Predicting Health Insurance Lead Conversion Using Machine Learning

Compiled by

Lovemore Rugare Madzara

Table of Contents

Abstract

Accurately identifying high-potential leads is critical for improving efficiency in health insurance sales and marketing. This project applies machine learning techniques to predict the likelihood of lead conversion using a structured health insurance lead dataset sourced from a publicly available repository (Kaggle, 2023). After data cleaning, exploratory analysis, and feature engineering, multiple classification models were evaluated, including a tuned Decision Tree and a Random Forest classifier. Model performance was assessed using metrics appropriate for imbalanced classification problems, with particular emphasis on the Area Under the Receiver Operating Characteristic Curve (AUC). Results show that while the Decision Tree provided interpretability, the Random Forest achieved superior discriminatory performance and more stable predictions. Threshold tuning was further applied to align model outputs with business objectives, improving the identification of potential converters. Overall, the findings demonstrate how machine learning-driven lead scoring can support more targeted outreach strategies and enhance decision-making in health insurance marketing.

1.0 Data Origin

The dataset used in this study was obtained from Kaggle, specifically the *Health Insurance Lead Prediction – Raw Data* dataset made publicly available by a community contributor. The dataset can be accessed at: Kaggle. This dataset is designed to support predictive modeling tasks related to health insurance lead conversion and reflects realistic insurance prospect information commonly encountered in marketing and sales analytics. As a publicly available dataset, it is appropriate for academic analysis and reproducible research.

1.1 Data Overview

The dataset contains customer-level observations, where each record represents an individual insurance lead. The target variable, Response, indicates whether a lead converted (1) or did not convert (0), framing the problem as a binary classification task. The dataset includes a combination of numerical and categorical features describing customer demographics and insurance related characteristics, such as

age ranges, recommended policy premium, policy type, policy duration, regional and city codes, and other categorical indicators.

1.2 Data Cleaning

The dataset was reviewed to assess data quality, including variable types, formatting consistency, and the presence of missing values. Features were examined to ensure suitability for analysis, and variables were appropriately categorized as numerical or categorical. Outliers were retained, as they represent realistic customer behavior in insurance contexts and may contain valuable predictive information.

1.3 Handling of Missing Values

An initial assessment identified missing values across several features. Records with missing target labels (Response) were excluded, as they cannot be used in supervised learning. For predictor variables, missing values were handled as part of the preprocessing and modeling pipeline, ensuring that all observations used for analysis contained complete feature information. This approach allowed the dataset to remain intact while supporting effective exploratory analysis and machine learning modeling.

2.0 Data Mutation and Feature Engineering

To enhance the analytical value of the dataset, several feature transformations and derived variables were created prior to modeling. These data mutations were designed to improve interpretability, capture non-linear relationships, and reflect meaningful business segments commonly used in insurance analytics.

Continuous variables such as age, policy premium, and policy duration were transformed into grouped or banded variables (e.g., age groups, premium bands, and duration bands) to better represent customer segments and reduce sensitivity to noise. Interaction features combining insurance type and policy category were also created to capture joint effects that may influence conversion behavior. These engineered features allowed for more informative exploratory analysis and supported models capable of learning complex interactions.

All feature engineering was completed prior to model training, ensuring that the transformed dataset accurately reflected customer-level characteristics while preserving the original unit of analysis.

2.1 One-Hot Encoding

Several variables in the dataset were categorical in nature and therefore unsuitable for direct use in machine learning models. To address this, categorical features were transformed using one-hot encoding as part of a preprocessing pipeline. One-hot encoding converts each categorical level into a binary indicator, allowing models to incorporate categorical information without imposing an artificial ordinal structure. All preprocessing steps were implemented using the scikit-learn framework, which supports pipeline based workflows that ensure consistent feature transformation during both training and evaluation phases and reduce the risk of data leakage (Pedregosa et al., 2011).

2.2 Exploratory Data Analysis:

2.2.1 Numerical Variables

Exploratory analysis of numerical variables was conducted to understand the distribution, range, and variability of key quantitative features related to lead conversion. Variables such as recommended policy premium, policy duration, and age-related measures were examined using summary statistics and visualizations including histograms and boxplots.

The distribution of policy premiums exhibited right skewness, indicating that most leads were associated with lower premium recommendations, with fewer high-premium offerings. This pattern is consistent with typical insurance pricing structures and supports the decision to retain outliers, as higher premiums may represent valid customer segments rather than data errors. Policy duration variables showed moderate dispersion, suggesting variation in how long prospects had been associated with prior policies.

Age related numerical variables displayed uneven distributions across their ranges, reinforcing the usefulness of transforming raw age values into age group categories for further analysis. Overall, numerical variables showed meaningful variability and

non-normal distributions, justifying the use of tree-based models that are robust to skewness and do not rely on strict distributional assumptions.

2.2.2 Categorical variables

Exploratory analysis of categorical variables was conducted to examine conversion patterns across different customer segments and policy characteristics. Key categorical features analyzed included insurance type, policy category, city and region codes, channel related variables, and other discrete customer attributes. Frequency distributions and grouped conversion rates were used to identify differences in lead behavior across categories.

Several categorical variables displayed substantial variation in conversion outcomes. Certain insurance types and policy categories were associated with noticeably higher conversion rates, suggesting that product characteristics play an important role in influencing customer decisions. Regional and city-level variables exhibited uneven distributions, reflecting differences in lead concentration across geographic areas rather than uniform customer representation.

Combined and interaction based categorical features further highlighted that conversion behavior varies across specific customer product combinations. These patterns supported the inclusion of categorical variables and interaction features in the modeling stage. Overall, categorical EDA revealed meaningful segmentation effects, reinforcing the need for models capable of capturing non-linear relationships and interactions, such as tree based classifiers.

2.2.3 Variable Combinations and Interactions

Beyond univariate and bivariate analysis, further exploratory analysis focused on examining combinations of variables to understand how multiple customer and policy characteristics jointly influence lead conversion behavior. This multivariate perspective is particularly important in insurance marketing contexts, where conversion decisions are rarely driven by a single factor in isolation.

Several variable combinations revealed meaningful interaction effects. Demographic attributes such as age segments were analyzed in conjunction with policy-related variables, including recommended policy premium and insurance type. These combined analyses suggested that conversion likelihood varies not only by age or

pricing independently, but also by how pricing levels align with specific customer age groups. For example, certain age segments exhibited higher conversion rates when associated with mid-range premium recommendations, while conversion rates declined for higher premiums within the same demographic groups. This pattern highlights the importance of aligning product offerings with demographic-specific affordability and preferences.

Additional combinations involving policy duration and insurance category further demonstrated that customer history and product structure jointly influence conversion outcomes. Leads associated with longer policy durations showed different conversion patterns depending on the recommended insurance type and policy category, suggesting that familiarity with insurance products may interact with product complexity or perceived value. These interaction effects would not be visible through single-variable analysis and underscore the importance of examining combined feature behavior.

Geographic variables were also analyzed in combination with policy attributes. While regional and city-level features individually reflected lead concentration rather than direct conversion drivers, their interaction with policy types revealed localized differences in product performance. Certain policy categories performed more effectively in specific regions or city segments, indicating that geographic context may influence how insurance products are received by prospective customers.

To further capture interaction effects, composite categorical features were created by combining insurance type and policy category. Analysis of these combined variables revealed distinct conversion patterns across specific customer–product pairings. These findings suggest that conversion behavior is influenced by nuanced combinations of product design and customer characteristics, rather than broad product categories alone.

3.0 Comparison of Supervised Machine Learning Models

To determine the most appropriate modeling approach for predicting health insurance lead conversion, multiple supervised machine learning models were

developed and systematically compared. The analysis focused on a Decision Tree classifier and a Random Forest classifier, as both models are well suited for tabular datasets containing a mixture of numerical and categorical variables and are capable of capturing non-linear relationships (Breiman, 2001; Hastie, Tibshirani, & Friedman, 2009).

The Decision Tree classifier was implemented as an interpretable baseline model. Hyperparameter tuning was conducted using cross-validation to control tree depth and splitting behavior in order to reduce overfitting. While the tuned Decision Tree provided clear insights into how individual features influenced predictions, its performance revealed important limitations. Specifically, the model exhibited conservative behavior, strongly favoring non-conversion predictions. This resulted in high accuracy driven primarily by correct identification of non-converting leads, but at the cost of a large number of false negatives. In the context of lead generation, this behavior is undesirable, as failing to identify potential converters represents a significant opportunity cost.

In contrast, the Random Forest classifier leveraged an ensemble learning approach by aggregating predictions from multiple decision trees. This structure reduces model variance and improves generalization by averaging over diverse decision boundaries (Breiman, 2001). As a result, the Random Forest demonstrated superior discriminatory performance, as reflected by a higher Area Under the Receiver Operating Characteristic Curve (AUC). Unlike the Decision Tree, which relied heavily on majority-class predictions, the Random Forest was more effective at ranking leads according to their likelihood of conversion.

Although overall accuracy between the two models was similar, accuracy alone was not considered a sufficient evaluation metric due to the imbalanced nature of the dataset. Instead, AUC was prioritized, as it provides a more robust measure of a model's ability to distinguish between converting and non-converting leads across different thresholds (Fawcett, 2006). Threshold tuning was further applied to the Random Forest model to align predictions with business objectives, increasing recall for converting leads while accepting a higher number of false positives.

Overall, the comparison demonstrated that while the Decision Tree offered transparency and interpretability, the Random Forest provided stronger predictive

performance and greater flexibility for lead prioritization. Based on these findings, the Random Forest model was selected as the final model for this project due to its superior ability to balance predictive accuracy with business relevance in health insurance lead generation.

4.0 Model Evaluation Metrics

Model performance was evaluated using metrics appropriate for imbalanced classification problems. Accuracy alone was not considered sufficient due to the uneven distribution of converting and non-converting leads. Instead, the Area Under the Receiver Operating Characteristic Curve (AUC) was emphasized, as it provides a robust measure of a model's ability to distinguish between classes across varying classification thresholds (Fawcett, 2006). Confusion matrices were additionally used to examine classification behavior and support threshold selection aligned with business objectives

5.0 Implications and Recommendations

The findings of this analysis demonstrate that machine learning–based lead scoring can provide meaningful support for decision making in health insurance marketing and sales. By ranking leads according to their predicted likelihood of conversion, the selected Random Forest model enables organizations to move beyond uniform outreach strategies toward more targeted and efficient engagement.

One key implication is the ability to prioritize sales and marketing resources. Rather than allocating equal effort to all leads, insurers can focus outreach on prospects with higher predicted conversion probabilities. This approach can improve response rates, reduce unnecessary contact with low potential leads, and increase overall return on marketing investment. The use of probability-based scores, rather than strict binary classifications, allows for flexible prioritization depending on available resources and campaign objectives.

The analysis also highlights the importance of pricing and product alignment. Feature importance results indicate that policy premium levels, policy characteristics, and age-related segments play a significant role in conversion outcomes. Insurers can use

these insights to tailor premium recommendations and product offerings to specific customer segments, potentially improving conversion rates by better matching customer expectations and affordability.

