HOW TO REDUCE CALL CENTER COSTS?

CLOVERSHIELD INSURANCE COMPANY

- Dameesh
- Greeshma
- Manidweep
- Saumya

METHODOLOGY

01

Data Preparation

- **Handled missing value**
  Number -> median
  Categorical value -> mode

- **Label encoding** converted text-based categories into numbers

- **Feature engineering** created new columns like Premium per Policy, Calls per Tenure

02

Target Variable & Cross-Validation

- **Features | Target Variables**

- **Stage 1** – Call count zero or not

- **Stage 2** – Actual number of calls

- **1000-Fold Cross Validation**

03

Modeling

- Binary Classification – Stage 1

- Regression Model – Stage 2

- Tuned on parameters like learning rate and regularization to prevent overfitting

METHODOLOGY

04

05

06

Feature Importance

- Identified the **7 most important features** that significantly influenced predictions
- Selected based on their **importance scores**, combining results from both models.

Validation & Testing

- **Root Mean Squared Error (RMSE)**: Measures prediction accuracy.
- **R^2 Score**: Shows how well the model captures variability in the data.
- These metrics helped us ensure the model was reliable and not overfitting.

Predictions

- After training, we used the models to predict call counts for the test dataset.
- To finalize, we made sure all predictions were non-negative and saved them for submission.

MODEL SELECTION

XGBoost

High performance, robust handling of various data types, provides clear feature importance, and effectively manages missing values.

CatBoost

Excellent for categorical data but slightly less optimal than XGBoost for our use case.

LightGBM

Generally, offers higher error rates in our tests, though efficient in large datasets.

Random Forest & Deep Learning

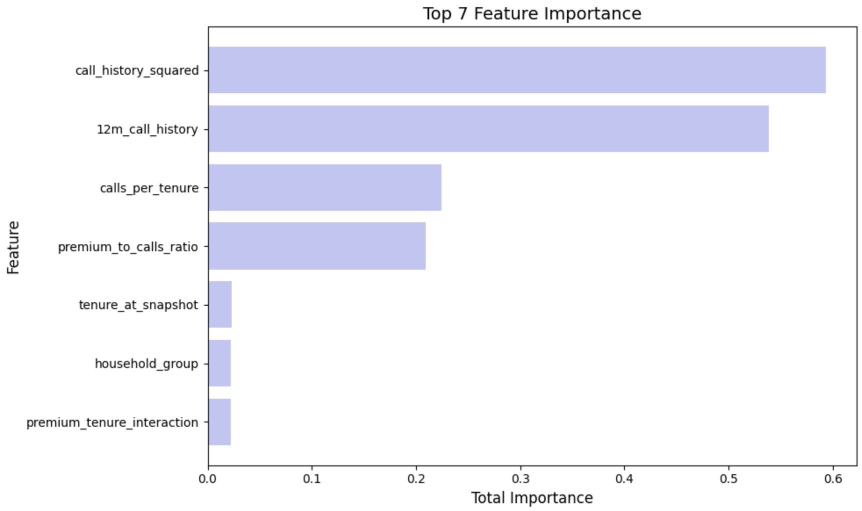
Powerful, but complexity can hinder interpretability.

24/7

XGBoost!

Superior performance and ability to deliver insights.

FEATURE ENGINEERING AND SELECTION



- **12m_call_history:** Indicates previous engagement patterns.
- **call_history_squared:** Captures non-linear effects of call history.
- **premium_to_calls_ratio:** Reflects the value of calls relative to premium amounts.
- **calls_per_tenure:** Measures call frequency relative to the duration of the policy.
- **tenure_at_snapshot:** Length of time the policy has been active.
- **premium_tenure_interaction:** Examines the relationship between premium and tenure.
- **ann_prm_amt:** Annualized premium amount influencing policyholder behavior.

MODEL EVALUATION

Root Mean Squared Error

Achieved an RMSE of 34.94, indicating strong predictive performance.

R^2 Value

Measures how well our model explains the variability in call counts; a higher value indicates a better fit.

Overfitting Assessment

Monitoring RMSE differences to ensure the model generalizes well to unseen data.

How will the predictions be useful?

