CROSS-SELLING OF INSURANCE PRODUCTS

BY: ARYAN DAHIYA

Abstract: *This study looks at selling insurance products across different lines using data on customer details and insurance info. I use several machine learning models, like Logistic Regression and Random Forest Classifier, to guess if a customer will say yes to offers for more insurance. I clean up the data to deal with missing info and group-based details, and I also make new features by putting number-based info into groups and turning group-based info into yes-no type data. To fix the problem of having too few "yes" responses compared to "no" responses, I use a method called SMOTE to even things out before splitting the data for training and testing. I check how well the models work by looking at how often they're right how precise they are how many of the real "yes" responses they catch, and a score that combines precision and recall. The Random Forest model got things right 78.87% of the time, was precise 72.49% of the time, caught 93.06% of the real "yes" responses, and had a combined score of 0.8149. The Logistic Regression model got things right 78.45% of the time, was precise 70.57% of the time, caught 97.62% of the real "yes" responses, and had a combined score of 0.8192. Even though the Random Forest model was a bit more accurate and precise, the Logistic Regression model did better overall because it caught more of the real "yes" responses and had a higher combined score. I use pictures like count plots and bar charts to show how the data is spread out and how well the models perform. This comprehensive analysis underscores the potential of machine learning models in enhancing cross-selling strategies within the insurance industry by identifying key customer characteristics and predicting their response behaviour more effectively.*

Key Words: Cross-selling, Logistic Regression, Random Forest Classifier, Data preprocessing, Feature engineering, SMOTE, Precision, Recall, F1-score, Customer demographics, Insurance industry, Predictive modeling, visualization, count plots

INTRODUCTION:

In the fiercely competitive insurance industry, businesses are always looking for new and creative ways to grow their clientele and boost profits. Selling extra goods or services to current clients, or cross-selling, became a key tactic for company expansion in this industry. This is where the power of data analytics and machine learning came into play. By leveraging these advanced techniques, I aimed to develop predictive models that could accurately identify potential customers for insurance among existing insurance policyholders. These models analyzed various customer attributes, including demographic information, policy details, and historical data, to discern patterns and predictors of interest in insurance.

The primary objective of this study was to create a robust framework for predicting cross-selling opportunities. To achieve this goal, I employed a range of machine learning techniques, including Logistic Regression and Random Forest algorithms. These models were trained on a comprehensive dataset of customer information, with careful attention paid to data preprocessing, feature engineering, and addressing class imbalance issues.

