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

Business Problem

As an employee of CloverShield Insurance company, you are tasked with addressing the challenge of reducing call center costs. Your business partners have requested the development of a predictive model that, based on the provided segmentation, forecasts the number of times a policyholder is likely to call. This model aims to optimize resource allocation and enhance cost-efficiency in call center operations.

To find all our works on this project go to this link https://github.com/maksudatoma/2024-Travelers-University-Modeling-Competition/tree/main

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.There are several variables where "call_counts(The number of call count generated by each policy) is the target variable. The other variables that were used as predicted variables areann_prm_amt(Annualized Premium Amount),bi_limit_group(Body injury limit group),channel(Distribution channel),newest_veh_age(The age of the newest vehicle insured on a policy (-20 represents non-auto or missing values)),geo_group(Indicates if the policyholder lives in a rural, urban, or suburban area),has_prior_carrier(Did the policyholder come from another carrier),home_lot_sq_footage(Square footage of the policyholder's home lot),household_group(The types of policy in household),household_policy_counts(Number of policies in the household),telematics_ind(Telematic indicator (0 represents auto missing values or didn't enroll and -2 represents non-auto)),digital_contacts_ind(An indicator to denote if the policy holder has opted into digital communication),12m_call_history(Past one year call count),tenure_at_snapshot(Policy active length in month),pay_type_code(Code indicating the payment method),acq_method(The acquisition method (Miss represents missing values)),trm_len_mo(Term length month),pol_edeliv_ind(An indicator for email delivery of documents (-2 represents missing values)),aproduct_sbtyp_grp(Product subtype group),product_sbtyp' (Product subtype)

Methodology

Imputation

Missing values in the dataset were imputed using the *Multivariate Imputation by Chained Equations (MICE)* package in R. MICE generates "plausible" synthetic values for incomplete columns based on the relationships with other variables in the dataset. The imputation process uses a Markov Chain Monte Carlo (MCMC) approach, specifically a technique known as *Gibbs sampling*, which iteratively updates missing values by sampling from conditional distributions of the observed data.

In this dataset, *four variables* contain missing values: acq_method(20%), newest_veh_age(72%), pol_edeliv_ind(1%), and telematics_ind(72%). Each variable was imputed using methods appropriate for its type:

1. acq_method: A **nominal variable** with four categories. Missing values were imputed using **polytomous logistic regression** (polyreg), which is designed for categorical variables with more than two levels.
2. newest_veh_age: A *numeric variable*. Missing values were imputed using **Predictive Mean Matching** (pmm), which ensures imputed values are plausible by selecting observed values close to the predicted mean.
3. pol_edeliv_ind and telematics_ind: Both are **binary variables**. Missing values were imputed using **logistic regression** (logreg), which models binary outcomes effectively.

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.

Overall, The dataset includes both categorical and numerical variables, and there are notable missing values in a few of the columns. The target variable, call_counts, is highly skewed and zero-inflated, necessitating the use of specific modeling techniques. Some numerical variables, such as home_lot_sq_footage and ann_prm_amt, contain large ranges and outliers, indicating that scaling or data transformation would be helpful.

The correlation heatmap shows that X12m_call_history has the strongest positive correlation (r≈0.28) with call_counts, making it the most important numeric predictor. Most other

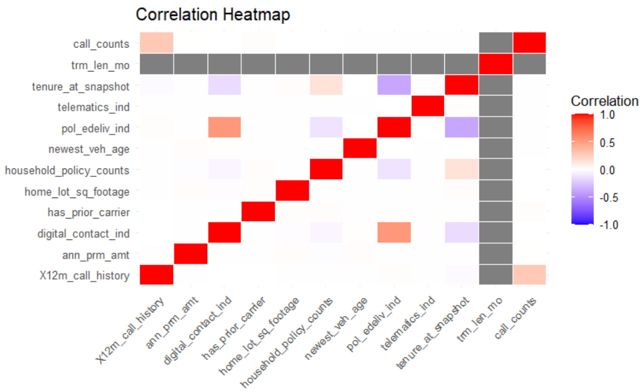


Figure 1: Heat Map

variables, such as ann_prm_amt, household_policy_counts, and home_lot_sq_footage, have weak or no significant correlations with the target variable, as indicated by grey cells. There are no strong negative correlations in the dataset. Overall, the relationships are mostly weak, suggesting that non-linear models or feature engineering may be needed to capture more complex interactions. The heatmap helps identify X12m_call_history as a key feature while others may contribute less linearly.

Result

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.

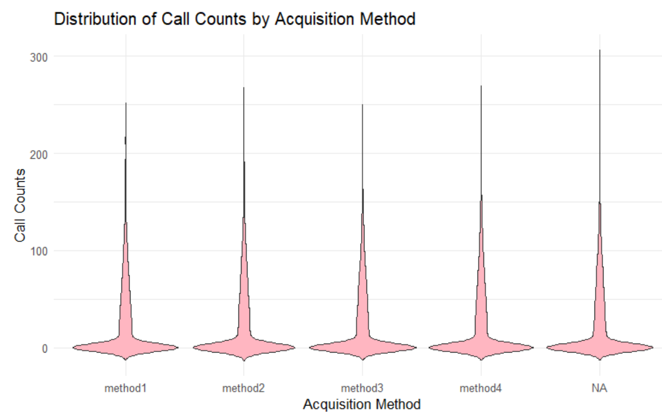


Figure 3: Violin Plot

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?

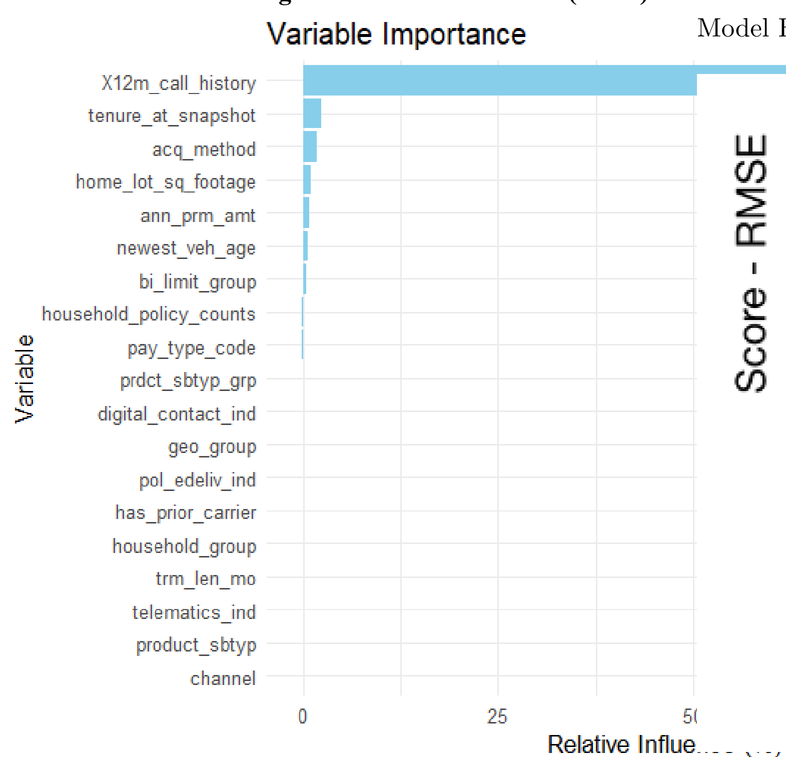
Model Selection

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)



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.

	var	rel.
X12m_call_history	X12m_call_history	92.989837
tenure_at_snapshot	tenure_at_snapshot	2.282196
acq_method	acq_method	1.668747
home_lot_sq_footage	home_lot_sq_footage	0.904520
ann_prm_amt	ann_prm_amt	0.861723
newest_veh_age	newest_veh_age	0.507674
bi_limit_group	bi_limit_group	0.392940
household_policy_counts	household_policy_counts	0.114731
pay_type_code	pay_type_code	0.112678
prdct_sbtyp_grp	prdct_sbtyp_grp	0.072202
digital_contact_ind	digital_contact_ind	0.040889
geo_group	geo_group	0.027534
pol_edeliv_ind	pol_edeliv_ind	0.014905
has_prior_carrier	has_prior_carrier	0.007238
household_group	household_group	0.002178
channel	channel	0.000000
product_sbtyp	product_sbtyp	0.000000
telematics_ind	telematics_ind	0.000000
trm_len_mo	trm_len_mo	0.000000

Figure 4: Important Variable

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

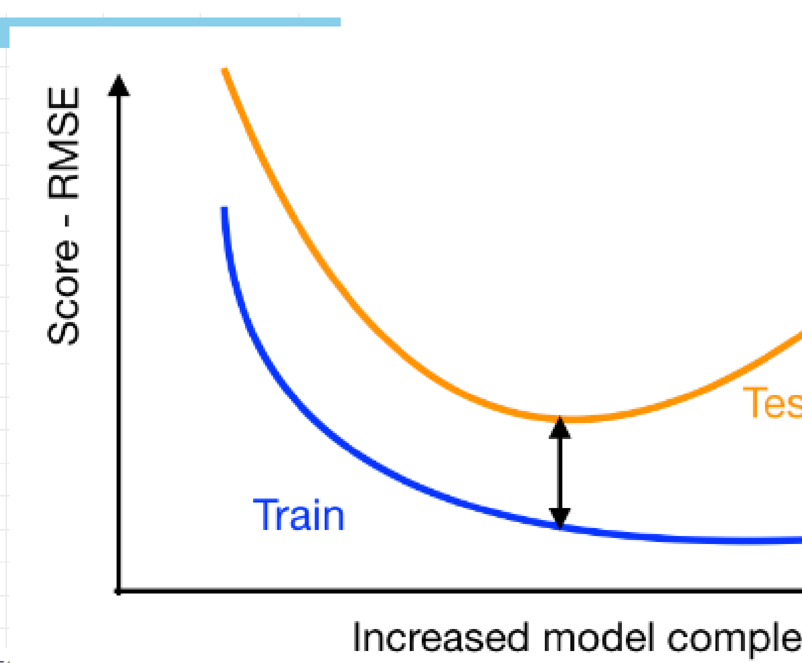


Figure 5: Train and Test RMSE Curves

- Train RMSE: 35.67179
- Test RMSE: 36.1742

Concerns

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

References

Data Preprocessing

Recoding

Exploratory Data Analysis

Results