UNIVERSITY OF SCIENCE AND TECHNOLOGY OF HANOI

UNDERGRADUATE SCHOOL

Fundamental of Data Science

PROJECT REPORT

By

BI12-365 Nguyễn Anh Quân

BI12-458 Phạm Xuân Trung

BI12-077 Hoàng Hà Đăng

Data Science

Title:

Data Science in Predicting Vehicle Insurance Cross-Sell for Health Insurance Customers

Hanoi, 2023

I. ABSTRACT

Cross-selling is a sales technique that involves selling additional products or services to an existing customer. It is a way to increase revenue from existing customers without having to acquire new ones. Cross-selling is often done by recommending complementary or related products to customers who have already purchased something from the company.

This report details the steps taken and the outcomes of developing a predictive model for cross-selling health insurance. Predicting whether health insurance purchasers will be interested in buying a vehicle insurance policy is the main objective. A variety of machine learning methods and data preparation procedures on the given dataset are utilized to accomplish this aim. Data exploration, preprocessing, feature engineering, model selection, assessment, and interpretation are all covered in the study.

II. INTRODUCTION

1. Background and Business Understanding

Cost selling is a sales strategy in which a company sells a product or service at a price that is lower than the cost to produce or acquire it. This strategy is often used to generate leads, increase market share, or clear out inventory. For example, in this case, a health insurance company has a large captive audience of potential customers, and they use the data to cross-sell car insurance to their existing customers.

Our client, an insurance provider, operates in a dynamic and competitive industry where understanding customer behavior is crucial. They offer health insurance policies, which are essential agreements that involve regular premium payments to provide financial protection against medical expenses arising from illnesses, accidents, or medical treatments. These policies offer individuals and organizations a financial safety net, ensuring they can access necessary medical care without bearing the full financial burden.

Additionally, our client is interested in expanding their offerings to include vehicle insurance, often referred to as car insurance, motor insurance, or auto insurance. Vehicle insurance plays a pivotal role in safeguarding various road vehicles, such as cars, trucks, and motorcycles, against unforeseen events. Its primary purpose is to provide financial protection against property damage or bodily injury resulting from accidents involving these vehicles. Moreover, it offers liability coverage for incidents that may occur while operating a vehicle, including damage to third parties.

The scope of vehicle insurance extends beyond accident-related expenses. It also encompasses protection against theft, damages caused by non-accident events like vandalism or natural disasters, and even collisions with stationary objects. It is essential to note that the specific terms and regulations governing vehicle insurance can vary significantly based on regional legal requirements.

Our customer wants to develop a solution in light of the variety of insurance products available and the possibility of cross-selling possibilities. The goal of this strategy is to locate current health insurance holders who could also be interested in buying auto insurance. Our client plans to improve their marketing and customer acquisition tactics by utilizing customer data and predictive analytics, which will eventually boost revenue and customer happiness.

In reality, this problem can be approached and solved using many different methods by experts, some approaches do not even require the help of data science, including traditional advertising, telemarketing, in-person sales meetings, and referral programs. Traditional advertising and telemarketing involve taking advantage of electronic devices or print media to raise awareness about vehicle insurance and its benefits. Hence, these methods can attract new customers who are interested and willing to pay for the products. Moreover, implementing referral programs that reward existing customers for referring customers who purchase traffic insurance is also a practical technique for referring to this content. However, the actions that are provided above may cost a substantial amount of time, money and labor to be resolved. Therefore, using machine learning models and data science methods to predict customers' ability to buy vehicle insurance based on health insurance data is considered an agile, modern, and highly effective solution.

2. Data Science-based approach process

Before implementing any particular scientific method, drawing out a process diagram clearly outlined would be essential. The figure below illustrates the complete stages of this project using data science and machine learning in building and selecting models for prediction.

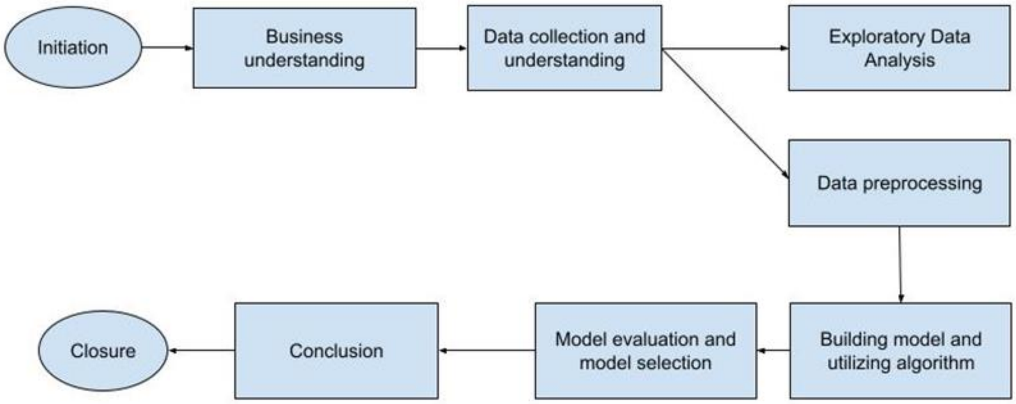

Initiation
Business understanding
Data collection and understanding
Exploratory Data Analysis
Data preprocessing
Building model and utilizing algorithm
Model evaluation and model selection
Conclusion
Closure

Figure 1 Data Science process

All the Data processing, visualizing, Model training and testing results will be achieved with the aid of Python code and necessary libraries.

3. Objectives

Build a prediction model to classify the response and find potential customers who will apply for vehicle insurance based on personal health insurance data.

Address the issue of the dataset's class imbalance, since a minority class may be interested in vehicle insurance.

Give the insurance company concrete insights to help them improve their consumer and marketing targeting methods.

4. Data Collection and Understanding

In the insurance industry, data collection is typically a result of systematic and structured data acquisition processes conducted by insurance companies. This data contains customer information, which is a crucial contributor for insurers to understand their policyholders better and tailor their services to meet individual needs while adhering to privacy and regulatory constraints. Therefore, performing a meticulous, professional data collection method will have a huge impact on the quality of expected outputs. To perform effective data gatherings, experts in practice have carried out numerous approaches, some remarkable ways that can be mentioned are:

Using application forms: When customers apply for insurance, they fill out application forms. These forms typically include personal information such as name, age, gender, address, and contact details.

Third-Party Data Sources: Insurance companies may also obtain information from third-party sources, such as credit bureaus, government agencies, and public records. This data can include credit scores, claims history, and driving records.

Social Media and Online Data: Some insurance companies use publicly available online data and social media profiles to gather additional information about potential customers.

There appears to be a boundless amount of collecting techniques, and operating them proficiently would greatly support the predictive models.

In this scenario, “Health Insurance Cross Sell Prediction” – a dataset from Kaggle that is collected from the health insurance owners to forecast the personal interest in vehicle insurance – is considered great material for reaching the objectives of this problem.

