Predicting Health Insurance Lead Conversion Using Machine Learning

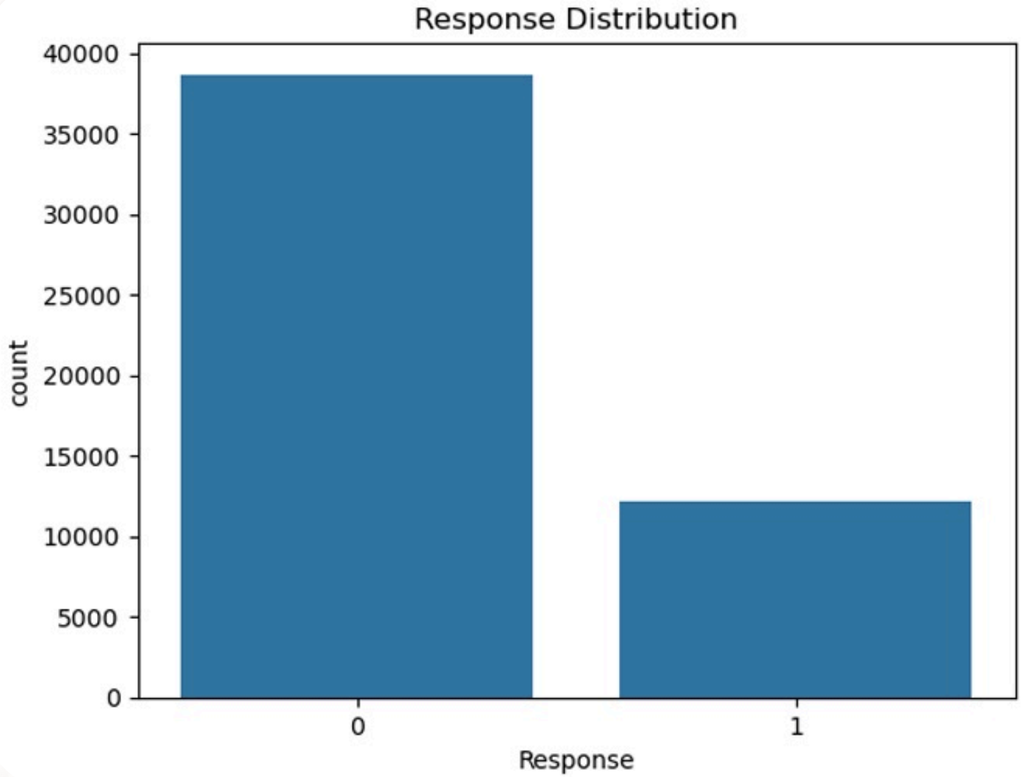
Rugare L. Madzara
MSc Applied AI and Business Analytics
GitHub link: https://github.com/rugaremadzara/Predicting-Health-Insurance-Lead-Conversion-Using-Machine-Learning

Problem statement

- Large volumes of leads
- Limited sales capacity
- Need to prioritize outreach efficiently

Data Overview

- Real-world insurance lead data (Kaggle)
- Customer demographics, policy attributes, pricing
- Binary outcome: conversion vs non-conversion



Business Objective

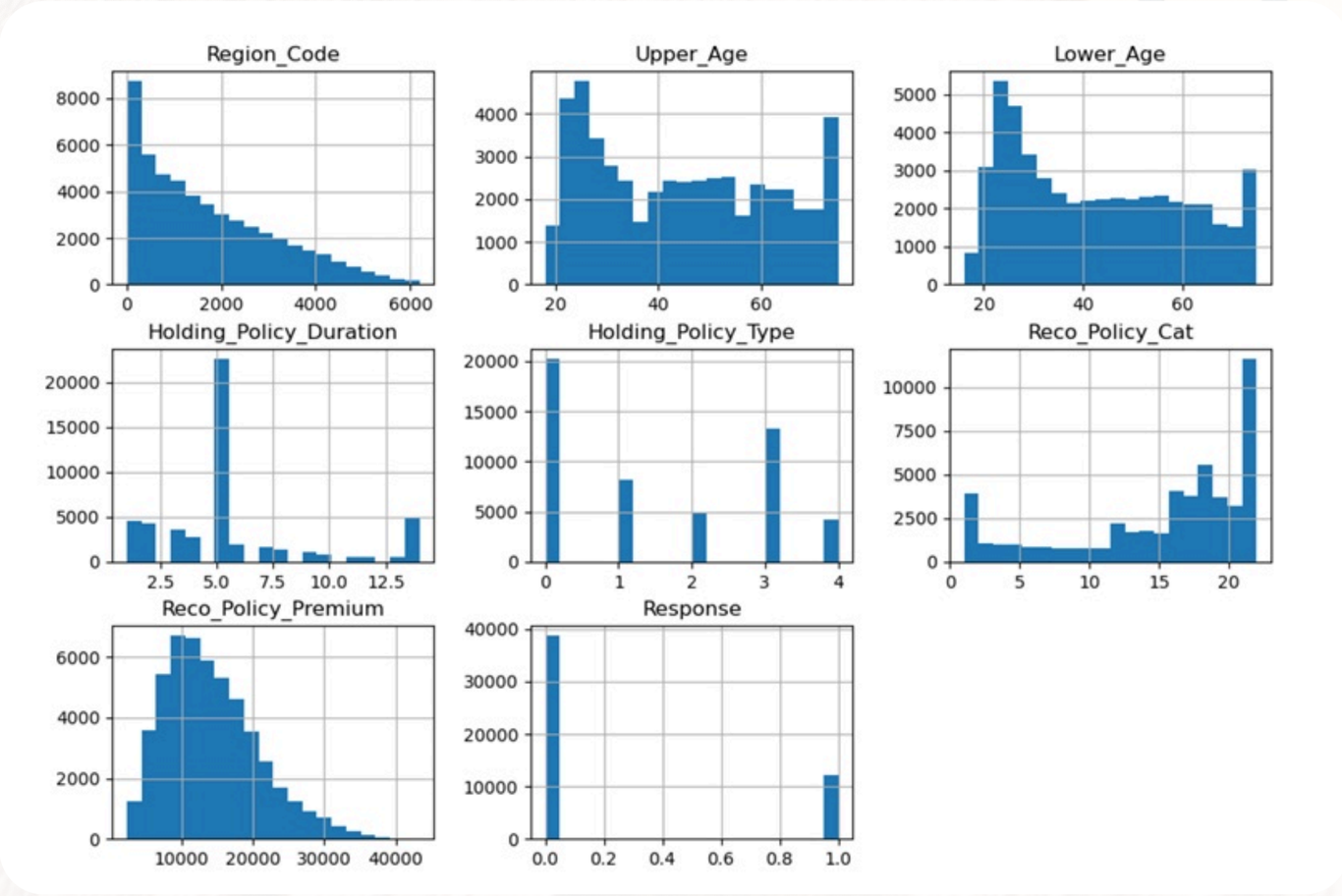
- Predict likelihood of lead conversion
- Rank leads by probability
- Support smarter outreach decisions

