Final Report

Maksuda Aktar Toma
Statistics Department,
University of Nebraska, Lincoln
mtoma2@huskers.unl.edu

Aarif Baksh
Statistics Department,
University of Nebraska, Lincoln
abaksh2@unl.edu

Abstract

This project aims to mitigate call center expenses for CloverShield Insurance by creating a predictive model to estimate the volume of calls (call_counts) made by policyholders. A dataset obtained from Kaggle, comprising 80,000 training samples and 20,000 validation samples, was utilized to examine several predictors, including annual premium amount, vehicle age, call history, and policyholder demographics. Data preprocessing encompassed addressing absent values, converting categorical variables, and analyzing patterns to establish a solid groundwork for modeling. Various statistical and machine learning methods were employed to determine the most significant predictors and to account for the zero-inflated and skewed characteristics of the target variable. The approach seeks to optimize resource allocation in call center operations, offering actionable insights to improve cost-efficiency.

For additional information and code implementation, please check- https://github.com/maksudatoma/2024-Travelers-University-Modeling-Competition/tree/main

1. Introduction

Minimizing call center expenses while preserving customer happiness is a significant challenge for CloverShield Insurance. A predictive model has been created to estimate the call volume (call_counts) a policyholder is expected to produce. Comprehending these patterns will facilitate enhanced resource allocation, augment operational efficiency, and diminish superfluous expenditures in call center operations.

The target variable, call_counts, denotes the quantity of calls made by a policyholder, whereas the independent variables are a combination of demographic, policy, and behavioral attributes. Principal predictors encompass X12m_call_history (previous year's call history), ann_prm_amt (annualized premium amount), newest_veh_age (age of the most recent insured vehicle), geo_group (policyholder's residential region), and digital_contacts_ind (digital communication indicator), among others.

The target variable has distinct attributes: it is zero-inflated, skewed, and count-based, presenting difficulties for conventional predictive models. Moreover, absent values exist in crucial predictors including newest_veh_age, telematics_ind, and acq_method. Consequently, meticulous data preprocessing, encompassing imputation and modification, was performed to guarantee data quality and reliability.

To ascertain the most precise and effective model, many methodologies were employed, including Gradient Boosting Machines (GBM), Random Forest, and Zero-Inflated Poisson (ZIP) models. The models were assessed on their capacity to encapsulate the diversity in call patterns and manage the intricacies of the target variable. The primary objective of this investigation is to furnish practical insights that empower CloverShield Insurance to allocate resources effectively, reduce call center expenses, and enhance customer service operations.

2. Methodology

2.1 Imputation

Missing values in the dataset were handled using the Multivariate Imputation by Chained Equations (MICE) package in R. MICE generates plausible synthetic values for incomplete columns by leveraging relationships with other variables through a Markov Chain Monte Carlo (MCMC) process, specifically utilizing Gibbs sampling. This iterative technique ensures missing values are updated based on the observed data's conditional distributions.

In this dataset, four variables contained missing values: acq_method (20%), newest_veh_age (72%), pol_edeliv_ind (1%), and telematics_ind (72%). Appropriate imputation methods were applied depending on the variable type:

- acq_method: A nominal variable with multiple categories; missing values were imputed using polytomous logistic regression (polyreg), suitable for unordered categorical variables with more than two levels.
- newest_veh_age: A numeric variable; imputed using Predictive Mean Matching (pmm), which preserves realistic values by selecting observed data close to the predicted mean.
- pol_edeliv_ind and telematics_ind: Binary variables; missing values were imputed using logistic regression (logreg), ideal for variables with two outcomes.

2.2 Zero Values

The target variable, call_counts, contains a significant proportion (50.18%) of zero values, indicating that many policyholders did not make any calls. This pattern highlights the need for specialized models such as Zero-Inflated Poisson (ZIP), which can effectively address zero-inflated and skewed count data.

Overall, the dataset combines categorical and numerical variables, with notable missingness in a few key columns. The target variable, call_counts, exhibits a heavily skewed distribution with a high frequency of zeros. Additionally, numerical predictors like home_lot_sq_footage and ann_prm_amt display wide ranges and outliers, suggesting that scaling or transformation may improve modeling performance.

2.3 Correlation Structure

The correlation matrix highlights that call_counts is the target variable, with X12m_call_history showing the strongest positive correlation (r≈0.28), suggesting that higher call history counts are associated with an increase in call counts. In contrast, other continuous variables, such as ann_prm_amt, home_lot_sq_footage, and telematics_ind, exhibit very weak correlations (r=0.001 to 0.005), indicating minimal linear relationships with the target variable. Variables like newest_veh_age and tenure_at_snapshot show negligible negative correlations. This lack of strong correlations suggests that most continuous predictors are not significant linear contributors to call_counts. However, these variables may still provide value when modeled using non-linear techniques, such as Gradient Boosting Machines (GBM) or Random Forest, or when interactions between variables are explored.

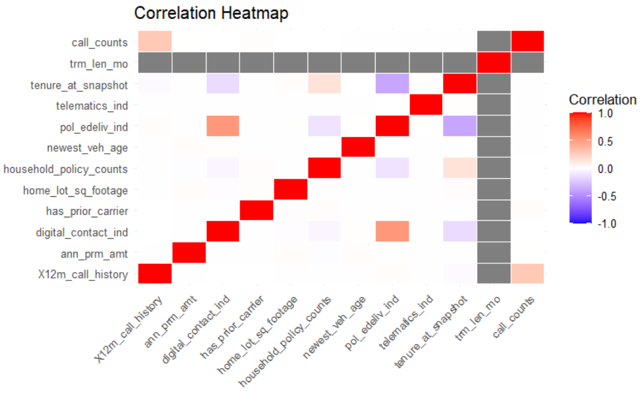
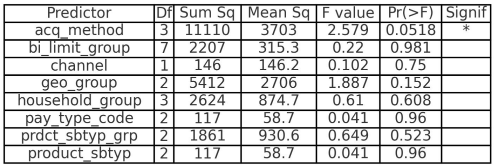


Figure 1: Heat Map

2.4 ANOVA Table Now we will see what's going on with the categorical variable through the ANOVA table.

The ANOVA results evaluate the effect of categorical variables on call_counts. Among the predictors, acq_method is marginally significant (p=0.0518), suggesting it may have a weak influence on call_counts. All other categorical variables, such as bi_limit_group, channel, and geo_group, have p-values greater than 0.1, indicating no statistically significant relationship with the target variable. Additionally, 16,066 rows were excluded due to missing data, which might affect the robustness of the results. It is recommended to focus on acq_method for further analysis and consider handling missing data to improve model accuracy.



Predictor	Df	Sum Sq	Mean Sq	F value	Pr(>F)	Signif
acq_method	3	11110	3703	2.579	0.0518	*
bi_limit_group	7	2207	315.3	0.22	0.981	
channel	1	146	146.2	0.102	0.75	
geo_group	2	5412	2706	1.887	0.152	
household_group	3	2624	874.7	0.61	0.608	
pay_type_code	2	117	58.7	0.041	0.96	
prdct_sbtyp_grp	2	1861	930.6	0.649	0.523	
product_sbtyp	2	117	58.7	0.041	0.96	

Figure 2: ANOVA result

The violin plot shows the distribution of call_counts across different acquisition methods (acq_method). All methods have a heavily skewed distribution, with most values near 0 and a few extreme outliers, indicating that the majority of customers make few calls. The distributions are nearly identical across all methods, including the NA category, suggesting that acq_method has minimal impact on call_counts. This aligns with the ANOVA results, where acq_method was marginally significant. Further analysis, such as handling outliers or exploring interactions with other variables, may provide additional insights.