The dataset contains 381109 records and 12 features, including one for target value and 11 for specific customer information. The table below shows a general description of this dataset:

Variable	Definition	Data Type
Id	Unique ID for the customer	Qualitative
Gender	Gender of the customer	Qualitative
Age	Age of the customer	Quantitative
Driving_License	0 : Customer does not have DL, 1 : Customer already has DL	Qualitative
Region_Code	Unique code for the region of the customer	Qualitative
Previously_Insured	1 : Customer already has Vehicle Insurance, 0 : Customer doesn't have Vehicle Insurance	Qualitative
Vehicle_Age	Represents the age of the customer's vehicle, typically categorized into groups like "1-2 Years," "< 1 Year," etc.	Qualitative
Vehicle_Damage	1 : Customer got his/her vehicle damaged in the past. 0 : Customer didn't get his/her vehicle damaged in the past.	Qualitative
Annual_Premium	The amount customer needs to pay as premium in the year (rupees)	Quantitative
Policy_Sales_Channel	Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc.	Qualitative
Vintage	Number of Days, Customer has been associated with the company	Quantitative
Response (target value)	1 : Customer is interested, 0 : Customer is not interested	Qualitative

Table 1 Data set feature description

5. Exploratory Data Analysis

The data contains no missing or duplicated values

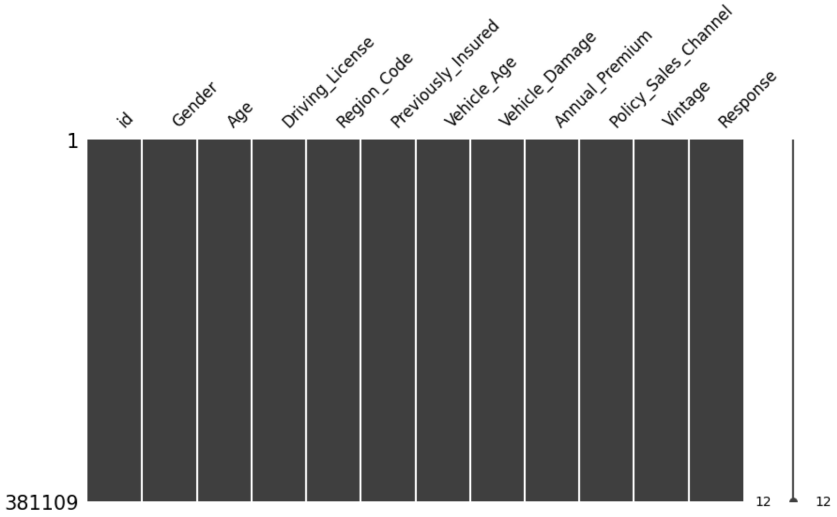


Figure 2 Counting Data set's record

To gain a comprehensive understanding of the content, data visualization should be used to reveal insightful statistics.

- **Response**

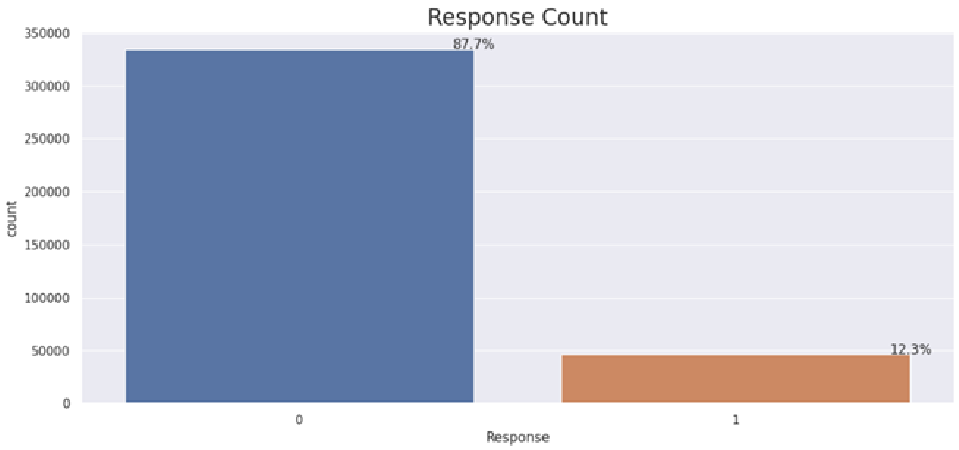


Figure 3 Distribution of customer response

The bar plot of Response feature values shows that 87.7% of customers are not interested in purchasing car insurance, while only 12.3% intend to buy it. As this feature is the label for the entire dataset, this indicates that the dataset is imbalanced, with a large ratio between the major class (not interested) and the minor class (interested). This can negatively impact the performance of machine learning models, leading to misleading accuracy metrics, poor generalization, and unfair predictions. Therefore, it is important to address data imbalance before performing prediction tasks.

- **Gender**

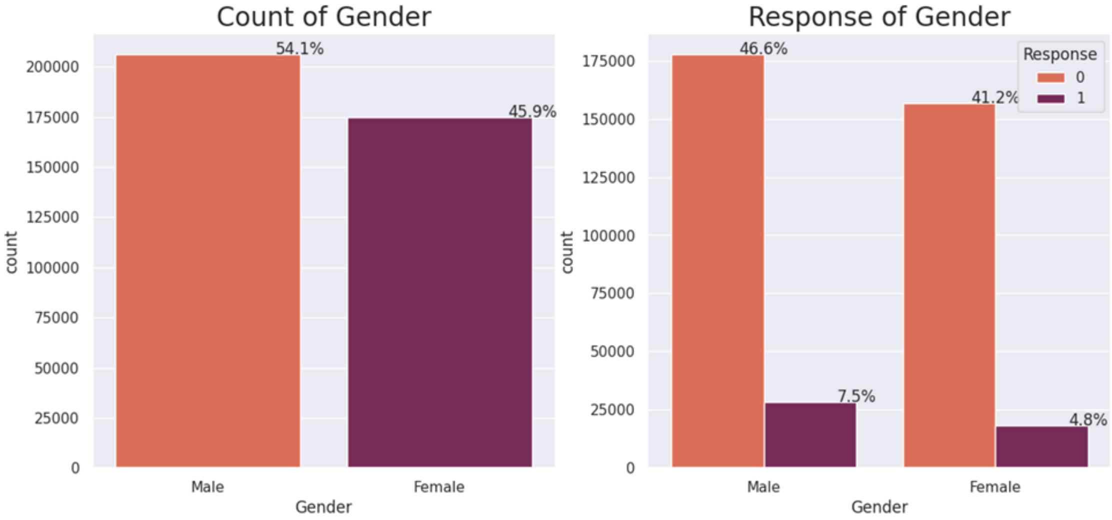


Figure 4 Distribution of Gender feature

The first graph shows that the genders are evenly distributed, with male and female customers accounting for approximately equal percentages of the dataset. However, the Response of Gender feature suggests that men are slightly more likely to purchase insurance than women, as the ratio of male subscribers to female subscribers is marginally higher.