Exploratory Data Analysis (EDA)

Understanding Our Customer Data

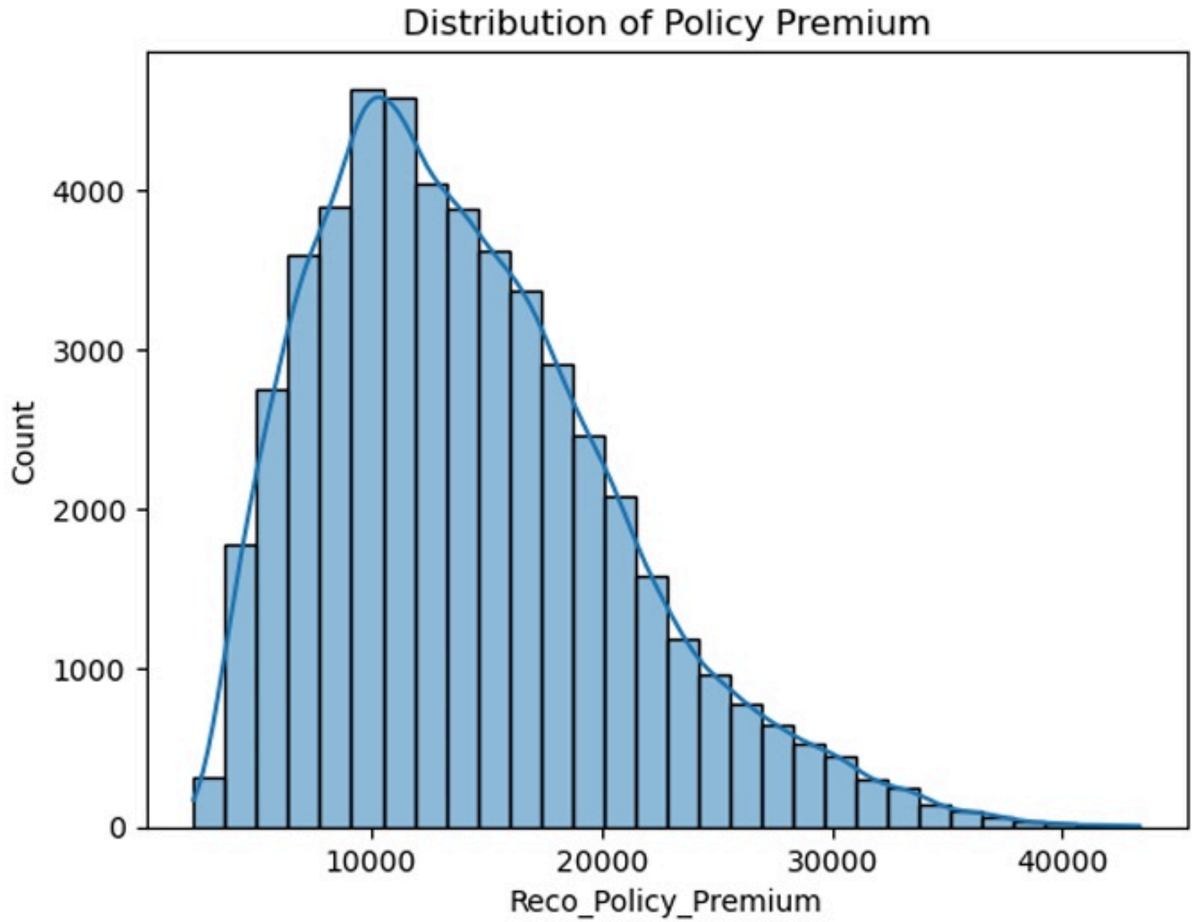
- Predict likelihood of lead conversion
- Rank leads Customer behavior varies widely across age, pricing, and policy duration
- Most customers do not convert, highlighting the need for prioritization
- Policy premiums and durations are not evenly distributed
- No single factor clearly determines conversion
- Decisions require combining multiple customer and policy signals
- by probability
- Support smarter outreach decisions

numerical variables



Exploratory Data Analysis (EDA)

