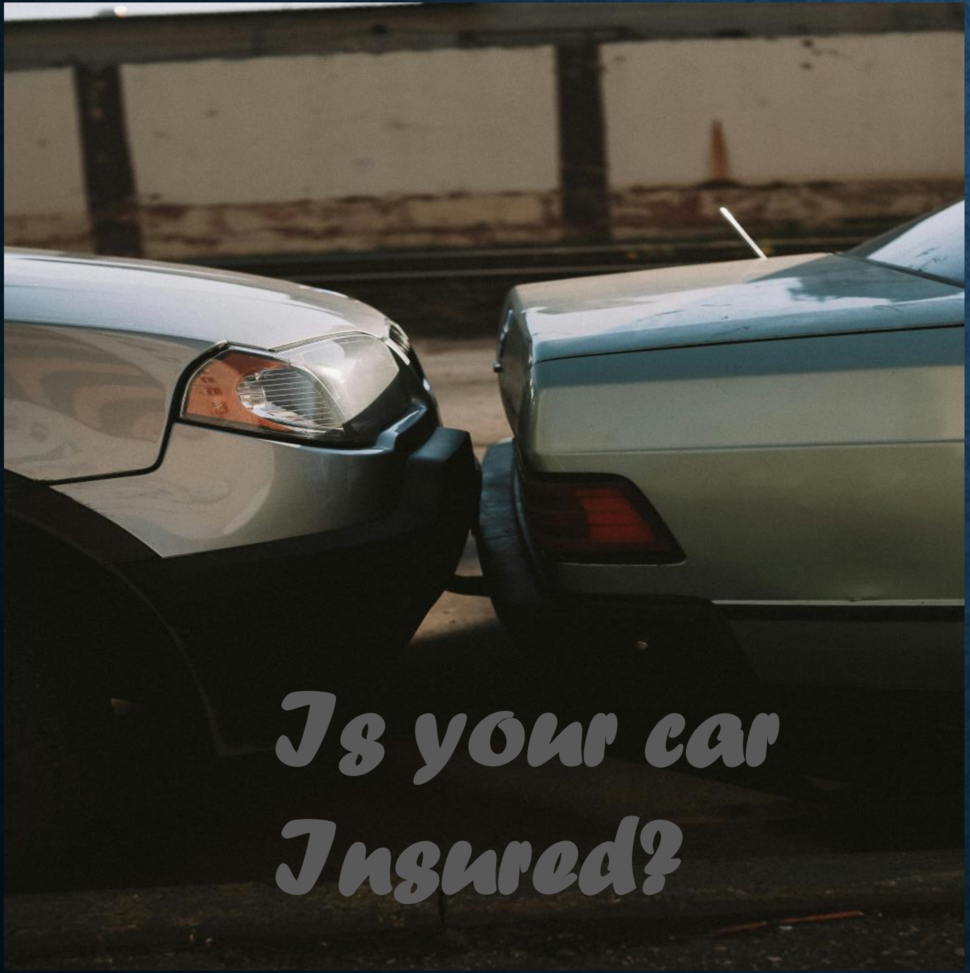
INSURANCE CROSS SELL PREDICTION

Project P.lan



1. Building a model to predict whether the policy holders from past year will be interested in Vehicle Insurance.

2. Provide business insights and recommendations to help the company design effective marketing campaigns

BUSINESS PROBLEM

- TruSecure insurance company is looking to expand its services by offering vehicle insurance policies. The company needs help in developing a predictive model to determine whether existing health insurance policyholders are likely to show interest in purchasing a vehicle insurance policy as well.