- **Driving License**

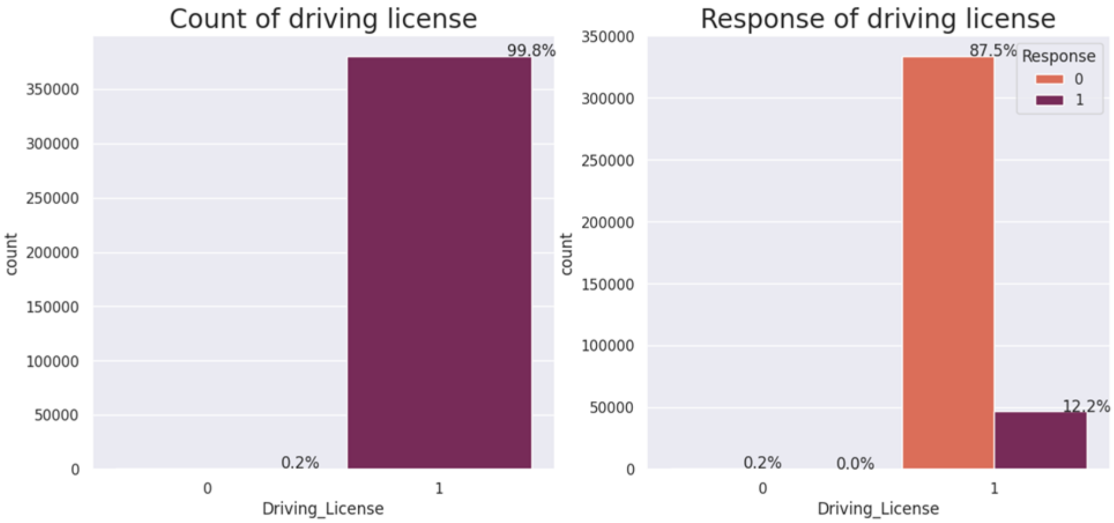


Figure 5 Distribution of Driving License feature

Figure 5 shows that 99.8% of customers have a driving license, which is overwhelmingly common. Since everyone must obtain a driving license before purchasing vehicle insurance, this feature has no effect on predicting the likelihood of purchase and should be dropped from the dataset.

- **Vehicle age**

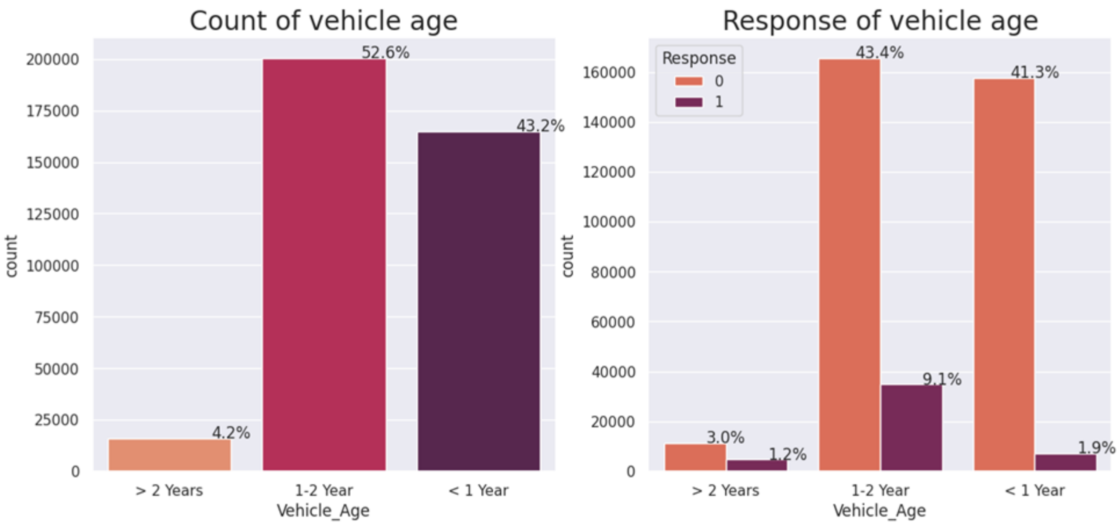


Figure 6 Distribution of vehicle age feature

The graph shows that 42.2% of cars are under a year old, 52.6% are between one and two years old, and 4.2% are more than two years old. Of the respondents, 1.2% are interested in buying insurance for cars that are more than two years old, 9.1% are interested in buying insurance for cars that are between one and two years old, and 1.9% are interested in buying insurance for cars that are under a year old. These statistics suggest that most people are aware of the importance of insurance and are willing to purchase it to reduce their risk as their car ages.

- **Annual Premium**

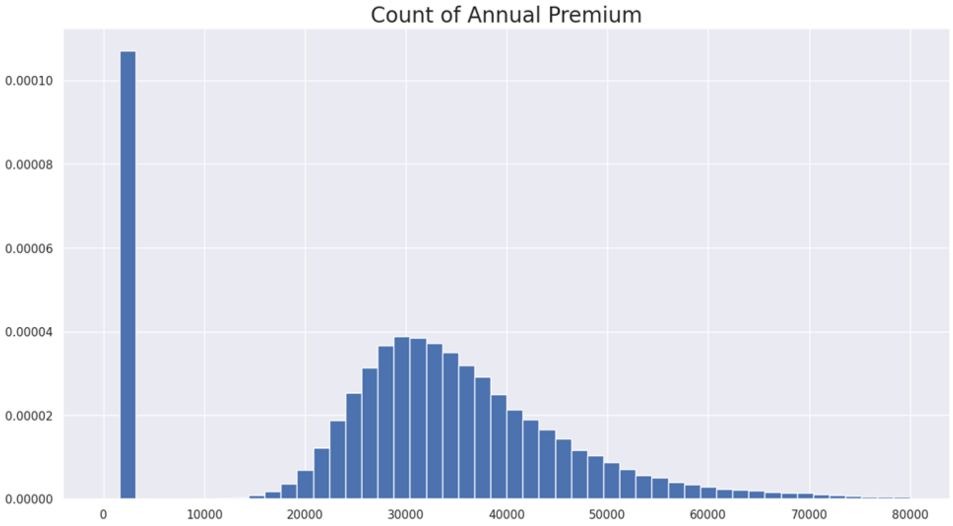


Figure 7 Distribution of Annual Premium feature

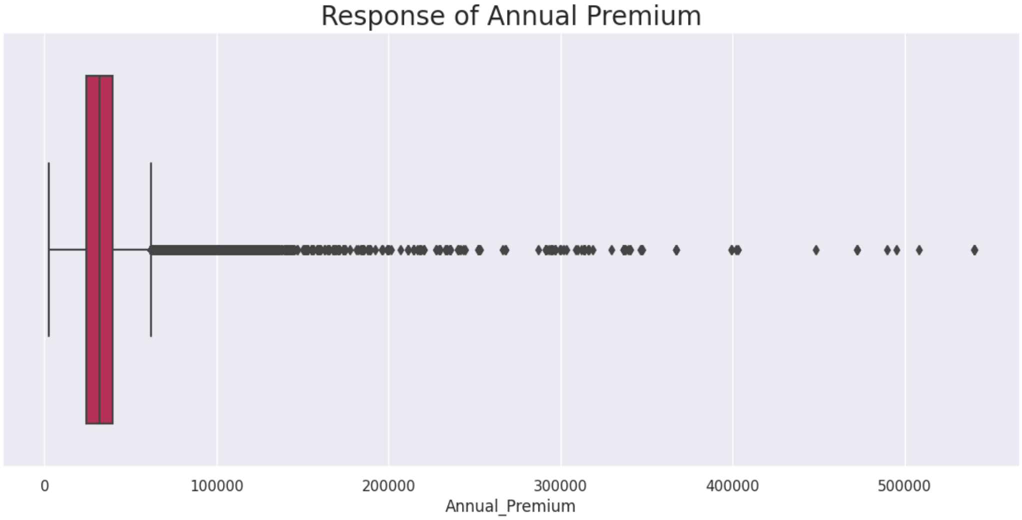


Figure 8 Box plot of Annual Premium feature

The premium cost distribution appears to be right-skewed, with most participants paying around 30,000 rupees for their previous health insurance. However, outliers may be detected and handled for this feature.