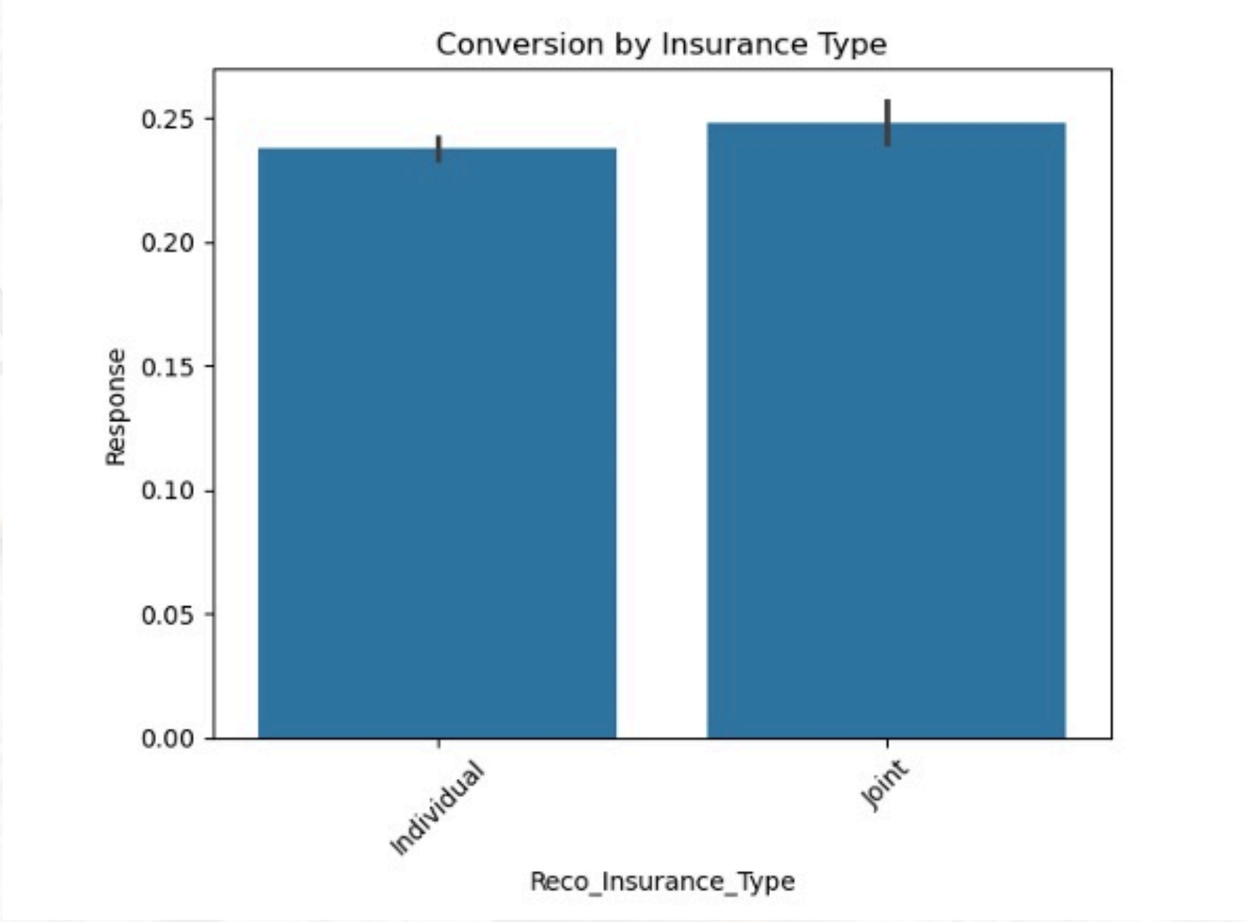
- Most customers fall into low to mid-range premium values
- A smaller group of leads have very high premiums
- Premium values are not evenly distributed
- Higher premiums do not automatically mean higher conversion
- Sales effort should reflect both likelihood to convert and deal value