Additionally, threshold tuning results suggest that business objectives should guide model deployment. Lowering the classification threshold increased the identification of converting leads at the expense of additional false positives. In a marketing context, this trade-off is often acceptable, as the cost of outreach is typically lower than the opportunity cost of missed conversions. Organizations can adjust thresholds dynamically based on campaign capacity, cost constraints, and desired recall levels.

First, insurers should implement probability-based lead ranking rather than binary acceptance or rejection of leads. Leads can be segmented into high-, medium-, and low priority groups based on predicted conversion probabilities. High-probability leads should receive immediate personalized follow-up by sales agents, while medium probability leads may be placed into targeted nurturing campaigns such as follow up emails or reminders. Low-probability leads can be deprioritized or revisited at a later stage, improving overall operational efficiency.

Second, organizations should adopt dynamic threshold tuning aligned with business objectives. For customer acquisition campaigns where maximizing conversions is critical, lower classification thresholds may be used to capture a broader set of potential buyers. Conversely, when sales capacity or operational costs are constrained, higher thresholds can be applied to focus only on the most promising leads. This flexibility allows the model to adapt to changing business conditions without retraining.

Third, insurers are encouraged to align pricing and product strategies with key customer segments. Since premium levels, policy categories, and age-related variables were found to be influential predictors, marketing teams can design differentiated offerings tailored to specific age groups and policy duration preferences. This segmentation-based approach can enhance perceived value and improve conversion likelihood.

Fourth, model interpretability should be incorporated into stakeholder communication and deployment decisions. While the Random Forest model demonstrated superior

predictive performance, the Decision Tree provides transparent decision rules that can help non-technical stakeholders understand the factors driving lead conversion. Using both models in tandem supports trust, compliance, and explainability in regulated environments such as insurance.

Organizations should establish a process for continuous monitoring and periodic model recalibration. Customer behavior, pricing structures, and market conditions may change over time, potentially reducing model effectiveness. Regular evaluation using updated data will ensure sustained predictive accuracy and long-term business value.

Finally, the interpretability provided by the Decision Tree analysis offers complementary value. While the Random Forest was selected as the final model, insights from the Decision Tree structure can help stakeholders understand key decision patterns and build trust in model-driven recommendations. Together, these models support both predictive performance and interpretability, enabling data-informed strategies that can enhance management and customer acquisition in the health insurance industry.

6.0 Conclusions

This project demonstrated the application of supervised machine learning techniques to predict health insurance lead conversion and support data-driven marketing decisions. Through comprehensive data cleaning, exploratory data analysis, feature engineering, and model evaluation, meaningful patterns in lead behavior were identified and translated into actionable insights.

Among the models evaluated, the Random Forest classifier proved most effective for this task. While a tuned Decision Tree offered interpretability, the Random Forest achieved superior discriminatory performance, as measured by AUC, and provided greater flexibility for lead prioritization. Threshold tuning further enhanced the model's practical value by aligning predictions with business objectives, allowing for improved identification of potential converters.

Beyond predictive performance, the analysis highlighted the importance of combining demographic information, policy characteristics, and pricing related

variables when assessing conversion likelihood. These findings underscore the value of using machine learning not as a replacement for business judgment, but as a decision support tool that enhances strategic planning and operational efficiency.

Overall, the results illustrate how machine learning–based lead scoring can be effectively integrated into health insurance marketing workflows to improve targeting, optimize resource allocation, and support more informed decision-making. Future work could extend this analysis by incorporating additional behavioral or temporal features to further refine prediction accuracy and adaptability.

7.0 REFERENCES

1.) Dataset:Kaggle. (2023). Health insurance lead prediction – raw data: Kaggle

2.) Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

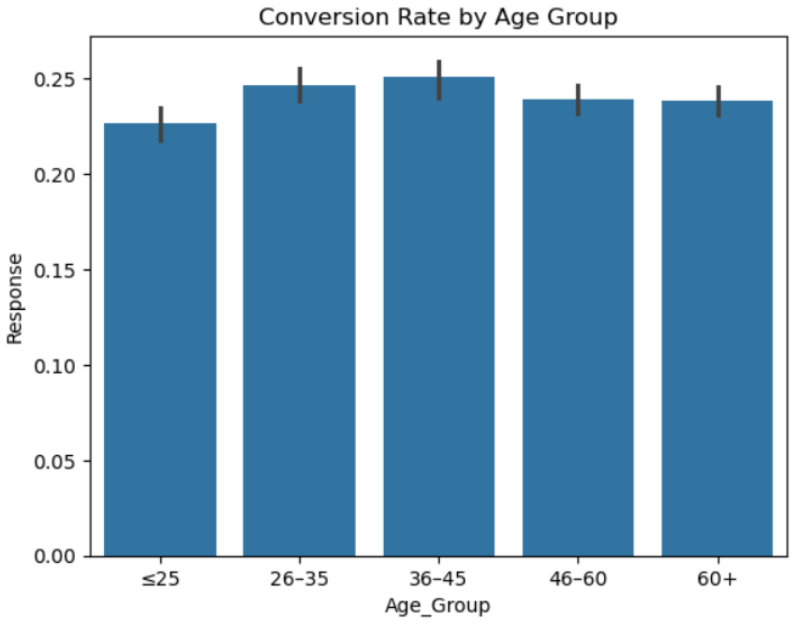
3.) Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer

4.) Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874.

5.) Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

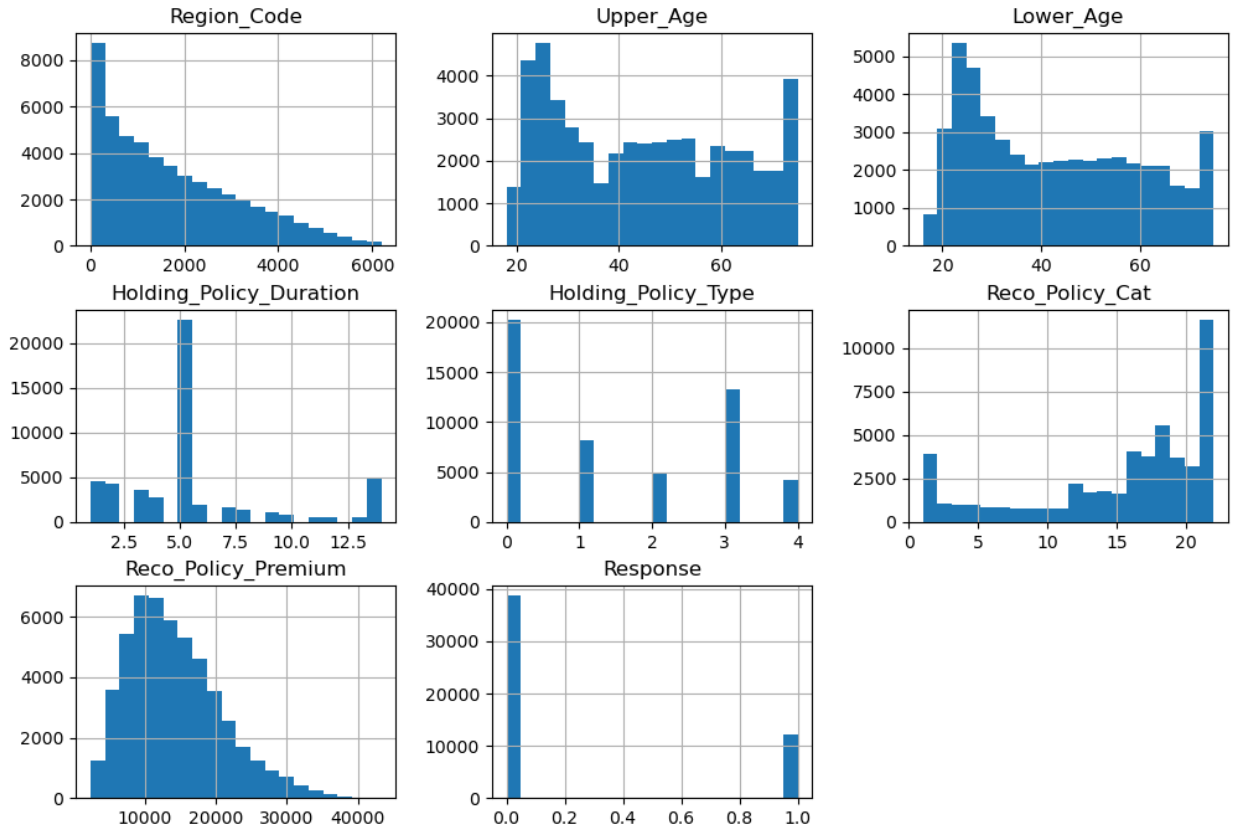
8.0 Appendices

Figure 1. Conversion Rate by Age Group



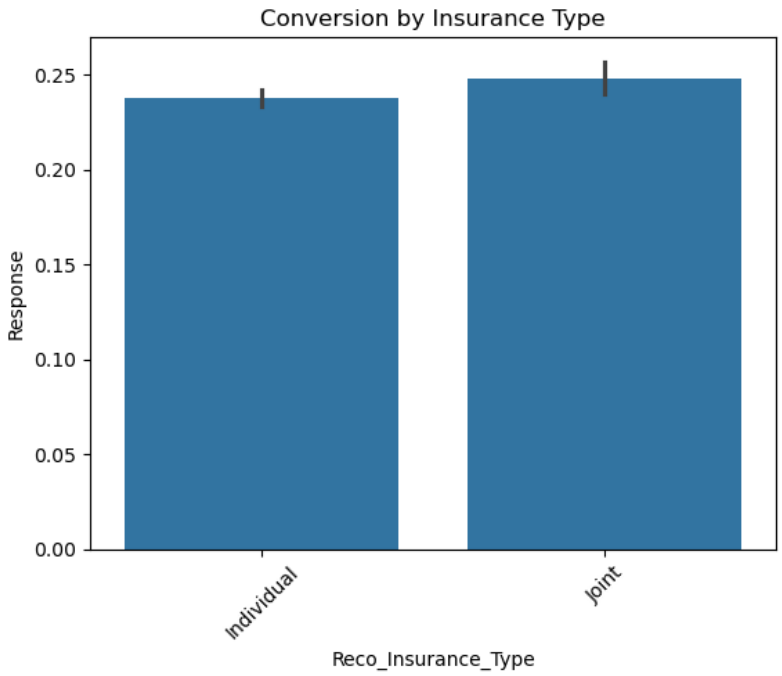
The bar chart displays the average lead conversion rate across five age groups. Conversion rates increase from the ≤25 group through the 36–45 group, which exhibits the highest conversion rate. After age 45, conversion rates decline slightly for the 46–60 group and remain relatively stable for customers aged 60 and above. Error bars indicate moderate variability but do not suggest extreme dispersion across groups. Leads aged 26–45 demonstrate the strongest propensity to convert, suggesting that mid-career individuals are the most responsive to health insurance offerings. In contrast, younger leads (≤25) show the lowest conversion rates, likely reflecting lower perceived need or limited purchasing power, while older segments (46+) may exhibit more cautious decision-making. From a business perspective, marketing resources and sales outreach may yield the highest returns when prioritized toward the 26–45 age segment, while alternative engagement strategies may be required to improve conversion among younger and older leads.

