Predicting Customer Cancellations in Insurance: A Data Science Journey

Understanding the Problem

Customer cancellations are a major concern for insurance companies. Predicting which customers are likely to cancel can help businesses take proactive measures.

- The project aimed to **predict cancellations using machine learning**.
- The dataset included **three customer statuses**:
  - **Active (84.6%)**
  - **Full-Term Cancellation (8.2%)**
  - **Mid-Term Cancellation (7.3%)**
- The imbalance made accurate prediction difficult.

Data Preparation and Cleaning

Before building models, the data had to be cleaned and structured:

- **Removed irrelevant data** and missing values.
- **Converted categorical data into numerical** using **Label Encoding** to maintain efficiency.
- **Checked correlations** to eliminate redundant features.
- **Created five new variables** to capture important business insights.

First Attempt: Multiclass Classification

The initial approach attempted to **predict all three customer statuses** in one step using:

- Logistic Regression
- Linear Regression
- Random Forest
- LightGBM
- XGBoost

However, this method **failed due to overfitting** and **poor recall scores**, leading to a new approach.

New Approach: Two-Step Modeling

To **improve accuracy**, the problem was split into two steps:

1. **Step 1: Predicting Customer Cancellation (Yes/No)**
   - Converted the problem into a **binary classification** (Active vs. Canceled).

    - Significantly **reduced class imbalance** and improved model performance.
2. **Step 2: Classifying the Cancellation Type**
    - For customers who canceled, another model predicted **Full-Term vs. Mid-Term cancellations**.

Step 1: Cancellation Prediction

Feature Selection

To choose the most important variables, I tested:
✔ Feature Importance Ranking
✔ SHAP Values
✔ Recursive Feature Elimination (RFE)
✔ Principal Component Analysis (PCA)

SHAP-based feature selection gave the best results.

Model Performance

LightGBM outperformed other models:

- **Precision:** 84% (No Cancellation), 82% (Cancellation)
- **ROC-AUC (Test Set):** 0.70
- **ROC-AUC (Holdout Set):** 0.65 (no overfitting)

Efforts to improve performance using **SMOTE, weighting, Lasso, and Ridge regularization** were unsuccessful. Instead, **creating business-driven variables** led to the best improvements.

Step 2: Cancellation Type Prediction

Challenges

- **Insufficient data for minority classes** (Mid-Term cancellations).
- **ZIP Code and Policy Effective Date** were the strongest predictors.
- **Precision remained low**:
    - **Full-Term Cancellation:** 63%
    - **Mid-Term Cancellation:** 59%

More data would be needed to **significantly improve Step 2 performance**.

Alternative Approach: Customer Segmentation

Due to time constraints (**only two months for the project**), I explored **K-Means Clustering** to segment customers based on:

- **Vehicle Age**

- **Insurance Premium Paid**
- **Customer Tenure**

This approach identified **four distinct customer segments**, providing useful business insights.

What Could Be Improved?

If given more time, I would:

📌 **Gather More Data** – Collect additional details on mid-term cancellations and external economic factors.

📌 **Improve Step 2 Model** – Test advanced balancing techniques and integrate both steps into a single predictive model.

📌 **Enhance Feature Engineering** – Develop more business-driven variables for better prediction.

Final Takeaways

This project reinforced key **machine learning and business insights**:

✅ **Binary classification worked better than multiclass prediction.**

✅ **LightGBM was the best-performing model.**

✅ **Feature engineering was more effective than resampling techniques.**

✅ **Customer segmentation provided additional strategic insights.**

While predicting cancellations remains complex, **machine learning can help businesses take proactive measures to retain customers**.