BIA | BOSTON INSTITUTE OF ANALYTICS ®

Insurance Policy Response Prediction

Presenter Name: Ishank Mittal

Agenda

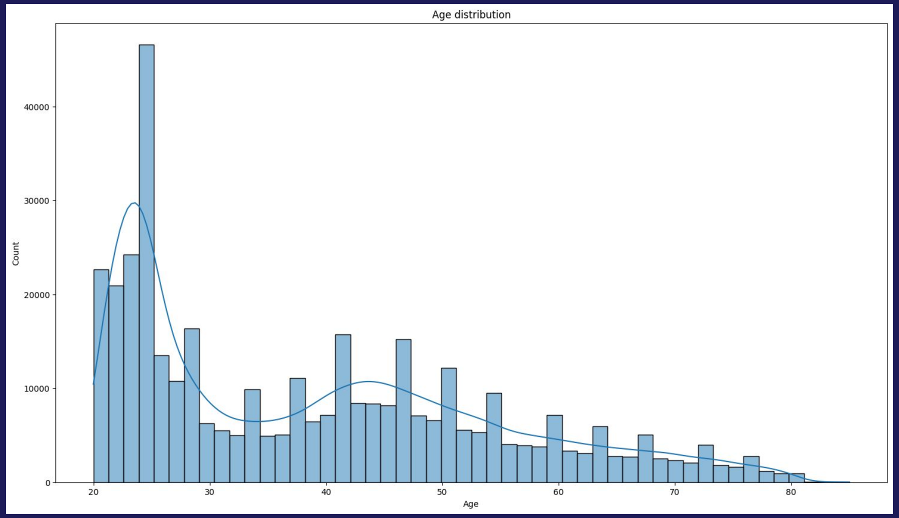
- Project Overview & Business Problem
- EDA Inferences: Distribution & Skewness
- EDA Inferences: Categorical Associations (Cramer's V)
- Data Preprocessing & Transformation Strategy
- Multicollinearity & Feature Engineering
- Model Selection Logic: Why Tree-Based Models?
- Handling Class Imbalance (SMOTEEK & Class Weights)
- Final Model Performance (AdaBoost)
- Conclusion & Questions

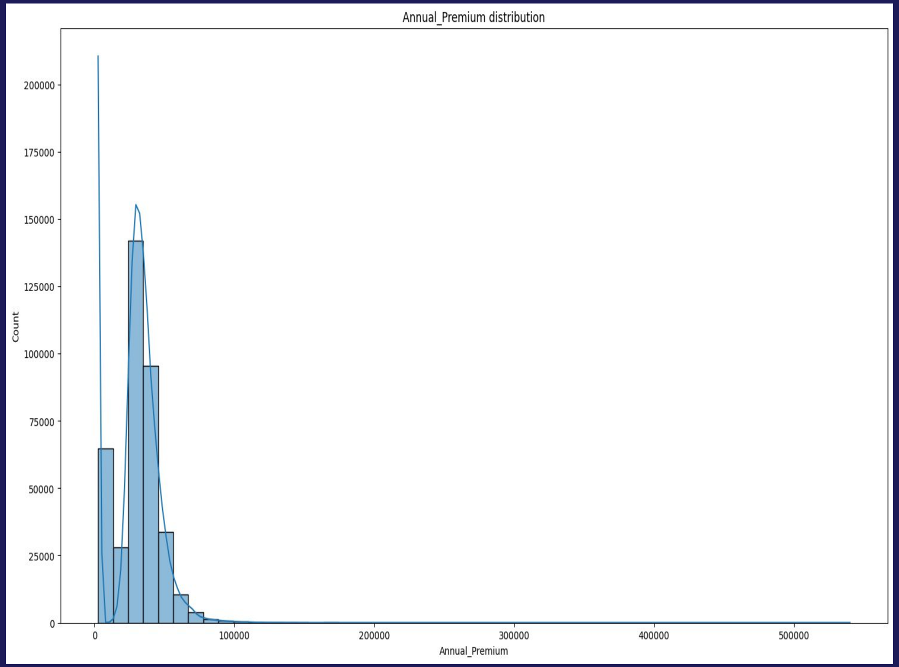
Project Overview & Problem Statement

- **Goal:** Predict customer responses to vehicle insurance offers to increase sales efficiency.
- **The Imbalance Challenge:** The dataset is highly imbalanced (~88% non-respondents); a naive model predicting all "0"s would still be 88% accurate but useless for sales.
- **Business Priority:** Focus on Recall. In insurance, it is better to contact someone uninterested (False Positive) than to miss a high-value interested customer (False Negative).
- **Critical Monitoring:** Use the Confusion Matrix to ensure the model actually learns positive response patterns rather than just predicting the majority class.