- To support this business goal, analyze customer data from the past year — including demographics, risk indicators, and policy history — to build a reliable model that predicts conversion likelihood. In addition to predictive modeling, provide business insights and strategic recommendations to help the company design effective marketing campaigns and ensure that policy offerings are inclusive, fair, and aligned with customer behavior

DATA OVERVIEW

The Dataset used in this project is Health Insurance Cross Sell Prediction dataset which is availale in kaggle. It contains important information about past clients details that may influence their decisions to purchasing vehicle insurance

Key Columns;

- Previously_Insured - Whether the client had been previously insured by our company. Yes =1 and No = 0
- Vehicle_Age - Whether vehicle age is above 2 years, between 1 and 2 years and below 1 year
- Vehicle_Damage - whether client had damaged vehicle Yes/No
- Annual_Premium - The total amount of money the client should pay per year.
- Policy_Sales_Channel - The channel used to engage and get response from clients. It is represented by anonimous code
- Vintage - The number of days the client has been registered in the company
- Response - This is the positive or negative response by the client. Positive response = 1 and negative response = 0

Note, the above are just but a fraction of columns, more columns and their decription are available in the notebook

PROJECT OBJECTIVES

1.Predict whether a customer will respond positively to an insurance offer

2. Identify Key Factors Influencing Purchase Decision

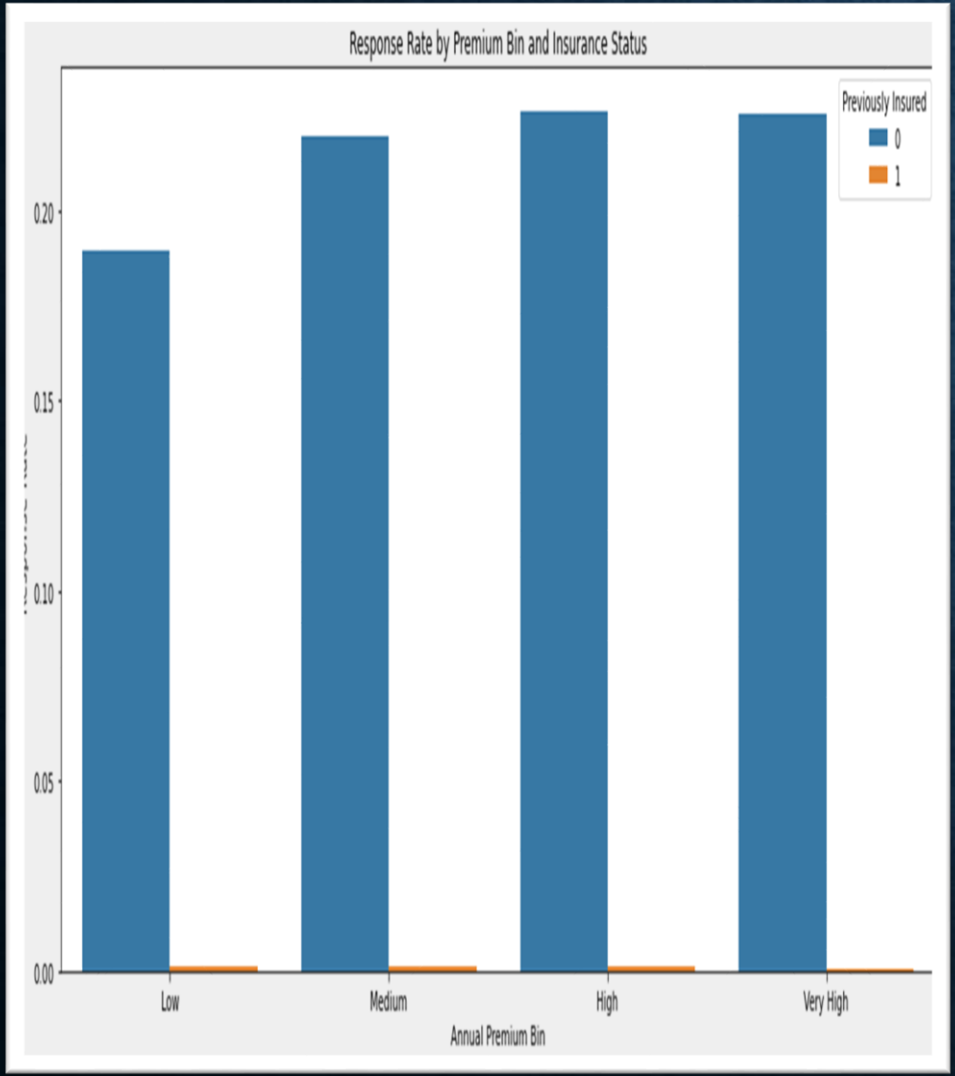
3.Segment Customers for Targeted Marketing
Group customers into meaningful segments based on age, risk indicators, and policy history to improve marketing efficiency