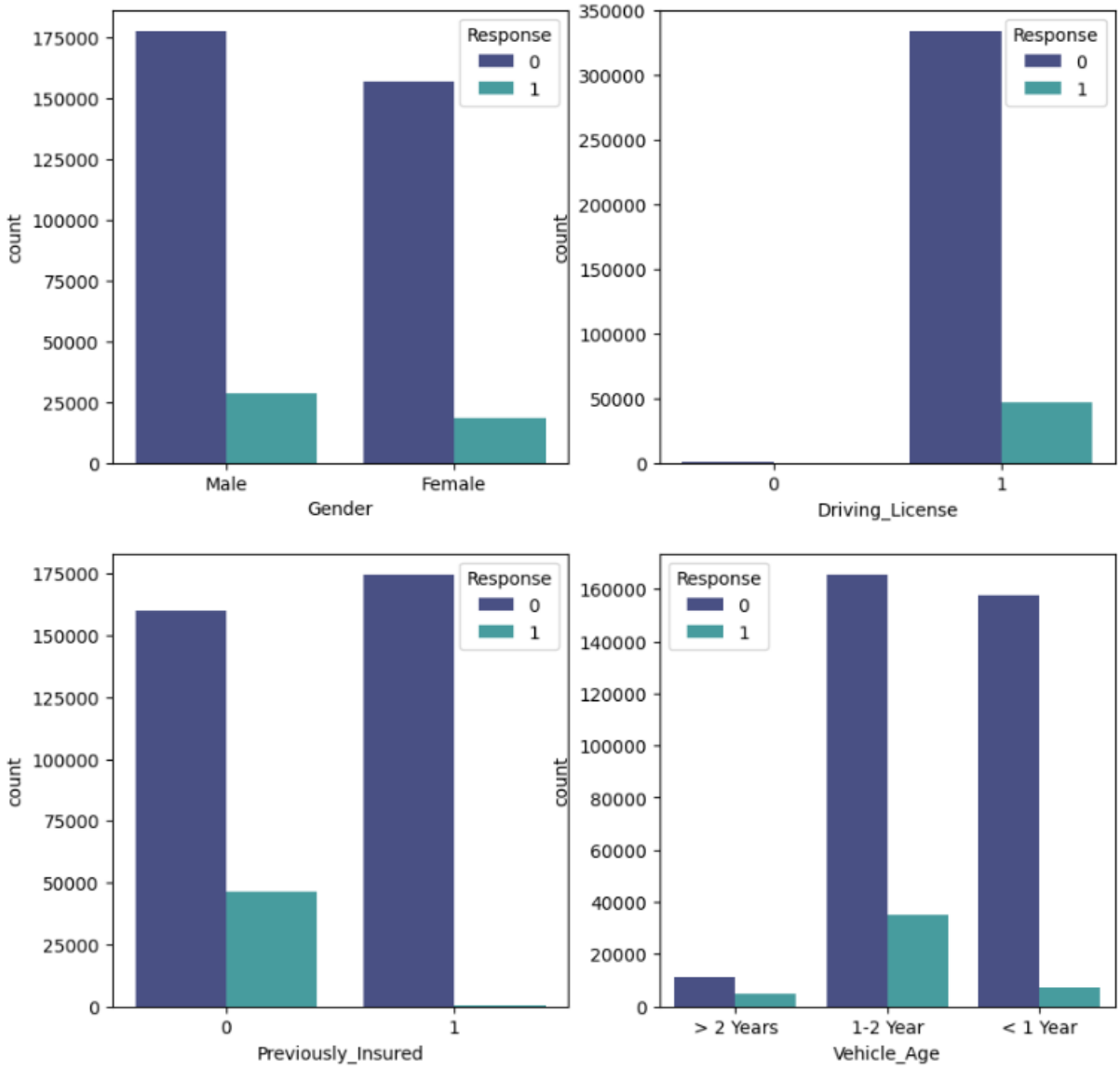
The outcomes of this study had the potential to significantly impact how insurance companies approached cross-selling, moving from broad-based strategies to data-driven, personalized marketing approaches. As I delved into the intricacies of the data and the performance of the models, I aimed to provide actionable insights that could be directly applied in real-world business scenarios, ultimately contributing to more efficient and effective cross-selling efforts in the insurance industry.

APPROACH AND OUTCOMES:

In this study, I took a step-by-step approach to create models that predict cross-selling of insurance products. I started by gathering and preparing a full dataset with info on customer demographics, policy details, and past data. To get the data ready, I dealt with missing values, turned categorical variables into codes, and made numerical features standard. I came up with new features by putting continuous variables into groups and changing categorical variables into dummy ones. The dataset had a problem: there were way fewer "yes" answers than "no" answers. To fix this, I used the Synthetic Minority Over-sampling Technique (SMOTE) to even things out.

The process of choosing features zeroed in on pinpointing crucial traits that relate to predicting cross-selling chances. These included things like age how old the vehicle is yearly premium, gender, if the vehicle had damage, whether the person has a driving license, and if they've had insurance before. After this, I divided the balanced dataset to create training and testing groups. I then used the training group to teach various machine learning models, like Logistic Regression and Random Forest classifiers. To figure out which model did the best job, I looked at things like how accurate it was how precise how well it could recall information, and its F1-score. To show how the data was spread out and how well the models worked, I used pictures like count plots and bar charts.