EDA: Univariate Analysis

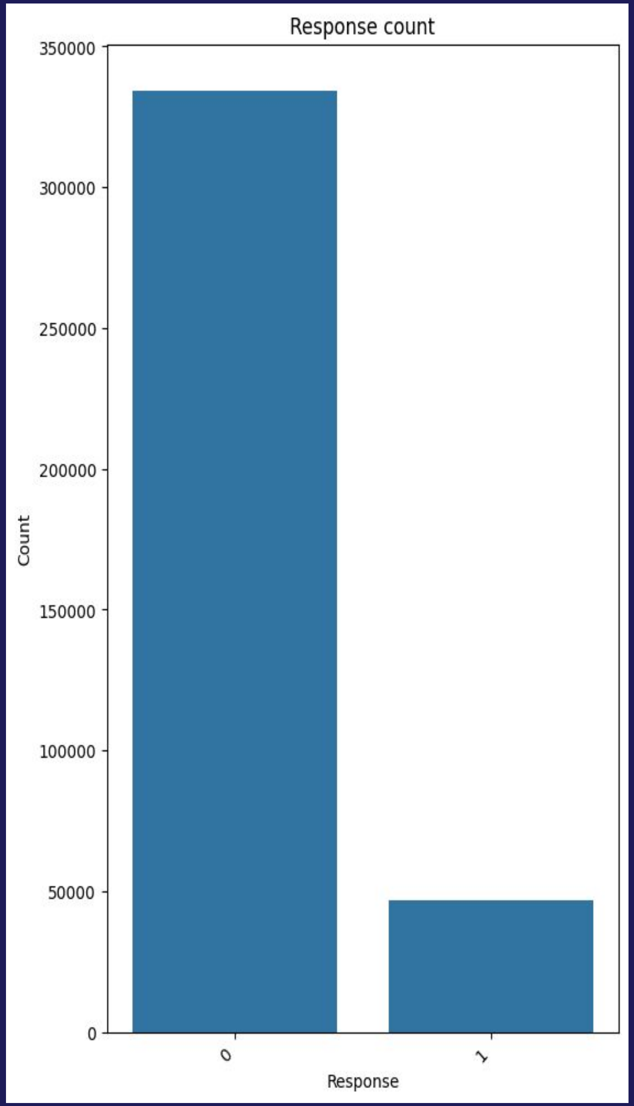
- Age is right-skewed with a strong concentration between 20–30 years
- The Vintage feature shows an approximately uniform distribution across customer tenure, indicating balanced representation of new and long-term customers.
- **Skewness & Kurtosis:** Annual Premium has a **Skewness** of 1.769 (Highly right-skewed) & **Kurtosis** of 34.103 (Extremely high, indicating many outliers).
- **Rule of Thumb:** Any feature with skewness > 1 or < -1 requires transformation.
- Annual Premium requires modern Box-Cox transformation and outlier capping to be useful for modeling.





EDA: Univariate Analysis

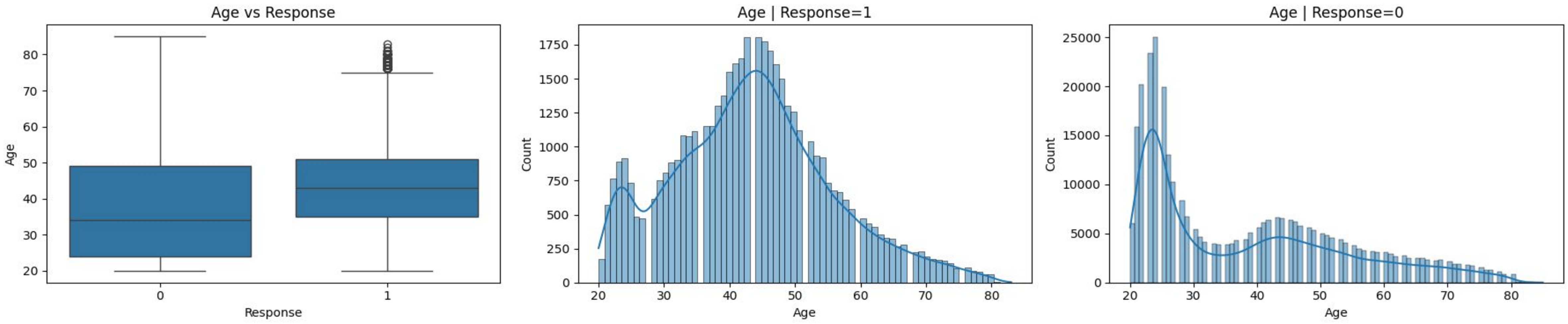
- Nearly all customers possess a driving license
- Positive respondents are overwhelmingly those with previous vehicle damage.
- Customers with vehicles < 1 year old have the lowest response rate
- **Class Imbalance:** Only about 12% of customers responded positively to the insurance offers. This confirmed that standard accuracy would be a misleading metric.
- While training, need to use oversampling techniques like smote