4. Provide Actionable Business Insights

Recommend data-driven strategies for improving cross-sell conversion rates

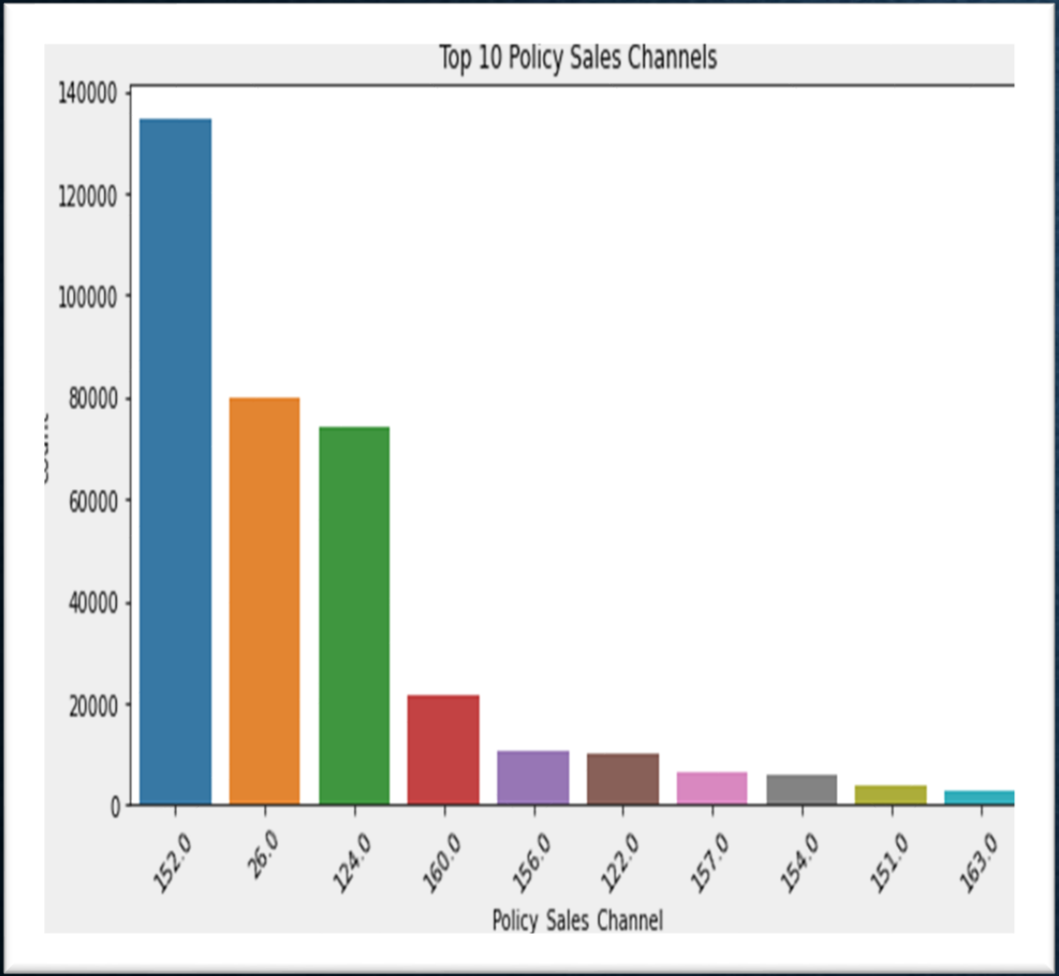
EXPLANATORY DATA ANALYSIS

Response Rate by Premium bins and Insurance status



- The plot shows four premium levels low, mediam,high and very high annual premium.
- The blue bars shows client had not been previously insured while the orange at the bottom shows the client had been previously insured.
- The height of the bar indicate level of reponse rate
- We see that those who have not been insured dominate the plot. Those who had been insured extreme low response.
- The company should focus marketing on those not currently insured, they’re much more receptive .

TOP 10 POLICY SALES CHANNELS



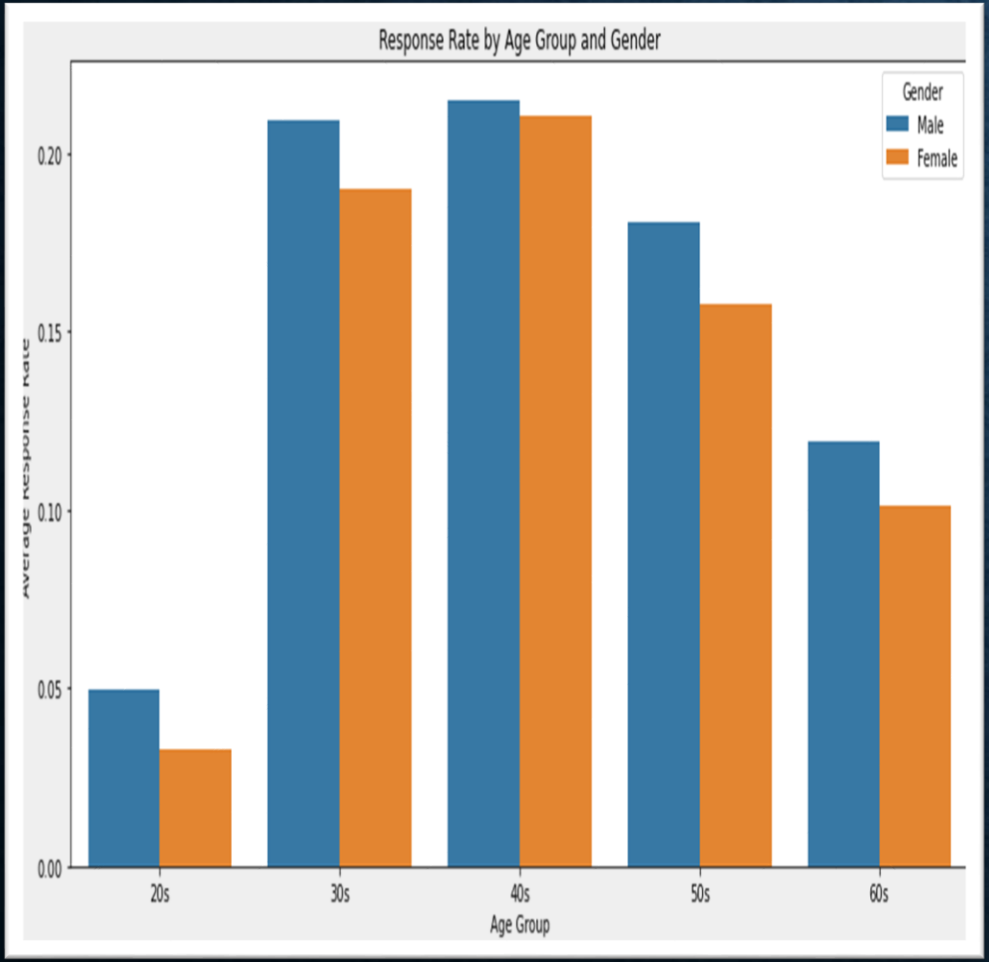
Hypothesis Testing to Check if there is Statistical difference in response rate across channels

Null Hypothesis (H_0)-The rate of response is the same across the top 10 policy sales channels.
Alternative Hypothesis (H_1)-At least one sales channel has a significantly different response rate.

P_value=0.0000
We Reject the Null Hypothesis
There is a statistically significant difference in response rates among the top 10 channels since our p value is less than our significance level of 0.05

Findings
Channel 152 is on the lead with high counts as well as higher response rate followed by Channel 26
This channels can be used for marketing in places whhere we have a lower response rate

RESPONSE RATE BY AGE GROUP AND GENDER



Middle-aged groups (30s to 50s) show higher response rates

- These age bins tend to have higher average response rates compared to the younger (20s) or older (60s) age groups.
- These individuals may be more financially stable and likely to consider purchasing insurance.

Gender differences are minimal.

- However, men dominate the response rate. Our plot shows minimal difference, gender may not be a strong differentiator in overall conversion.

20s and 60s have the lowest response rates

- Customers under 30 or over 60 are less likely to respond positively.
- Younger individuals may feel less need for insurance.
- Older individuals may already be insured or face premium/eligibility barriers.

MODELING

Objective

To predict whether previous policy holders will also
Vehicle Insurance provided by the company

Data Preprocessing

1. Encoding Categorical Variables

- Cleaned and standardized string values in Gender, Vehicle_Damage, and Vehicle_
- Used **OneHotEncoding** for nominal features (Gender, Vehicle_Damage)
- Applied **OrdinalEncoding** to Vehicle_Age using logical age order (<1 Year < 1-2

2. Feature Scaling

- Scaled continuous variables like Age, Annual_Premium, Vintage, etc.
- Used **StandardScaler** to normalize feature ranges for model compatibility

3. ColumnTransformer Pipeline

- Combined all preprocessing steps in a single pipeline to maintain consistency and

MODEL PERFORMANCE COMPARISON

Model	Accuracy	Class 1 Recall	Class 1 Precision	F1 Score (Class 1)	ROC AUC
Logistic Regression	0.88	0.00	0.40	0.00	0.8342
XGBoost	0.72	0.90	0.29	0.44	0.8558
Random Forest	0.69	0.94	0.28	0.43	0.8548
LightGBM	0.70	0.93	0.28	0.43	0.8578

From the above, predicting who will respond positively is more important than just getting high overall accuracy. Therefore, we will keep a close eye on;

•Model with High recall and precision on class 1 (the minority class).

•Good AUC (how well the model separates the classes).

RECOMMENDED MODEL BASED ON PERFORANCE

LightGBM

•Highest **ROC AUC** = 0.8578 → best class separation overall.

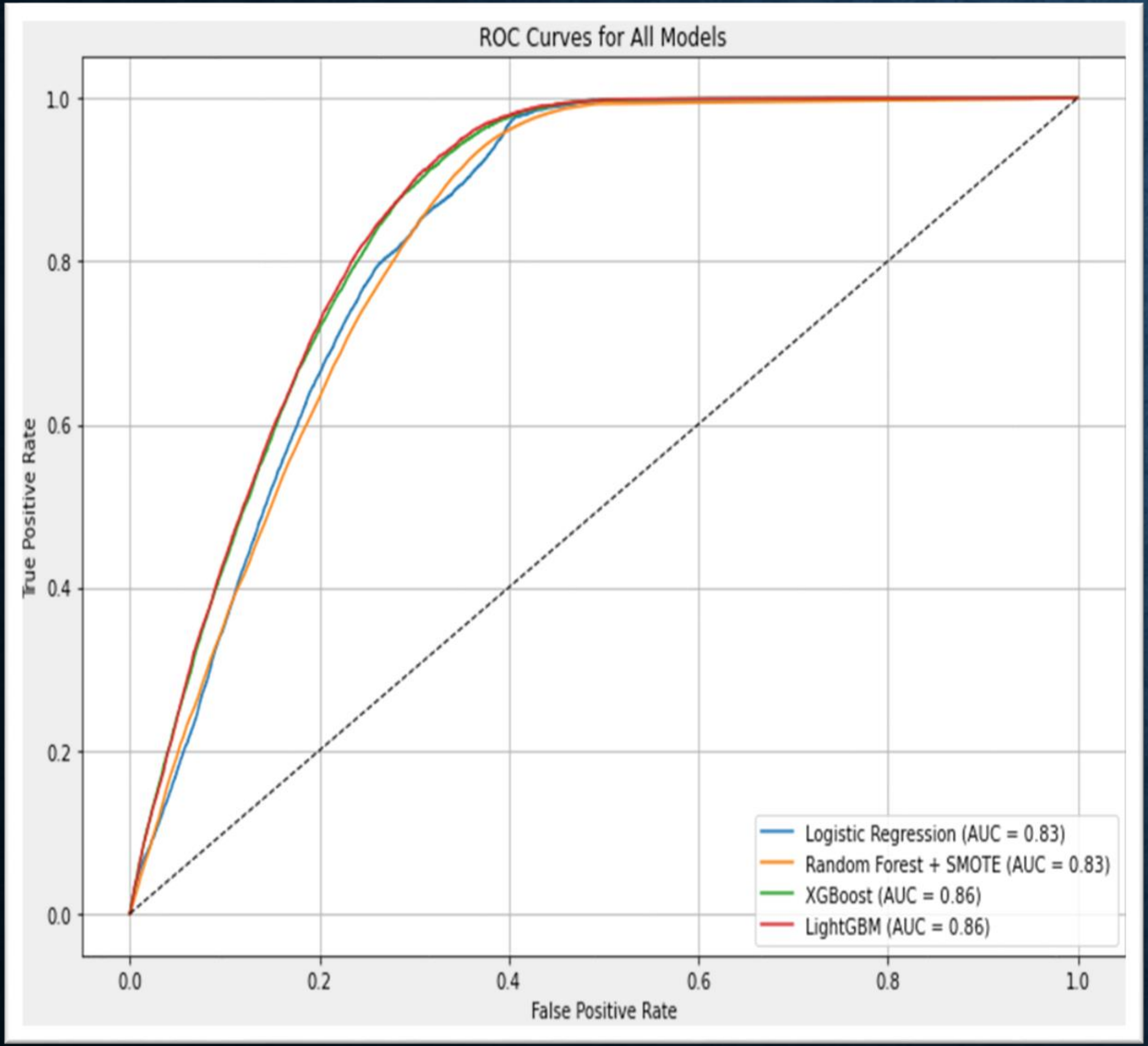
•Highest recall on class **1** = 0.93 → captures most positive responders.

•F1-score and precision similar to XGBoost and Random Forest → fair balance.

•Good overall **trade-off between performance and interpretability**.

•Can be made faster and lighter in deployment.

ROC CURVES FOR ALL MODELS



LightGBM had the **highest ROC AUC** (≈ 0.858), meaning it is slightly better at ranking customers correctly than the other models.

CONCLUSION

While all models were fairly strong, LightGBM demonstrated the best balance between correctly identifying potential responders and minimizing false positives.

This makes it the most suitable model for deployment, especially if our goal is to **target the right** customers effectively without overwhelming resources.

BUSINESS RECOMMENDATIONS

1.Use LightGBM model for production deployment since it offers better performance with fewer resources and handles imbalanced data well.

2. Focus personalized campaigns on mid-aged groups (30–50s). Also consider education campaigns or beginner-friendly policies for younger groups.

3. Prioritize the top-performing channels such as channel 152 and 26 in underperforming regions so as to improve marketing. However ,enhance agent training and lead quality in high-traffic channels like 123 and 43.

4. Target Uninsured Customer since they show significantly higher response rates. Focus marketing campaigns and Allocating higher outreach budget to this segment will result to a better ROI.

NEXT STEPS

1. Model Deployment

- Deploy the **LightGBM model** into production to predict customer responses in real-time.
- Integrate the model with the marketing system to **prioritize leads** based on response probability.

2. A/B Testing of Campaigns

- Run A/B tests using the model to **compare targeted marketing (model-based)** vs **broad campaigns**.
- Measure **conversion rates, cost per acquisition (CPA)**, and ROI improvements.

CONCLUSION

- This project has provided valuable insights into customer response behavior and identified the key drivers behind successful insurance policy conversions.

- Through predictive modeling, we uncovered that LightGBM **model** emerged as the best-performing algorithm, with strong recall for the positive class and the highest ROC AUC, making it well-suited for real-world deployment.

- With the model in place, the organization can now proactively target high-value leads, reduce acquisition costs, and improve policy uptake rates.

ANY QUESTIONS

Thank You