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`

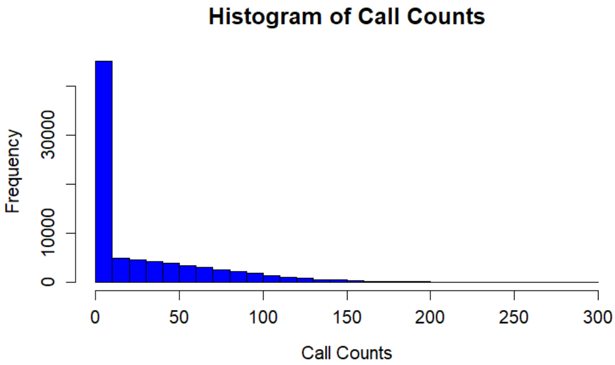


Figure 1: Fig-1: Call Counts Distribution

Observations:

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

Missing Data Summary

Variable	Number of missing values
acq_method	16,066 (20.08%)
newest_veh_age	58,015 (72%)
pol_edeliv_ind	838
telematics_ind	58,015 (72%)

We performed imputation on the dataset to fill in the missing values, ensuring that our analysis is based on complete data and providing more accurate insights for decision-making.

Data Cleaning and Missing Value Handling

Cleaning Data: The dataset was cleaned by replacing placeholders like -2, -20, and “missing” with proper markers for missing values, and converting text-based columns into categories for easier analysis.

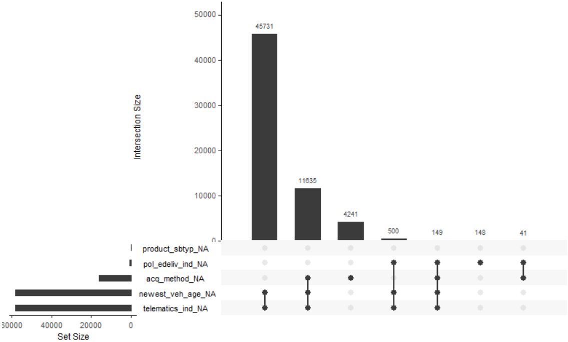
Imputation:

- Missing data was handled through “imputation,” which intelligently fills gaps based on patterns, generating five dataset versions and selecting the most consistent one for analysis.
- The system predicts missing values based on patterns, such as similar characteristics for numbers, “yes/no” guesses for binary data, and the best-fitting category for grouped data, ensuring the dataset’s structure is preserved

Missing Value Display-1

The UpSet Plot visualizes missing data patterns across variables, with newest_veh_age and telematics_ind having the highest missingness.

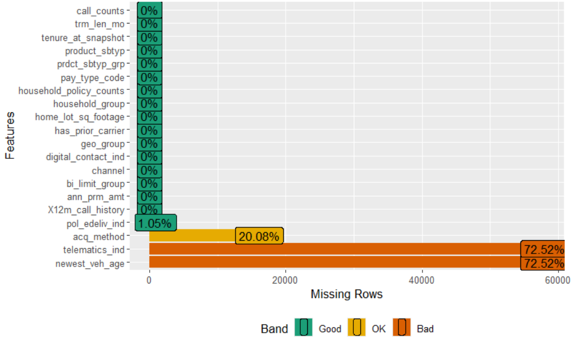
- Most rows (45,731) have missing values only in newest_veh_age, while overlapping missingness across multiple variables is less common.
- This suggests prioritizing simple imputation for isolated missingness and predictive methods for overlapping patterns.



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%).

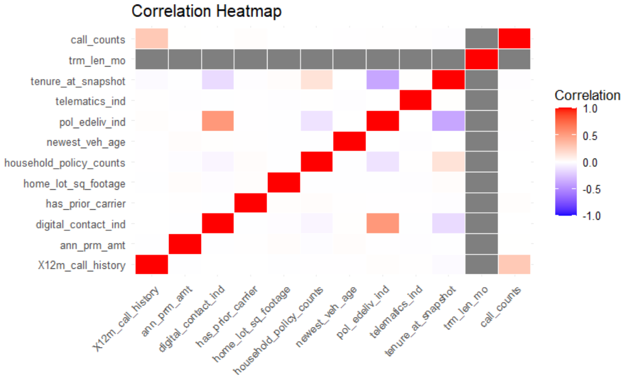


Zero Values

- About half of the customers (50.18%) in the dataset didn't make any calls, which is important because it means most of the data has a lot of zeros.
- To understand this better, we might need special tools that can deal with situations where many people don't take an action, like making a call.In that case we might use models like Zero-Inflated Poisson (ZIP) to handle the high frequency of zeros
- Some columns in the data, like how much people pay for premiums or the size of their property, have very big differences or unusual 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.