BIA BOSTON INSTITUTE OF ANALYTICS

EDA: Demographic Analysis

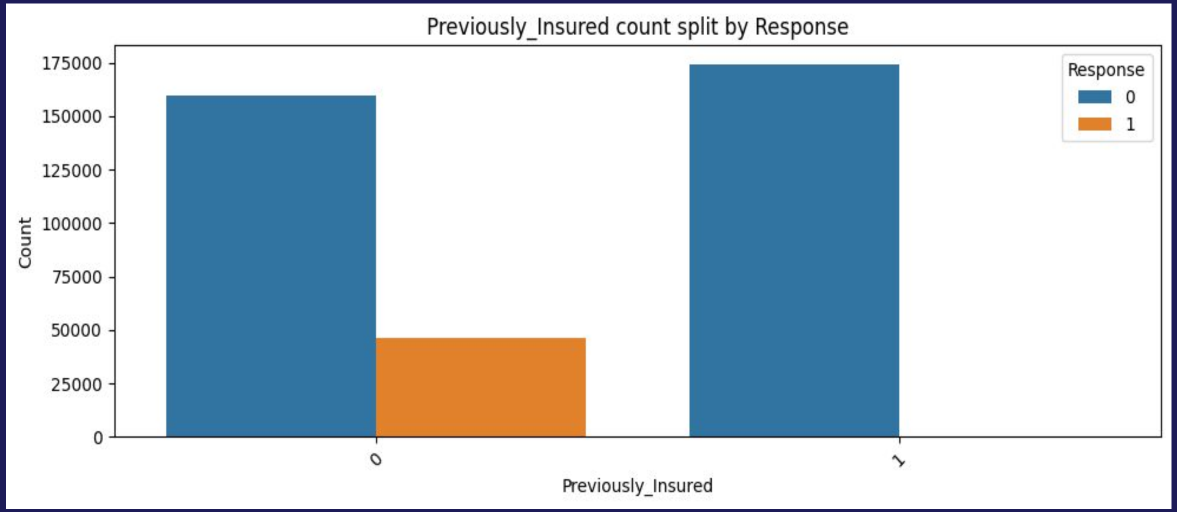
- Younger people (around age 20–30) are much less likely to respond. The "sweet spot" for a positive response is the middle-aged demographic, peaking around 40–50 years old.
- A higher volume of positive responses is observed among Male subjects compared to Female subjects.
- The vast majority of the population possesses a license; this variable shows minimal variance between response classes.

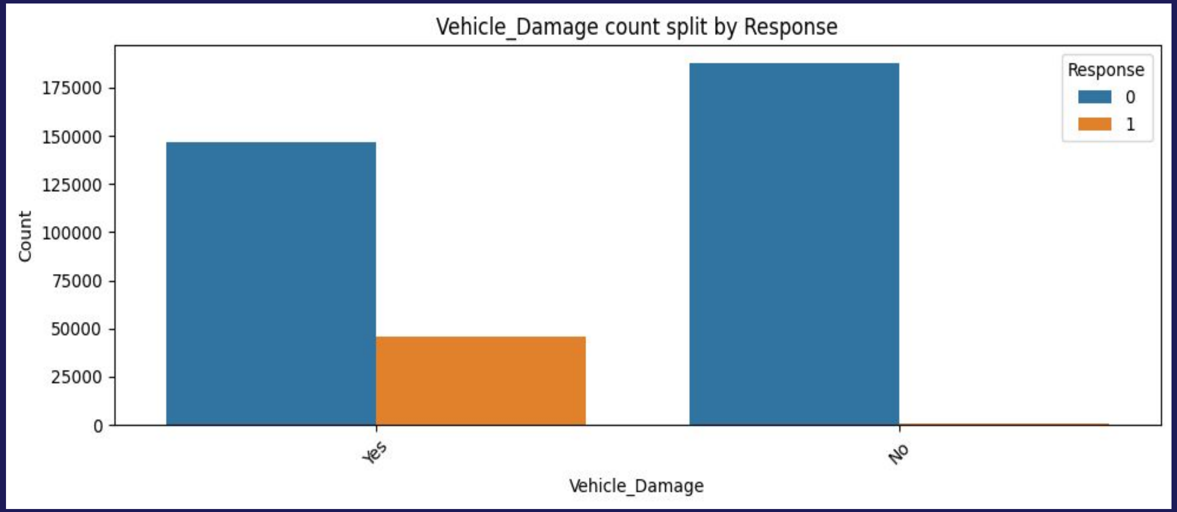


BIA BOSTON INSTITUTE OF ANALYTICS ®

EDA: Vehicle & Insurance History

- **Vehicle Damage:** Positive responses are almost exclusively concentrated among subjects with a history of prior vehicle damage.
- **Previously Insured:** There is a strong negative correlation between being previously insured and providing a positive response. Most "Yes" responses come from uninsured subjects.
- **Vehicle Age:** Subjects with vehicles aged 1–2 years show the highest proportion of positive responses. Vehicles < 1 year show the lowest.

