# PREDICTIVE RISK MODELING FOR LIFE INSURANCE

Yenying Chen

# OVERVIEW

**DATA PREPARATION**

CLEANING,
HANDLING MISSING VALUES,
STANDARDIZATION

**MODELING & EVALUATION**

FEATURE ENGINEERING,
MODEL PERFORMANCE,
EVALUATION METRICS

**KEY INSIGHTS & STRATEGIES**

BUSINESS INSIGHTS,
ACTIONABLE RECOMMENDATIONS,
FUTURE ENHANCEMENTS

# OBJECTIVES

- Improve risk assessment accuracy
- Optimize the underwriting process to increase the proportion of automated underwriting
- Reduce long-term claim risks caused by underwriting errors
- Balance risk management and premium revenue

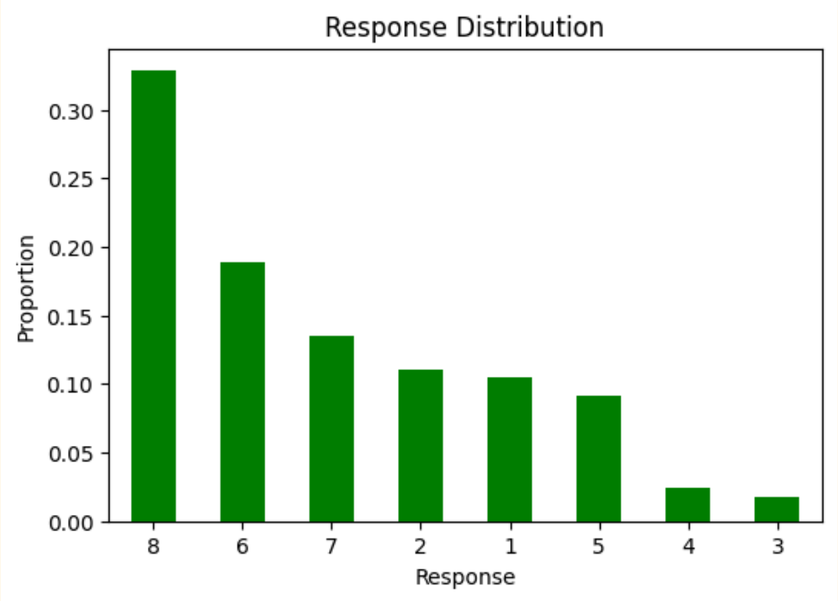
# DATASET

Prudential Life Insurance Assessment
train: 59,381 train rows, 128 features
test: 19,765 test rows, 127 features
imbalanced dataset

# DATA CLEANING & FEATURE ENGINEERING

- **Data Cleaning**
  - Remove features with more than 50% missing values
  - Fill missing values using median/mode
  - Numerical Variables: Standardization
  - Categorical Variables: Label encoding

- **Feature Transformation and Engineering**
  - Add interaction features to strengthen relationships
  - Merge multiple indicators to compute a composite risk score



# MODEL PERFORMANCE & EVALUATION

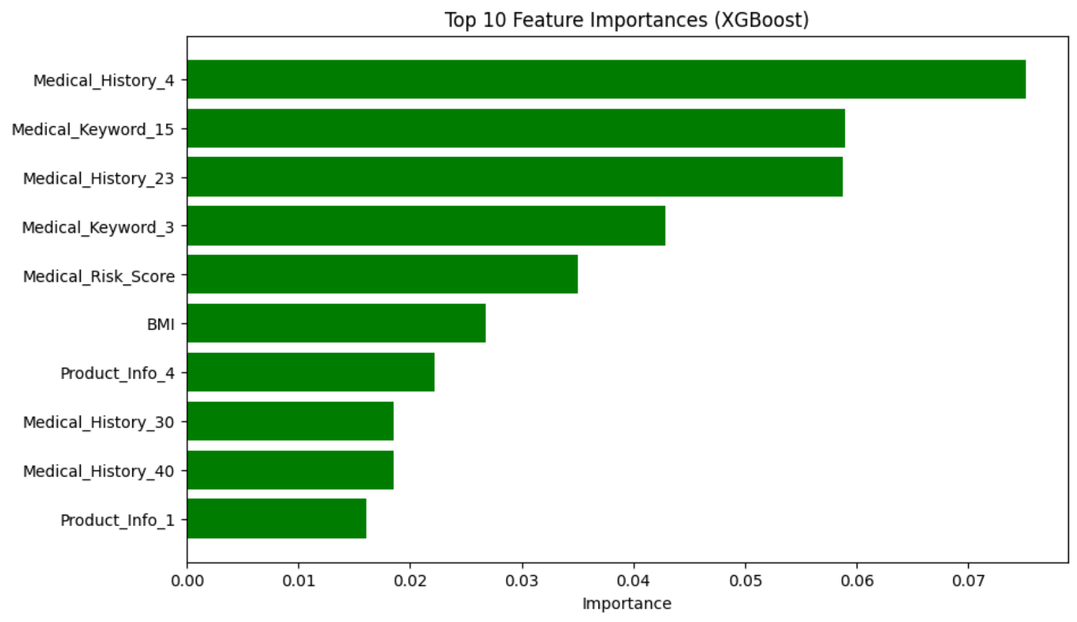
| Metric / Model | Logistic Regression | Random Forest | XGBoost | XGBoost_best |
|---|---|---|---|---|
| **Balanced Accuracy** | 0.286 | 0.318 | 0.381 | 0.380 |
| **ROC-AUC** | 0.801 | 0.825 | 0.842 | 0.847 |
| **Confusion Matrix - Recall** (Classes 6–7) | 71.13% | 73.92% | 73.44% | 73.82% |