Vehicle Damage

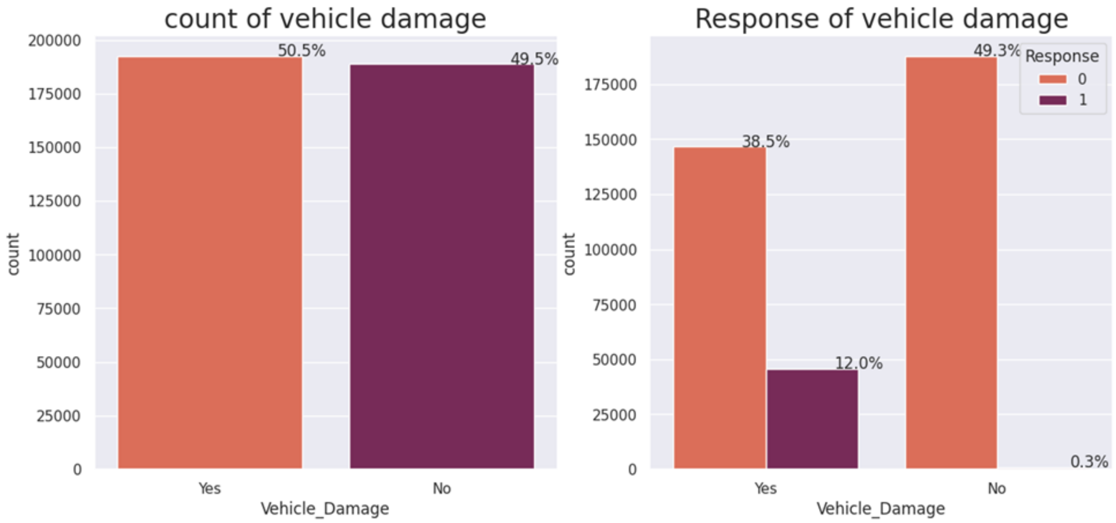


Figure 9 Distribution of Vehicle Damage feature

The Vehicle Damage plots outline nearly half of the cars and motos have all broken down at least once, and 12% of those people want to acquire vehicle insurance, while the others seem to be not interested. This means that it is common for experienced individuals to focus more on future difficulties, and this has a significant impact on the problem's desired outcome.

- **Age**

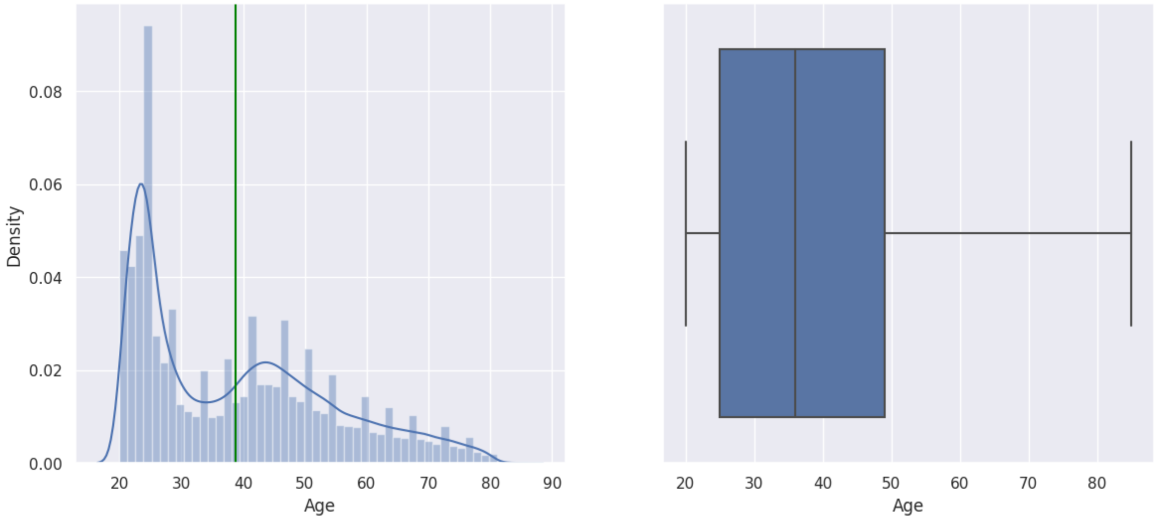


Figure 10 Distribution and box plot of Age feature

Figure 10 shows that the population is relatively young, with the largest number of people in the 20-30 age range. The number of people gradually declines with age, with the fewest people in the 80+ age range. The box plot also supports that there are no outliers in the Age feature.