EXPLORATORY DATA ANALYSIS:



1. Gender Plot:

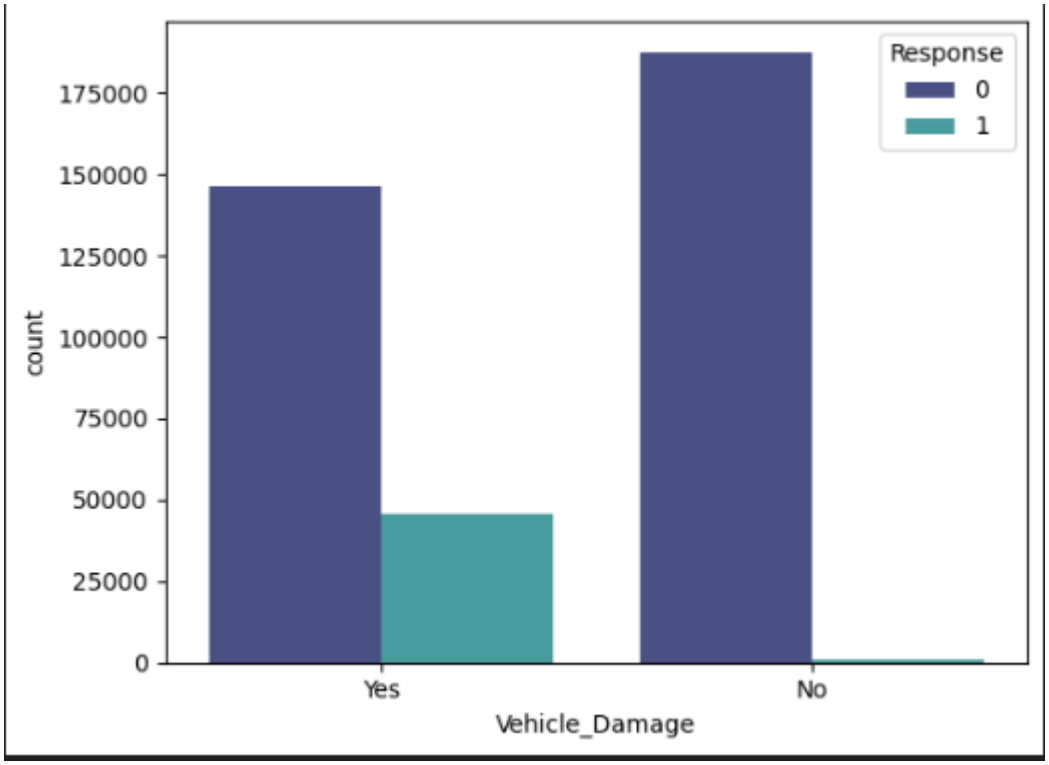
This plot shows the distribution of responses by gender. Both males and females have a higher count of '0' responses (indicating no interest in insurance) compared to '1' responses (indicating interest). However, the proportion of interested customers seems slightly higher for males than females. This insight could be valuable for tailoring marketing strategies based on gender.

2. Driving License Plot:
This plot demonstrates the relationship between having a driving license and interest in insurance. Almost all respondents have a driving license (value 1), which makes sense as it's typically required for vehicle insurance. The proportion of interested customers (response 1) is relatively small compared to uninterested ones, indicating a potential challenge in cross-selling.

3. Previously Insured Plot:
This graph shows how previous insurance history affects interest in new insurance. Customers who weren't previously insured (0) show a higher proportion of interest in new insurance compared to those who were previously insured (1). This suggests that customers without prior insurance might be more promising targets for cross-selling efforts.

4. Vehicle Age Plot:
This plot illustrates how the age of the vehicle relates to interest in insurance. Interestingly, customers with vehicles 1-2 years old show the highest count of both interested and uninterested responses. Vehicles less than 1 year old have the lowest overall count but a relatively high proportion of interested customers. This information could help in targeting customers based on their vehicle age.