# MISCLASSIFICATION FINANCIAL IMPACT

| Metric / Model | Logistic Regression | Random Forest | XGBoost | XGBoost_best |
|---|---|---|---|---|
| High-risk misclassified cases (Class 6-7 misclassified as 0-5) | 3,766 | 3,367 | 3,727 | 3,740 |
| Estimated future claims loss ($50,000/person) (thousands) | $188,300 | $168,350 | $186,350 | $187,000 |
| Revenue loss due to customer churn (10-year LTV) (thousands) | $4,033 | $4,614 | $3,647 | $3,790 |
| Total financial impact (Claims + Churn) (thousands) | $192,333 | $172,964 | $189,997 | $190,790 |

# KEY RISK DRIVERS & BUSINESS APPLICATIONS

- Medical History is a key factor in underwriting decisions, highlighting the need for improved completeness and accuracy of health data.
- Using BMI and medical history analysis to enhance policyholder risk segmentation.
- Developing more refined premium adjustment strategies for different risk groups.
- Monitoring policyholder behavior patterns to improve fraud detection accuracy.



# FUTURE OPTIMIZATION DIRECTIONS

**Operational Applications & Continuous Monitoring**

- Deploy models in underwriting systems for real-time risk assessment
- Dynamically adjust premium pricing and automated underwriting decisions based on risk evaluation results
- Flag high-risk cases to ensure manual review is more focused
- Establish data monitoring mechanisms, regularly adjusting decision thresholds to adapt to market changes (e.g., health data, medical trends)

**Model Optimization**

- Improve recall for low-risk customers to reduce wrongful underwriting rejections
- Apply ensemble learning to further optimize prediction capabilities
- Continuously fine-tune hyperparameters to enhance model accuracy
- Periodically retrain the model to maintain underwriting decision stability and financial control

# THANK YOU

## Q&A

# APPENDIX: COMBINED RECALL FOR CLASS 6 & 7

**WEIGHTED RECALL= (TP6+TP7)/(TP6+FN6+TP7+FN7)**

- LOGISTIC REGRESSION

(2013 + 17560)/(2013 + 2099 + 17560 + 1929) = 71.13%

- RANDOM FOREST

(2192 + 18149)/(2192 + 2395 + 18149 + 1340) = 73.92%

- XGBOOST

(3059 + 17148)/(3059 + 2096 + 17148 + 2341) = 73.44%

- XGBOOST_BEST

(3056 + 17257)(3056 + 2053 + 17257 + 2232) = 73.82%

# APPENDIX: FINANCIAL IMPACT CALCULATION

**HIGH-RISK MISCLASSIFICATION (ACTUAL CLASS 6-7, PREDICTED AS 0-5)**

- ACTUAL CLASS 6 : 277 + 147 + 2 + 12 + 122 + 1,539 = 2,099
- ACTUAL CLASS 7：149 + 81 + 2 + 17 + 48 + 1,370 = 1,667
- TOTAL MISCLASSIFIED CASES：2,099 + 1,667 = 3,766
- COST CALCULATION：3,766 × $50,000 = $188,300,000

**LOW-RISK MISCLASSIFICATION (ACTUAL CLASS 0-1, PREDICTED AS 4-7):**

- ACTUAL CLASS 0：378 + 1,309 + 550 + 1,514 = 3,751
- ACTUAL CLASS 1：809 + 1,521 + 512 + 1,472 = 4,314
- TOTAL MISCLASSIFIED CASES：3,751 + 4,314 = 8,065
- CUSTOMER CHURN (10% CHURN RATE)：8,065 × 0.10 = 807
- CLV LOSS：807 × $500 × 10 = $4,033,000

**TOTAL：$188,300,000 + $4,033,000 = $192,333,000**