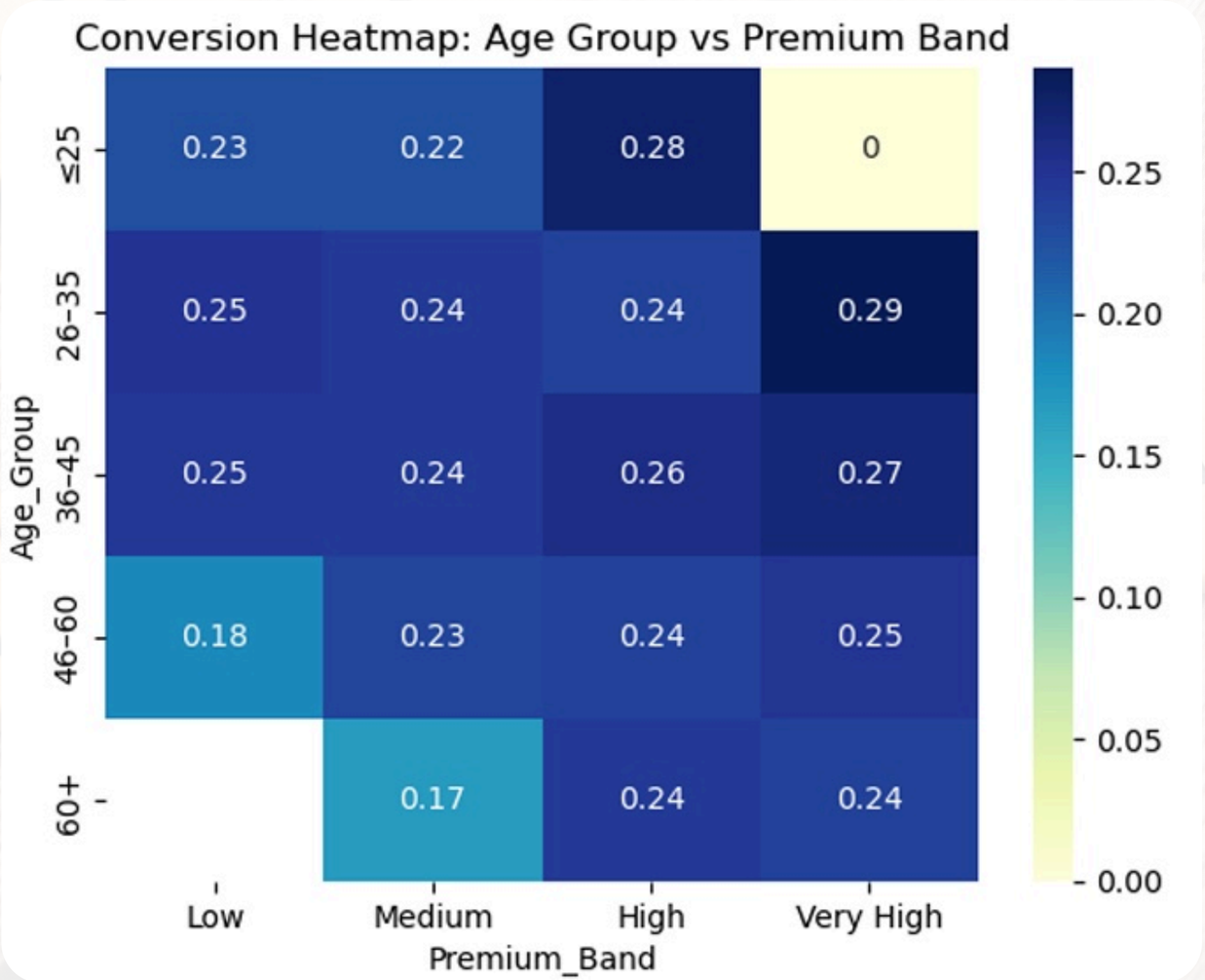
Exploratory Data Analysis (EDA)

Categorical variables

- Joint insurance policies show slightly higher conversion rates
- Individual policies still represent a large volume of conversions
- Insurance type alone does not strongly determine conversion
- Customer context matters more than product type alone
- Opportunity to tailor sales messaging by policy type



