2024 Travelers University Modeling Competition: CloverShield Insurance Company Modeling Problem

LightGBM Modeling Approach

Oluwafunmibi Omotayo Fasanya,

December 5, 2024

Problem Statement

Objective:

- Develop a predictive model to forecast policyholder call frequency (*call_counts*) for CloverShield Insurance, based on customer and policy data.

Goal:

- Reduce call center costs by optimizing resource allocation and improving efficiency through customer segmentation.

Dataset:

- The training data has 21 columns with 80,000 rows while the test dataset has 20 columns (No "Call Counts" column) with 20,000 rows.

Missing Values

Approach:

- Converted the values (-20, -2) to NA in the following columns:
  - Age of the newest vehicle insured on a policy (`newest_veh_age`)
  - Telematic indicator (`telematics_ind`)
  - Email delivery of documents (`pol_edeliv_ind`)
- Applied the same process to both the test and training datasets.

id	X12m_call_history	acq_method	ann_prm_amt	bi_limit_group
0.00000	0.00000	0.00000	0.00000	0.00000
channel	digital_contact_ind	geo_group	has_prior_carrier	home_lot_sq_footage
0.00000	0.00000	0.00000	0.00000	0.00000
household_group	household_policy_counts	newest_veh_age	pay_type_code	pol_edeliv_ind
0.00000	0.00000	72.51875	0.00000	1.04750
prdct_sbtyp_grp	product_sbtyp	telematics_ind	tenure_at_snapshot	trm_len_mo
0.00000	0.00000	72.51875	0.00000	0.00000
call_counts				
0.00000				

Percentage of missing variable in Train Data.

Variable Selection

Handling Missing Values:

- Identified variables with high percentages of missing values and flagged them for imputation.
- Created missing value indicator variables to capture potential predictive information from missingness
- Imputation of Missing Values:
  - Continuous variables: Median imputation (Robust to outliers).
  - Categorical variables: Mode imputation (Most likely category).

Variable Selection

Correlation Analysis:

- Evaluated relationships with the target variable (*call_counts*):
  - **Continuous variables:** Pearson correlation coefficient.
  - **Categorical variables:** Point-Biserial Correlation to assess the relationship between binary predictors and *call_counts*.
- *X12m_call_history* showed the highest positive correlation (0.28) with *call_counts* among continuous variables.

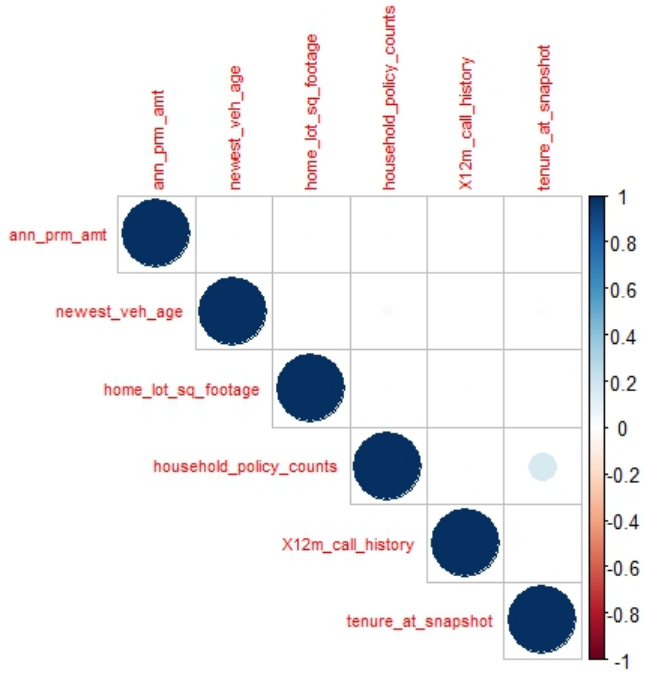
ann_prm_amt	newest_veh_age	home_lot_sq_footage	household_policy_counts	X12m_call_history
0.0009293953	-0.0030184309	0.0009486643	-0.0033470952	0.2799527640
tenure_at_snapshot				
-0.0014746341				

Asscoiation between the continuous variables and call count

- All of the categorical explanatory variable have very weak associations with call count.

