Surabaya, 13 January 2025

# Optimizing Reinsurance Decisions with Machine Learning

## Cost Efficiency and Risk Management

Muhammad Fa'iz Ismail
JCDS -- 0412

# Introduction

## Problem Statement

- **TravelSafe Insurance** faces high reinsurance costs due to reinsuring all policies **without** assessing risk levels.
- This **"one-size-fits-all"** approach leads to **inefficiencies**, such as large claims being uncovered and low-risk policies being unnecessarily reinsured.

## Goals

- **Predict** the **likelihood** of a policy filing a claim to focus reinsurance on high-risk policies.
- **Identify factors** driving claim likelihood to improve reinsurance strategies and risk management.

## Analytical Approach

- **Analyze historical** claims data to identify patterns and **build a classification model** to predict claim likelihood, enabling optimized reinsurance strategies.

# Why Use Machine Learning

## More Accurate For Prediction

- Can handle historical pattern analysis of millions of claims data
- Identify hidden patterns that are difficult for humans to see
- Real-time prediction updates

## Identification of Risk Factors

- Analyze hundreds of variables simultaneously
- Rank factors by importance
- Real-time prediction updates
- Actionable insights

## Business Benefits

**Cost** ↓
Reduction of unnecessary reinsurance costs

**Efficiency** ↑
Reduction of unnecessary reinsurance costs

**Profit** ↑
Better risk portfolio

# Metric Evaluation

## Business Metric

Reinsurance Efficiency Ratio **(RER)**

$$RER = \frac{Total\ Reinsurance\ Cost + Total\ Uncovered\ Claims}{Total\ Premiums}$$

**Interpretation**

- **Low RER** : Reinsurance risk management is more efficient.
- **High RER** : Too many large claims are not covered by reinsurance.

## Machine Learning Metrics

- **Type 1 Error** (False Positive):
  - Predicting claims when there are no claims.
  - Impact: Reinsurance costs increase.
- **Type 2 Error** (False Negative):
  - Predicting no claims when claims occur.
  - Impact: Large financial risk due to claims not being reinsured.
- **ROC-AUC** as a Key Metric, minimize False Negatives with also controlled False Positives

# Data Understanding

## Data Overview

- 44.238 travel insurance customers
- 10 columns (policies, sales, claims)

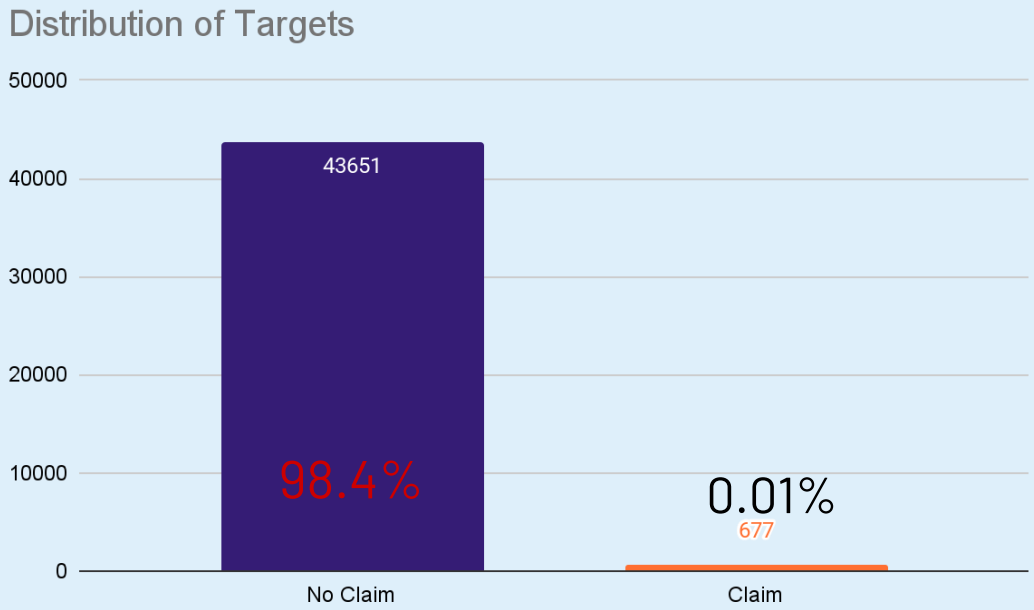
## Categories

- Customer Demographics:
  - Gender, Age, Destination
- Agency
  - Agency (name), Agency Type,
- Policy Details:
  - Duration, Product Name, Net Sales, Commission
- Distribution:
  - Channel and Agency Type Analysis
- Claims:
  - Binary indicator for modeling

## Data Features

- **Categorical Features:**
  - Agency: [EPX, CWT, C2B, JZI, SSI, LWC, RAB, TST, JWT, KML, ART, CCR, CSR, CBH, TTW, ADM]
  - Agency Type [Travel Agency, Airlines]
  - Distribution Channel [Online, Offline]
  - Product Name [Cancellation Plan, Value Plan, Gold Plan, etc.]
  - Destination [158 Countries]
- **Numerical Features:**
  - Net Sales [-5573 to 61.0]
  - Commission [0 to 263.5]
  - Duration [0 to 503]
  - Age [1-118 Years Old]
- **Target Variable:**
  - Claim [Yes, No]

# Exploratory Data Target (Claim)




- Standard evaluation metrics such as accuracy may not be relevant
- Special techniques are needed to handle imbalanced data
- Models can be biased towards the majority class (No Claim)

# Data Cleaning

**1. Missing Values**

- From the dataset, it was found that there were quite a lot of missing values. There were 31,647 data in the Gender column, this covers about 71% of the dataset, while All other columns have no missing values (0%).
- Plus there is no relationship between gender variables with other variable
- Finally, the deletion of the 'Gender' column was carried out, so we lost 1 feature

**2. Duplicated Values**

- Dataset contain 5004 rows containing duplicate values. The next step is to remove it because in Machine Learning modeling, duplicate data can cause bias in the model and slow down computation
- So now the dataset contain 39324 dataset with 9 feature and 1 target

**3. Identify Spelling Errors**

- Destination has quite a large number, it is very possible that there are errors in the spelling of country names.

# Data Cleaning

**4. Identify Anomaly Values**

4.1 Check Distribution (Numerical Variable)

- Filtering anomalous data, namely data in the Duration column of more than 4000 days and Age of less than 100
- Long-term visas such as work/study visas are usually a maximum of 2-5 years.(Reference: Schengen Visa Code)
- Most travel insurances have a maximum age limit of 75-85 years. Some companies offer up to 90 years with higher premiums (Reference: AIG Travel Guard and Travelex Insurance have a maximum age limit of 85 years)

4.2 Check Cardinality (Categorical Variable)

- The 'Claim' column, which is defined as the target, is very unbalanced, so a handling strategy is needed. No -> 0.982, Yes -> 0.017
- The 'Destination' column has too many unique values so that binning can be done.

# Data Transformation

**1. Binning (Grouping)**

- **Destination** Column:
  Simplify by grouping destinations into regions:
  0 = Asia, 1 = Africa, 2 = Europe, 3 = North America, 4 = South America, 5 = Oceania.
- **Age** Column:
  Group into age categories:
  0 = Kid (0–17), 1 = Young (18–30), 2 = Mature (31–50), 3 = Senior (51–60), 4 = Old (60+).
- **Duration** Column:
  Categorize trip durations:
  0 = <1 week, 1 = 1–2 weeks, 2 = 2–3 weeks, 3 = 3–4 weeks, 4 = 1–2 months, 5 = 2–3 months, 6 = >3 months.

**2. Scaling**

- **Net Sales** and **Commission (in value)** Column
- Apply **RobustScaler** or **PowerTransformer** due to non-normal distribution.

**3. Encoding**

- **One-Hot Encoding** (Both contain 2 unique values) :
  **Agency Type** Column
  **Distribution Channel** Column
- **Binary Encoding** (Contain multiple unique values.)
  **Agency** Column
  **Product Name** Column

# Benchmarking

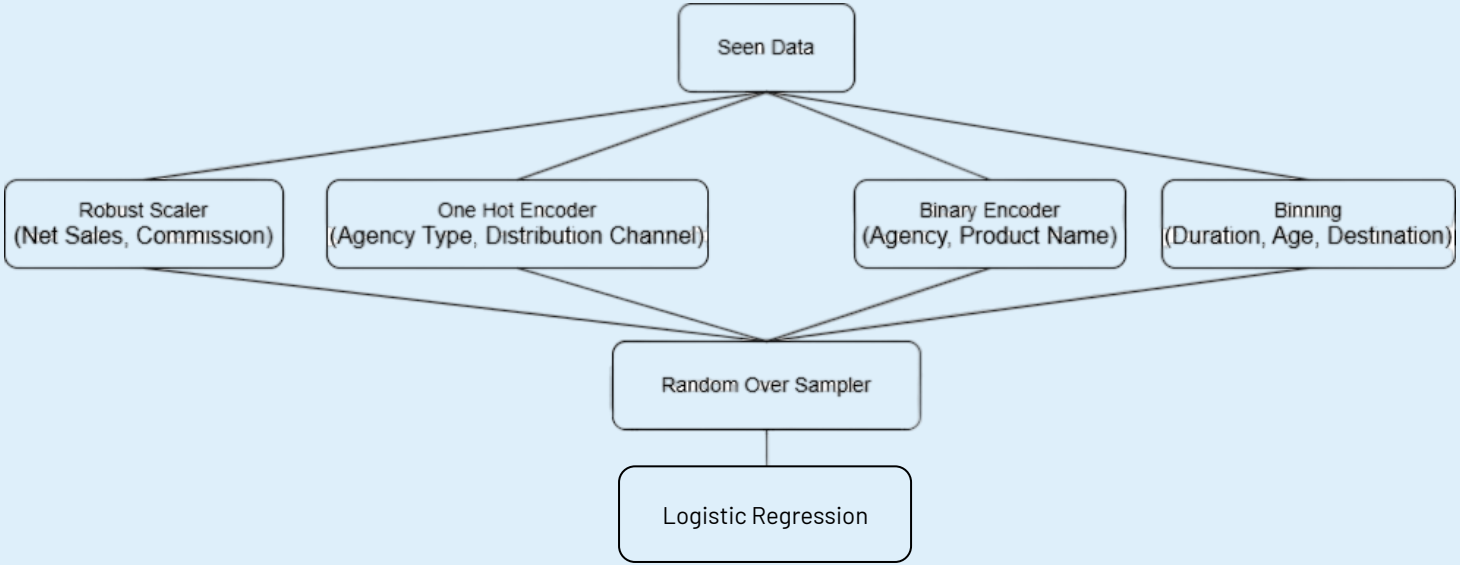
## Resampling

- **Oversampling**
  - Random Over Sampler
  - SMOTE
  - ADASYN
- **Undersampling**
  - NearMiss

## Modelling

- **Base Model**
  - Logistic Regression
  - KNN
- **Tree Based Model**
  - Decision Tree
  - Random Forest
- **Boosting Model**
  - XGBClassifier
  - LGBMClassifier
  - AdaBoost
  - GBC
- **Specific Imbalance Model**
  - Balanced Random Forest Classifier
  - Easy Ensemble Classifier
  - RUS Boost Classifier

# Best Model



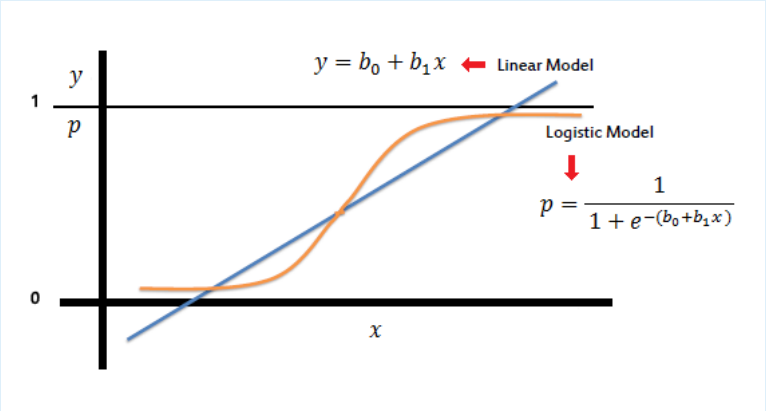

**ROC-AUC Score : 0.801 → Training dataset Stratified K-Fold**
**P-Value : 0,0000**

# What is Logistic Regression?

- Logistic Regression is a machine learning method used to classify data into two categories, like "yes" or "no".
- It works by calculating the probability of something happening, using a formula called the sigmoid function to give results between 0 and 1.

# How it Works?

- If the probability is 0.5 or higher, the result is classified as "yes"; otherwise, it's "no".
- It uses a mathematical formula to adjust and improve predictions based on the data.



# Hyperparameter Tunning

In the hyperparameter tuning process, a parameter search space is defined to optimize the classifier model's performance. The parameters being tested are:

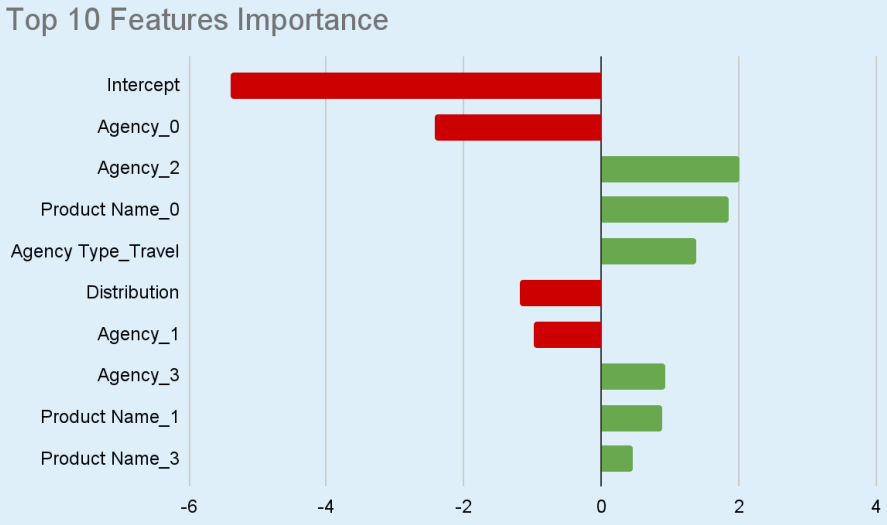
- **Regularization**
  - **C**: The regularization parameter with values **[0.001, 0.01, 0.1, 1, 10, 100]**, controlling the strength of regularization (smaller values mean stronger regularization).
  - **Penalty**: The type of penalty applied to the model's coefficients: **l1, l2, or elastic net**.
- **Solver Algorithm**
  - Different optimization algorithms (**lbfgs, liblinear, newton-cg, newton-cholesky, sag, and saga**) are tested to find the optimal parameters.
- **Maximum Iterations**
  - **max_iter**: The maximum number of iterations for model training, tested at **[100, 200, 300, 500].**
- **Multi-Class Scheme**
  - **multi_class**: Specifies the approach for multi-class classification **(auto, ovr, or multinomial).**
- **Class Weights**
  - **class_weight:** Handles data imbalance using the **balanced** scheme.

# Tunning Result

| | Training Dataset | Testing Dataset |
|---|---|---|
| Before Tunning | 0.8168 | 0.8095 |
| After Tunning | 0.8168 | 0.8091 |

# Feature Significant

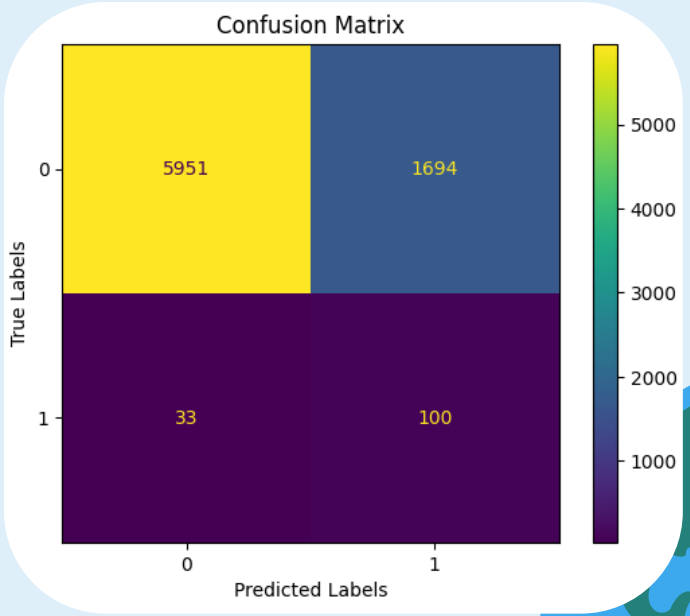
- All features in the dataset have a significant effect on the model.
- The P-Value  of all features < 0.05, which means the features are statistically significant at the 95% confidence level.



# Business Calculation with New Data

$$RER = \frac{Total\ Reinsurance\ Cost + Total\ Uncovered\ Claims}{Total\ Premiums}$$

- Assumption :
  - Average premium : Rp. 291.411
  - Average claim = Rp. 5.000.000
- Before using model
  - RER = ((5941 + 33 + 1694 + 100) + 0 ) / ((5941 + 33 + 1694 + 100) *5.000.000) = **0.3**
  - Loss = (5941 + 33 + 1694 + 100) *5.000.000) = **Rp.679.978.427**
- After using model
  - RER = ((1694 + 100) + 33 * 5.000.000 ) / ((1694 + 100) *5.000.000) + 0 = **0.14**
  - Loss = (694 + 100) *5.000.000) + (33 * 5.000.000) = **Rp.321.837.400**



# Conclusion

## Model

- Logistic Regression + Random OverSampler is the best combination, achieving a ROC-AUC of 0.800
- Top 10 significant features:
  - Agency_0: Largest negative influence on claims.
  - Agency_2, Product Name_0: Strong positive contributors.
  - Agency Type_Travel Agency, Distribution Channel_Online: Highlight key behavioral insights.

## Business

- Cost Savings: IDR 357,966,181 saved compared to operations without the model.
- Efficiency: Reinsurance Efficiency Ratio (RER) reduced from 30% to 14.21%, improving risk management.

# Recommendation

## Model

- **Increase data** for the "Yes" class to reduce imbalance and improve model robustness.
- **Data Quality**: Fix spelling inconsistencies in categorical features (e.g., Destination, Product Name).
- **Optimization**: Fine-tune prediction thresholds and retrain the model regularly with updated data.
- **Feature Expansion**: Include additional features or external data sources for better accuracy.

## Business

- The model can be **applied** on a limited basis (pilot project) with **strict** monitoring.
- **Monitor the prediction results** from this subset and use real-time data to refine the model if necessary.

# Thanks