Logistic Regression for Insurance Cross-Sell Prediction

Abhishek Aern

Student ID – 010946935

Executive Summary

Problem Statement and Hypothesis

The objective of this analysis is to predict policyholders' interest in vehicle insurance using logistic regression. The analysis aims to uncover significant relationships between predictor variables and the likelihood of customers being interested in purchasing vehicle insurance.

Research Question – How accurately can we predict whether a policyholder will be interested in vehicle insurance using Logistic Regression Machine Learning Classifier?

Null hypothesis (H_0) There is no significant relationship between the predictor variables and policyholders' interest in vehicle insurance when using Logistic Regression.

Alternate Hypothesis-(H_1) There is a significant relationship between the predictor variables and policyholders' interest in vehicle insurance when using Logistic Regression with model accuracy greater than 80%.

Summary of the data analysis process

Data Overview

The dataset consisted of various quantitative and categorical variables, including customer ID, gender, age, driving license status, region code, vehicle age, vehicle damage history, annual premium, policy sales channel, vintage (customer tenure), and the response variable indicating customer interest in vehicle insurance (Kumar, Health Insurance Cross Sell Prediction, 2020).

Feature Name	Type	Description
Id	Quantitative	The unique ID assigned to each customer
Gender	Categorical	Gender of the customer, male or female
Age	Quantitative	customer's age in years
Driving_License	Categorical	Indicating whether the customer has a driving license or not
Region_Code	Quantitative	unique code for the region where the customer is located
Previously_Insured	Categorical	Indicating whether the customer already has vehicle insurance or not
Vehicle_Age	Categorical	Age of the customer's vehicle
Vehicle_Damage	Categorical	Indicating whether the customer's vehicle has been damaged in the past or not
Annual_Premium	Quantitative	Amount the customer needs to pay as a premium for insurance in a year

Policy_Sales_Channel	Quantitative	Representing an anonymized code for the channel used to reach out to the customer for insurance purposes
Vintage	Quantitative	Number of days the customer has been associated with the insurance company
Response	Categorical	This is the target variable indicating whether the customer's interested in vehicle insurance or not. 1 (customer is interested) or 0 (customer is not interested)

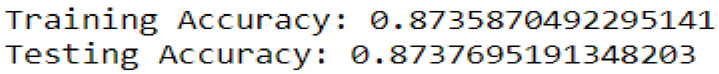
Analysis Approach

The data analysis process involved importing relevant libraries for data manipulation and preprocessing, loading the dataset from Kaggle, and conducting exploratory data analysis (EDA) to gain insights into the dataset. Null values, missing values, and duplicates were examined, and outliers were identified using histograms and boxplots. The meaningless "id" column was removed, and categorical variables were encoded. Multicollinearity was tested using correlation analysis, and outliers in the "Annual_Premium" variable were handled using the Interquartile Range method. The dataset was then split into training and testing sets for model training and evaluation (Bhor, 2021).

Logistic regression was used as the main analysis technique to model the relationship between predictor variables and the response variable. The analysis approach employed a train and test split methodology, where the dataset was divided into a training set (80% of the data) and a test set (20% of the data). Logistic regression was applied to the training set to build a predictive model, which was then evaluated using the test set to assess its performance on unseen data. Performance evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC are utilized to assess the model's effectiveness.

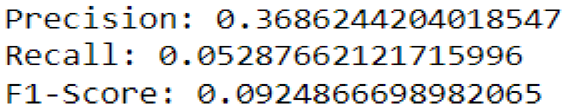
Model Performance

The results of the analysis indicated that the logistic regression model achieved an overall accuracy of approximately 87%. This implies that the model was able to correctly predict the policyholders' interest in vehicle insurance in 87% of the cases in both the training and testing datasets which is greater than the threshold of 80% stated in the alternate hypothesis.



```
Training Accuracy: 0.8735870492295141
Testing Accuracy: 0.8737695191348203
```