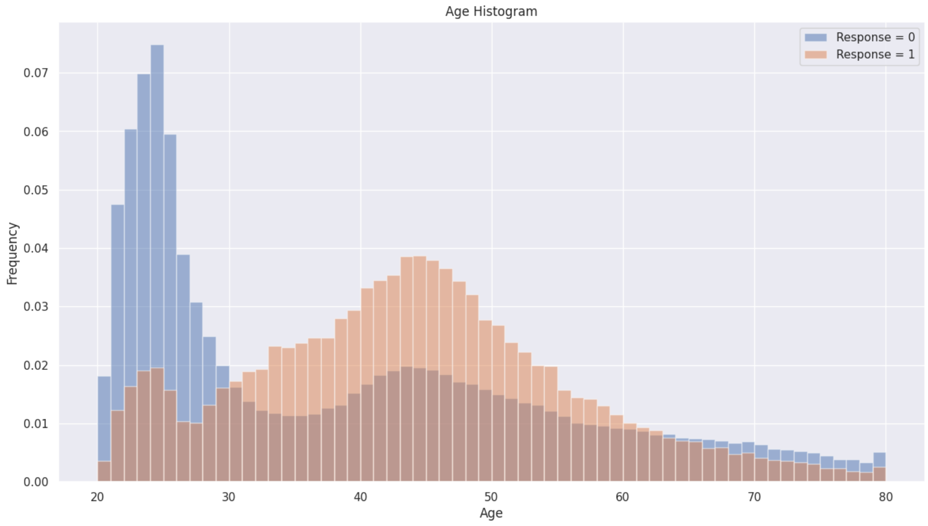
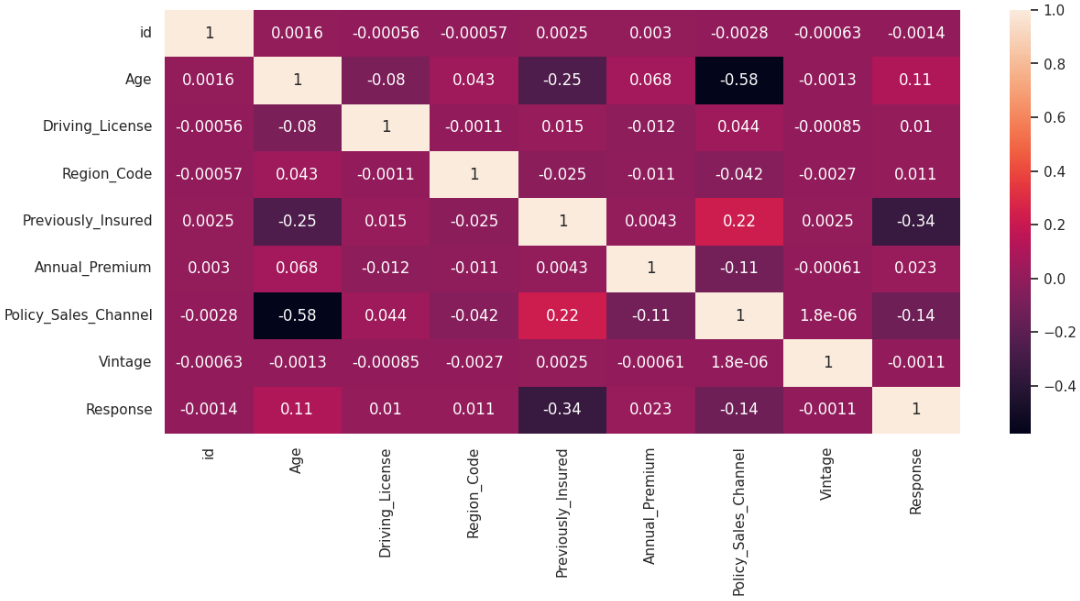


Figure 11 Histogram of Age feature based on Response

Furthermore, the graph illustrates the distribution of disagreement and acceptance across different age groups. It is evident that people aged 30 to 50 are more likely to purchase insurance. This is likely because car and motorbike owners in this age group are more likely to have modern and luxury vehicles, which may require more expensive repairs and maintenance in the event of an accident.

- **Correlation Matrix**



	id	Age	Driving_License	Region_Code	Previously_Insured	Annual_Premium	Policy_Sales_Channel	Vintage	Response
id	1	0.0016	-0.00056	-0.00057	0.0025	0.003	-0.0028	-0.00063	-0.0014
Age	0.0016	1	-0.08	0.043	-0.25	0.068	-0.58	-0.0013	0.11
Driving_License	-0.00056	-0.08	1	-0.0011	0.015	-0.012	0.044	-0.00085	0.01
Region_Code	-0.00057	0.043	-0.0011	1	-0.025	-0.011	-0.042	-0.0027	0.011
Previously_Insured	0.0025	-0.25	0.015	-0.025	1	0.0043	0.22	0.0025	-0.34
Annual_Premium	0.003	0.068	-0.012	-0.011	0.0043	1	-0.11	-0.00061	0.023
Policy_Sales_Channel	-0.0028	-0.58	0.044	-0.042	0.22	-0.11	1	1.8e-06	-0.14
Vintage	-0.00063	-0.0013	-0.00085	-0.0027	0.0025	-0.00061	1.8e-06	1	-0.0011
Response	-0.0014	0.11	0.01	0.011	-0.34	0.023	-0.14	-0.0011	1

Figure 12 Correlation Matrix

Correlation matrix's analysis of the data features reveals several key insights about the relationships between the independent variables and the target variable, response. For example, Driving License, Region Code, and Vintage have negligible correlations with

Response, indicating that they are not significant predictors of customer decisions. However, there is a moderate negative correlation between previously insured and response, suggesting that people who have been previously insured are less likely to purchase vehicle insurance again. Age also has a slight positive correlation with response, indicating that older customers may be more likely to purchase vehicle insurance.

It is important to note that the correlation matrix only shows linear relationships between variables. This means that it cannot capture any non-linear relationships or interactions between variables. Additionally, the correlation matrix does not identify strongly correlated pairs of variables. Therefore, dropping features based on correlation matrix analysis is not recommended in this scenario.

III. METHOD AND IMPLEMENTATION

1. Data Preprocessing

a. Handling missing and duplicated values

After detecting no missing or duplicated values, no techniques are implied in this aspect.

b. Feature eliminating

The Id column only provides a unique identifier for each customer in the record and has no bearing on the insurance process. Additionally, as discussed above, it is impossible to provide vehicle insurance to someone without a driver's license. Therefore, the Id and driving license columns are dropped before performing any other operations.

c. Outlier detection and cleaning

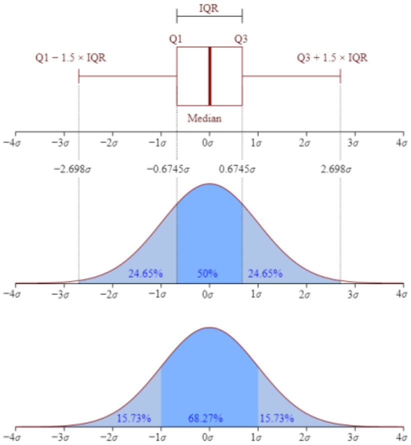


Figure 13 Outlier detection method

The interquartile range (IQR) is a common method for detecting outliers in data. Outliers are defined as observations that fall below the first quartile (Q1) minus 1.5 times the IQR or above the third quartile (Q3) plus 1.5 times the IQR.

A large number of outliers were detected in the Annual Premium feature. To address this, a method was used to eliminate 10,320 outliers.

d. Categorical Value Encoder

The Ordinal Encoder method is implied to transform categorical(qualitative) data into numerical data.

e. Prepare data for training models

In this step, the dataset is split into a training set (80%) and a test set (20%).

As discussed above, the dataset is imbalanced. Therefore, a random oversampling method is used to address this issue. Random oversampling is a technique for balancing imbalanced datasets by duplicating examples from the minority class. This creates a new training set with an equal number of examples from both classes, which allows for more effective insights to be drawn from the training set.

2. Model Building

The classification models are built in order to predict customers whether they are interested in purchasing vehicle insurance and determine the features that affect their decision. In this project, we implemented 5 models: Logistic Regression, Random Forest and XGBClassifier, GradientBoostingClassifier, and Gaussian Naive Bayes in order to compare the performance between models. By doing that, we can achieve insights into how these models work and behave.

a. Logistic Regression

Logistic Regression is a statistical model used to predict the probability of a binary outcome based on one or more predictor variables. It uses the logistic function to transform the probability of an event into a linear combination of the predictor variables. The coefficients of the predictor variables are estimated by maximizing the likelihood function, which measures how well the model fits the data.