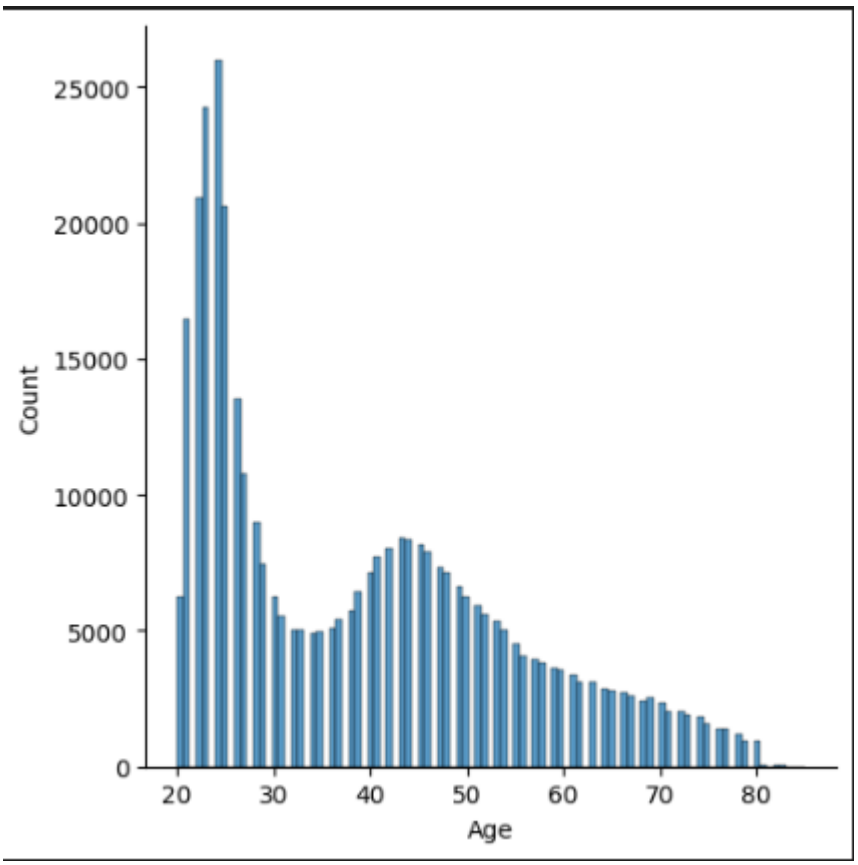
The visualizations indicate that male customers are slightly more likely to own vehicles and purchase insurance compared to females. Additionally, customers with driving licenses are more inclined to opt for insurance than those without licenses. Furthermore, the analysis shows that customers who already possess insurance are less likely to convert, suggesting a preference among customers to hold only one insurance policy.



his graph illustrates the relationship between vehicle damage history and interest in vehicle insurance. Let's break it down in the context of our cross-selling analysis:

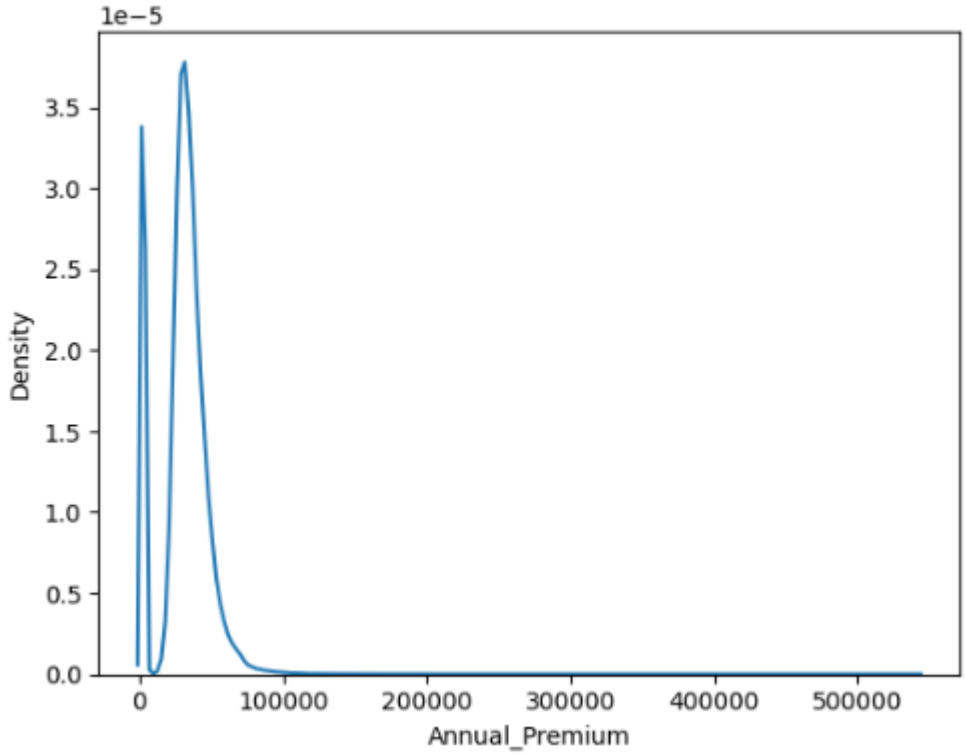
1. Vehicle Damage "Yes":
    - For customers with a history of vehicle damage, there's a notable number of both interested (1) and uninterested (0) responses.
    - The proportion of interested customers seems higher for those with vehicle damage compared to those without.
2. Vehicle Damage "No":
    - There's a significantly larger number of customers without vehicle damage history.
    - Among these customers, the vast majority are not interested in vehicle insurance (0 response).
    - There's a very small proportion of interested customers in this category.

This visualization adds valuable context to our cross-selling analysis, highlighting how past experiences (like vehicle damage) can significantly influence a customer's interest in insurance products. It underscores the importance of segmenting customers based on their history and tailoring our approach accordingly.



A number of significant revelations were made by the cross-selling analysis of auto insurance based on the client age distribution graph. The x-axis covered a wide range of adult ages, from roughly 20 to 80 years old. There were fewer consumers in the older age groups due to the right-skewed distribution, which peaked in the mid-20s to early-30s and gradually declined in frequency as age grew. This implied a greater group of younger clients, which might make them the main focus of cross-selling initiatives. It's possible that different marketing approaches were needed for different age groups, and that insurance products had to be customized to meet the needs of both younger and older clients. The age distribution suggested that, although non-linear

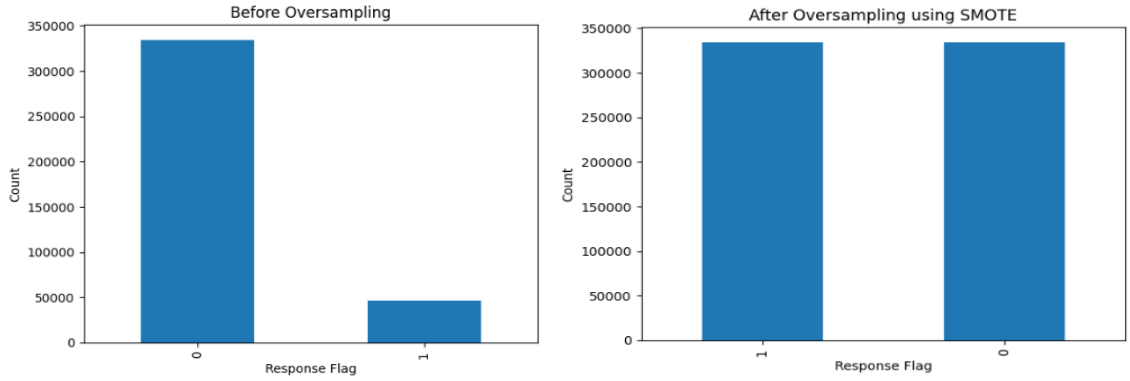
connections between age and insurance interest might not always emerge, age was likely to be a key element in prediction models. This background was helpful in directing the modeling methodology, analyzing the findings, and formulating focused plans for tailoring insurance proposals according to age groups.