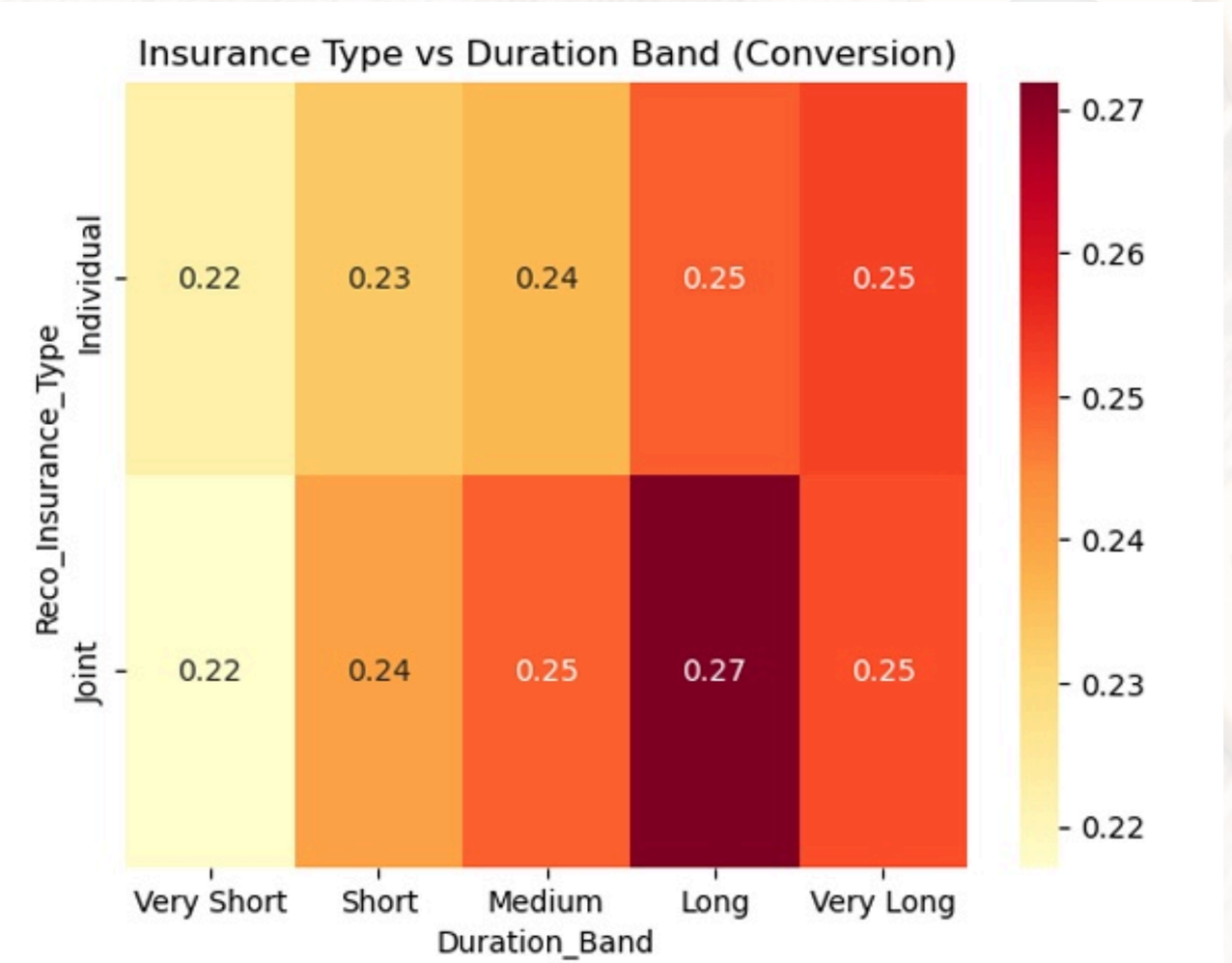
Variable Interactions



- Conversion depends on age and premium together, not individually
- Middle-age groups (26–45) convert best at higher premium levels
- Younger customers (≤25) respond better to mid-to-high premiums
- Older customers (60+) show lower conversion at low premiums
- One-size-fits-all pricing strategies miss conversion opportunities

Variable Interactions

- Conversion improves as policy duration increases
- Joint insurance plans outperform individual plans
- Long and very long durations show highest conversion
- Short-term policies convert the least
- Customers value long-term security, not short-term savings



Feature Engineering

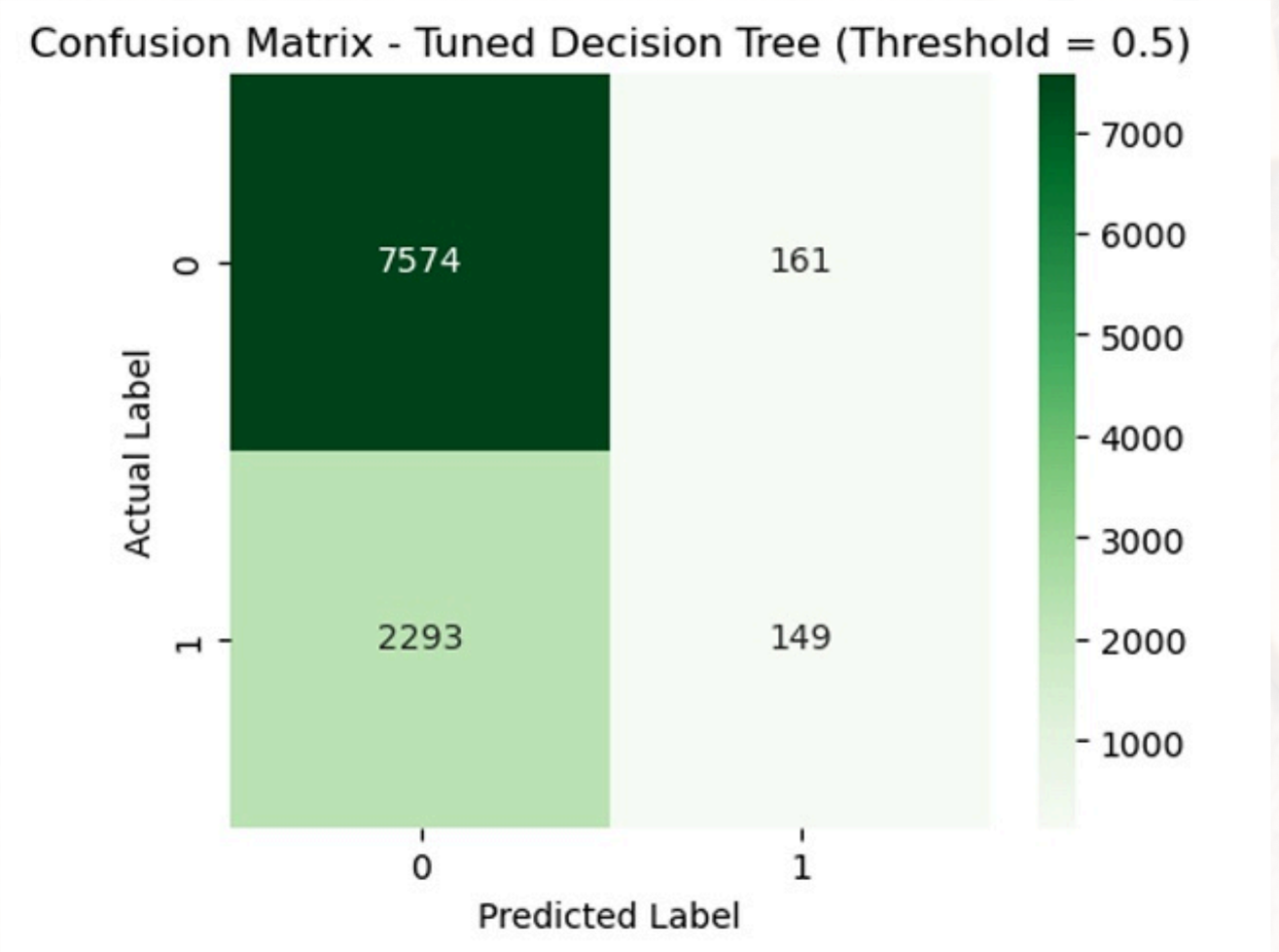
- Age groups and age ranges
- Premium bands
- Policy duration bands
- Combined categorical features