Feature	Vehicle Damage	Region Code	Previously Insured	Age	Vintage
Coefficient	1.060	-0.009	-2.030	-0.081	-0.007
Feature	Gender	Annual Premium	Vehicle Age	Policy Sales Channel	
Coefficient	0.065	0.021	-0.121	-0.181	

Table 2 Feature importance rate of Logistic Regression

The logistic regression model results show that the coefficient for the previously insured feature is negative (-2.030), indicating that customers who are previously insured are significantly less likely to purchase insurance. The coefficient for the vehicle damage feature is positive (1.060), indicating that customers with damaged vehicles are more likely to be interested in insurance. The coefficient values for the other features are low, suggesting that they are not significant factors in the model.

Based on these results, insurers can target customers who are not previously insured and focus on customers with damaged vehicles, as these customers are more likely to be interested in purchasing insurance.

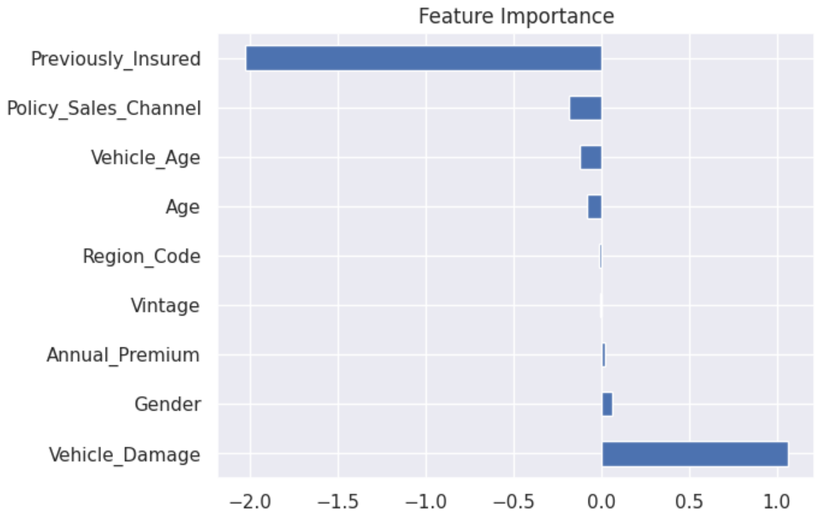


Figure 14 Feature importance visualization

b. Random Forest

A random forest model is a machine learning algorithm that combines the predictions of multiple decision trees to produce a more accurate prediction. By following the large number principle, a random forest classifier can reduce error and improve performance compared to a single classifier.

In this project, the random forest algorithm was trained and optimized for the number of tree parameters. The model was built by changing the n-estimator parameter in increments of 50, from 50 to 200. The results showed that the optimal number of trees was 50.

Feature	Vehicle Damage	Region Code	Previously Insured	Age	Vintage
Importance	0.420	0.004	0.386	0.093	0.000
Feature	Gender	Annual Premium	Vehicle Age	Policy Sales Channel	
Importance	0.001	0.002	0.036	0.053	

Table 3 Feature importance rate of Random Forest

The table above shows the feature importance rates for predicting which features have the greatest impact on customer decisions. Vehicle damage has the highest importance rate (0.420), followed by previously insured (0.386) and age (0.093). This indicates that vehicle damage is the most important factor in determining whether a customer is interested in purchasing insurance, followed by whether the customer is previously insured and their age.

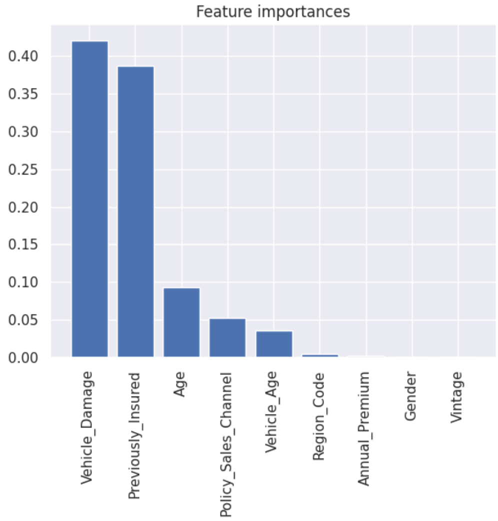


Figure 15 Feature importance visualization

c. XGB Classifier

The XGBoost Classifier model is a machine learning algorithm that extends the gradient boosting algorithm. It builds an ensemble of weak prediction models, typically simple decision trees, to create a more accurate prediction model. It uses gradient boosting to iteratively add new models to the ensemble, with each new model correcting the errors of the previous models. The XGBoost Classifier model also includes regularization techniques to prevent overfitting and improve generalization performance.

Feature	Vehicle Damage	Region Code	Previously Insured	Age	Vintage
Importance	0.481	0.003	0.488	0.008	0.002
Feature	Gender	Annual Premium	Vehicle Age	Policy Sales Channel	
Importance	0.002	0.002	0.005	0.004	

Table 4 Feature importance rate of XGB Classifier

The table above shows the feature importance scores for predicting which features have the greatest impact on customer decisions. Previously insured has the highest importance score (0.488), followed by vehicle damage (0.481). This indicates that reviously insured is the most important factor in determining whether a customer is interested in purchasing insurance, followed by vehicle damage.

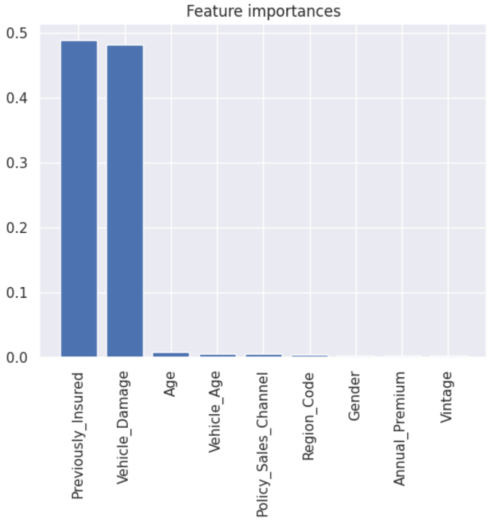


Figure 16 Feature importance rate of XGB Classifier

d. Gradient Boosting

Gradient boosting is a machine learning algorithm that builds an ensemble of weak prediction models, typically simple decision trees, to create a prediction model. It can be used for both regression and classification tasks. Gradient boosting is similar to other boosting methods in that it combines weak learners into a single strong learner in an iterative fashion. However, it differs from other boosting methods by allowing optimization of an arbitrary differentiable loss function.

Feature	Vehicle Damage	Region Code	Previously Insured	Age	Vintage
Importance	0.759	0.007	0.124	0.07	0.000
Feature	Gender	Annual Premium	Vehicle Age	Policy Sales Channel	
Importance	0.000	0.001	0.007	0.021	

Table 5 Feature importance rate of Gradient Boosting

According to the table, Vehicle Damage is by far the most important feature, with an extremely high feature importance score (0.759). This indicates that the presence or absence of vehicle damage is the most critical predictor for insurance-related decisions. Customers with vehicle damage are overwhelmingly more likely to be interested in insurance. Other features have less impact on insurance-related decisions and can be considered to a lesser extent.