Figure 2. Distribution of Numerical Variables



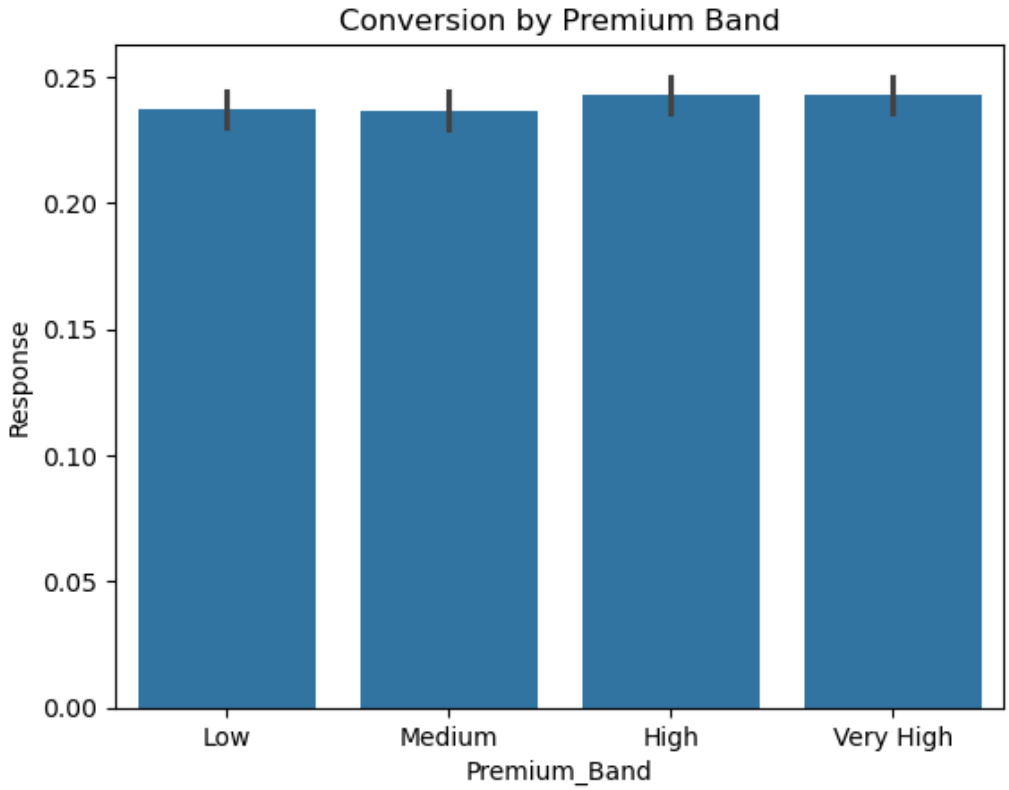
Most numerical features display non-normal and skewed distributions, particularly policy premium and holding duration, indicating substantial variability in customer profiles and policy characteristics. The right-skew in premium values suggests that most leads are associated with moderately priced policies, with fewer high-premium offerings. Age variables show broader dispersion, reflecting a diverse customer base across working age and senior segments. The pronounced imbalance in the Response variable confirms that conversions are relatively rare events, justifying the use of evaluation metrics such as ROC–AUC and threshold tuning rather than accuracy alone. These distributional patterns support the use of tree based models, which are robust to skewed features and do not assume linear relationships.

Figure 3. Conversion Rate by Insurance Type



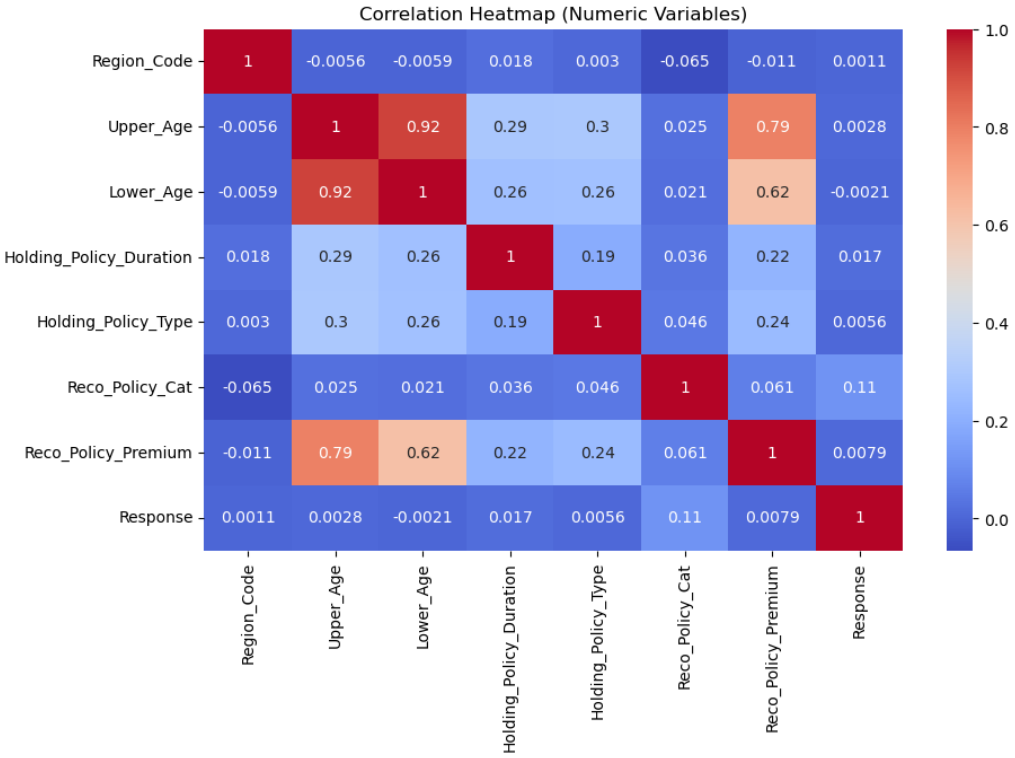
Leads recommended joint insurance policies demonstrate a marginally higher likelihood of conversion compared to those offered individual plans. This pattern may reflect stronger perceived value among households or couples who view joint coverage as more cost-effective or comprehensive. From a business perspective, emphasizing joint policy offerings during lead engagement, particularly for eligible customer segments may improve overall conversion performance. However, the relatively small difference also indicates that insurance type alone is not a dominant driver of conversion and should be considered in combination with demographic and policy attributes.

Figure 4. Conversion Rate by Insurance Type



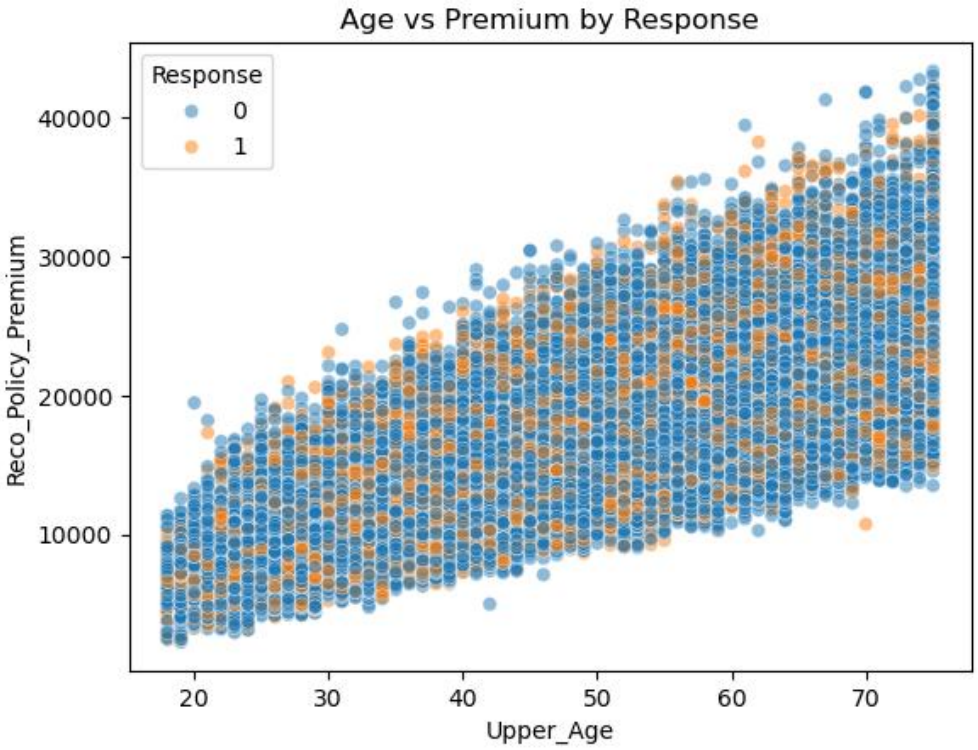
The relatively consistent conversion rates across premium bands suggest that premium amount alone does not strongly deter lead conversion. The modest increase in conversion among higher premium bands may indicate that customers engaging with higher-priced policies are more informed or have a stronger intent to purchase. From a business standpoint, this finding implies that focusing solely on lower-priced offerings may not maximize conversions, and that higher premium products can be marketed effectively when aligned with customer needs and perceived value.

Figure 5. Correlation Heatmap of Numerical Variables



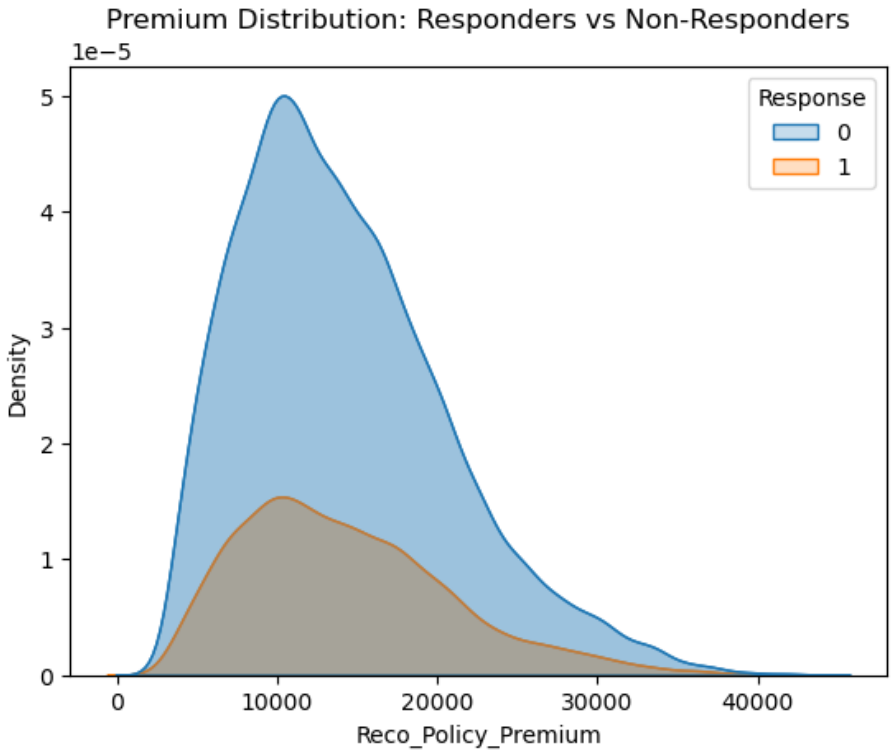
The strong correlation between Upper_Age and Lower_Age reflects their structural relationship as age bounds rather than independent predictors. Recommended policy premium shows a notable positive association with age, suggesting that older leads are typically offered higher-premium policies. Importantly, correlations between individual numerical features and the Response variable are uniformly weak, indicating that lead conversion is not driven by any single numeric attribute in isolation. This reinforces the need for multivariate and non-linear modeling approaches, such as Random Forests and Decision Trees, which can capture complex interactions that are not evident through linear correlation analysis alone.