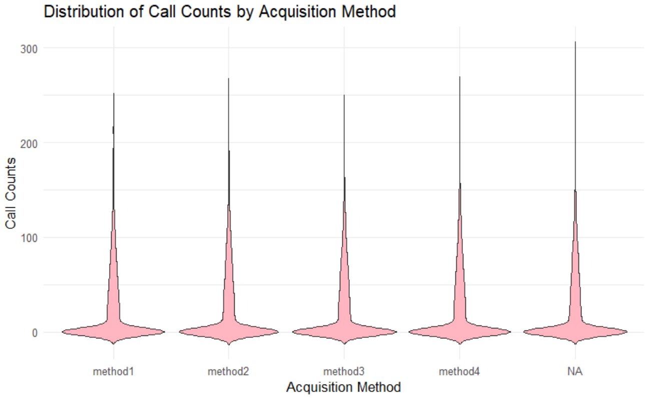
Correlation Matrix

- 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.



Models

Models	Test RMSE	Status
Gradient Boosted Machine (GBM)	36.1614	Tried
Random Forest	36.30212	Tried
Zero-Inflated Poisson (ZIP)	36.61514	Tried
Zero-Inflated Negative Binomial (ZINB)	36.85568	Tried
Hurdle	-	Considered
Two-Part Model	-	Considered

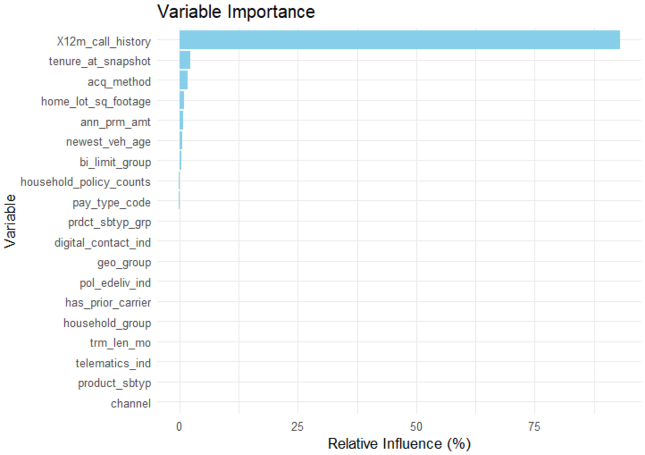
Model Selection

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)



An initial GBM was run with all the variables, and then a subset of 3 variables was selected from the variable importance plot, and another gbm model was run with those three variables.

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 4: Important Variable

- 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 Importance Plot: 36.1614
- Limitation: Variable importance does not specify the relationship between the predictors and call_counts

Model Evaluation

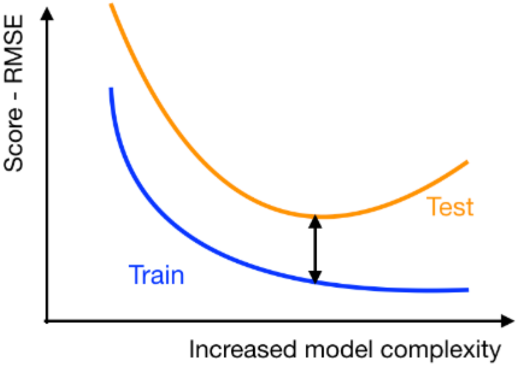


Figure 5: Train and Test RMSE Curves

- Train RMSE: 35.67179
- Test RMSE: 36.1742

Concerns

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.

Recommendations

To improve the model's performance, we recommend using a grid search for hyperparameter optimization. This method systematically explores a range of parameter combinations to identify the optimal values, resulting in a more accurate and reliable model. With better computing power, implementing a grid search would be feasible and could significantly enhance the model's predictive capability.

Thank You

Question?