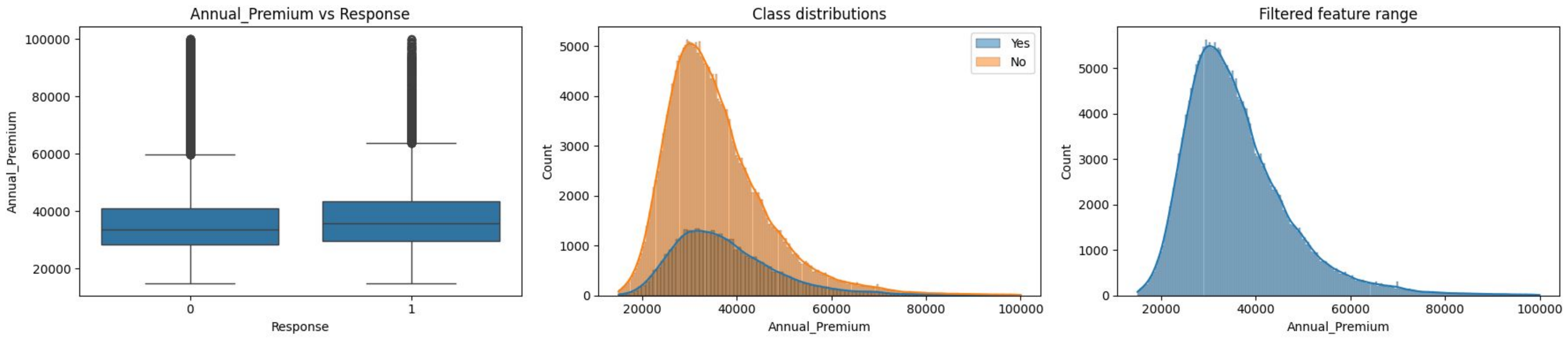




BIA BOSTON INSTITUTE OF ANALYTICS

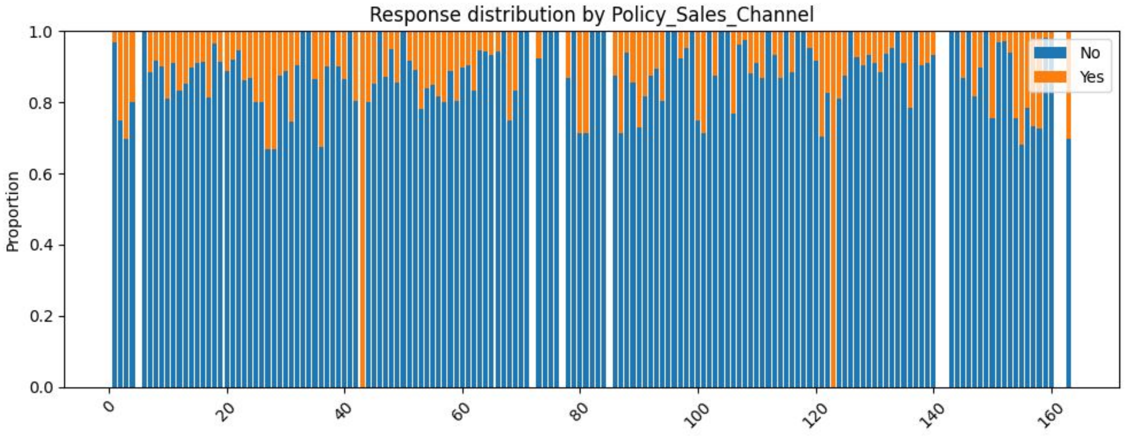
EDA: Policy & Engagement Metrics

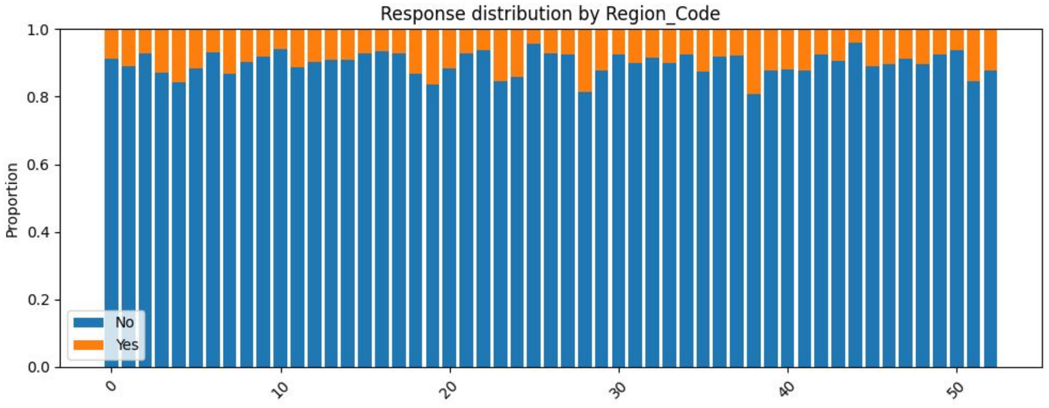
- **Annual Premium:** The distribution of premium costs is statistically similar for both response classes, characterized by a right-skew and significant outliers.
- **Policy Sales Channel:** Response rates vary significantly across different Channel IDs, with specific channels yielding higher proportions of "Yes" responses.
- **Region Code:** Distribution of responses is non-uniform across geographic region codes.
- **Vintage:** The duration of customer tenure (Vintage) shows a uniform distribution across both response categories, indicating no correlation with the target variable.

