Final Report

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

Aarif Baksh
Statistics Department,
University of Nebraska, Lincoln
abaksh2@huskers.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 missing values through Multivariate Imputation by Chained Equations (MICE). Various statistical and machine learning methods were employed, with Gradient Boosted Machine (GBM) emerging as the best performer with a test RMSE of 36.1614, followed closely by Random Forest (RMSE = 36.30212). The final GBM model, trained on the top three predictors—12-month call history, tenure at snapshot, and acquisition method—achieved an accuracy of 0.24601 on the Kaggle validation set. While this accuracy indicates room for improvement, the model provides a foundation for optimizing resource allocation in call center operations. Future enhancements could include expanded hyperparameter tuning and exploration of alternative modeling approaches to better capture the zero-inflated and skewed characteristics of the target variable.

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

1. Introduction

CloverShield is a leading insurance company committed to delivering reliable coverage and exceptional customer support. A key challenge for the organization lies in managing and reducing call center costs while maintaining high service standards. To address this, a predictive model has been created to estimate the call volume (call_counts) a policyholder is expected to make. 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), acq_method (acquisition method), 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, imputation was performed to address missing values and ensure the dataset's reliability.

The analysis considered Gradient Boosting Machines (GBM), Random Forest, and Zero-Inflated models, evaluating their ability to capture the diverse call patterns and handle the complexities of the target variable. The primary goal was to provide actionable insights that enable CloverShield Insurance to optimize resource allocation, minimize call center costs, and improve overall customer service operations.

2. Methodology

2.1 Dataset Structure

Overall, the dataset contains 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.

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.

2.3 Numeric Predictors: Correlation Structure

The correlation matrix shows that X12m_call_history is most strongly correlated with call_counts ($r \approx 0.28$), 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 call_counts. 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.

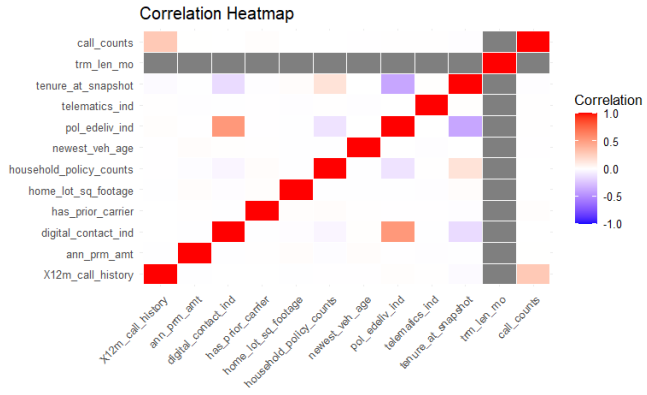

Figure 1: Heat Map of correlation matrix

2.4 Categorical Predictors: ANOVA Table

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

ANOVA evaluates 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.

Given that the ANOVA table indicates acq_method is marginally significant, further analysis was conducted to explore its impact. 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,

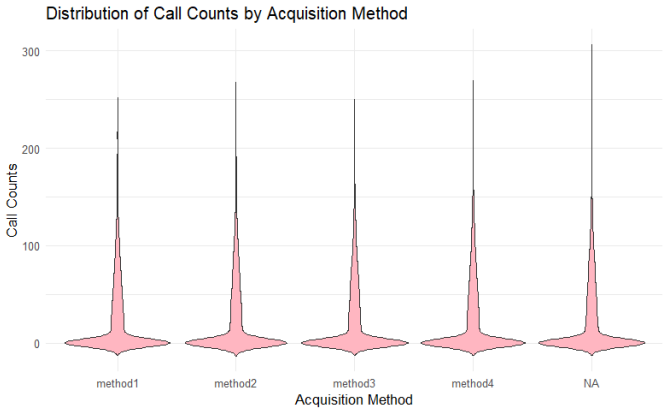

Figure 3: Violin Plot

where acq_method was marginally significant. Further analysis, such as handling outliers or exploring interactions with other variables, may provide additional insights.

2.5 Imputation

Missing values in the dataset were handled using the Multivariate Imputation(Troyanskaya et al. 2001) 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.

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 1000 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 50. 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 n.trees= 1100, shrinkage= 0.03, interaction.depth = 7 and n.minobsinnode= 30.

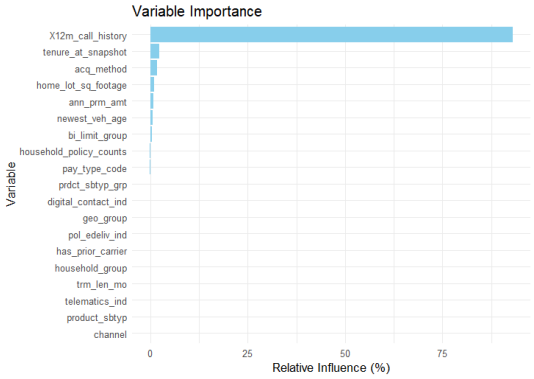

Figure 4: Variable Importance Plot

	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: Variable Importance Table

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.

3.1.1 Gradient Boosted Machines (GBM) and Random Forests

Gradient Boosted Machines (GBMs) and Random Forests were applied to identify the most effective predictors from the dataset (Friedman 2001; Stekhoven and Bühlmann 2012). Using subsets of the data, the top 3 to top 10 variables—selected based on variable importance plots—were utilized for model training on a randomly selected one-third portion of the dataset (training set). The performance of these models was evaluated on the remaining one-third of the dataset (test set). Among the trained models, the GBM and Random Forest models that incorporated only the top three predictors—**12m_call_history**, **tenure_at_snapshot**, and **acq_method**—achieved the best results, yielding the lowest Root Mean Squared Error (RMSE) values on the test set. Specifically, the GBM achieved an RMSE of 36.1614, while the Random Forest recorded an RMSE of 36.30212.

The GBM model, which emerged as the top performer, was optimized through cross-validation on the training set, utilizing the parameters: n.trees = 1200, shrinkage = 0.02, interaction.depth = 2, and n.minobsinnode = 20. Similarly, the Random Forest model underwent cross-validation to determine its optimal configuration, with n.trees = 50 and nodesize = 20 (Breiman 2001). These results underscore the importance of carefully selecting key predictors and tuning hyperparameters to achieve robust predictive performance.

3.1.2 Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB)

Zero-inflated models were built to predict call counts using key predictor variables. The training was conducted on the same one-third of the dataset as used for the GBM and Random Forest models, with the remaining one-third reserved for testing. The count component of the model included **X12m_call_history**, **tenure_at_snapshot**, and **acq_method**, while the zero-inflation component incorporated **X12m_call_history** and **ann_prm_amt**. Two variations of the model were tested: one using the Poisson distribution and the other using the Negative Binomial distribution. On the test set, the Poisson-based model achieved a test RMSE of 36.61514, while the Negative Binomial model recorded a slightly higher RMSE of 36.85568.

The zero-inflated model includes a count and a zero-inflation component. For the count component, *X12m_call_history,* ***tenure_at_snapshot, and*** *acq_method* were used. *X12m_call_history* predicts future calls based

on past behavior, *tenure_at_snapshot* reflects the policyholder's relationship duration, and *acq_method* captures call behavior differences by acquisition type. The zero-inflation component uses *X12m_call_history* and *ann_prm_amt. A lack of calls in **X12m_call_history* indicates a higher likelihood of structural zeros, while *ann_prm_amt* (annual premium amount) reflects engagement, with extreme values potentially linked to zero calls.

3.1.3 Test RMSE

The predict function was used to generate predicted values for each method, and the test RMSE was subsequently calculated using the formula: Test RMSE $= \sqrt{\frac{\sum(y_i - \hat{y_i})}{n}}$ where y_i is the observed count and $\hat{y_i}$ is the predicted count and n is the total number of observations in the test set.

3.2 Model Result

The Gradient Boosted Machine (GBM) attained the lowest test RMSE of 36.1614, signifying superior predictive accuracy compared to other evaluated models, with Random Forest closely trailing at a test RMSE of 36.30212. The Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) models resulted in marginally higher test 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

3.3 Final Model Selection

The final model was selected based on test RMSE, with the Gradient Boosted Machine (GBM) chosen as the best-performing model. To assess whether the model was appropriately fitted, a comparison was made between the test RMSE and the training RMSE. The test RMSE was 36.1614, while the training RMSE was 35.63423. The close proximity of these values indicates that the model generalizes well to unseen data, suggesting it is neither underfitted nor overfitted. This means that the model is able to capture the underlying patterns in the data without overly relying on the training dataset or losing predictive accuracy on the test set.

When tested on the Kaggle validation set, the model achieved an accuracy score of 0.24601. This low score reveals the model's difficulty in predicting outcomes for new, unseen data. The challenge may lie in the complexity of the underlying patterns, or the model might be missing key relationships within the dataset that are crucial for accurate predictions.

4. Concerns

A key limitation was the restricted scope of the grid search for hyperparameter tuning; the grid was not broad enough to explore a wider range of parameters due to the computational constraints posed by the large dataset. Additionally, attempts to include more variables in the Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) models resulted in convergence failures, again due to limited computing power.With better computing power, implementing a wider grid search would be feasible and could significantly enhance the model's predictive capability.

5. Recommendations

To improve the predictive accuracy of the models, future efforts should focus on expanding hyperparameter tuning with a broader grid search to explore a wider range of parameter combinations. This can be facilitated by leveraging more robust computing resources to handle the computational demands of larger datasets and complex models. Additionally, incorporating advanced modeling approaches, such as ensemble methods or deep learning architectures, may better capture the intricate patterns in the data. Addressing the limitations of the Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) models by testing alternative zero-inflation techniques or feature engineering strategies could further enhance predictive performance.

6. Conclusion

While our predictive model provides a starting point to address CloverShield Insurance's call center challenges, its current accuracy of 25% highlights opportunities for improvement. Moving forward, we can refine the model through enhanced computing resources, and employing a wider grid search. Additionally, exploring a different set of models may better capture the unique characteristics of the data. These efforts will help us build a more accurate and reliable forecasting tool to optimize resource allocation and reduce costs effectively.

7. References

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

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.

Troyanskaya, Olga, Michael Cantor, Gavin Sherlock, Patrick Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. 2001. "Missing Value Estimation Methods for DNA Microarrays." *Bioinformatics* 17 (6): 520–25. https://doi.org/10.1093/bioinformatics/17.6.520.