Figure 6. Relationship Between Age, Premium, and Conversion Outcome



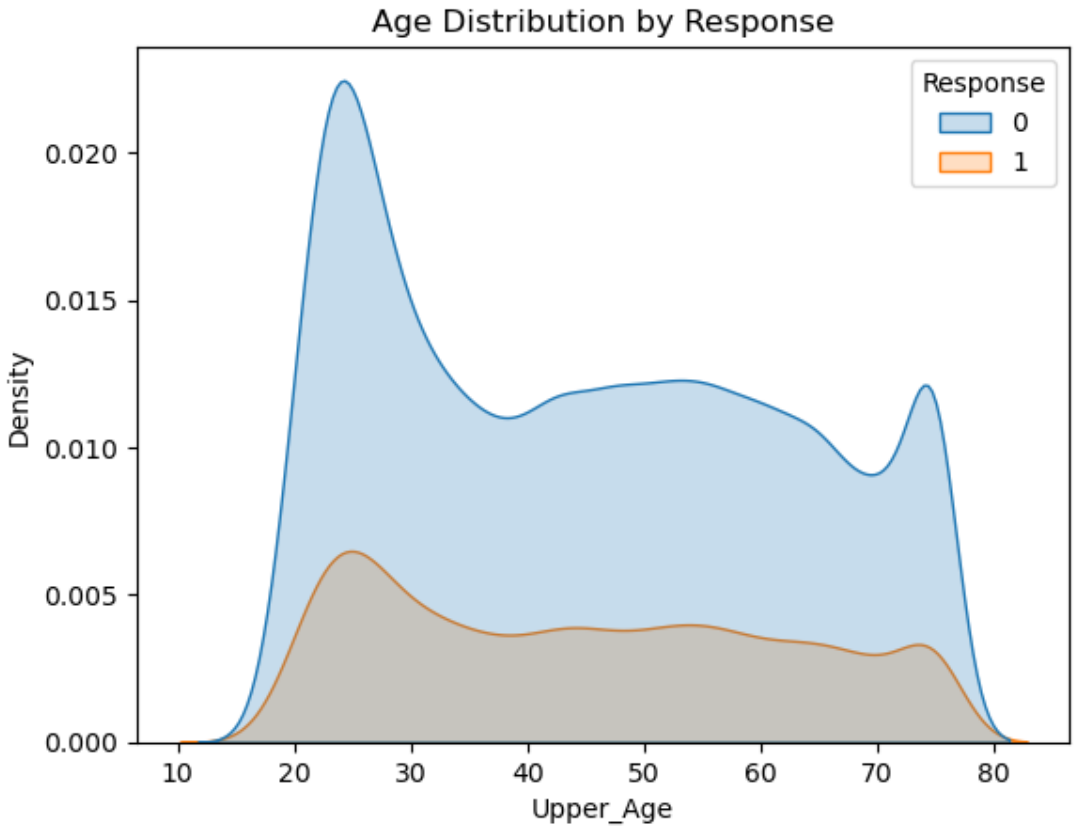
The visualization confirms a strong positive relationship between age and recommended premium, indicating that older leads are consistently offered higher-priced policies. However, the extensive overlap between converted and non-converted leads suggests that neither age nor premium alone is sufficient to distinguish conversion outcomes. This pattern indicates that lead conversion is influenced by interactions among multiple variables, rather than simple linear effects. As a result, nonlinear, interaction aware models such as Random Forests and Decision Trees are better suited to capturing the underlying decision structure than single-variable or linear approaches.

Figure 7. Premium Distribution for Responders and Non-Responders



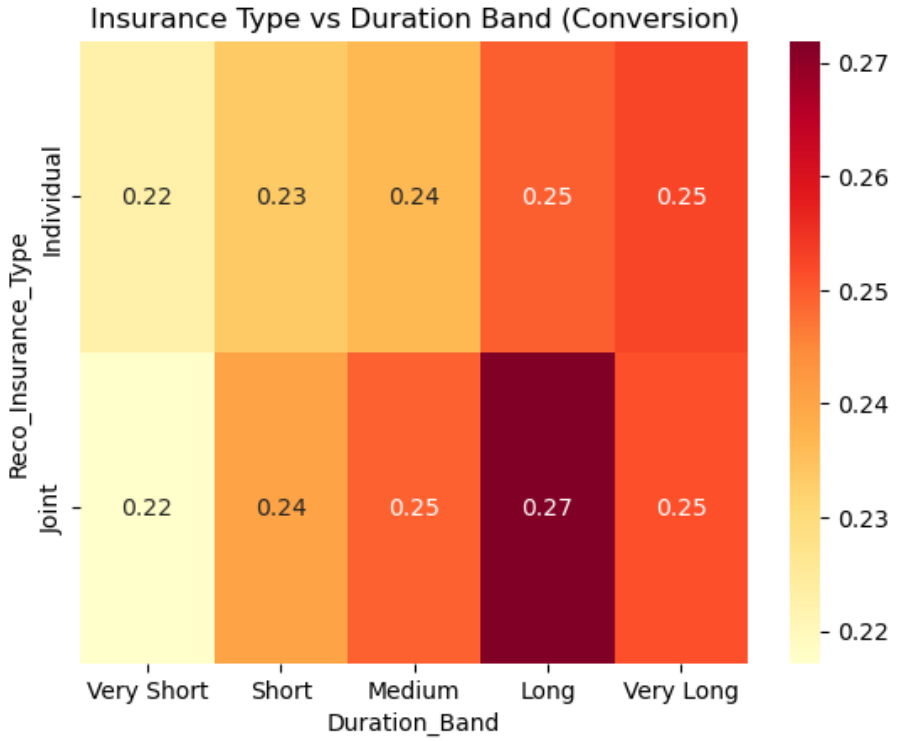
The significant overlap between responder and non-responder premium distributions indicates that policy premium alone is not a strong discriminator of conversion behavior. While responders appear marginally more represented in mid-to-higher premium ranges, the difference is not pronounced enough to support premium based segmentation in isolation. This finding reinforces earlier correlation and interaction analyses, suggesting that conversion outcomes are driven by combinations of demographic and policy attributes rather than price alone. Consequently, predictive models must rely on multivariate patterns rather than simple threshold based pricing rules.

Figure 8. Age Distribution by Conversion Outcome



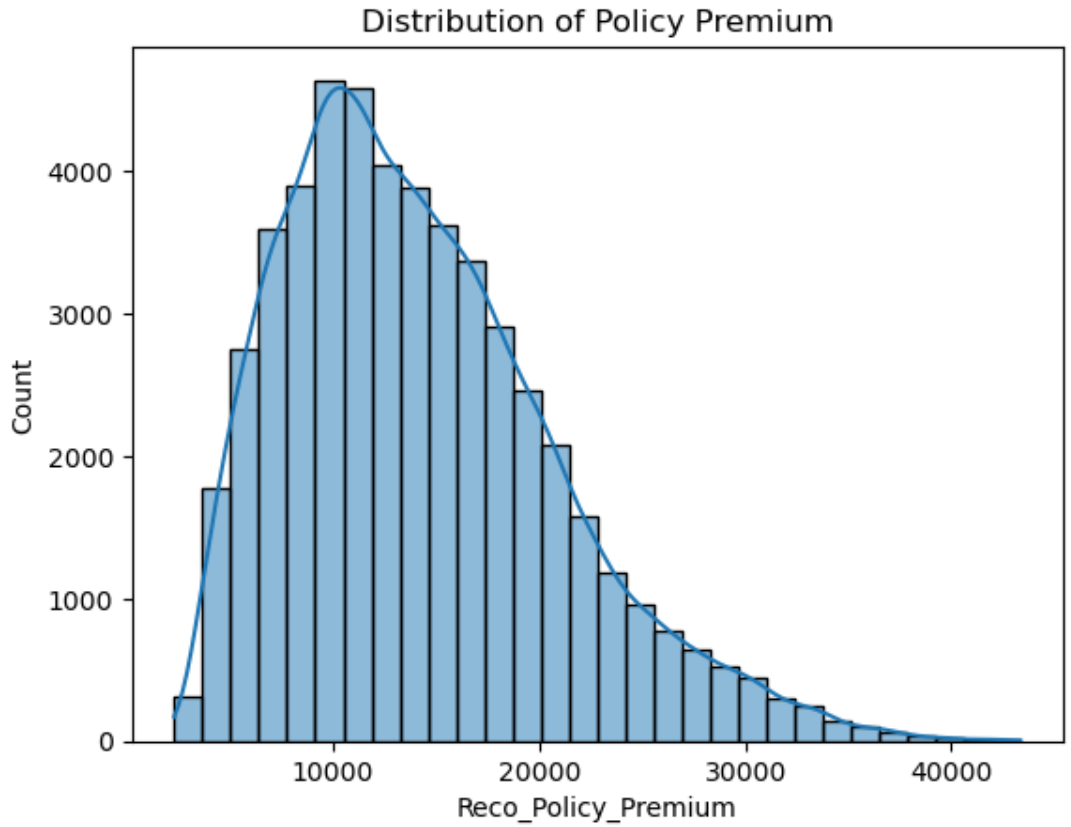
The substantial overlap between the age distributions of responders and non-responders indicates that age alone is not a decisive factor in lead conversion. While minor variations in density are visible across certain age ranges, no clear age threshold distinctly separates converted from non-converted leads. This reinforces earlier findings that demographic variables must be evaluated in conjunction with policy characteristics and other contextual factors, supporting the use of multivariate models rather than age-based segmentation rules.