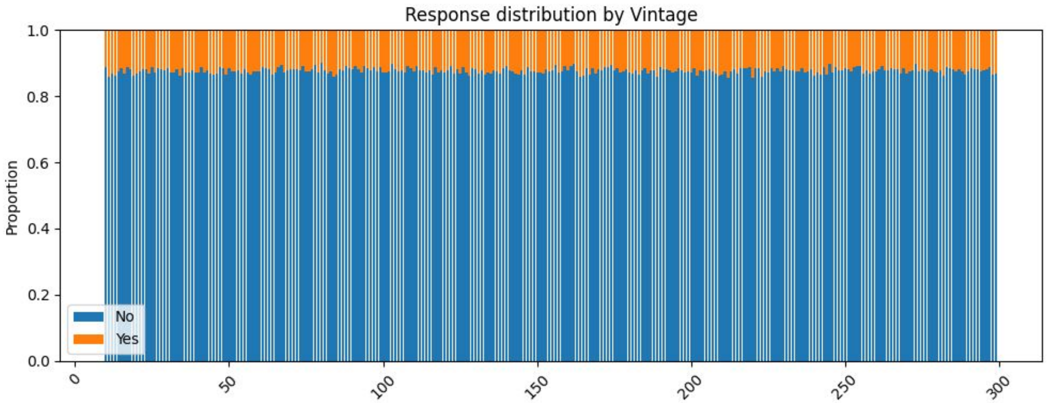


BIA BOSTON INSTITUTE OF ANALYTICS ®

EDA: Policy & Engagement Metrics



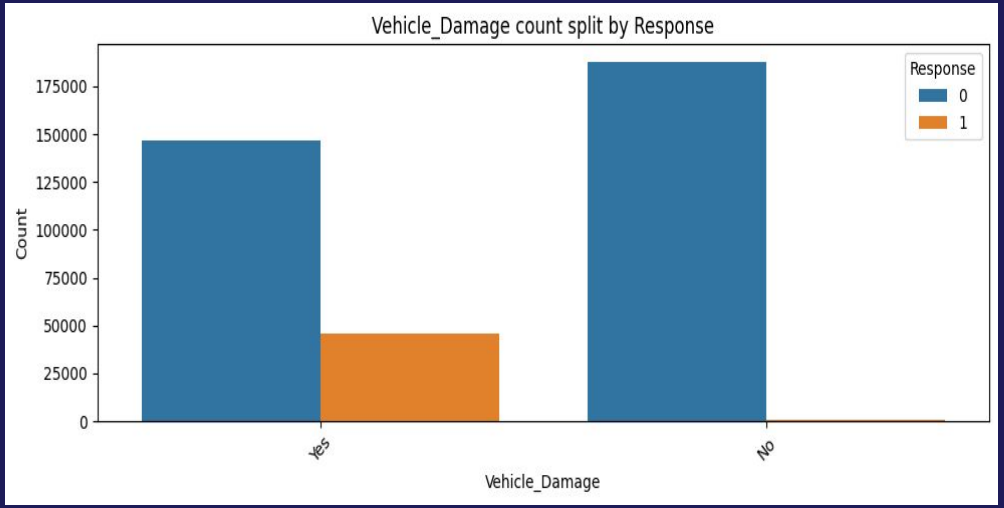


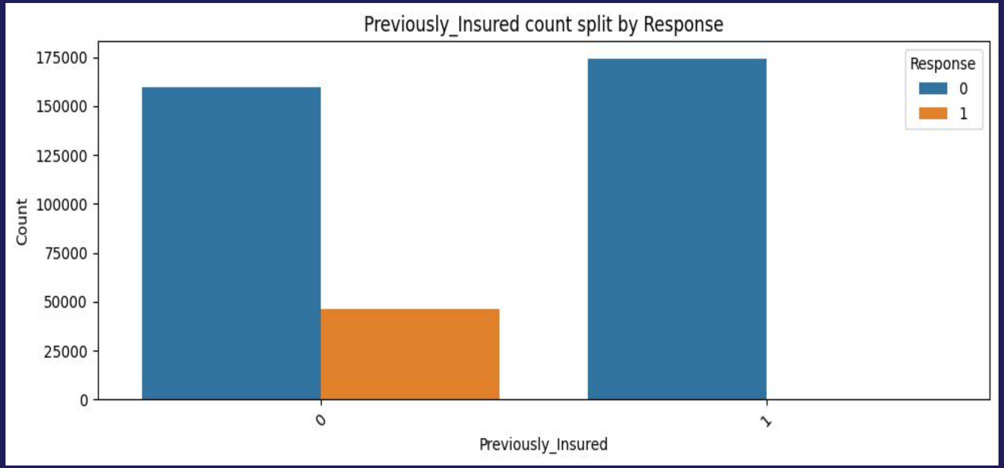


BOSTON INSTITUTE OF ANALYTICS

EDA: Bivariate Analysis

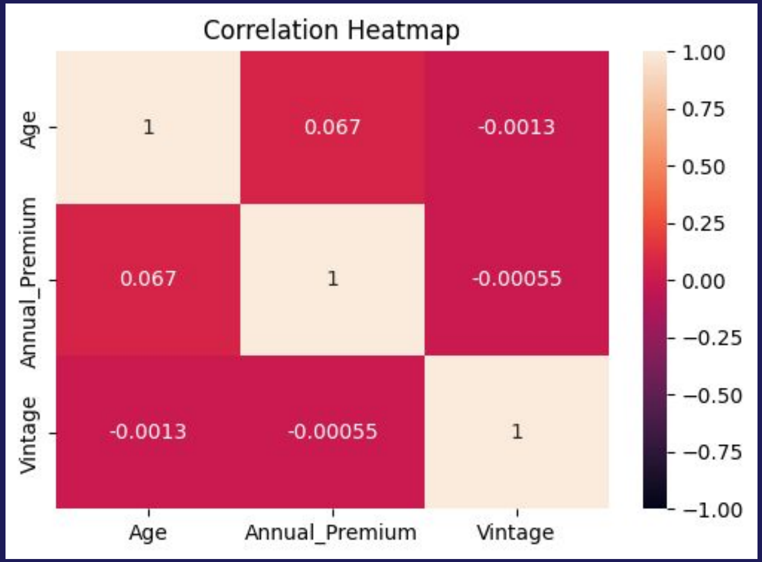
- Customers who have had previous vehicle damage are far more likely to respond positively.
- Customers who do not already have vehicle insurance are the primary respondents.
- Owners of vehicles 1–2 years old (and to a lesser extent, >2 years) are much more interested. Owners of new vehicles (<1 year) show very low interest.