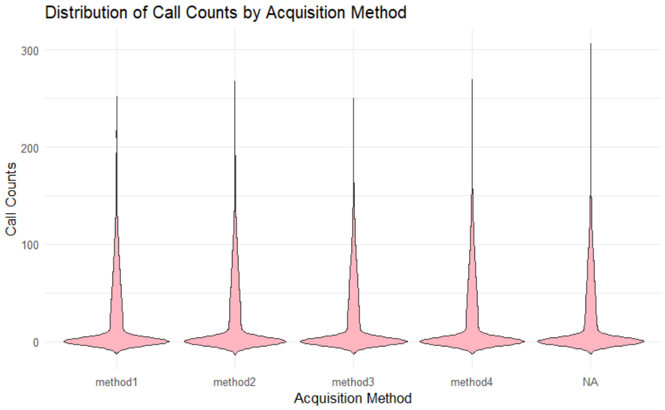


Figure 3: Violin Plot

3. Result

3.1 Model Selection and Hyperparameter Tuning

An initial GBM model was built using all predictors and one-thrid of the data in the training dataset. Since call_counts is a count variable the Poisson distribution was used. Repeated cross-validation was implemented through trainControl, using 5-fold cross-validation repeated 3 times and the model performance was measured using **Root Mean Square Error (RMSE)**. A grid search for hyperparameter tuning was conducted using tuneGrid, varying the number of trees (n.trees) from 100 to 1500 in increments of 100 and the learning rate (shrinkage) from 0.01 to 0.10 in increments of 0.01. Additionally, the interaction.depth was tuned between 2 and 10, and the **minimum number of observations in terminal nodes** (n.minobsinnode) was adjusted between 10 and 100. The parameter bag.fraction was set to 1, ensuring that all data were used in each boosting iteration.

The parameters for this model that resulted in the lowest RMSE (RMSE = 36.1742) were determined to be n.trees= 1000, shrinkage= 0.02, interaction.depth = 7 and n.minobsinnode= 20.

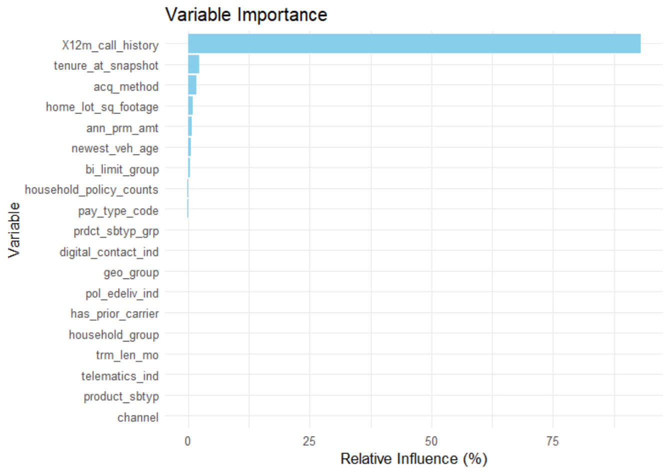
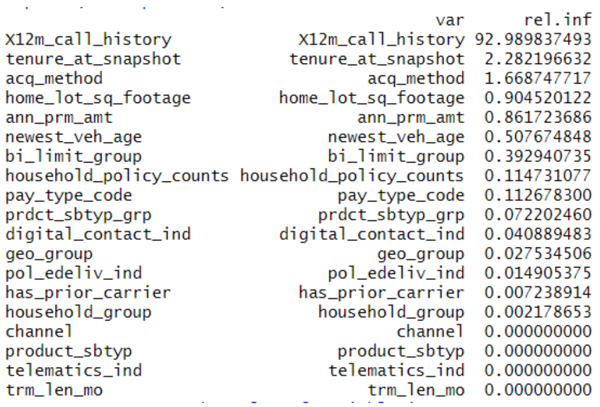