Variable Selection

Multicollinearity Check:



Correlation analysis between the continuous explanatory variables.

Categorical Encoding:

- Transformed categorical variables using one-hot encoding (dummy variables) to ensure compatibility with proposed model (LightGBM).

Response Variable Summary

Key Insights:

- **Response Variable:** `call_counts`
  - Represents the count of calls for each customer.
- **Percentage of Zero Values:**
  - Approximately 50.18% of customers have zero call counts in the training dataset.

```
Percentage of zero values in call_counts: 50.18 %
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    0.00    0.00   25.91   44.00  294.00
```

Distribution indicates potential sparsity in the response variable.

Methods Considered

Tree-Based Models:

- **Random Forest:** Captures non-linear relationships and interactions robustly.
- **XGBoost:** Effective gradient boosting algorithm for structured data.
- **Light Gradient Boosting Machine(LightGBM):** Efficient histogram-based decision tree with lower memory consumption.

Zero-Inflated Models:

- **Zero-Inflated Poisson (ZIP):** Addresses excess zeros in count data.
- **Zero-Inflated Negative Binomial (ZINB):** Handles overdispersion and excess zeros.
- **Hurdle Model:** A two-part model designed to separately handle the zero and non-zero counts

Chosen Method: LightGBM

LightGBM is a gradient-boosting framework that integrates decision trees (weak learners) sequentially using boosting.

- Boosting is an ensemble learning that trains weak learners iteratively on the dataset and combines them to form a strong learner

