INSURANCE CROSS-SELL PREDICTION

Rahull Borana
Simran Kaur
Aditya Kumar
Manideep Telukuntla
Daniel Lievano

PROJECT GOAL

To enhance the client penetration of an insurance company who just introduced a new segment of vehicle insurance by means of a predictive model of cross-sell analysis to effectively target potential clients and improve decision-making regarding vehicle insurance sales.

PROBLEM OVERVIEW

Namaste Insurance, a prominent health insurance provider, is venturing into vehicle insurance, leveraging its established market presence for cross-selling insurance products

Namaste Insurance is dedicated to enriching client engagement by harnessing the potential of a predictive model, aimed at discerning customers' inclination towards vehicle insurance purchase. Through a meticulously designed three phase methodology encompassing data refinement, in-depth exploratory analysis, and strategic model selection, we are poised to elevate our sales approach. This initiative not only promises to forecast customer behavior but also empowers our sales team to efficiently pinpoint potential clients by uncovering the pivotal factors influencing their vehicle insurance decisions.

PROJECT MOTIVATION

The primary motivation behind this project is to capitalize on Namaste Insurance's well-established position in the market and its recent expansion into the vehicle insurance segment. By leveraging data-driven insights through a predictive cross-sell model, the company aims to maximize its market penetration and revenue potential. This initiative not only enhances customer targeting and engagement but also streamlines decision-making processes, ensuring that the company remains at the forefront of the insurance industry's evolution and continues to provide tailored solutions to its clients.

Namaste Insurance provides a comprehensive set of customer features, including gender, age, driver's license status, region, previous insurance history, vehicle age, vehicle damage, outreach channel, duration of association, and response status for vehicle insurance purchases.

EXPLORATORY DATA ANALYSIS

Through extensive exploratory data analysis, we delved into diverse variables, uncovering insights into customer interest in vehicle insurance. Our aim was to shape an ideal customer persona with key attributes, yielding valuable guidance for our client's marketing strategy. Beginning with correlation matrix analysis, we identified pivotal variables—previous insurance, accident history, and vehicle age—significantly influencing the response variable. These factors, meticulously examined in our EDA research, hold utmost importance in gauging consumer interest. While noting correlations of lesser significance, such as vehicle age and consumer age, we now delve into defining the optimal customer persona and their defining traits.

Demographics: Notably, older people displayed a distinct preference for vehicle insurance. By targeting this demographic, our client can effectively expand reach and engagement, optimizing marketing strategies.

Customer experience: A noteworthy trend is the appeal of untapped new consumers without prior car insurance. To enhance consumer acquisition, channel efforts towards this market segment. Moreover, targeting "new vehicle owners" presents a promising avenue to cultivate positive experiences and favor our client's offerings.

Regional Patterns: Surprisingly, no significant trends emerged across diverse regions, but minor uptake variation was observed in the West zone.

Channel Preferences: Our exploration uncovered remarkable preferences among consumers who are called and direct-mailed to. Additionally, customers that were digitally targeted demonstrated an encouraging curiosity, offering a practical way to engage with new customers.

SOLUTIONS & INSIGHTS

Overview of Models:

- **Naive Bayes:** In our Multinomial Naive Bayes model, we applied binning to continuous variables for natural implementation. With an accuracy score of 0.764, precision of 0.305, and recall of 0.727, it

offered tailor-made results for our business problem, prioritizing low false negatives. This model allows us to identify potential insurance buyers more effectively.

- **Random Forest:** The initial Random Forest model with 200 estimators and 5 features achieved an accuracy of 0.8697. However, it had a low recall of 0.0993, meaning it missed a significant number of actual insurance buyers. To optimize, we adjusted the threshold to 0.19, resulting in improved recall (0.7543). While this may slightly increase the false positives, the trade-off ensures a substantial reduction in missed genuine buyers, aligning better with our business goals.