Modeling Strategy

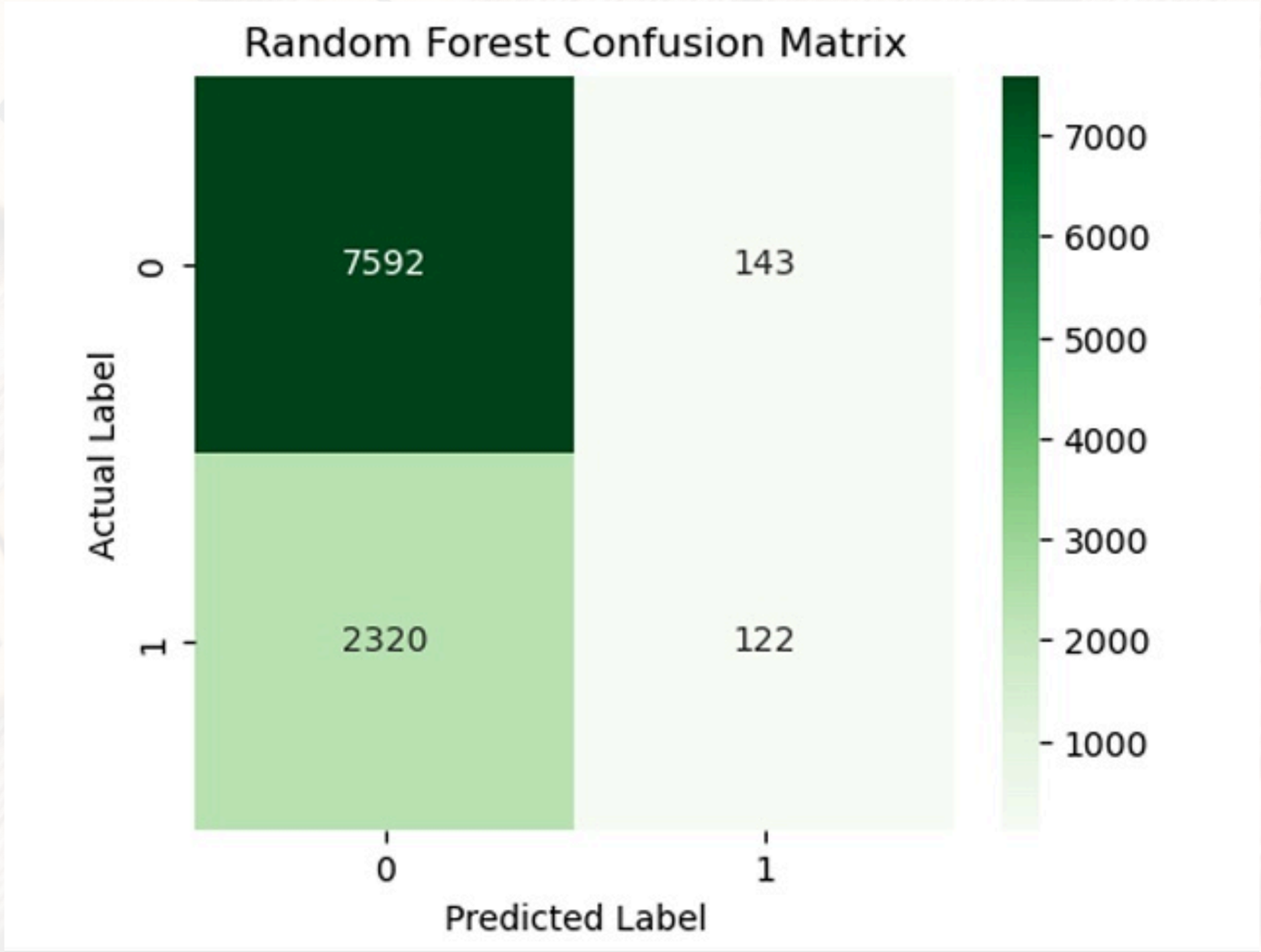
- Compared multiple models for reliability
- Evaluated how well models separate high- and low-quality leads
- Focused on ranking leads, not just yes/no predictions
- Used metrics aligned with sales decision-making

Decision Tree Classifier with Hyper parameter tuning

- Simple, rule-based model
- Optimized to avoid overfitting
- Easy to interpret decision logic
- Used as a benchmark and explanation tool