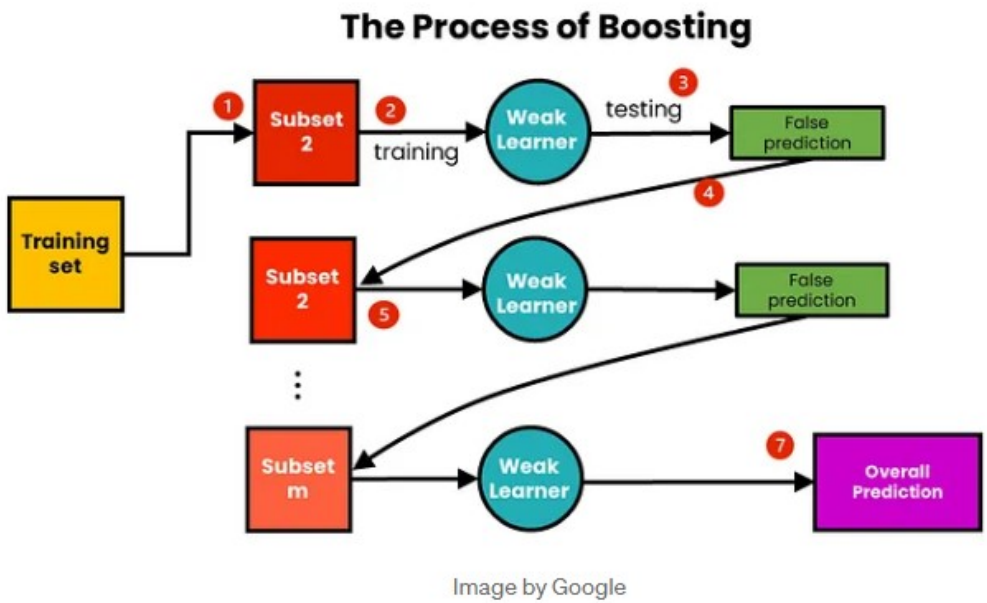


Image by Google

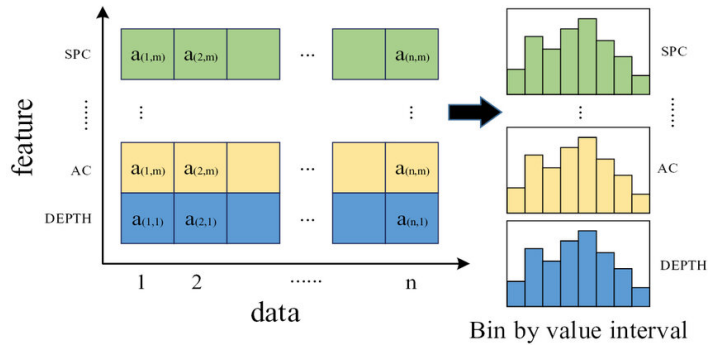
Each weak learner learn from the mistakes of the previous learners leading to a more accurate and robust final model

Chosen Method: LightGBM

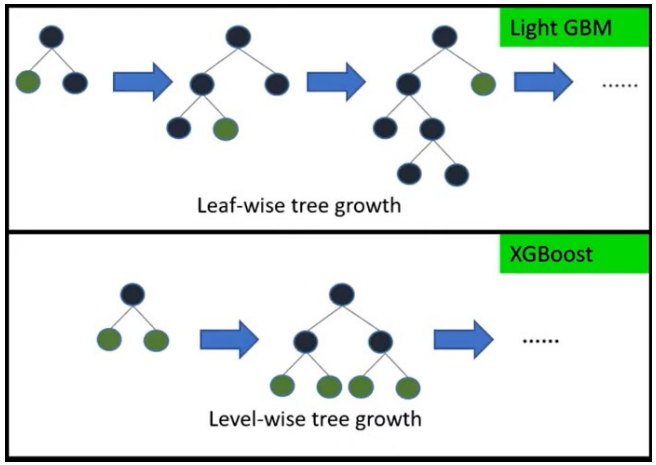
LightGBM Overview:

- Utilizes a histogram-based approach for faster training.
- Grows trees leaf-wise, which helps minimize loss more effectively.

Histogram-based Approach:



Leaf-wise Tree Growth:



LightGBM: How It Works

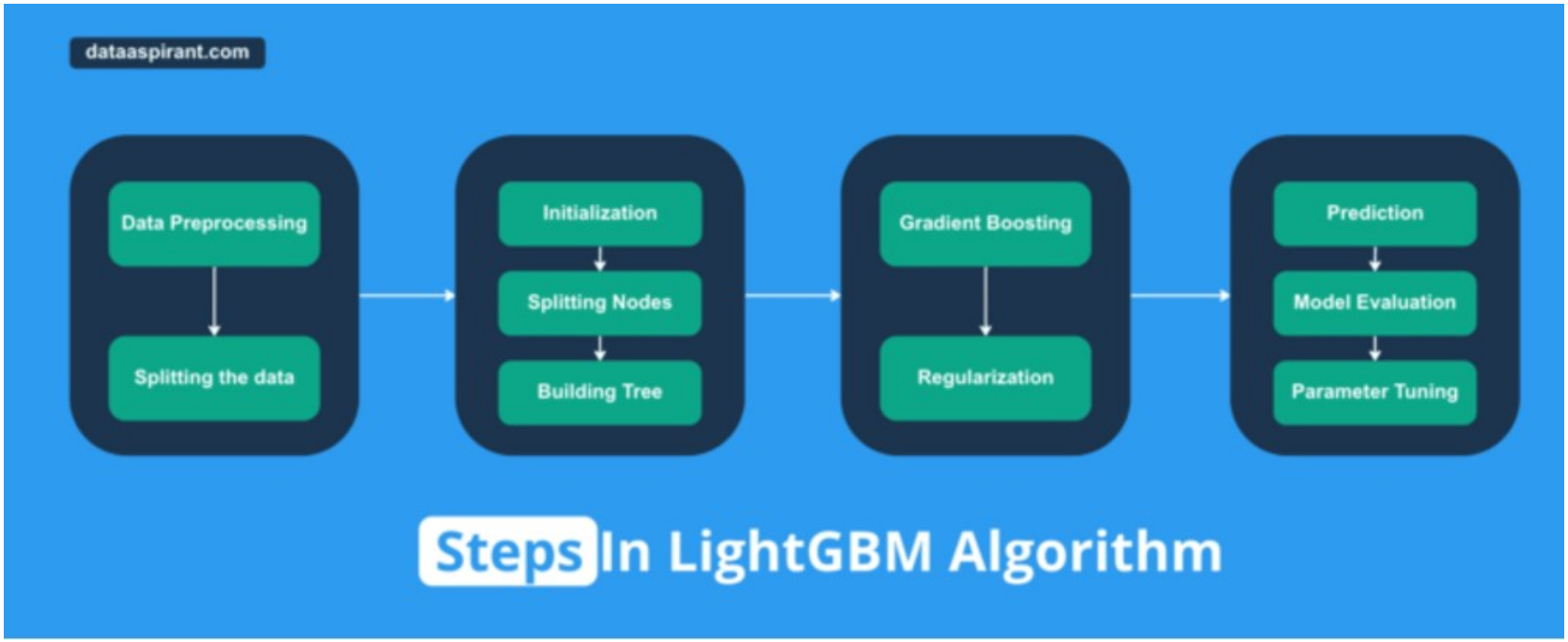
Step-by-Step Process:

- Starts with a single decision tree predicting the target variable based on input features.
- Iteratively adds more decision trees, each correcting the errors of the previous tree.

Key Features of LightGBM:

- **Decision Tree Learning:** Forms the backbone of the algorithm.
- **Histogram-Based Approach:** Accelerates the training process by binning continuous data.
- **Leaf-Wise Tree Growth:** Minimizes loss more effectively than level-wise growth.
- **Regularization:** Prevents overfitting and improves model generalization.

General Workflow of LightGBM



Model Evaluation

Metrics:

- RMSE, MAE: Measure prediction accuracy.
- Poisson log-likelihood: Validates count data assumptions.
- R-squared: Goodness-of-fit.

Validation Approach:

- Train-Valid split (80-20).
- Early stopping to avoid overfitting.

Feature Importance:

- LightGBM's feature importance analysis to identify influential predictors.

Hyperparameter Optimization: Steps Performed

Parameter Grid:

- Key Parameters: `num_leaves`, `learning_rate`, `feature_fraction`, `min_data_in_leaf`, `max_depth`.

Cross-Validation:

- 5-fold cross-validation using `lgb.cv`.
- Evaluation metric: Poisson loss.
- Early Stopping: Halted training if no improvement was observed within 50 rounds.

Selection of Best Parameters:

- Tracked the lowest cross-validation score to select optimal parameters.

Results: Best Parameters Identified

Optimal Parameters:

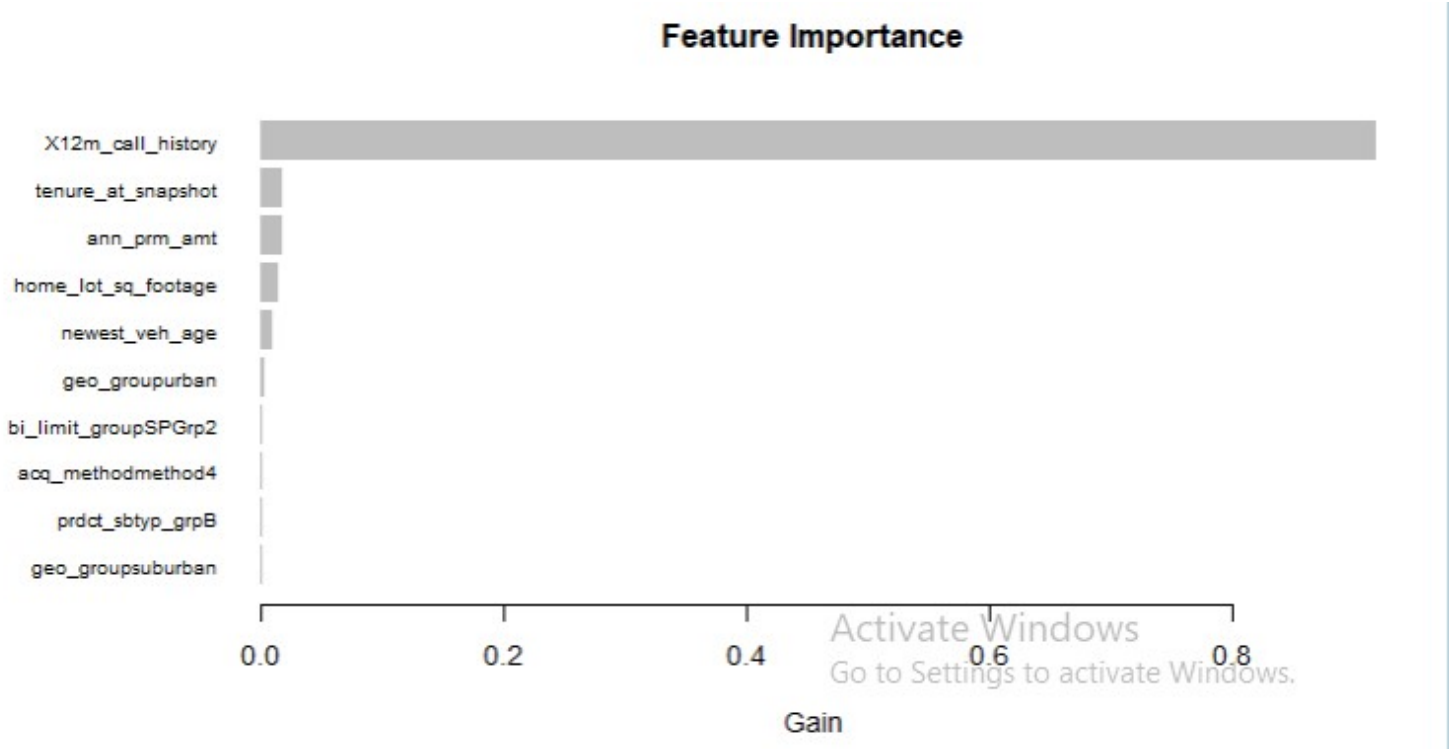
- num_leaves: 127
- learning_rate: 0.1
- feature_fraction: 0.9
- min_data_in_leaf: 100
- max_depth: 5

Training Performance Metrics

- RMSE: 35.68386
- MAE: 27.53179
- Poisson Log-Loss: 61.21314
- R^2: 0.1124

Feature Importance: Top Contributing Variables

Visualization:



Potentially Useful Variables

Demographic Factors:

- Education level, marital status, employment status.

Customer Behavioral Indicators:

- Payment history and risk profile, Duration of loyalty with the company, Frequency of customer service interactions.

Policy and Service Features:

- Insurance add-ons or optional coverages, Payment frequency (e.g., monthly or annually), Policyholder location.

Conclusion

Summary:

- LightGBM was selected for its efficiency, accuracy, and suitability for count data.
- Key predictors such as Past one year call count, Policy active length in month, Annualized Premium Amount, Square footage of the policyholder's home lot, age of the newest vehicle insured on a policy were instrumental in explaining call counts.
- Model evaluation metrics confirmed the robustness of the approach.

Future Work:

- Explore hyperparameter optimization (e.g., grid search).
- Incorporate other predictive features.