- **Gradient Boosting (Best Model):** The Gradient Boosting model, with 400 estimators and max depth of 3, had an accuracy of 87.77%. But its low recall (0.7%) raised concerns as it overlooked 99.3% of actual insurance buyers. To address this, we optimized the threshold at 0.19, significantly improving recall (75%), and strengthening alignment with our primary goal of capturing potential buyers. Although this resulted in a slightly reduced accuracy (78.05%) and precision (32.65%), the model now effectively targets genuine insurance buyers.

- **Logistic Regression:** The logistic regression model is utilizing the 'newton-cholesky' solver, the central aspect of the code involves the classification process based on a predefined threshold value of 0.22. Samples where the predicted positive class probability surpasses this threshold are assigned to class 1, while others are categorized as class 0. The selection of this threshold influences the balance between Accuracy and Precision, which is 76.2% and 74.4% respectively.

Key Influencing Features:

- Vehicle Damage (0.6235): This feature stands out significantly, exerting the largest influence on a customer's interest in purchasing insurance. Customers with vehicle damage are more likely to be interested in insurance offerings.
- Age (0.1767): The age of the customer emerges as the second most influential factor. Distinct age groups exhibit varying propensities to buy insurance, driven by factors like risk appetite, awareness, or past experiences.
- Previously Insured (0.1067): Whether a customer has prior insurance experience holds notable importance. Those with previous insurance may seek better options or possess greater knowledge about insurance benefits.

Moderately Influencing Features:

- Features such as Vehicle Age, Policy Channels (e.g., insurance agents, brokers, online marketplaces), Annual Premium, and Region demonstrate moderate influence.
- The age of the vehicle, the communication channels for insurance, premium amounts, and the geographic region of the customer have noticeable, albeit not dominant, effects on the buying decision.

CONCLUSION :

After carefully studying the business problem in hand by performing EDA, comparing models, and improving the accuracy of our models, we have the following insights and recommendations.

Insights and Key Recommendations:

Marketing and Engagement:

- Focus marketing efforts on individuals with vehicle damage to tap into a receptive audience, as they are more likely to show interest in purchasing insurance.
- Implement age-specific marketing campaigns to resonate with age groups that exhibit a higher propensity to buy insurance.

Engagement Channels:

- Strengthen partnerships with insurance agents and brokers while bolstering online marketplace presence. These channels hold moderate influence and present lucrative opportunities for customer engagement.

Customer Insights:

- Investigate the underlying reasons behind regional disparities in insurance interest, such as the North or West, to inform targeted marketing strategies. Understanding regional preferences will optimize marketing efforts in specific areas.
- Gain deep insights into the preferences of previously insured customers to tailor offerings that align with their unique needs and preferences, ultimately leading to improved conversion rates.

By incorporating these recommendations, our client can optimize their marketing initiatives, amplify customer outreach, and establish a stronger foothold in the dynamic vehicle insurance market. This multi-pronged approach will undoubtedly boost customer interest and maximize business growth.

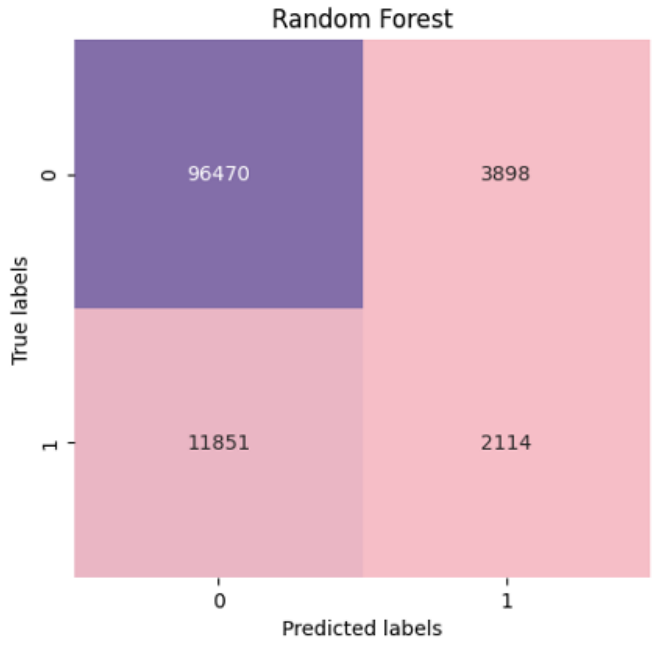
APPENDIX:

The customer features provided by "Namaste Insurance" are the following:

- Gender
- Age
- Driver License (0 : Customer does not have DL, 1 : Customer already has DL)
- Region
- Previously Insured (1 : Customer has had vehicle insurance in the past, 0 : Customer doesn't have had vehicle insurance in the past)
- Vehicle Age
- Vehicle Damage (1 : Customer got his/her vehicle damaged in the past. 0 : Customer didn't get his/her vehicle damaged in the past).
- Channel of outreaching to the customer.
- Number of days, Customer has been associated with the company
- Response (1: Customer bought vehicle insurance 0: Customer has not bought vehicle insurance).

Random Forest:

- Parameters chosen by 10 fold cross validation: Number of estimators = 200, Maximum features = 5
- Accuracy score: 0.8697, Recall: 0.0993, Precision: 0.35



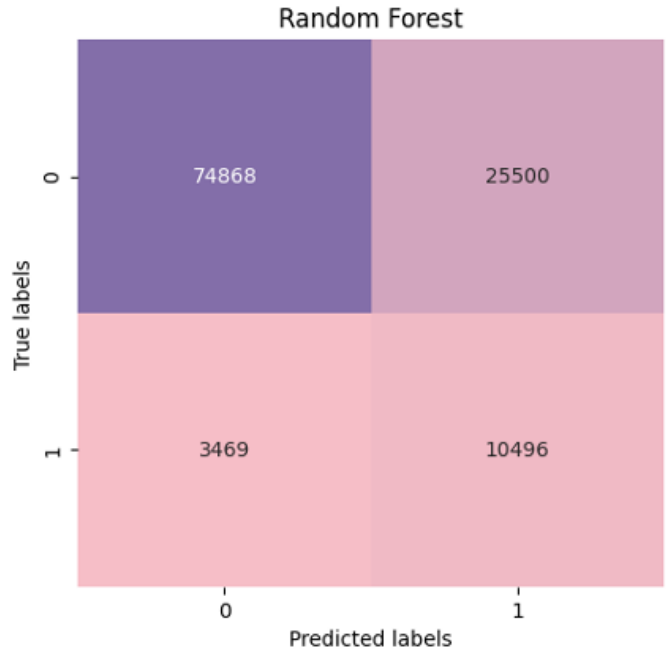
While the initial model showed a relatively high accuracy, the main concern is to minimize False Negatives. In the context of predicting whether a customer bought vehicle insurance or not, False Negatives represent cases where the model incorrectly predicts that a customer did not buy insurance when they actually did. This could lead to missed opportunities for targeting potential customers.

After adjusting the threshold:

- Selected Threshold: 0.19
- Selected True Positive Rate (TPR) or Recall: 0.7543

After adjusting the threshold based on the true positive rate, the model's performance improved significantly in terms of Recall and True Positive Rate while making a trade-off in overall accuracy.

- Accuracy score: 0.7587, Recall: 0.7543, Precision: 0.3036



The adjusted model achieved a higher Recall and True Positive Rate, meaning it is better at correctly predicting customers who bought the vehicle insurance. This reduces the number of False Negatives, ensuring that the model identifies more potential customers who purchased the insurance. However, the Precision decreased, indicating that the model may predict more False Positives (customers predicted as having bought insurance, but they have not). This could lead to some additional efforts or costs in targeting these false positive customers.

In summary, the adjusted Random Forest model is more suitable for the business problem of predicting customers who bought vehicle insurance, as it reduces the chances of missing potential customers who purchased the insurance while slightly sacrificing overall accuracy and increasing the number of false positives. The trade-off ensures that the model effectively targets interested customers, maximizing the chances of successful insurance sales.