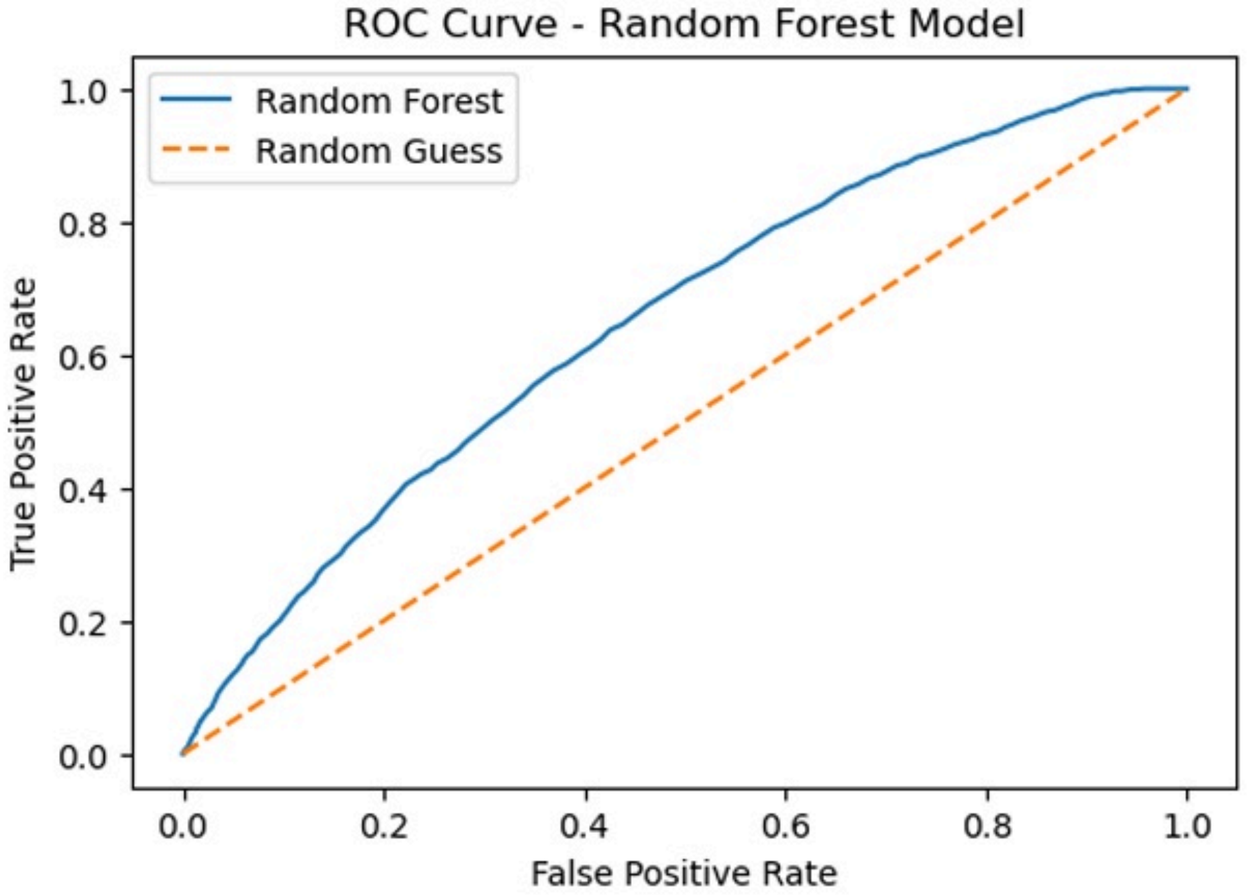


Random Forest



- Ensemble model combining multiple decision trees
- Strong performance in ranking leads by conversion likelihood
- Handles complex interactions between customer and policy factors
- Selected as the final model for lead scoring

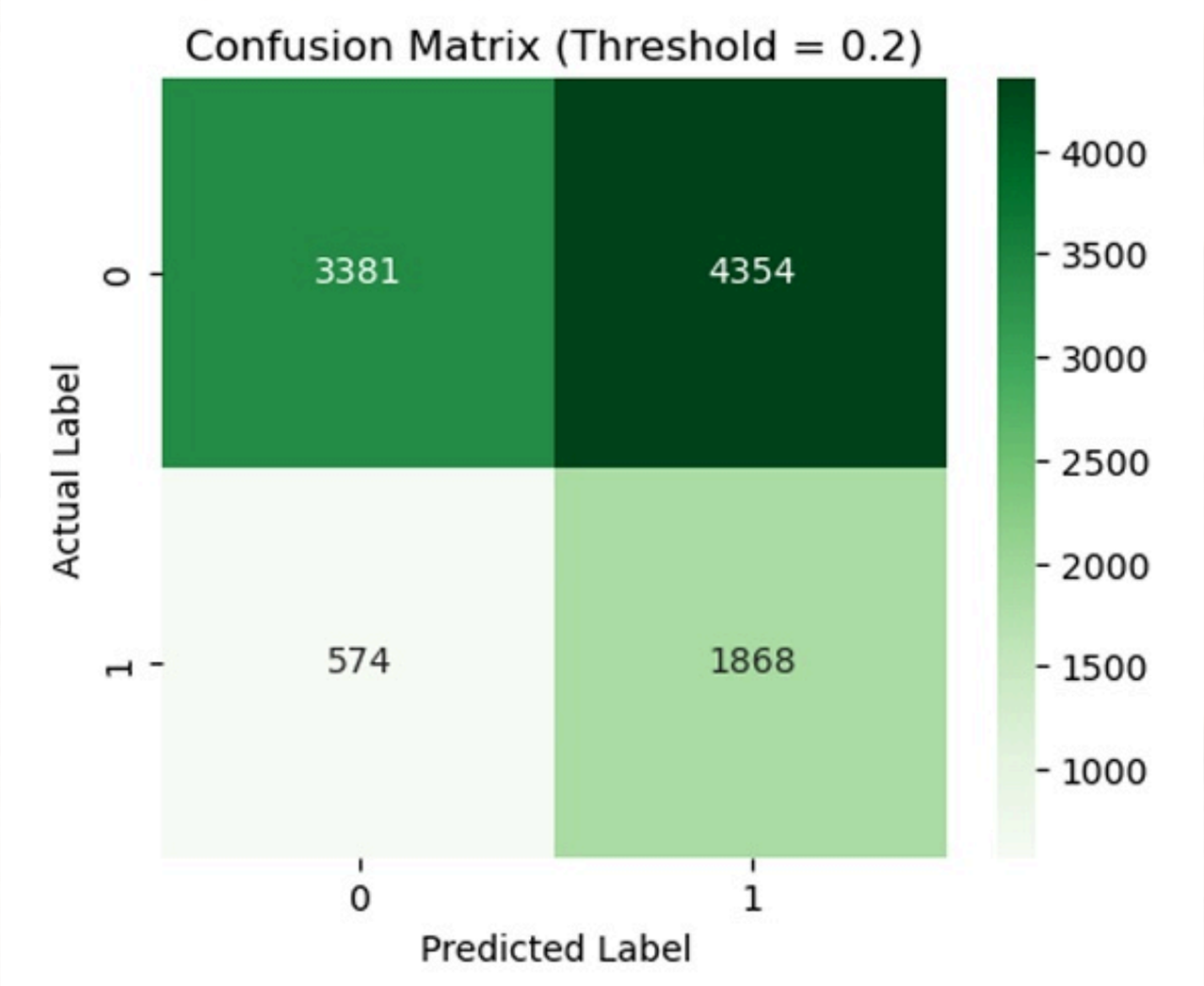
ROC Curve (Model Quality)



- Measures how well the model distinguishes converters from non-converters
- Compares model performance to random guessing
- Higher curve = better lead ranking ability
- Useful for selecting decision thresholds

Threshold Tuning

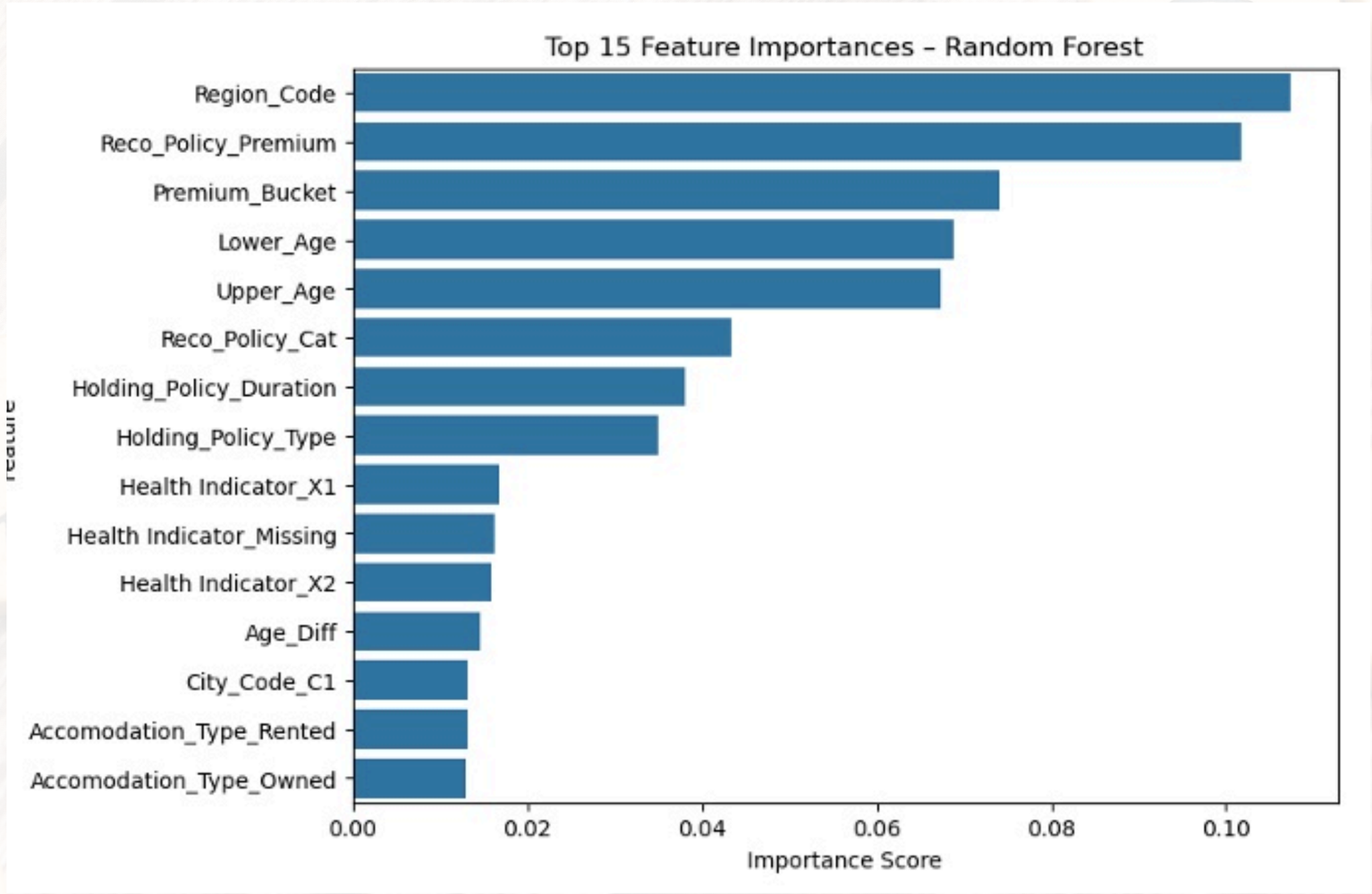
- Model outputs probabilities, not just yes/no
- Threshold controls how aggressive outreach is
- Lower threshold → more leads contacted
- Trade-off between missed opportunities and extra outreach



Feature Importance

Top drivers include:

- Policy premium level
- Premium pricing bucket
- Customer age segments
- Geographic region
- Policy duration and type



Business Implications

- Improved lead prioritization
- More efficient sales effort
- Reduced wasted outreach
- Data-driven campaign planning

Recommendations

- Deploy probability-based lead scoring
- Adjust thresholds by campaign goals
- Focus sales on top-ranked leads
- Use interpretable models to build trust

GAP ANALYSIS

- CRM integration
- Real-time lead scoring
- Model retraining with live data
- A/B testing outreach strategies

Conclusions

- ML supports decisions, not replaces people
- Business context drives model value
- Open for discussion