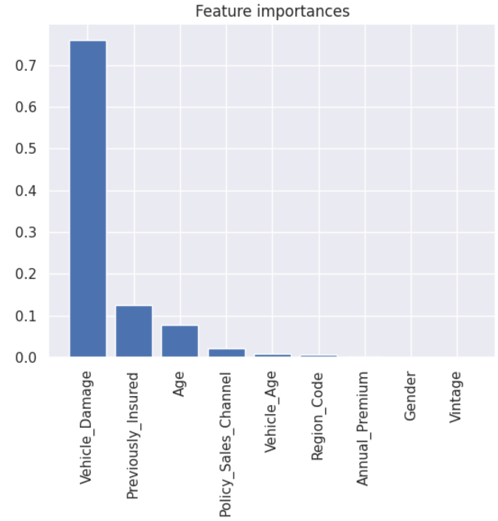


Figure 17 Feature importance visualization

e. Gaussian Naïve Bayes classifier

Gaussian Naive Bayes models are a variant of the Naive Bayes algorithm that assume that the features are normally distributed. They are probabilistic algorithms used for classification tasks, where the goal is to predict the class of a given data point based on its features. The algorithm calculates the probability of each class given the input features and selects the class with the highest probability as the predicted class.

However, Gaussian Naive Bayes models do not offer an intrinsic method to evaluate feature importance. Naive Bayes methods work by determining the conditional and unconditional probabilities associated with the features and predicting the class with the highest probability. Therefore, no coefficients are computed or associated with the features used to train the model. However, we can calculate permutation importance for Gaussian Naive Bayes models, which is not a direct measure of feature importance but rather a measure of how much a feature's permutation affects the model's performance.

Permutation feature importance is defined as the decrease in a model score when a single feature value is randomly shuffled. This procedure breaks the relationship between the feature and the target, so the drop in the model score indicates how much the model depends on the feature.

Feature	Vehicle Damage	Region Code	Previously Insured	Age	Vintage
Permutation Importance	-0.89	0.0	-0.096	0.0	0.0
Feature	Gender	Annual Premium	Vehicle Age	Policy Sales Channel	
Permutation Importance	0.0	0.0	0.0	0.0	

Table 6 Permutation Importance generated by Gaussian Naive Bayes

The permutation importance scores for the Naive Bayes model were -0.096 for the Previously Insured feature and -0.89 for the Vehicle Damage feature. However, Naive Bayes is not typically used for feature importance analysis, so this result may not be conclusive. Other features all had a permutation score of 0.000, suggesting that the model is not strongly influenced by any of these features.

In summary, the permutation importance scores obtained from the Naive Bayes model may not provide meaningful insights about feature importance. This is because the model's simplicity and independence assumptions make it less suitable for this type of analysis. For feature importance analysis, models like Random Forest, XGBoost, or Gradient Boosting are better suited for the task.

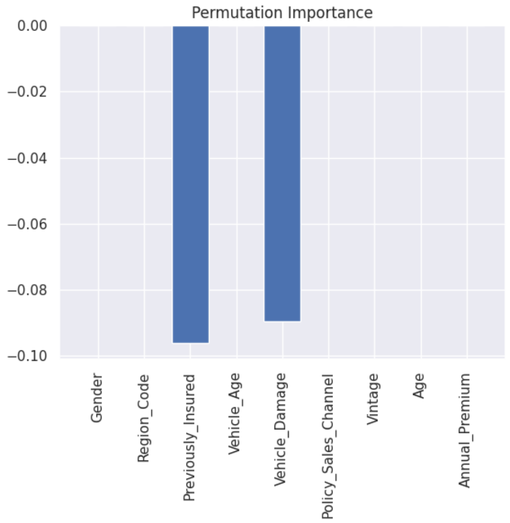


Figure 18 Feature importance visualization

IV. EVALUATION AND PERFORMANCE

1. Model Evaluation

The classification models are built in order to predict customer interest in vehicle insurance rates and determine what features affect this. In this study, there are some particular models built using Logistic Regression, Random Forest and XGBClassifier, GradientBoostingClassifier, Gaussian Naive Bayes evaluated by Precision, Recall, Area under the Curves (AUC), and StratifiedKfold cross-validation to assess the performance of a machine learning model and to estimate its generalization ability. The value of k was 10

a. Logistic Regression

The result evaluation is based on the predictive results and actual labels through the confusion matrices of Table 7. The confusion matrix helps evaluate how the model performed with values equivalent to true positive, true negative, false positive, and false negative, respectively 8770, 38634, 26539, and 215 with 24.8% precision and 97.6% of recall. Moreover, the confusion matrix is also used to depict AUC results. In detail, the parts related to true positive and false positive are taken to illustrate figure 19. As can be seen in this figure, the perfect classifier is the 1.0 point, with the curve nearly approaching this perfect classification leading to the result of 78.4% of accuracy. This means that the model built using the logistic regression algorithm obtains a good predictor. On the other hand, another procedure used is k-fold cross-validation with the three metrics above, which resulted in a recall of 97.6%, a precision of 70.5%, and an AUC score of 82%.

		Actual	
		Interested	**Not interested**
Predict	**Interested**	8770	26539
	Not interested	215	38634

Table 7 Confusion matrix evaluating Logistic Regression model

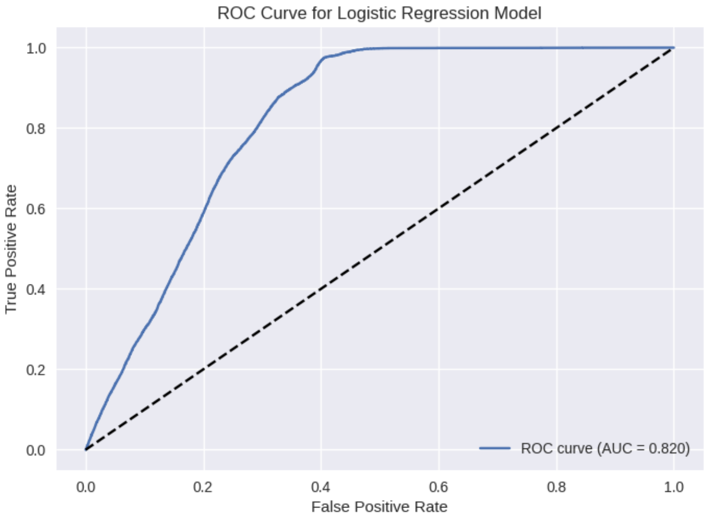


Figure 19 AUC figure to evaluate the Logistic Regression model

Procedure	Score
Cross validate with precision	70.5%
Cross validate with recall	97.6%
Cross validate with AUC_score	82%

Table 8 The procedures used to evaluate the model and their scores

b. Random Forest

The result evaluation is based on the predictive results and actual labels through confusion matrices of table 9. The confusion matrix helps evaluate how the model performed with values equivalent to true positive, true negative, false positive, and false negative, respectively 8415, 42716, 22457, and 570, with 27.2% of precision and 93.7%% of recall. Moreover, the confusion matrix is also used to depict AUC results. In detail, the parts related to true positive and false positive are taken to illustrate figure 20. As can be seen in this figure, the perfect classifier is the 1.0 point, with the curve nearly approaching this perfect classification leading to the result of 79.6% of accuracy. This means that the model built using the random forest algorithm obtains a good predictor. On the other hand, another procedure used is k-fold cross-validation with the three metrics above, which resulted in a recall of 99.9%, a precision of 92.1%, and an AUC score of 85.4%.

