictor of call_counts (r 0.28), while most other variables show weak or no correlations. There are no strong negative relationships, and overall correlations are weak. This suggests the need for non-linear models or feature engineering to capture complex interactions.



Call_counts distribution with significant predictor

The violin plot reveals a heavily skewed distribution of call_counts across all acq_method categories, with most values near 0 and a few outliers. The similar distributions across methods, including the NA category, suggest minimal impact of acq_method on call_counts. This aligns with ANOVA results showing marginal significance, warranting further analysis of outliers or interactions.



Models

Models	**Status**
Gradient Boosted Machine (GBM)	Tried
Zero Inflated Poission (ZIP)	Tried
Zero Inflated Negative Binomial (ZINB)	Tried
Random Forest	Tried
THurdle	Considered
Two-Part Model	Considered

Model Comparison

1. **Gradient Boosting Machine (GBM)**
   - Test RMSE: 36.1614
   - Key predictor: `X12m_call_history`
2. **Random Forest**

- Test RMSE: 36.30212

3. **Zero-Inflated Poisson (ZIP)**
   - Test RMSE: 36.61514
4. **Zero-Inflated Negative Binomial (ZINB)**
   - Test RMSE: 36.85568

Model Selection

Gradient Boosting Machine (GBM)

- Test RMSE: 36.1614
- Best Performing Model
- **Parameter Tuning**: Trial and Error
- **Challenge**: Dataset was too large for hyperparameter tuning

Variable Selection

Gradient Boosting Machine (GBM)

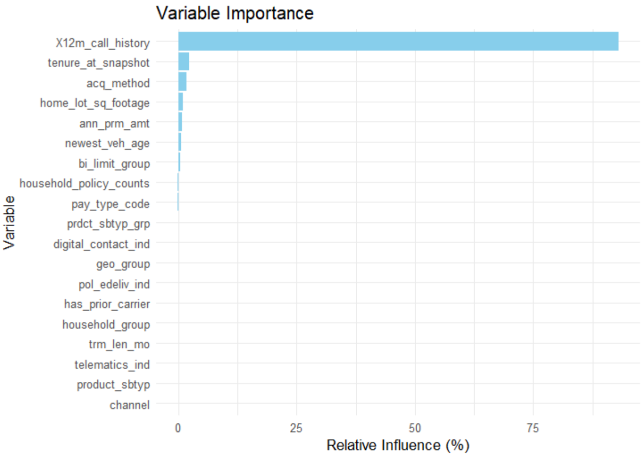


Figure 2: Fig-5: Variable Importance Plot

- An initial GBM was run with all the variables, and then a

Variable Selection

	var	rel.inf
X12m_call_history	X12m_call_history	92.989837493
tenure_at_snapshot	tenure_at_snapshot	2.282196632
acq_method	acq_method	1.668747717
home_lot_sq_footage	home_lot_sq_footage	0.904520122
ann_prm_amt	ann_prm_amt	0.861723686
newest_veh_age	newest_veh_age	0.507674848
bi_limit_group	bi_limit_group	0.392940735
household_policy_counts	household_policy_counts	0.114731077
pay_type_code	pay_type_code	0.112678300
prdct_sbtyp_grp	prdct_sbtyp_grp	0.072202460
digital_contact_ind	digital_contact_ind	0.040889483
geo_group	geo_group	0.027534506
pol_edeliv_ind	pol_edeliv_ind	0.014905375
has_prior_carrier	has_prior_carrier	0.007238914
household_group	household_group	0.002178653
channel	channel	0.000000000
product_sbtyp	product_sbtyp	0.000000000
telematics_ind	telematics_ind	0.000000000
trm_len_mo	trm_len_mo	0.000000000

Figure 3: Fig-6: Variable Importance

- Most Important Variables: X12m_call_history, tenure_at_snapshot, and acq_method
- Test RMSE for Model with all variables: 36.1742
- Test RMSE for Model with 3 variables selected from Variable

Model Evaluation

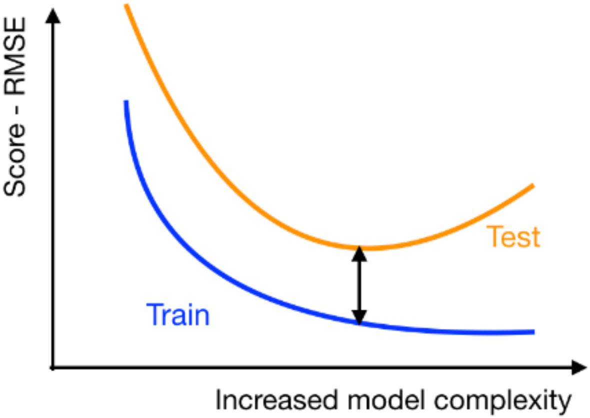


Figure 4: Fig-6: Train and Test RMSE Curves

- Train RMSE: 35.67179
- Test RMSE: 36.1742

Concerns

1. The model is likely sub-optimal, as it struggled to achieve a good accuracy score (on the validation set) and the parameters were tuned through trial and error instead of using a grid search to find the optimal values.