2024 Travelers University Modeling Competition

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

2024-12-05

Introduction

The data obtained from Kaggle, is split into two parts: training data and validation data. In the validation data, the target variable, call_counts, is omitted. The training dataset contains 80,000 samples, and the validation dataset contains 20,000 samples. We will be mostly interested in `call_counts`, `12m_call_history`, `ann_prm_amt`, `newest_veh_age`, `home_lot_sq_footage`, `digital_contacts_ind`, `has_prior_carrier`and so on.

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.

Distribution of `call_counts`

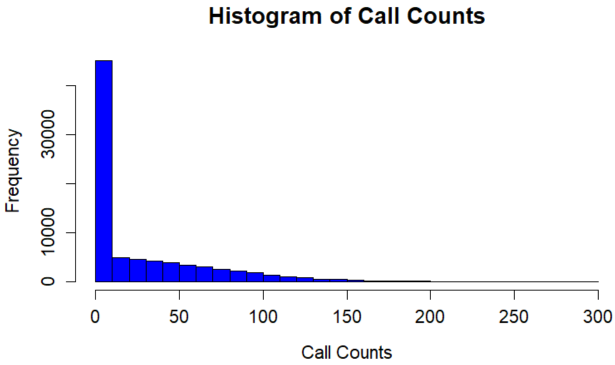


Figure 1: Fig-1: Call Counts Distribution

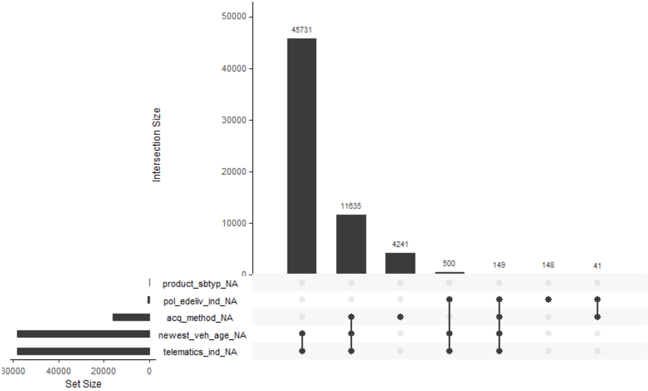
This graph shows that the response variable is rightly skewed.

Missing Data Summary

Variable	Missing (%)
telematics_ind	72%
newest_veh_age	72%

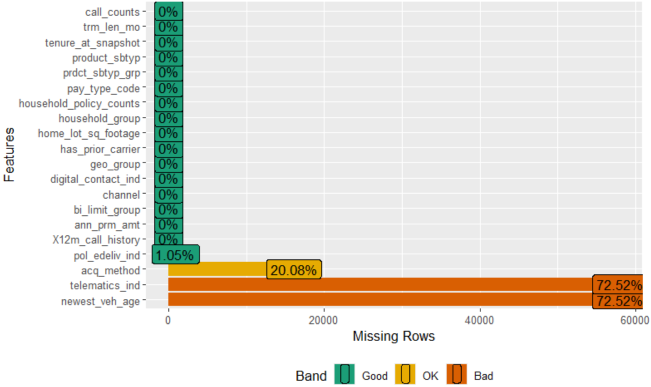
Missing Value display-1

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.



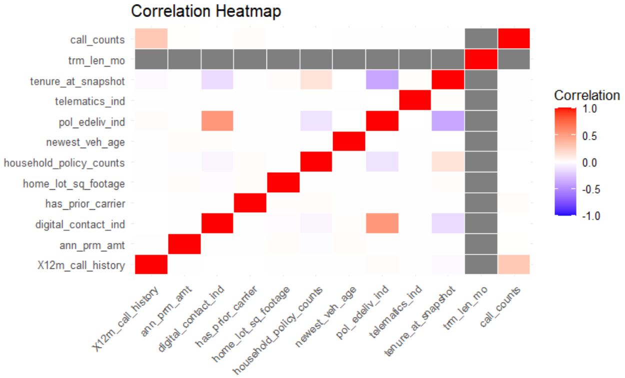
Missing Value display-2

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%).



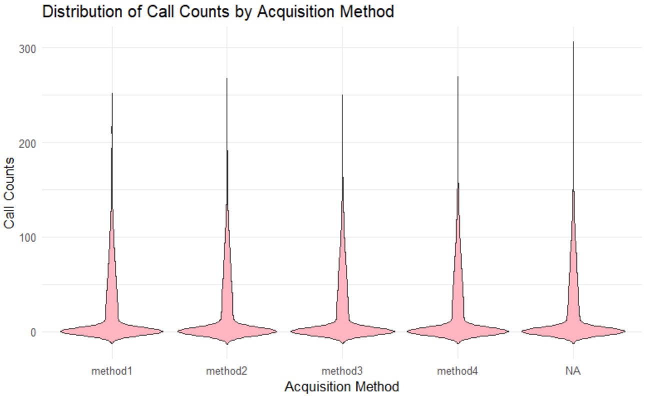
Correlation Matrix

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.



Several Models

We'll show here GBM, ZIP model

[Can you include those?]

Model Comparison

1. **Gradient Boosting Machine (GBM)**
   - RMSE: 36.06
   - Key predictor: `X12m_call_history`
2. **Zero-Inflated Poisson (ZIP)**
   - RMSE: 36.53
   - Suitable for zero-inflated data.