Figure 9. Conversion Rates by Insurance Type and Policy Duration Band



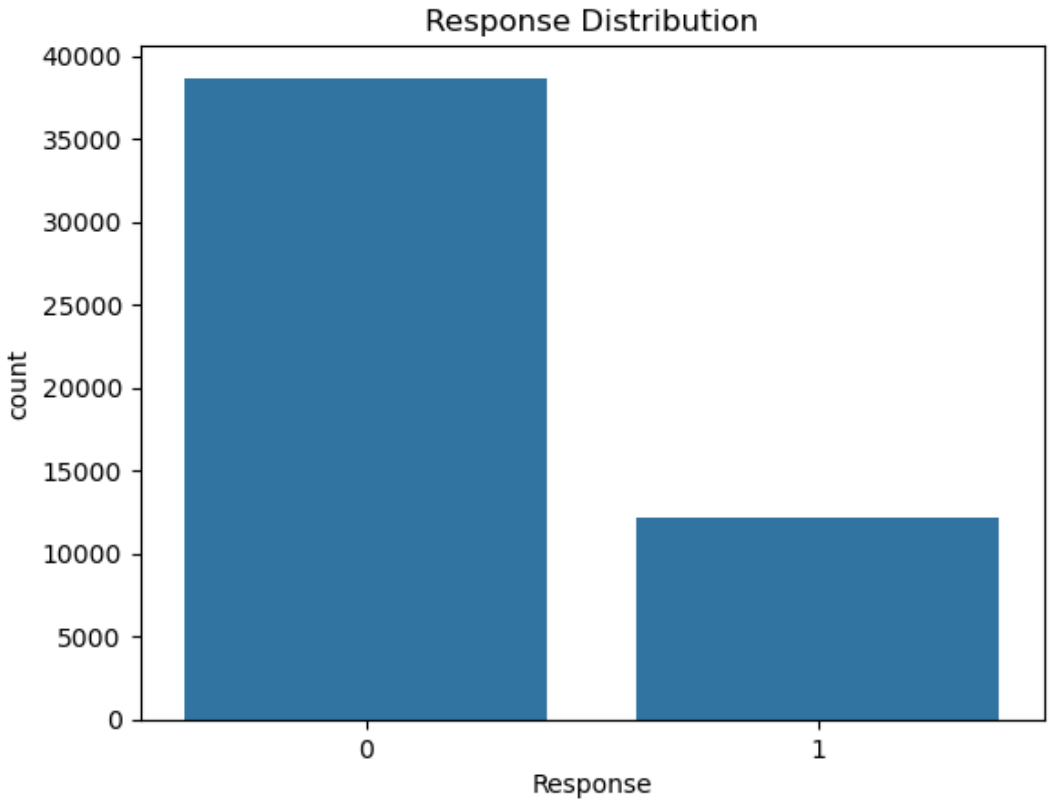
The results indicate that longer policy durations are associated with higher conversion likelihood, suggesting that customers considering extended coverage periods may exhibit stronger purchase intent. Additionally, joint insurance policies outperform individual policies across nearly all duration bands, reinforcing earlier findings that bundled or shared coverage options may be perceived as offering greater value. From a business perspective, targeting leads with longer intended coverage durations and promoting joint policy options may improve conversion performance. However, the gradual nature of the increase suggests that duration and insurance type should be leveraged together with other features rather than used as standalone decision rules.

Figure 10. Distribution of Recommended Policy Premium



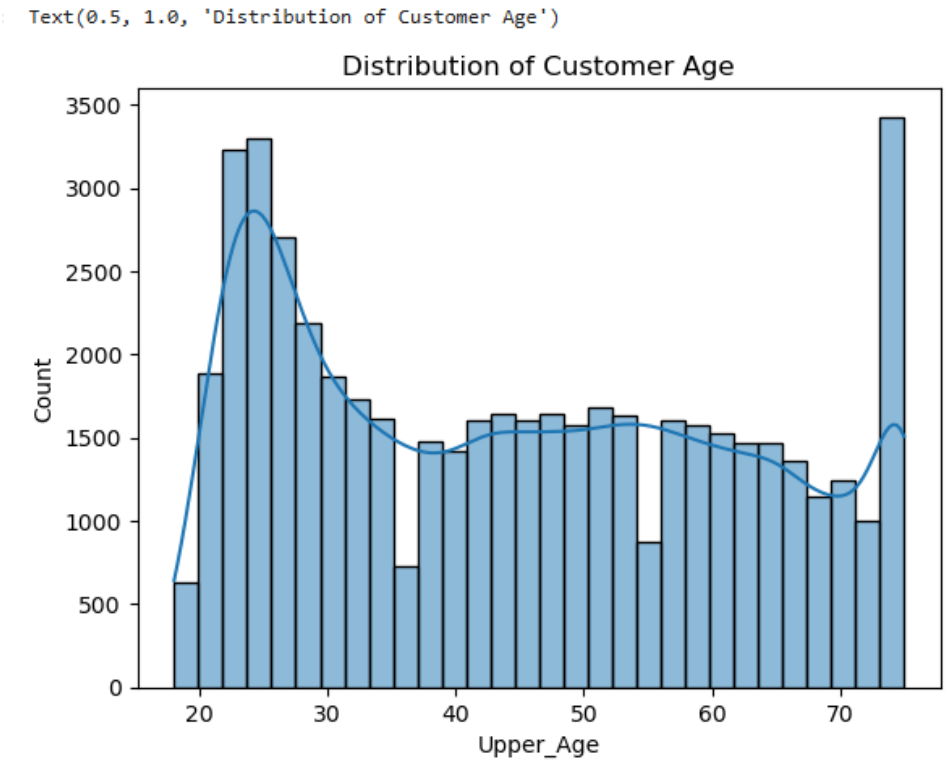
The right-skewed distribution indicates that most leads are offered moderately priced insurance policies, with progressively fewer high-premium recommendations. This suggests a pricing strategy oriented towards affordability for most customers, while still accommodating higher-value offerings for specific segments. The skewness and presence of extreme values further justify the use of tree based models, which are robust to non-normal distributions and less sensitive to outliers compared to linear models. From an analytical perspective, this distribution explains why premium amount alone shows limited linear correlation with conversion, despite remaining an important component in multivariate decision making.

Figure 11. Distribution of Lead Conversion Outcome



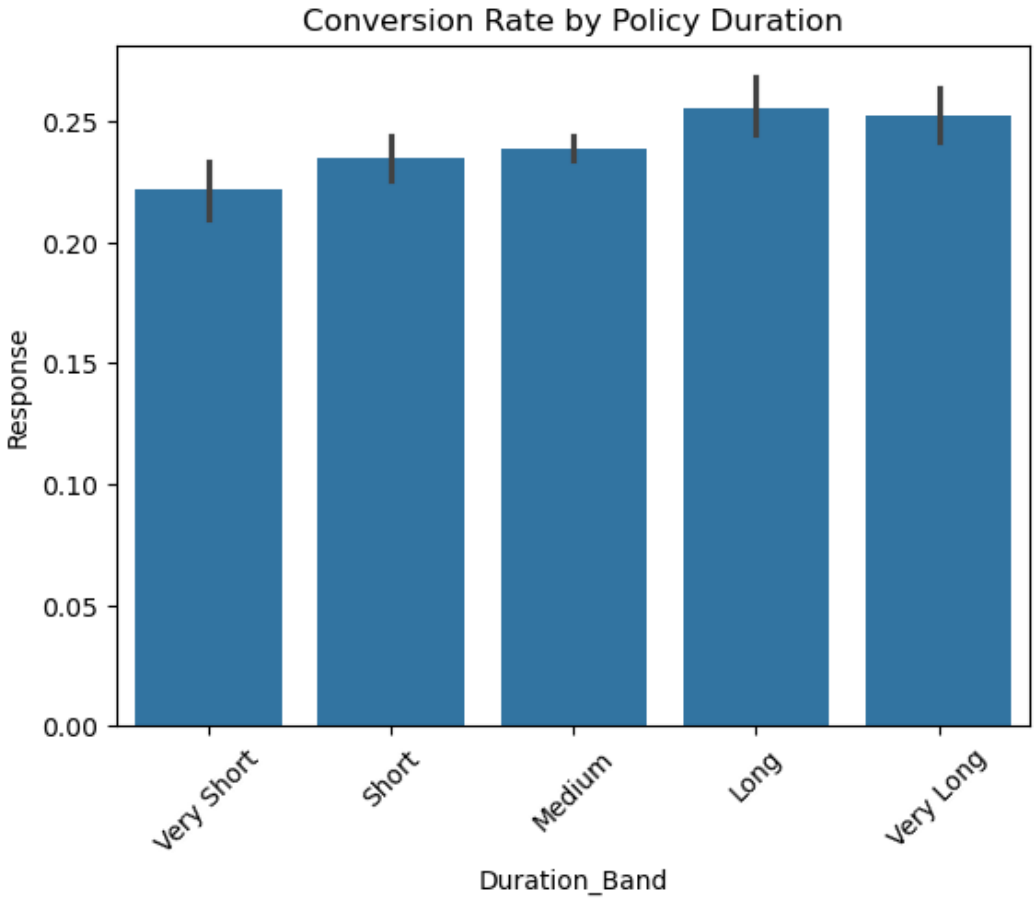
The response distribution reveals a clear class imbalance, with non-conversions significantly outnumbering conversions. This imbalance has important implications for model development and evaluation. Relying solely on accuracy could produce misleading results by favoring the dominant class, underscoring the need for alternative evaluation metrics such as ROC–AUC and confusion matrices. Additionally, the imbalance justifies the use of techniques such as class weighting and probability threshold tuning to better capture conversion behavior and improve the identification of high potential leads.

Figure 12. Distribution of Customer Age



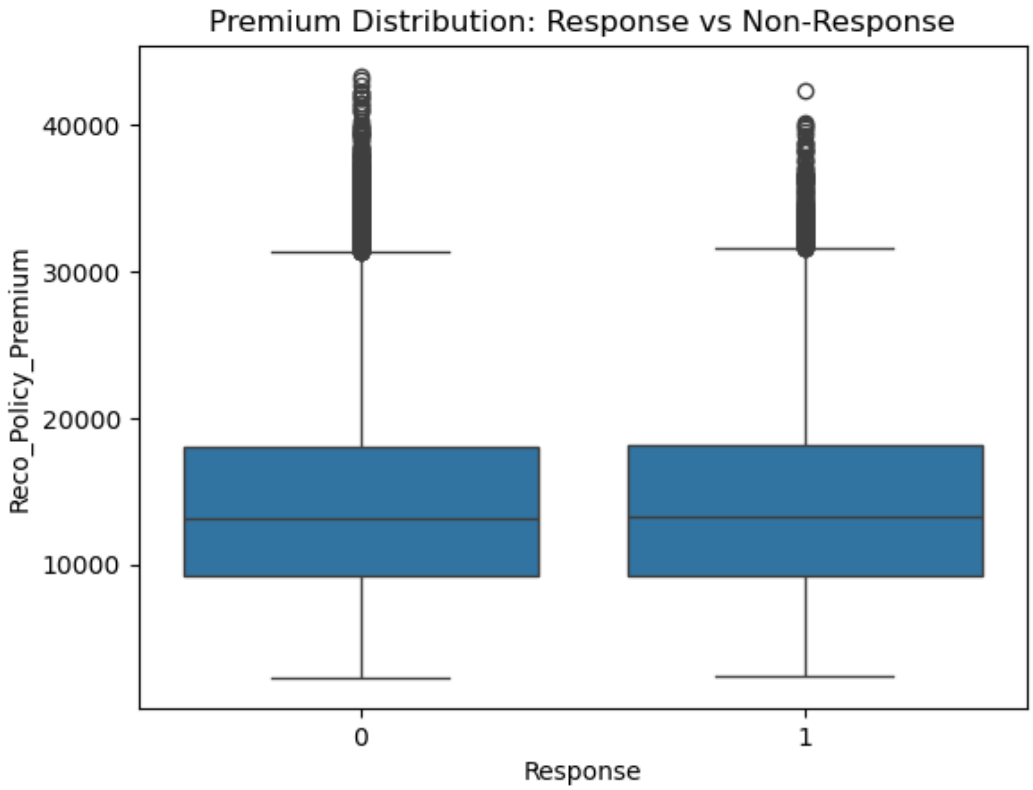
The multi-modal age distribution suggests that the dataset comprises distinct age segments with potentially different insurance needs and purchasing behaviors. Younger customers may represent early-stage insurance adoption, while older customers may be seeking coverage adjustments related to life stage or health considerations. This diversity in age profiles reinforces the importance of segment-aware modeling and supports earlier findings that age interacts with policy attributes rather than acting alone to influence conversion outcomes. Consequently, age is best leveraged as part of a broader multivariate feature set rather than as a standalone predictor.