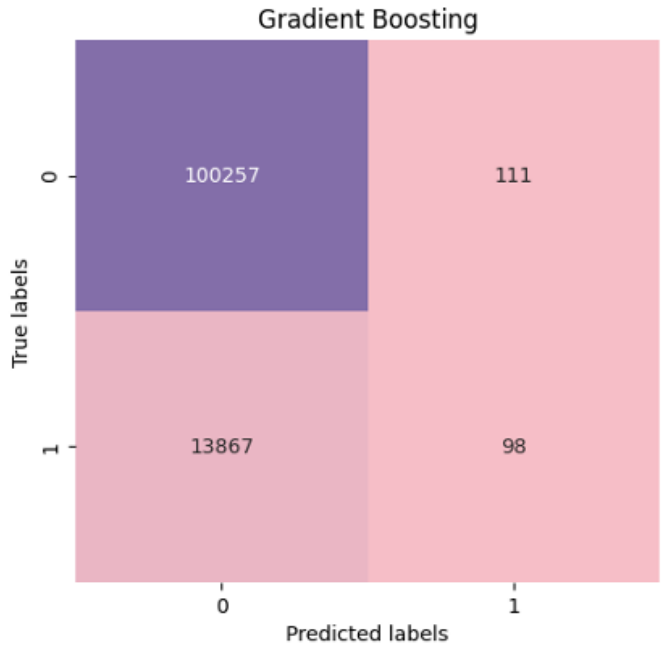
Gradient Boosting:

Initial Gradient Boosting Model:

- Parameters chosen by 10 fold cross validation: Number of estimators = 400, Maximum Depth = 3

Performance Metrics:

- Accuracy: 87.77%
  - While the model shows high accuracy, this metric alone can be misleading especially when our primary focus is on reducing False Negatives.
- Recall: 0.7%
  - Given our business context, this value is problematic. It indicates that the model misses about 99.3% of the customers who actually bought insurance.



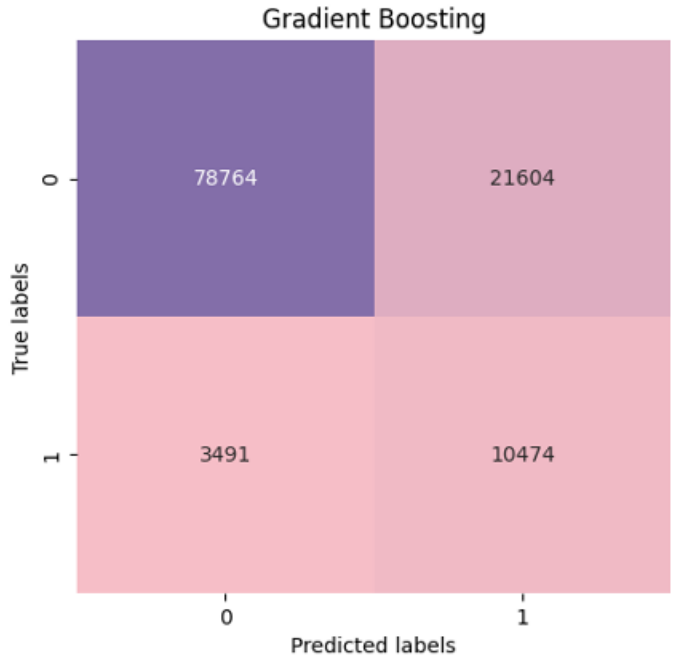
Adjusted Gradient Boosting Model (with threshold adjustment):

- Parameters:

    - Threshold: 0.19
- ROC Metrics:
    - False Positive Rate (FPR): 24.07%
    - True Positive Rate (TPR) or Recall: 75.43%

Performance Metrics:

- Accuracy: 78.05%
    - While there's a decrease in accuracy compared to the initial model, this adjusted accuracy might be acceptable if it means significantly reducing False Negatives.
- Recall: 75.00%
    - This is a vast improvement and aligns more closely with our business needs. The model can now correctly identify 75% of the customers who actually bought insurance, hence reducing the chances of missing out on interested customers.
- Precision: 32.65%
    - Precision has taken a hit, which means the model might predict more customers as being interested in buying insurance than those who actually are. But this trade-off is understandable given our priority to minimize False Negatives.