Figure 4: Important Variable



```
                                            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 5: Important Variable

Variable selection for the final models was conducted using the variable importance plot generated from a GBM trained on all available predictors. The importance scores provided insights into the relative contribution of each variable to the model's predictions.

The results revealed that 12-month call history (12m_call_history) is the most significant predictor, with an importance score of 92.98. This aligns with the earlier observed correlation, highlighting a customer's call history in the past 12 months as the strongest determinant of future call volumes. Following this, tenure at snapshot (tenure_at_snapshot) and acquisition method (acq_method) ranked second and third, with importance scores of 2.28 and 1.66, respectively.

Variables such as pay_type_code (0.11) and digital_contact_ind (0.04) showed minimal importance, contributing little to the model's predictive power. Other variables, including product_sbtyp, telematics_ind, and trm_len_mo, had importance scores of 0.00, indicating no measurable influence on call volume predictions. It is important to note that variable importance scores do not provide information on the direction or nature of the relationships (linear or nonlinear) between predictors and the target variable. Additionally, variables with zero importance may still play indirect roles or contribute to interactions with other predictors.

To simplify the model, Gradient Boosted Machines (GBMs) were built using the top 3 to top 10 variables identified from the variable importance plot. Among these, the model utilizing only the top three variables—**12m_call_history**, **tenure_at_snapshot**, and **acq_method**—achieved the lowest test RMSE (RMSE = 36.1614) within this group. This result not only highlights the predictive strength of these three variables but also justifies the selection of a smaller, more interpretable model without sacrificing accuracy.

[I HAVE TO ADD THE DETAILS OF THE OTHER GBM, ZIP, ZINB and RANDOM FOREST HERE]

Model Result

The Gradient Boosted Machine (GBM) attained the lowest RMSE of 36.1614, signifying superior predictive accuracy compared to other evaluated models, with Random Forest closely trailing at an RMSE of 36.30212. The Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) models exhibited elevated RMSE values, signifying diminished accuracy. The Hurdle and Two-Part Models were contemplated but remain untested, allowing for future assessment. Gradient Boosting Machine (GBM) and Random Forest have the highest performance according to Root Mean Square Error (RMSE). Additional evaluation of the Hurdle and Two-Part Models may yield chances for enhancing forecasts.

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

I HAVE STOPPED HERE. CAN YOU WRITE NEXT PROCEDURE IN DETAILS? NEED TO FOCUS ON

THESE 3. In the Method section describe the technical details of the steps you had taken. techincal description of imputation. If you are using GLM, what are the models for Bernoulli section and the Count section. If you are using RF, what is the node cost function, stopping rule, etc.

4. In the result section offer all model comparison result. Describe if you were doing a full cross-validation or a single hold out test. if you are using single hold out, why is that appropriate?
5. How you are you handling multiple tuning parameters that you obtain in each fold of CV?

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)

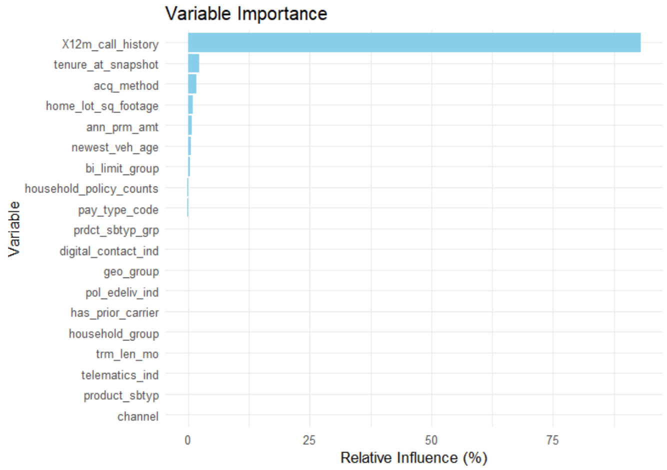


Figure 6: Variable Importance Plot

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

- 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

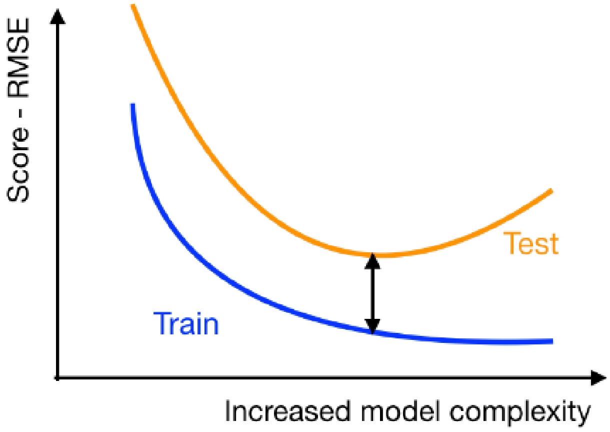


Figure 7: Train and Test RMSE Curves

- Train RMSE: 35.67179
- Test RMSE: 36.1742

Concerns

The model is likely sub-optimal, as it struggled to achieve a good accuracy score (about 25% 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

With better computing power, implementing a grid search would be feasible and could significantly enhance the model's predictive capability.

The MissForest method (Stekhoven and Bühlmann 2012) uses Random Forests for non-parametric imputation. Gradient Boosting Machines (GBM) have also been widely explored for predictive modeling (Friedman 2001). Stacking, a form of ensemble learning, is discussed in Boehmke's work (Boehmke 2021).

References

Boehmke, Bradley. 2021. "Stacking Models in Machine Learning." 2021. https://bradleyboehmke.github.io/HOML/stacking.html.

Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics* 29 (5): 1189–1232. https://doi.org/10.1214/aos/1013203451.

Stekhoven, Daniel J, and Peter Bühlmann. 2012. "MissForest—Non-Parametric Missing Value Imputation for Mixed-Type Data." *Bioinformatics* 28 (1): 112–18. https://doi.org/10.1093/bioinformatics/btr597.