Figure 13. Conversion Rate by Policy Duration



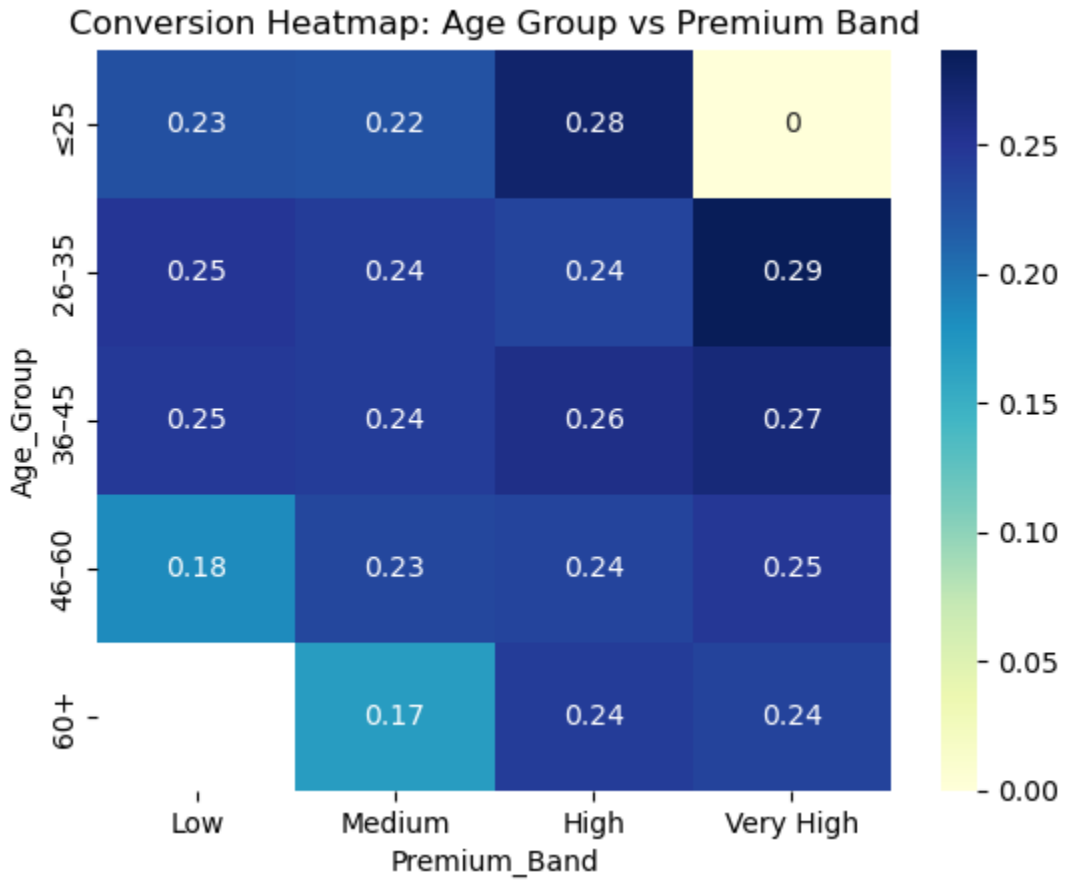
The upward trend suggests that leads considering longer policy durations are more likely to convert, potentially reflecting higher purchase intent or greater perceived value in extended coverage. Short-duration policies exhibit the lowest conversion rates, indicating possible hesitancy or lower commitment among these leads. From a business perspective, this finding implies that identifying and prioritizing leads interested in longer coverage periods may improve conversion outcomes. Analytically, the monotonic pattern supports the inclusion of policy duration as an important predictor and helps explain why duration related features contribute meaningfully within tree-based models.

Figure 14. Policy Premium Distribution by Conversion Outcome



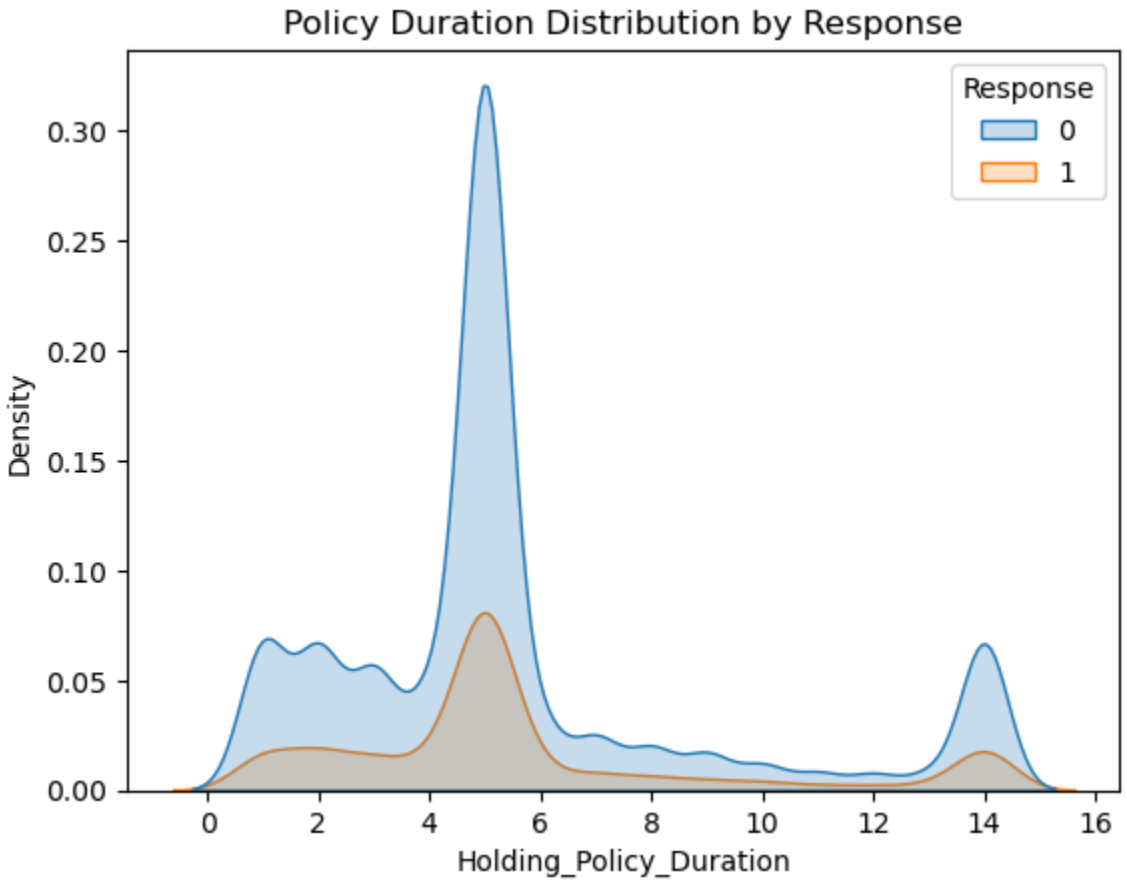
The close similarity between the premium distributions of responders and non-responders indicates that policy premium alone does not strongly differentiate conversion outcomes. Although converted leads show a marginally higher median premium, the extensive overlap suggests that premium level is not a decisive factor in isolation. The presence of many high-premium outliers in both groups further supports the decision not to remove outliers and highlights why tree-based models rather than linear or threshold-based approaches are appropriate for capturing complex, interaction-driven patterns in the data.