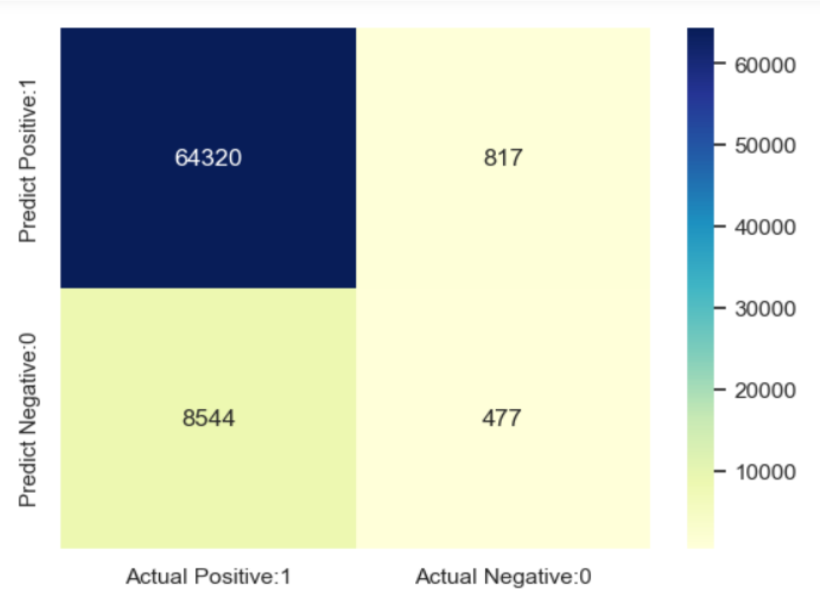
The precision, recall, and F1-score of the model were also calculated to be 0.3686, 0.0528, and 0.0924, respectively, providing insights into its predictive capability. The low values for precision, recall, and F1-score indicate that the model's performance in correctly identifying interested customers (positive class) is relatively low (Bhor, 2021).



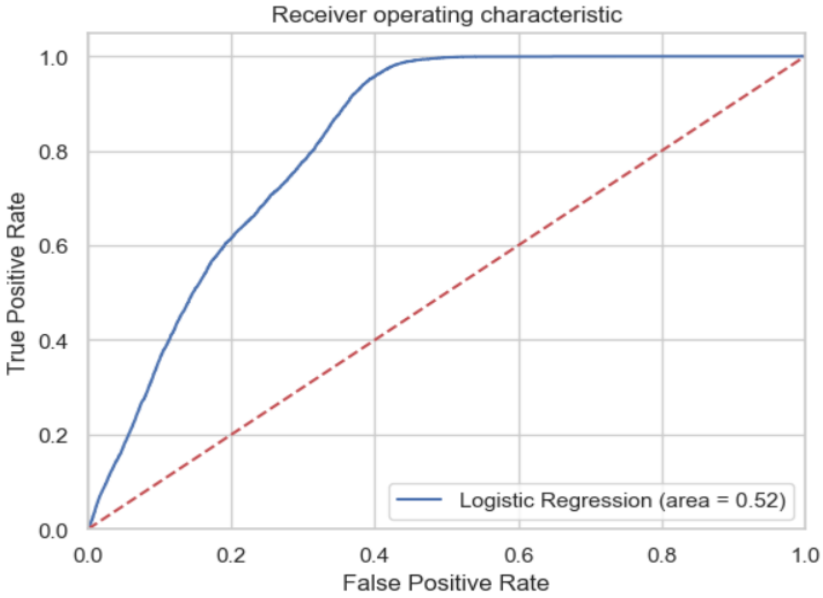
```
Precision: 0.3686244204018547
Recall: 0.05287662121715996
F1-Score: 0.0924866698982065
```

The confusion matrix was used to evaluate the performance of the model in terms of true positives, true negatives, false positives, and false negatives. This information helped in understanding the model's ability to correctly classify individuals based on their interest in vehicle insurance.

As per this matrix, the relatively higher number of false negatives (8,544) suggests that the model is more likely to miss predicting individuals who are actually interested in vehicle insurance. This could potentially result in missed opportunities for targeting and engaging potential customers.



One other important metric used in the evaluation was the ROC-AUC score, which measures the model's ability to discriminate between positive and negative classes. The ROC-AUC score obtained was approximately 0.52, indicating that the model's predictions are no better than random chance.