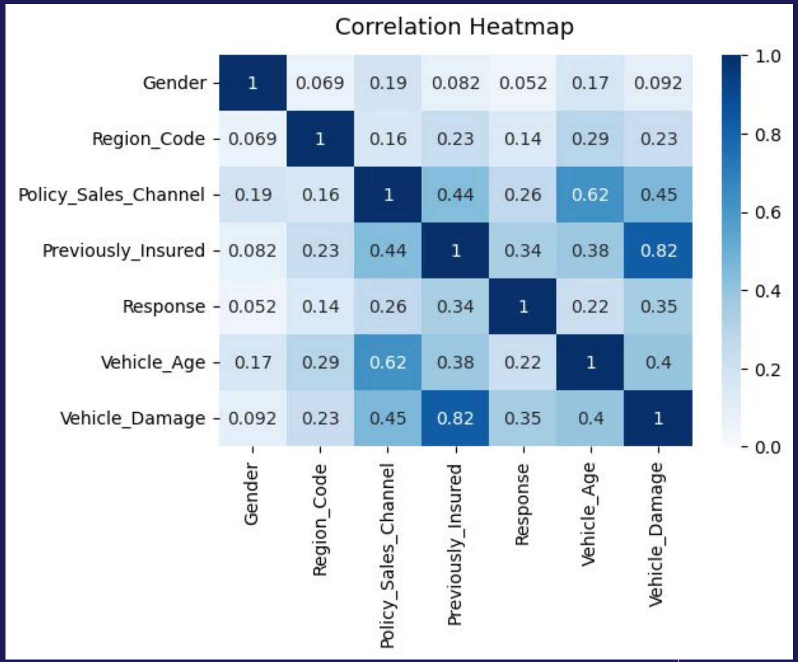




EDA: Multivariate Analysis

- All numerical features exhibit very low linear inter-dependency, with all non-diagonal coefficients falling below
- Vehicle_Damage (0.35) and Previously_Insured (0.34) show the strongest association with the Response variable.
- A very high correlation exists between Vehicle_Damage and Previously_Insured (0.82).
- Vehicle_Age shows a significant association with Policy_Sales_Channel (0.62).
- Gender shows the lowest correlation with the Response variable (0.052).

