2024 Travelers University Modeling Competition: CloverShield Insurance Company Modeling Problem

LightGBM Modeling Approach

Oluwafunmibi Omotayo Fasanya, Augustine Kena Adjei

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.

Variable Selection Process

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 (e.g., customers without telematics data may have unique behaviors).
- Imputation techniques:
  - **Continuous variables:** Replaced missing values with the median, which is robust to outliers.
  - **Categorical variables:** Replaced missing values with the mode, representing the most likely category.

Variable Selection Process

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 correlation (0.28) with *call_counts* among continuous variables.
- Others variables has a very low association (less than 0.1) with call count.

Variable Selection Process

Categorical Encoding:

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

Multicollinearity Check:

- Examined correlation matrix for continuous variables to identify and address any strong correlations among predictors.

Methods Considered

Tree-Based Models:

- **Random Forest:** Captures non-linear relationships and interactions robustly.
- **XGBoost:** Effective gradient boosting algorithm for structured data.
- **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 (1/2)

Overview of LightGBM:

- LightGBM is a gradient-boosting framework that integrates decision trees (weak learners) sequentially using boosting.
- Unlike traditional methods, it grows trees leaf-wise:
  - Selects the leaf with the maximum delta loss to grow, minimizing training loss efficiently.
- Shares key advantages with XGBoost:
  - Regularization, sparse optimization, bagging, parallel training, and early stopping.
  - Flexibility to handle multiple loss functions, including Poisson for count data.

Chosen Method: LightGBM (2/2)

Why LightGBM?

- Designed for high-speed computation and low memory usage.
- Handles large datasets and categorical variables effectively.

Advanced Techniques:

- **Leaf-wise growth:** Focuses on growing deeper trees in areas requiring finer granularity for better accuracy.
- **GOSS (Gradient-based One-Side Sampling):** Accelerates training by prioritizing instances with larger gradient values.
- **Automatic binning:** Reduces memory consumption and improves efficiency by binning continuous features.

Model Evaluation

Metrics:

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

Validation Approach:

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

Feature Importance:

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

Potentially Useful Variables (Not in the Dataset)

Demographic Factors:

- Education level, marital status, employment status.
- Local economic conditions.

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).

Geographical and Economic Indicators:

- Policyholder location.
- Regional economic trends.

Conclusion

Summary:

- LightGBM was selected for its efficiency, accuracy, and suitability for count data.
- Key predictors such as *X12m_call_history* and behavioral factors 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 external features (e.g., market trends).