Figure 15. Conversion Rates by Age Group and Premium Band



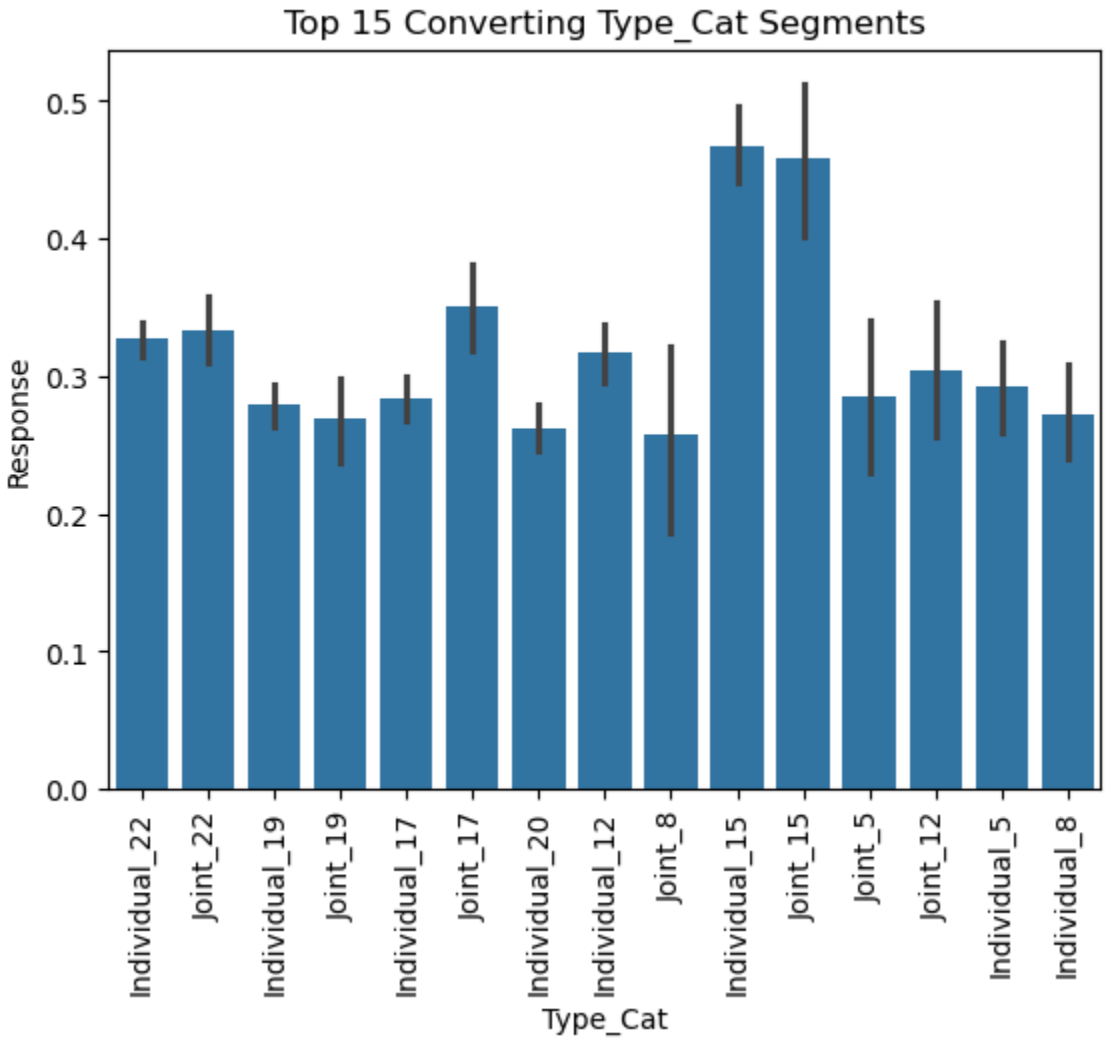
The heatmap reveals a clear interaction effect between age and premium level, indicating that conversion behavior cannot be explained by either factor independently. Leads aged 26–45 consistently demonstrate higher conversion rates, particularly for high and very high premium bands, suggesting stronger purchasing intent and financial readiness within this segment. In contrast, younger leads (≤25) show limited responsiveness to very high premiums, while older leads (60+) convert more selectively depending on premium level. These patterns reinforce the value of segment-specific pricing and targeting strategies and help explain why tree-based models, which capture interactions between variables, outperform simpler linear approaches in predicting lead conversion.

Figure 16. Policy Duration Distribution by Conversion Outcome



While both responders and non-responders are concentrated around similar policy duration ranges, converted leads display relatively higher density at moderate to longer holding durations, suggesting stronger commitment among these customers. Short-duration policies are associated with lower conversion likelihood, indicating weaker purchase intent or exploratory behavior. The overlapping yet distinct density patterns highlight that policy duration contributes to conversion prediction primarily through interaction with other variables, rather than acting as a standalone determinant. This further supports the use of tree-based models, which can exploit such interaction effects effectively.

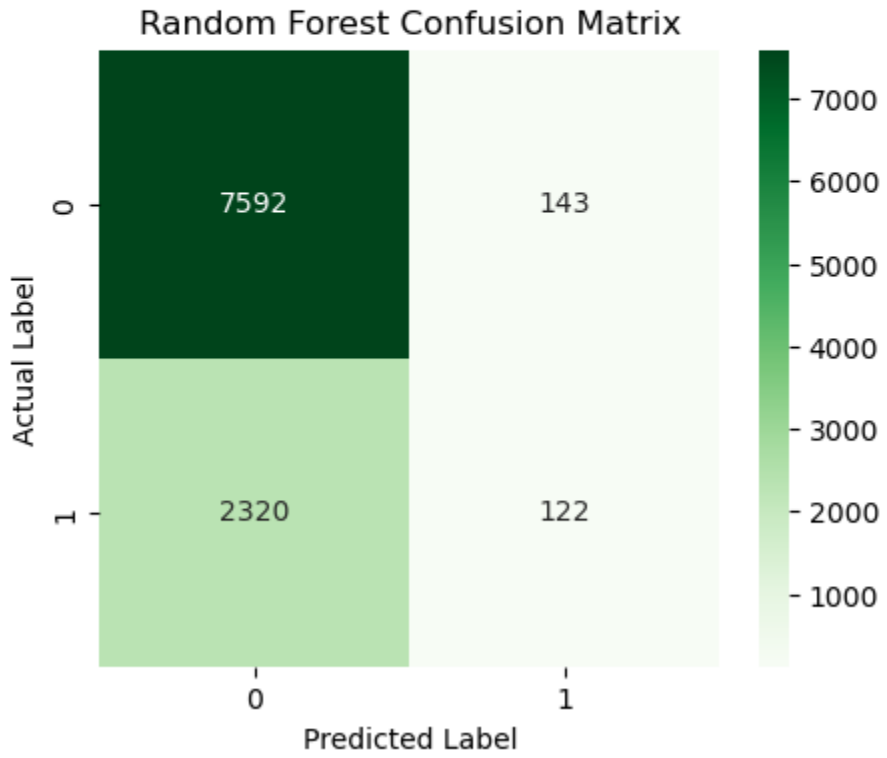
Figure 17. Top Converting Type–Category Segments



The results highlight specific type–category combinations that consistently outperform others, demonstrating that conversion behavior is strongly influenced by the interaction between insurance structure and policy category. Notably, several joint policy combinations appear among the highest-performing segments, reinforcing earlier findings that bundled coverage options may offer greater perceived value. From a business perspective, these high-performing segments represent prime candidates for targeted marketing, prioritized lead scoring, and tailored sales messaging. Analytically, this segmentation underscores why interaction-aware models such as Random Forests and Decision Trees outperform simpler approaches, as they can capture these compound effects that are invisible in single-variable analyses.

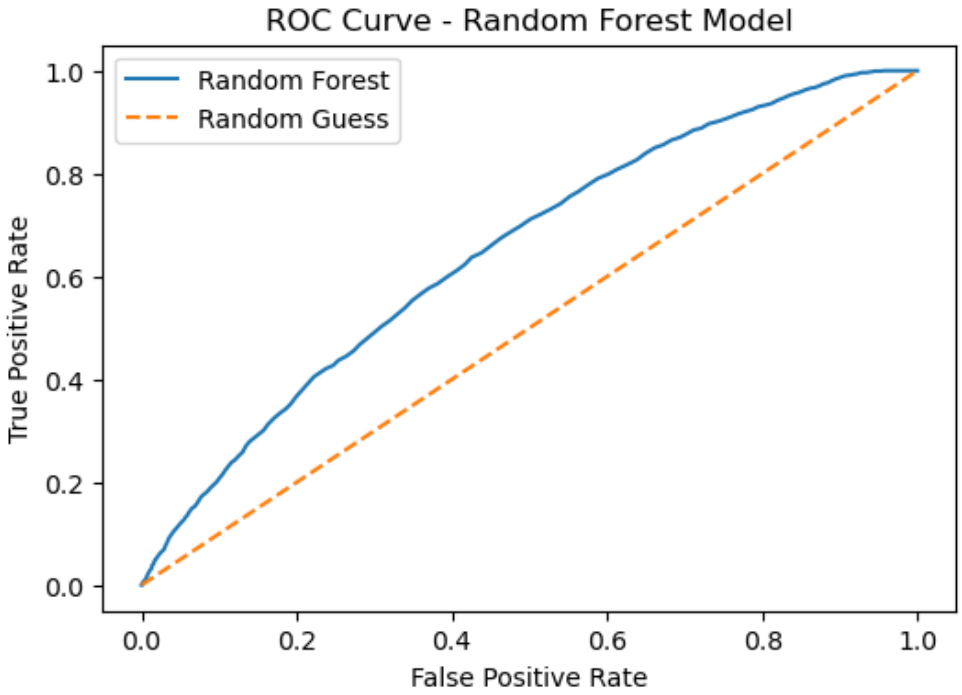
Figure 18. Random Forest Confusion Matrix

```
# Plot confusion matrix
plt.figure(figsize=(5,4))
cm = confusion_matrix(y_test, y_pred)
sns.heatmap(cm, annot=True, fmt='d', cmap='Greens')
plt.xlabel("Predicted Label")
plt.ylabel("Actual Label")
plt.title("Random Forest Confusion Matrix")
plt.show()
```



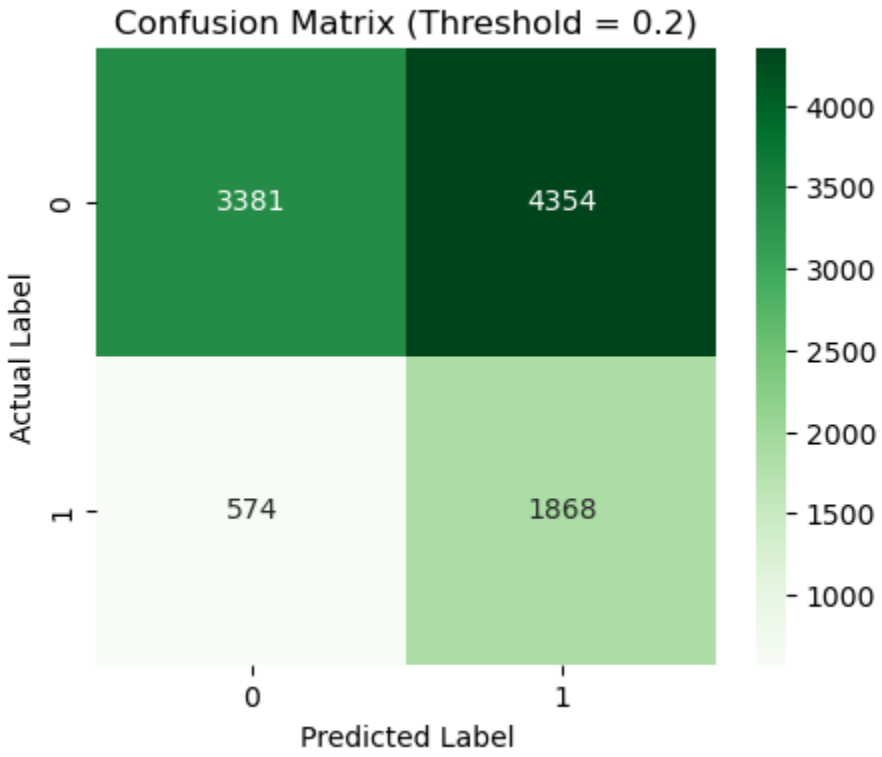
The confusion matrix highlights the impact of class imbalance on model performance. While the Random Forest model demonstrates strong ability to identify non-converting leads, it is more conservative in predicting conversions, resulting in a higher number of false negatives. This behavior is typical in imbalanced classification settings and aligns with the model's moderate ROC–AUC score. From a business perspective, this suggests that the model is effective at filtering out low probability leads but may miss some potential converters. This trade-off motivated the use of probability-based evaluation and threshold tuning to better balance precision and recall depending on business objectives.

Figure 19. ROC Curve – Random Forest Model



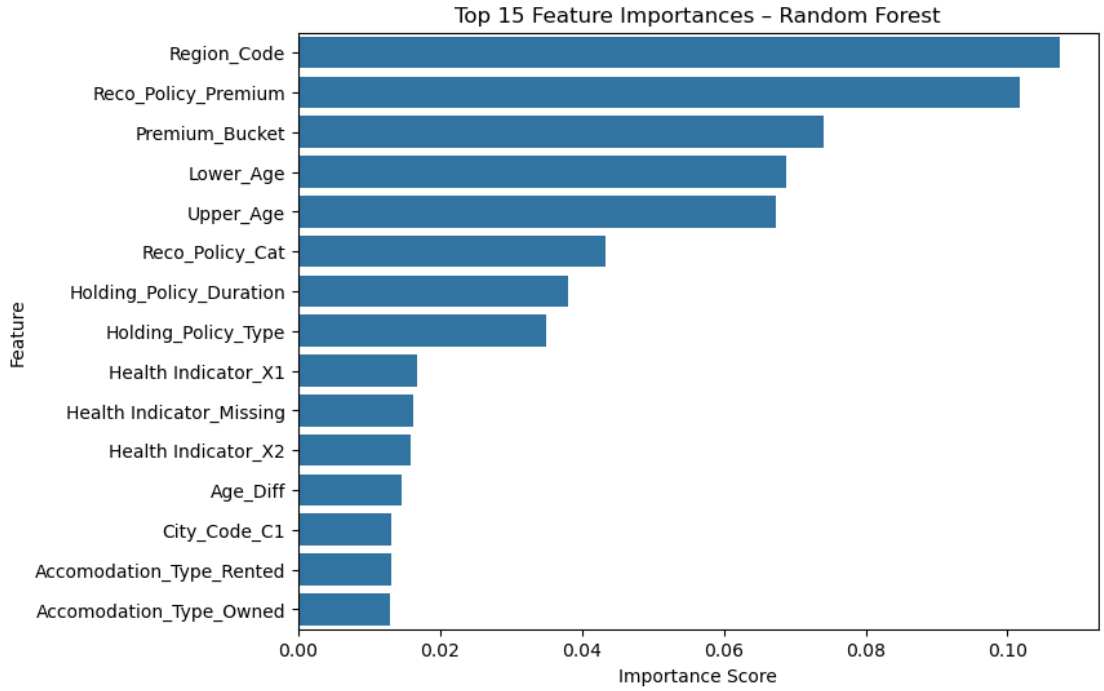
The shape of the ROC curve shows that the Random Forest model has moderate discriminatory power in separating converting from non-converting insurance leads. An AUC of approximately 0.65 suggests the model performs meaningfully better than random guessing but is not a perfect classifier. Importantly, the curve rises steadily rather than sharply, indicating that gains in recall for converters come with increasing false positives. From a business standpoint, this supports the use of probability based decision thresholds rather than a fixed 0.5 cutoff, allowing stakeholders to balance missed opportunities against outreach costs depending on campaign objectives.

Figure 20. Confusion Matrix at Custom Threshold (0.2)



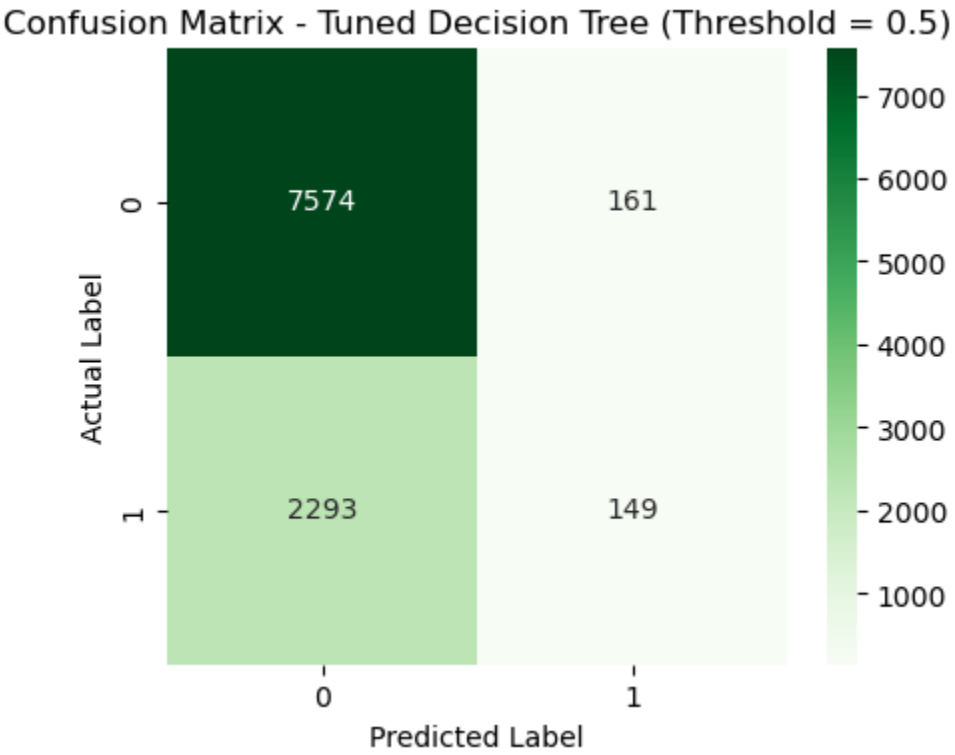
Lowering the classification threshold to 0.2 significantly shifts the model toward a high-recall strategy, allowing it to correctly identify a much larger proportion of actual converters (1,868 true positives) while substantially reducing missed opportunities (574 false negatives). This improvement comes at the cost of a sharp increase in false positives (4,354), meaning more non-converting leads are flagged as likely converters. From a business perspective, this trade-off is often acceptable in insurance marketing contexts where the cost of missing a genuinely interested customer outweighs the cost of additional outreach. The result highlights why threshold tuning matters: the model is not just a predictive tool but a decision-support system that can be calibrated to match campaign objectives, budget constraints, and risk tolerance, making it directly actionable for real-world lead prioritization and sales strategies.

Figure 21. Feature Importance



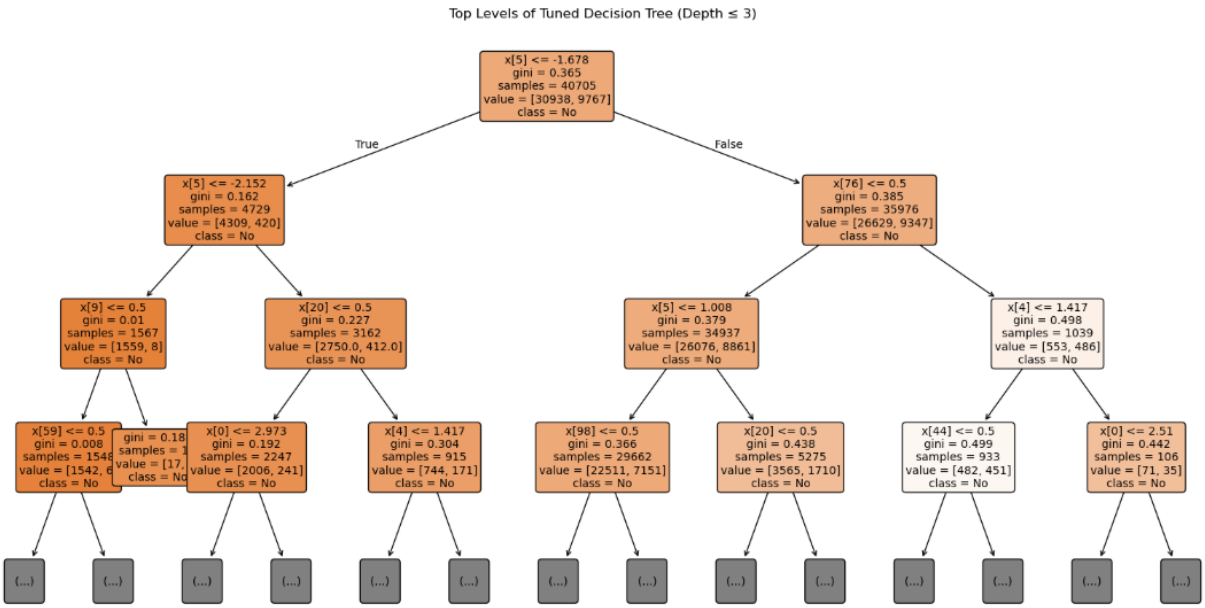
The feature importance results show that geographic and pricing-related variables are the strongest drivers of conversion, with Region_Code and Reco_Policy_Premium emerging as the most influential predictors. This suggests that customer response behavior is highly location-sensitive and closely tied to the recommended premium amount, reinforcing the importance of region-specific pricing and targeting strategies. Age-related variables (Upper_Age, Lower_Age, and Age_Diff) also play a meaningful role, indicating that life stage and age bands influence insurance purchase decisions. Policy characteristics such as premium bucket, policy category, holding duration, and policy type further contribute, highlighting that both affordability and product structure matter in conversion outcomes. From a business standpoint, these insights matter because they point directly to levers the company can control regional segmentation, premium optimization, and tailored product recommendations allowing marketing and sales teams to prioritize high-impact segments, personalize outreach, and improve campaign efficiency rather than relying on broad, uniform targeting

Figure 22. Confusion Matrix – Tuned Decision Tree (Threshold = 0.5)



The tuned Decision Tree model shows strong performance in identifying non-responding customers, correctly classifying 7,574 non-conversions, but performs poorly in detecting actual responders, with only 149 true positives compared to 2,293 false negatives. This indicates that the model is highly conservative and biased toward predicting non-conversion, likely due to class imbalance and the default probability threshold of 0.5. From a business perspective, this means the model would miss a large proportion of customers who are actually willing to convert, resulting in lost sales opportunities and inefficient targeting. While the Decision Tree offers interpretability, its low recall for responders makes it unsuitable as a primary decision tool for lead prioritization. This highlights why threshold tuning and alternative models, such as Random Forest, are more appropriate when the business objective prioritizes capturing potential converters over minimizing false positives.

Figure 23. Top Levels of the Tuned Decision Tree (Maximum Depth = 3)



This decision tree visualization shows the most influential decision rules learned by the tuned Decision Tree model, limited to the top three levels for interpretability. The root and early splits are dominated by transformed numeric features (e.g., standardized age and premium-related variables), indicating that age-related attributes and pricing signals are the primary drivers of conversion decisions. Most nodes still predict the “No” (non-conversion) class, which aligns with the underlying class imbalance in the dataset and reflects the real-world challenge of converting insurance leads. The progressive reduction in Gini impurity across branches demonstrates that the model is effectively segmenting customers into more homogeneous risk groups as depth increases. From a business perspective, this tree provides actionable rules that can be translated into targeting strategies for example, identifying age premium combinations where conversion likelihood improves while also justifying why tree-based models are valuable for explainability in regulated domains like insurance.