The "Annual_Premium" variable from the dataset is plotted as a kernel density estimation (KDE) in the grpah, which reveals numerous important findings. The distribution had a long tail that extended to the right and was strongly biased to the right. Near the left side of the plot, I noticed a steep peak that suggested a lot of premiums were concentrated at lower levels. The x-axis displayed the annual premium range, which was probably expressed in local money, from 0 to roughly 500,000. Density was represented by the y-axis, which was scaled to 10^-5. A much smaller secondary peak that I saw was located just to the right of the primary peak, which may indicate a bimodal distribution or a shared premium tier. While most premiums were quite moderate, there were some very high premiums, but they were far less common, as shown by the long tail

to the right. This distribution was useful for analyzing pricing strategies or client categories within the insurance industry represented by this data, as it indicated that the majority of insurance policies in the dataset had relatively low annual premiums, with a few high-value outliers.

MODEL TRAINING:



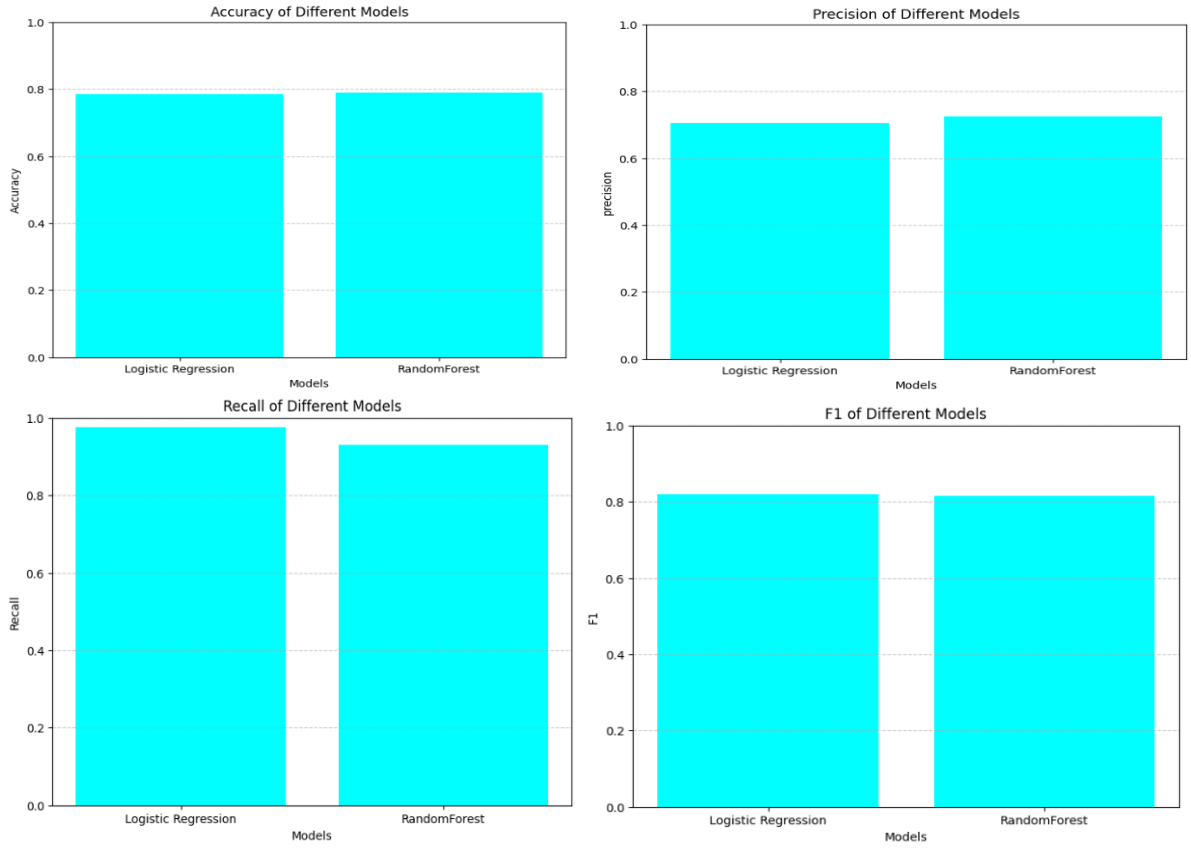
I used the train-test split method in this investigation to assess how well the machine learning models performed using the cross-selling insurance dataset. In order to resolve the class imbalance in the dataset, I first used the Synthetic Minority Over-sampling Technique (SMOTE), which produced a balanced set of features and target variables. I used the train_test_split function from the sklearn.model_selection module to split the oversampled data into training and testing sets in order to further evaluate the model's performance. I specifically divided the data into subsets for training and testing, allocating 70% of the data for training and 30% for testing. The balance attained through SMOTE was preserved thanks to this stratified split, which made sure that the goal variable, "Response," was distributed evenly throughout both subsets. I made sure the results could be repeated by using a random seed (random_state= 42). The models were fitted to the training set, which helped them identify patterns and connections in the data. The testing set functioned as an independent dataset to assess the models'