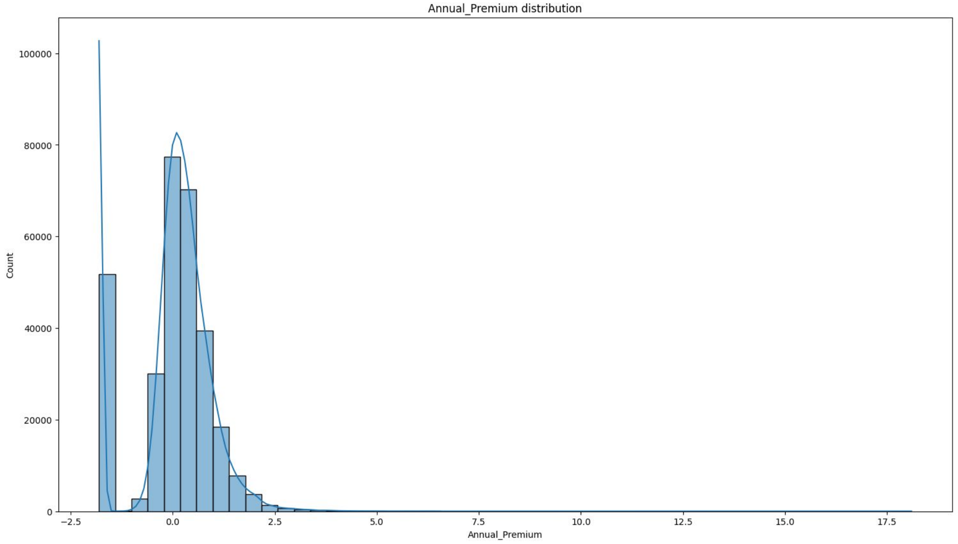


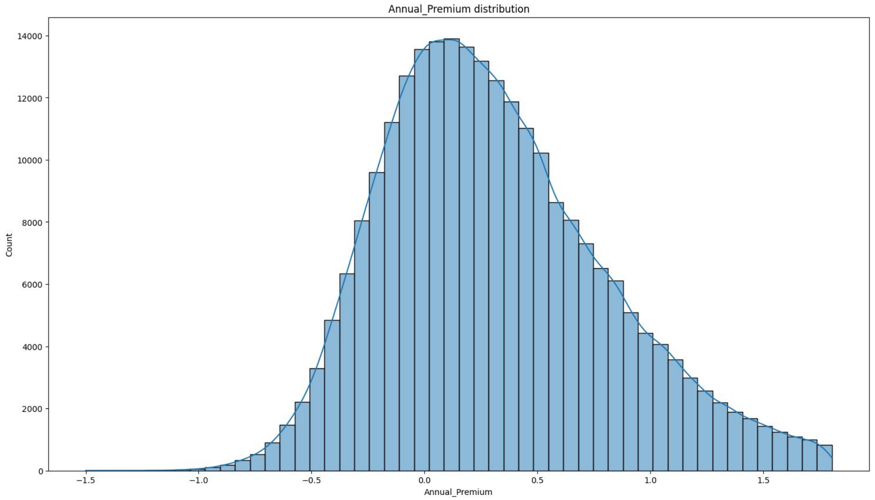


BIA BOSTON INSTITUTE OF ANALYTICS

Data Preprocessing & Feature Engineering

- Split the dataset into test and training datasets
- Applied modern Box-Cox to Annual_Premium to normalize distribution. And capped the feature using IQR method.
- Used binary encoding to transform columns with binary features and frequency encoded Policy sales channels and region code as they have multiple values.
- Scaled the features





Data Preprocessing & Feature Engineering

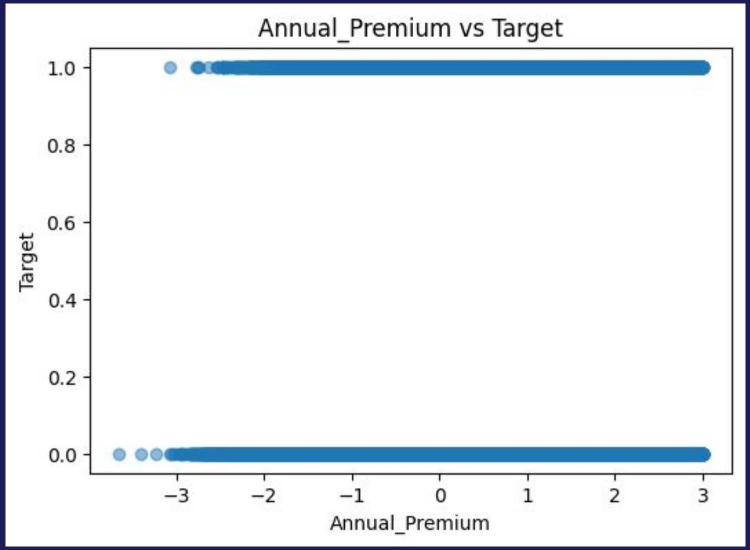
- Variance Inflation Factor (VIF) is a statistical measure used to detect the presence and severity of multicollinearity in a regression analysis.
- Multicollinearity occurs when independent variables (features) are highly correlated with each other rather than being truly independent.
- All features exhibit a VIF below the common threshold of 5 (or 10), suggesting that multicollinearity is within an acceptable range for most modeling purposes.
-

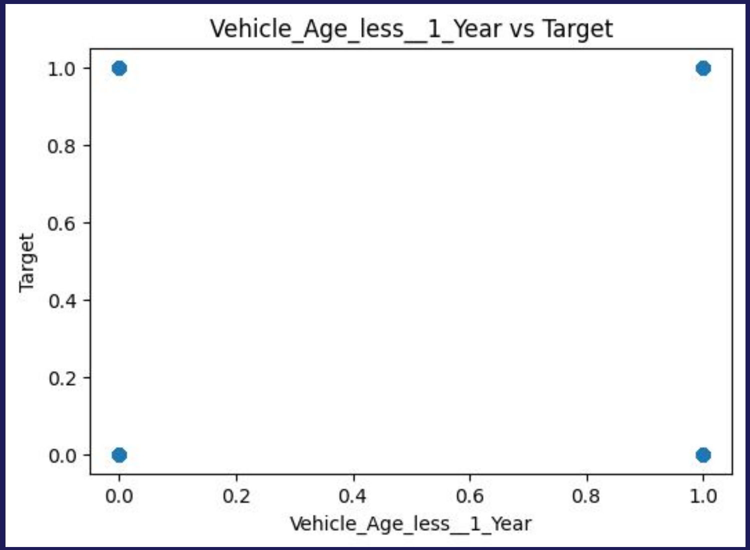
Feature	VIF
Vehicle_Age_< 1 Year	3.567543
Vehicle_Damage	3.526181
Previously_Insured	3.481877
Age	2.971635
Policy_Sales_Channel	1.485093
Region_Code	1.335274
Annual_Premium	1.183715
Vehicle_Age_> 2 Years	1.090315
Gender	1.033519
Driving_License	1.008248
Vintage	1.000056

BIA BOSTON INSTITUTE OF ANALYTICS

Model Selection Logic

- **Non-Linear Relationships:** Significant class overlap across all features indicates that response behavior is driven by non-linear decision boundaries.
- **Interaction-Driven Signal:** Linear models fail to capture complex dependencies, such as the combined effect of Age and Vehicle Damage.
- **Weak Correlation:** Target variables show low linear correlation with individual features, making standard regression less effective.
- **Tree-Based Advantage:** Algorithms like Random Forest or XGBoost naturally handle skewed data and the high volume of outliers found in Annual Premium.
- **Automatic Feature Learning:** Tree models capture complex patterns without the need for manual feature engineering.