1 Resource Allocation

Anticipate call volume to predict staffing

2 Potential Fraud Detection

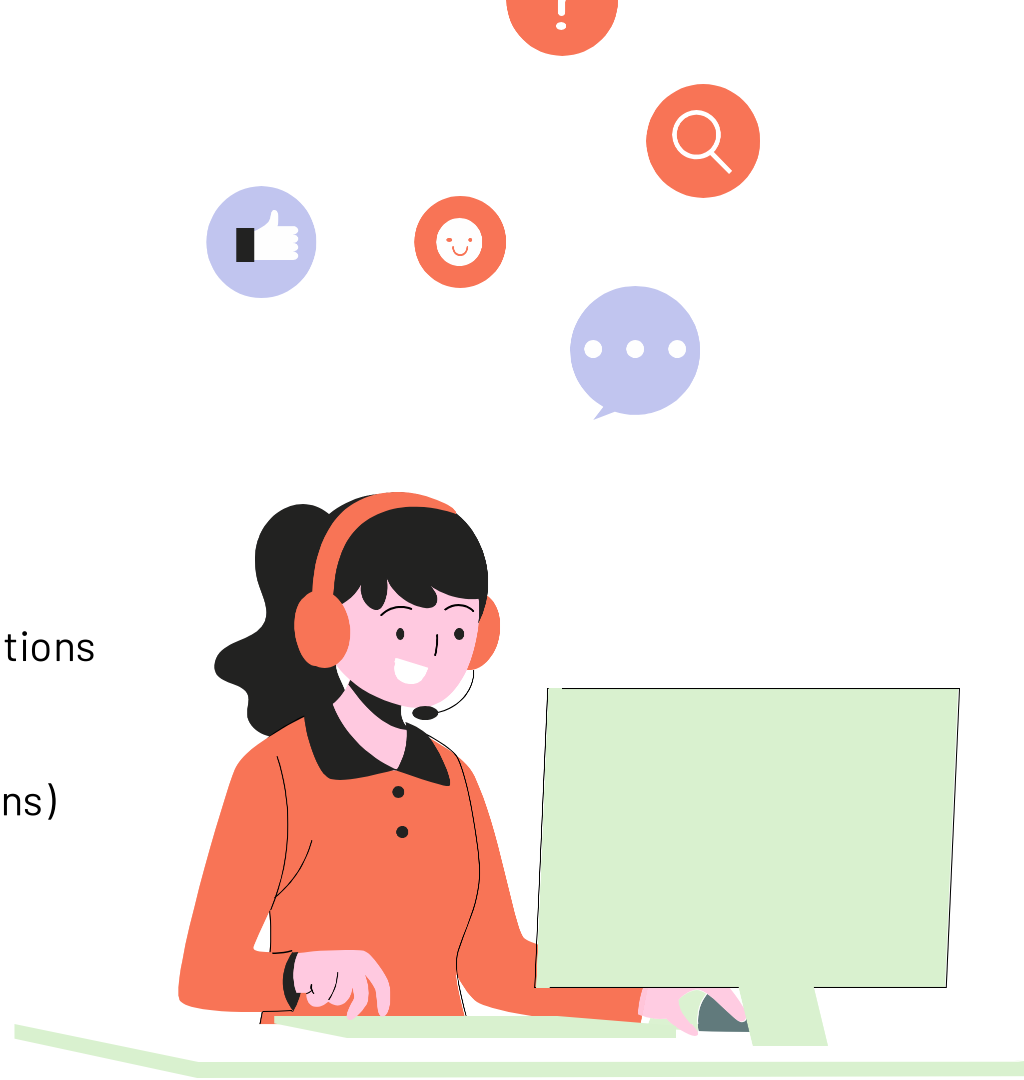
Unusual spikes can indicate fraudulent activity.

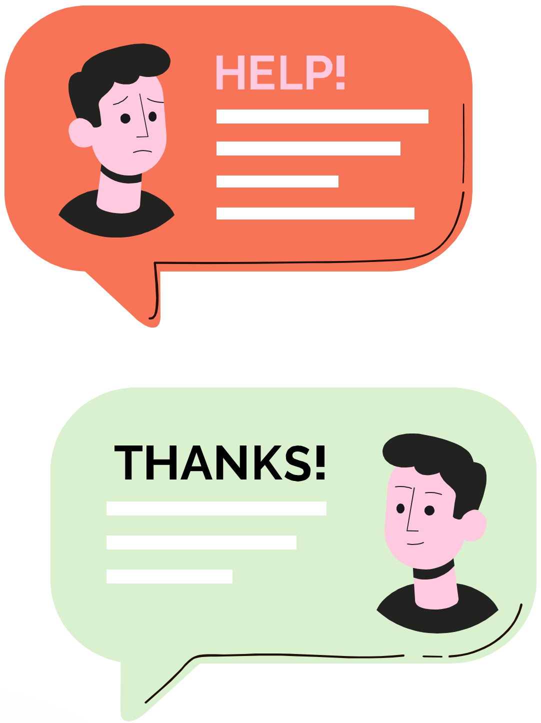
3 Business Planning

It can help in setting targets, managing budgets and making strategic decisions.

Other variables that might be useful

- Age of the policyholder
- Preferred language
- Nature of prior calls
- Seasonal Peaks-Local weather conditions
- Customer satisfaction score
- External Factors (Marketing campaigns)





Questions about the data

- Is the skewness in the target variable intentional?
- How is the target variable defined? Is it based on history, or any measure used?
- Are thy any known issues with the dataset like data entry errors?

THANK YOU!

Questions?