Key Findings

Based on the Logistic regression summary -

- Gender, Age, Driving license status, Previously insured status, Vehicle age, Vehicle damage history, and Policy sales channel all exhibited significant relationships with customers' interest in vehicle insurance.
- The annual premium has a minimal impact on customer interest in vehicle insurance, with a very small coefficient. This indicates that the price of the premium does not strongly influence customer decision-making regarding insurance.
- The policy sales channel has a slight negative impact on customer interest. Insurance company should consider optimizing their sales channels or exploring alternative approaches to improve customer response.

Based on the Visualizations –

- Age is a key factor influencing customer response to vehicle insurance, with customers between 31 and 50 years old showing higher interest.
- Gender does not significantly impact customer response.
- Possessing a driver's license does not strongly influence interest, indicating it is not a decisive factor.
- Customers without prior insurance coverage are more likely to be interested, suggesting targeting this group can yield better results.
- Vehicle age, particularly vehicles aged 1-2 years, generates more interest.
- The presence of vehicle damage increases customer interest, allowing for targeted marketing towards individuals with past damage or accidents.

Limitations

The presence of imbalanced classes in the target variable (Response) can pose challenges. In this analysis, there is a significant imbalance between positive and negative responses, which can lead to biased model performance. It is important to be cautious of the potential impact of class imbalance and consider techniques like oversampling, undersampling, or using appropriate evaluation metrics to address this issue.

The dataset used in the analysis may not include all the relevant features that could influence the prediction of policyholders' interest in vehicle insurance. There might be other variables, such as socio-economic factors, location-based information, or external factors, that could significantly impact customer behavior.

Recommendations

Based on the analysis results, it is recommended to target marketing efforts towards specific customer segments identified as having a higher interest in vehicle insurance, such as customers between the ages of 31 and 50, those without prior insurance, and those with a vehicle damage history. Further exploration and refinement of the policy sales channel strategy could help improve customer targeting and increase conversion rates. Collecting additional data on external factors that may influence customer interest, such as income levels, regional factors, or customer preferences, could enhance the accuracy of the predictive model.

Address the imbalance in the target variable (Response) by employing techniques such as oversampling the minority class or adjusting class weights during model training. This will help mitigate any biases in the model's performance and improve overall predictive accuracy.

Also, While logistic regression provides interpretable coefficient estimates, it may not capture complex nonlinear relationships or interactions between variables. Other machine learning models, such as decision trees or ensemble methods, could be explored to capture more intricate patterns in the data.

By implementing these proposed actions, the insurance company can optimize its marketing strategies, target the right customer segments, and improve the overall effectiveness of its vehicle insurance campaigns.

Expected Benefits

Here are some of the expected benefits of the analysis and findings :

Enhanced Marketing Strategies: By understanding the factors that influence customers' interest in vehicle insurance, companies can tailor their marketing strategies to target specific customer segments. This targeted approach can result in higher conversion rates, increased sales, and improved ROI on marketing efforts.

Improved Customer Segmentation: This analysis helped in identifying customer segments that are more likely to show interest in vehicle insurance. This information can be used to create customized products, offers, and communication strategies that resonate with the needs and preferences of different customer groups. Improved segmentation can lead to more personalized and effective marketing campaigns.

Data-Driven Decision-Making: The analysis provides a data-driven foundation for decision-making. It enables companies to make informed decisions about product development, pricing strategies, marketing campaigns, and customer relationship management based on the identified relationships between variables and customer interest.

Quantifying the expected benefits will require a deeper understanding of the company's specific context, such as the current conversion rates, average revenue per customer, marketing and operational costs, and market size. The dataset used for this study does not provide any such details to conduct a detailed financial analysis. Overall, all the benefits mentioned above can lead to business growth, profitability, and long-term success.

References

Kumar, A. (2020). Health Insurance Cross Sell Prediction.
https://www.kaggle.com/datasets/anmolkumar/health-insurance-cross-sell-prediction

Bhor, Y. (2021, September 29). Guide for building an End-to-End Logistic Regression Model.
https://www.analyticsvidhya.com/blog/2021/09/guide-for-building-an-end-to-end-logistic-regression-model/