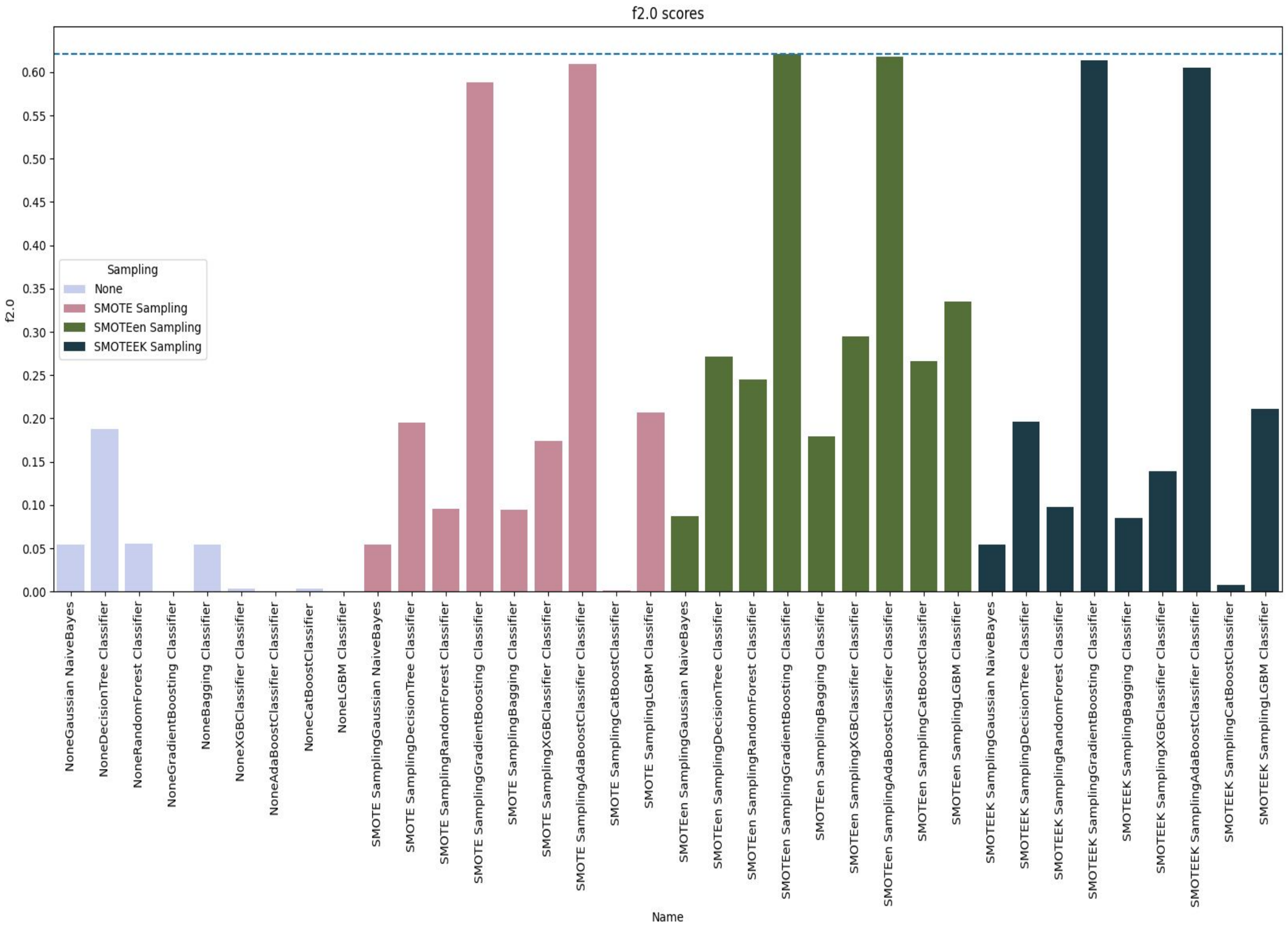




BOSTON INSTITUTE OF ANALYTICS

Imbalance Handling & Evaluation Strategy

- **SMOTEEK/SMOTEEN:** Synthetic oversampling combined with cleaning techniques to clarify the decision boundary.
- **Metric Focus:** Prioritized the F2-measure to put more weight on minimizing False Negatives (missing potential customers).
- **Best Performers:** Top three pipelines included SMOTEEN/SMOTEEK sampled AdaBoost and Gradient Boosting classifiers.



Final Performance – AdaBoost + SMOTEEK

- **Training Results:** The fine-tuned AdaBoost model on the SMOTEEK dataset was identified as the final model.
- **Performance Summary:**
  - **Accuracy:** 0.7929
  - **Recall (Risk Class):** 0.7015— successfully identifying over 70% of interested leads.
  - **F1-Macro:** 0.6604
- **Classification Insight:** While precision for the interested class is 0.33, the high recall ensures maximum business opportunity capture

Conclusion & Business Impact

- **Strategic Result:** The model effectively filters the customer base to a high-probability subset, capturing 70% of all potential responses.
- **Optimization**: Targeted marketing can now focus on customers with previous vehicle damage and specific vehicle age ranges to maximize conversion.+
- **Final Output:** Model saved as best_model_AdaBoost_SMOTEEK.pkl for production use, with all experiments logged via MLflow.

Questions ?

Thank You!