performance because it wasn't used during training. This method helped avoid overfitting by giving a trustworthy assessment of each model's performance on hypothetical data, revealing information about how well it can generalize.

Two primary classification algorithms were implemented and evaluated in the code:

1. **Logistic Regression:** This model is suitable for binary classification problems like bankruptcy prediction. It estimates the probability of a company going bankrupt based on the input features.

2. **Random Forest:** Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive accuracy. It's robust to overfitting and can handle various data types. Random Forest is known for its versatility and performance in various classification tasks.



The evaluation graphs making comparison between the performances of Logistic Regression and Random Forest models reveal a more complex analysis of their effectiveness in the insurance cross-sell prediction task. Both models have similar high accuracy rates (around 0.78-0.79), suggesting that they predict correctly whether customers would respond to cross-sell offers with a probability of 78-79%. However, it is worth mentioning that Random Forest is slightly better than Logistic Regression with precision of about 0.73 against only 0.71 for the latter; thus, it means that when predicting positive responses - Random Forest has marginally higher accuracy than its counterpart. On the other hand, recall is greater for Logistic Regression achieving approximately 0.97 vis-a-vis Random Forest's 0.93 and hence has an edge over the other in terms of she being able to recognize real disturbingly high percentage responses while also avoiding missing out on a number of them. The F1 scores are not much different for both models at around 0.82 which indicates a balanced measure between precision and recall values The ultimate conclusion drawn from the analysis suggests that indeed both models do perform well but with regard to minimizing false negatives – Logistic Regression would be ideal while for precision – Random Forest remains supreme; therefore deciding between these two will depend on if our concern lies more with reducing False Negatives versus False Positives within the context of insurance cross selling.

CONCLUSION:

In conclusion, the predictive nature of Logistic Regression and Random Forest models shows strong responses from customers to insurance cross-sell offers. With Logistic Regression, recall is excellent thus giving it a superior edge in identifying true positive responses which are vital in capturing and retaining as many clients as possible. Conversely, Random Forest is slightly more precise since when it predicts a positive response, there is an assurance of

reliability. The similarities found in both their F1 scores suggest balanced precision-recall performances for the two models. As such, decisions regarding use of either Logistic Regression or Random Forest depend on specific business priorities: to maximize the identification of potential customers (favoring Logistic Regression) as opposed to improving reliability of positive response predictions (favoring Random Forest). Therefore, this will serve to refine cross-selling initiatives thereby becoming efficient in engaging potential insurance clients.