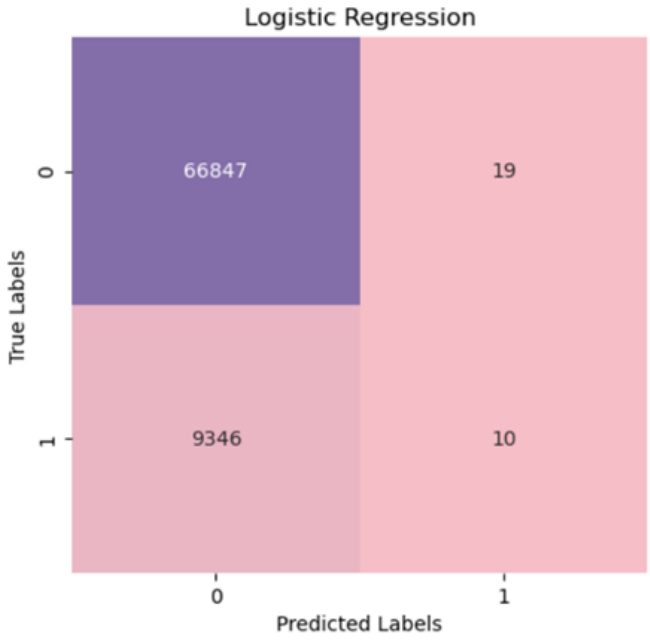


Logistic Regression model:

We faced the same issue of very less accuracy here as well. So we added the predictive probability threshold, determined it by using the ROC curve, and

the Precision-Recall curve. An optimized solver was chosen using trial and error.

- Parameters: solver = 'newton-cholesky'', threshold = 0.22



- Accuracy : An accuracy of 76.2 % was achieved from the model, after adding the threshold of 0.22

- Recall : Our Recall greatly increased from 0.1 % to 74.4 % after adding the threshold which is way better than the original.

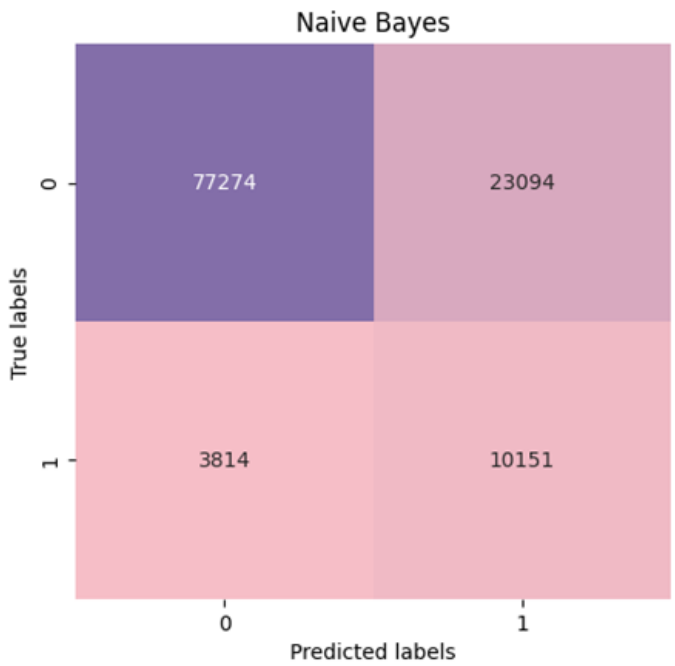
Given the importance of minimizing False Negatives in our business scenario, the adjusted model is more aligned with our requirements. By adjusting the threshold, we've managed to greatly reduce the chances of incorrectly predicting that a customer isn't interested when they actually are. However, this has come at the expense of accuracy and precision.

Naive Bayes:

Naive Bayes was not so Naive in our case, and it did not require the threshold tuning to give us the desired accuracy and recall values.

Below is the Multinomial Naive Bayes Implementation:

- Bins created for continuous variables: Age, Vehicle Age, Annual Premium, Vintage.



- Results: Accuracy score: 0.764, Recall: 0.727, Precision: 0.305.
- Tailor-made results for recall score without hyperparameter tuning.
- Focus on reducing false negatives to improve model performance.

Feature Importance

This graph displays the top 5 Features Impacting the Insurance Purchase Decisions: