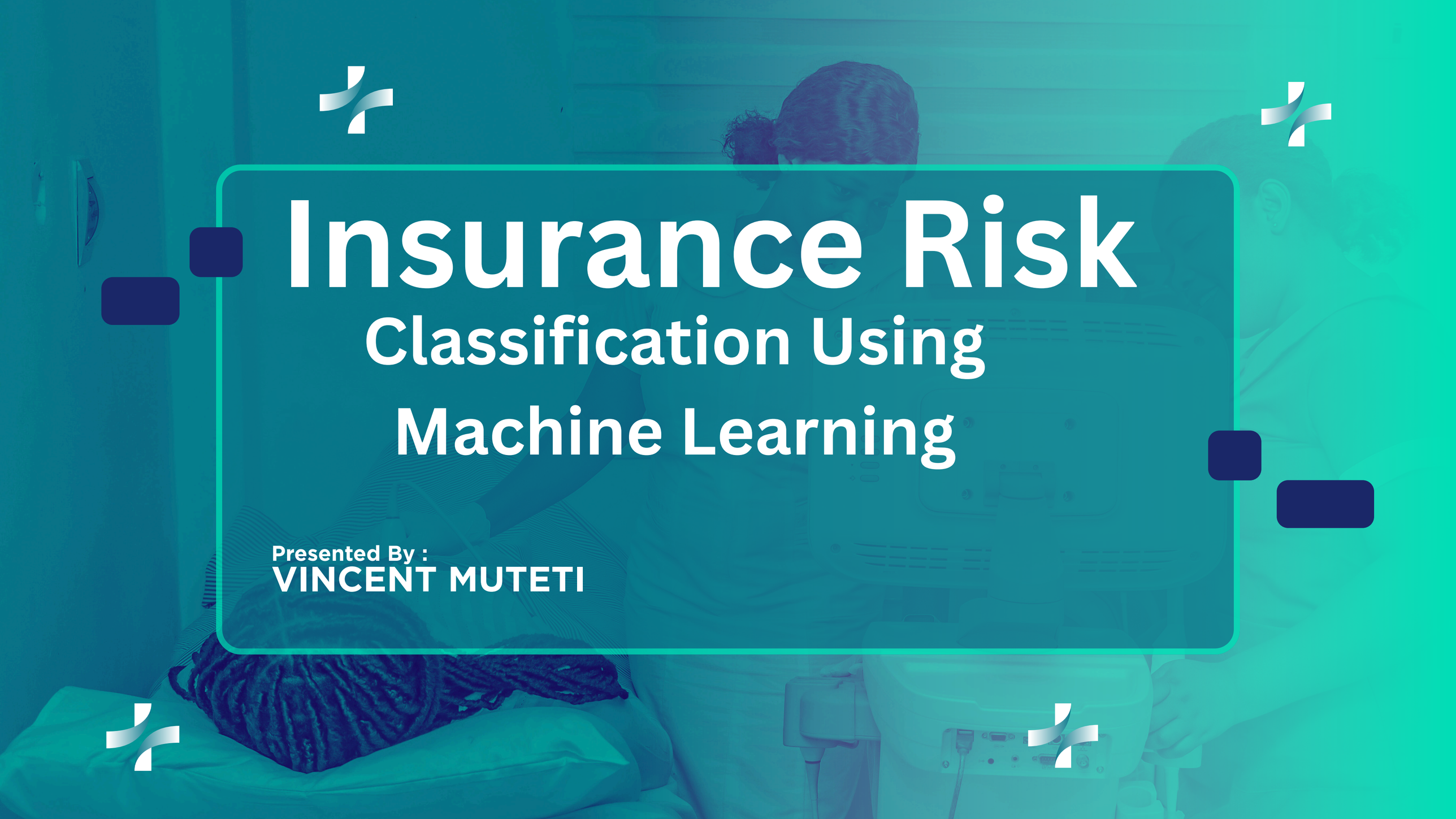


Insurance Risk Classification Using Machine Learning

Presented By :
VINCENT MUTETI

Business Context

Why This Problem Matters

Rising healthcare costs strain insurance providers

A small group of high-risk individuals drives a large share of costs

Late identification leads to reactive and expensive care

Early risk detection enables better cost control and care planning

Business Problem Statement

Problem Definition

Insurance providers struggle to identify high-risk individuals early

Reactive care leads to inefficient resource allocation

Manual or rule-based risk identification is insufficient

Can we accurately predict high-risk policyholders using historical data?

Business Objective

- Predict whether an insured individual is high-risk
- Enable:
  - Proactive care management programs
  - Risk-based pricing strategies
  - Reduction of avoidable medical costs

Dataset Overview

Data Description

- 100,000 insured individuals
- Features include:
  - Demographics
  - Lifestyle behaviors
  - Medical history
  - Insurance plan details
  - Healthcare utilization
- Target variable: High-risk status (Binary)

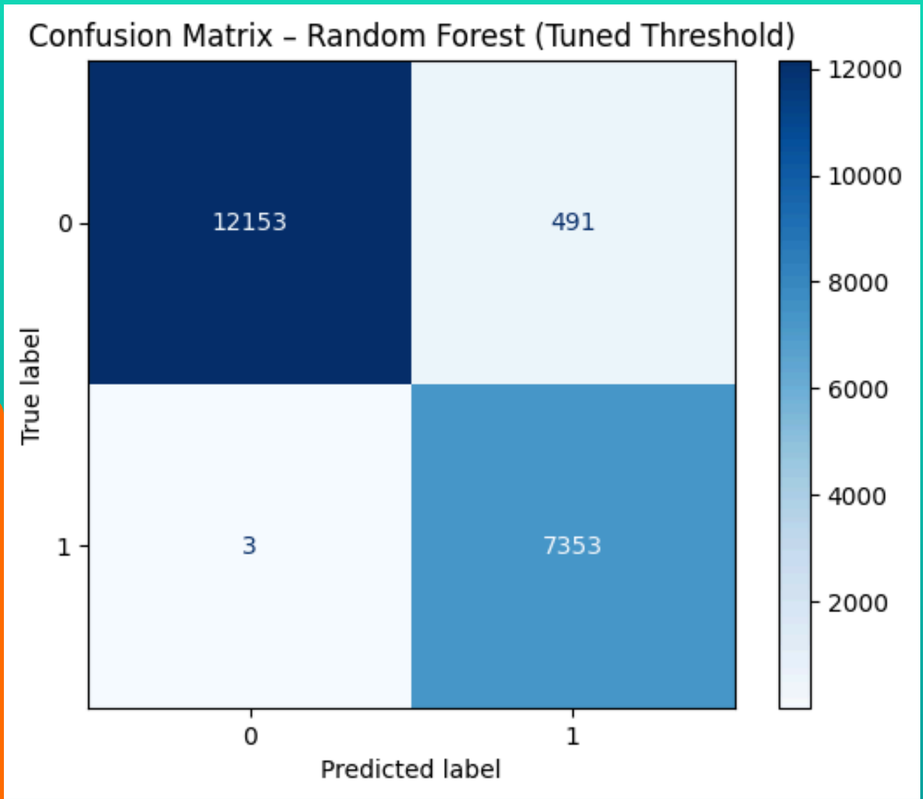
Exploratory Data Analysis

Key EDA Insights

- High-risk individuals show:
- Higher medical costs and utilization
- Greater prevalence of chronic conditions
- More frequent procedures and hospital visits
- Clear separation between high-risk and non–high-risk groups

Data Preprocessing

- Removed unique identifiers
- Handled missing values:
- Numerical: median imputation
- Categorical: most frequent value
- One-hot encoding for categorical variables
- Feature scaling for numerical variables
- Stratified train–test split
- Pipeline used to prevent data leakage



Baseline Model – Logistic Regression

- Logistic Regression used as interpretable baseline
- Accuracy: ~97%
- Recall (High Risk): ~97%
- Very few high-risk cases missed

High recall aligns with business goal of minimizing missed high-risk individuals

Model Interpretation

- Chronic disease burden
- Prior hospitalizations
- HbA1c and BMI
- Smoking status
- Healthcare utilization metrics

✓ Findings align with medical risk assessment standards

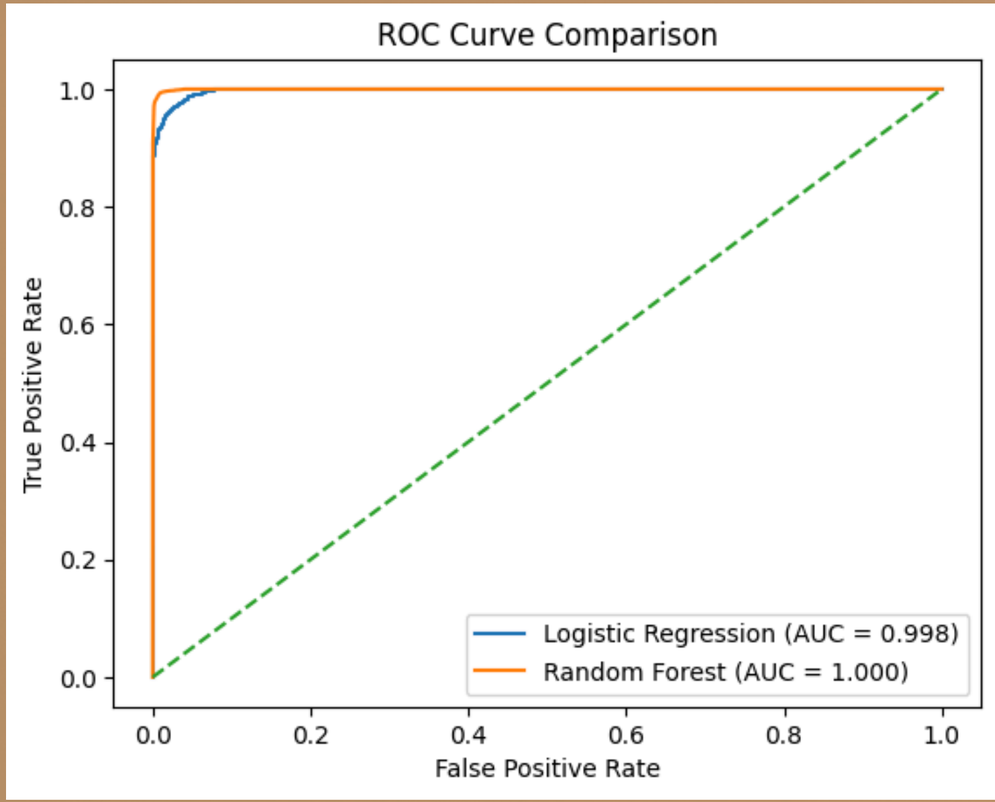
✓ Model is transparent and trustworthy

MODEL Comparison

Logistic Regression vs Random Forest

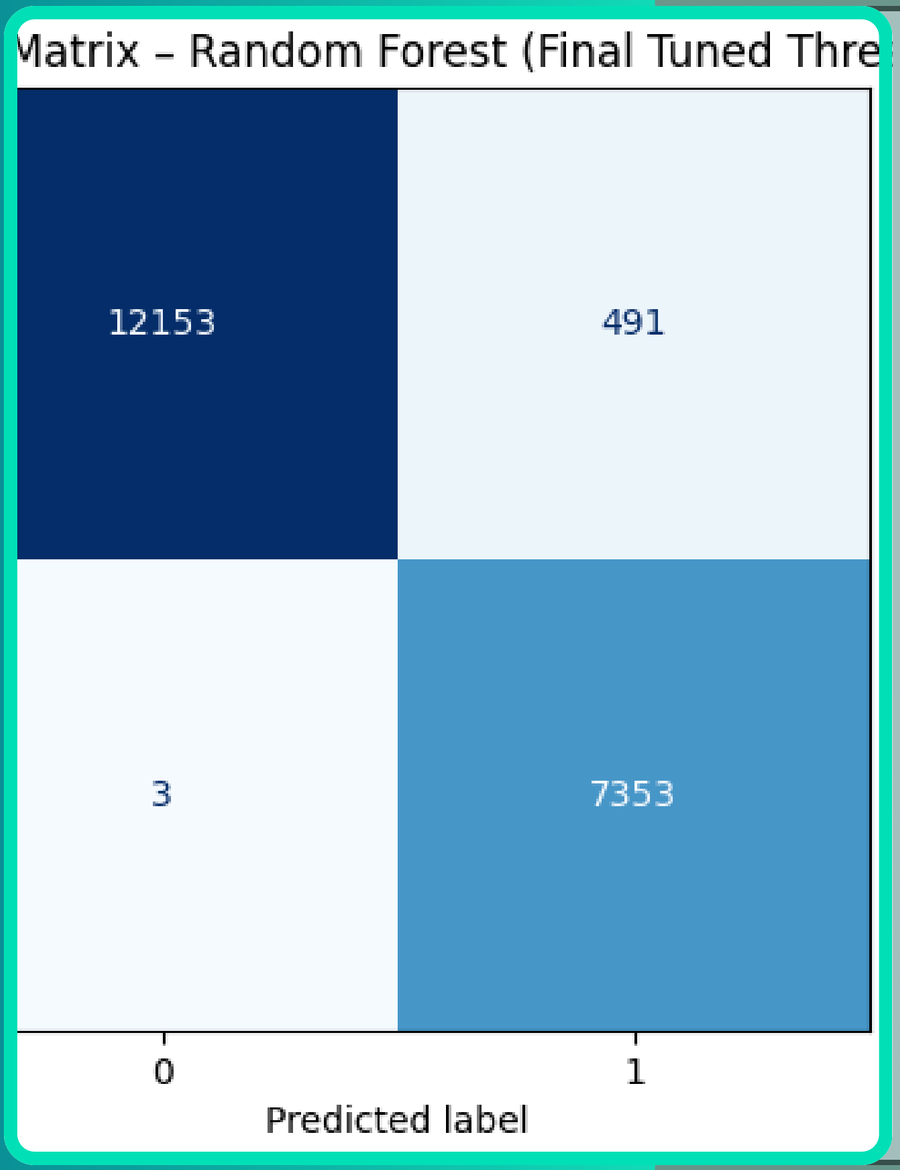
- Random Forest captures non-linear relationships
- Slight performance improvement over Logistic Regression
- Logistic Regression:
  - Highly interpretable
- Random Forest:
  - Better classification balance

ROC–AUC Performance



Model Discrimination Power

- Logistic Regression AUC: 0.998
- Random Forest AUC: 0.9997
- Both models show excellent ability to separate risk classes
- Random Forest performs marginally better



Threshold Optimization

- Aligning Model with Business Priorities
- Adjusted classification threshold to 0.4
- Increased recall for high-risk individuals
- Reduced false negatives
- Slight increase in false positives (acceptable trade-off)
- ✓ Early identification supports proactive care and cost containment

Final Model & Recommendations

- Final Decision & Business Impact
- Selected Model:
- Random Forest with tuned threshold (0.4)
- Key Metrics:
- ROC–AUC: 0.9997
- Recall (High Risk): ~99%
- Minimal false negatives

- Recommendations:
- Deploy model for early risk detection
- Integrate with care management workflows
- Use predictions for targeted interventions and pricing strategies