		Actual	
		Interested	**Not interested**
Predict	**Interested**	8415	22457
	Not interested	570	42716

Table 9 Confusion matrix evaluating Random Forest model

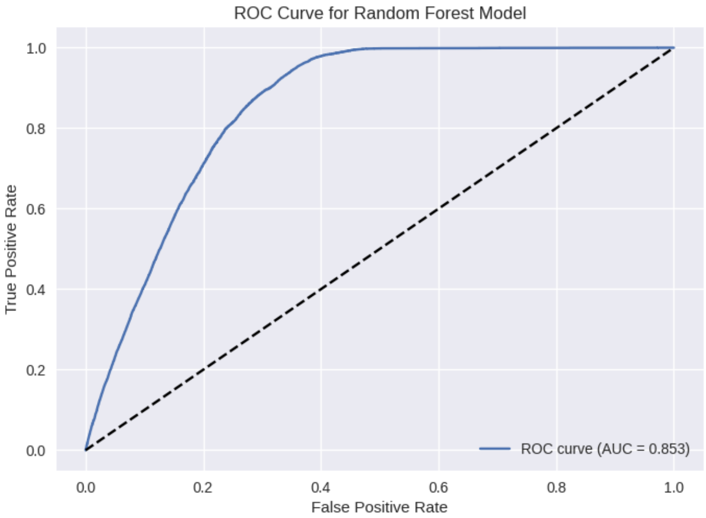


Figure 20 AUC figure to evaluate the Random Forest model

Procedure	Score
cross-validate with precision	92.1%
cross-validate with recall	99.9%
cross-validate with AUC score	85.4%

Table 10 The procedures used to evaluate the model and their scores

c. XGB Classifier

The outcome evaluation is according to the prediction results and actual labels using a confusion matrix represented in table 11. This evaluates how the XGB Classifier algorithm executes with values corresponding to true positive, true negative, false positive, and false negative, which are respectively 8097, 45025, 20148, and 888 leading to the recall is 90.1%, precision is 28.7% . Then further, AUC findings are captured based on the confusion matrix. To demonstrate the chart, the elements relevant to true positive and false positive are in figure 21. yielding an accuracy of 79.6%. This suggests that the random forest algorithm produced a decent prediction, along with the outcome of the K-fold cross-validation with the three metrics above, which resulted in a recall of 94.2%, a precision of 75.3% and an AUC score of 87.3.

		Actual	
		Interested	Not interested
Predict	Interested	8097	20148
	Not interested	888	45025

Table 11 Confusion matrix evaluating XGB Classifier

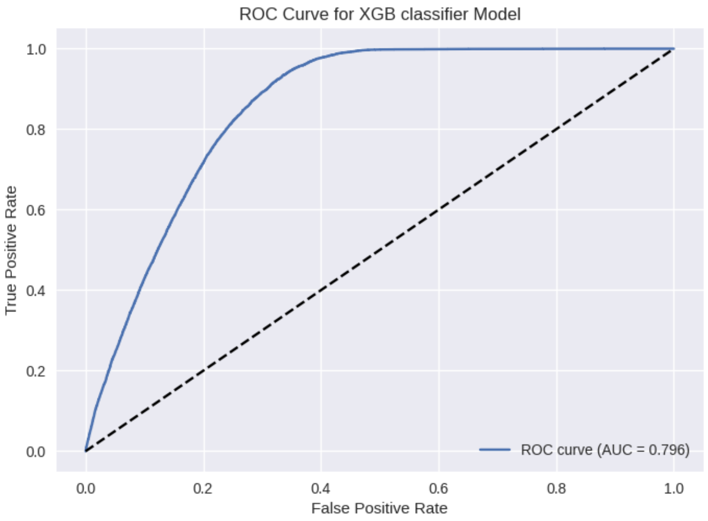


Figure 21 AUC figure to evaluate the XGB Classifier model

Procedure	Score
cross-validate with precision	75.3%
cross-validate with recall	94.2%
cross-validate with AUC score	87.3%

Table 12 The procedures used to evaluate the model and their scores

d. Gradient Boosting Classifier

The evaluation of the outcomes according to the predictive and actual results shown through the confusion tables. The values comparable to true positive, true negative, false positive, and false negative respectively include 8339, 43579, 21594 and 646, resulting in the precision is 27.8%, the recall is 92.8%. Furthermore, the confusion matrix is used to illustrate AUC values, with the accuracy of 79.8%.On the other hand, another procedure used is k-fold cross validation with the three metrics above, which resulted in a recall of 92.8%, a precision of 73.6%, and an AUC score of 85.7%.

		Actual	
		Interested	**Not interested**
Predict	**Interested**	8339	21594
	Not interested	646	43579

Table 13 Confusion matrix evaluating Gradient Boosting Classifier model

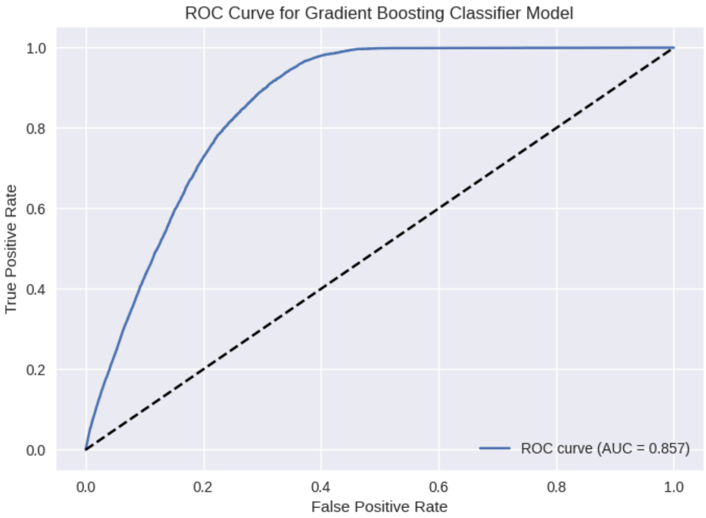


Figure 22 AUC figure to evaluate the Gradient Boosting Classifier model

Procedure	Score
cross-validate with precision	73.6%
cross-validate with recall	92.8%
cross-validate with AUC	85.7%

Table 14 The procedures used to evaluate the model and their scores

e. Gaussian Naïve Bayes classifier

The performance of the model was evaluated by comparing its predictions to the actual results, as shown in the confusion table. The confusion table showed that the model had 8770 true positives, 38634 true negatives, 26539 false positives, and 215 false negatives. This resulted in an overall precision of 24.8% and an overall recall of 97.6%. Additionally, the confusion matrix was used to calculate the model's AUC score, which was 78.4%. On the other hand, another procedure used is k-fold cross validation with the three metrics above, which resulted in a recall of 97.7%, a precision of 70.5%, and an AUC score of 82.5%.

		Actual	
		Interested	**Not interested**
Predict	**Interested**	8770	26539
	Not interested	215	38634

Table 15 Confusion matrix evaluating Gaussian Naïve Bayes model

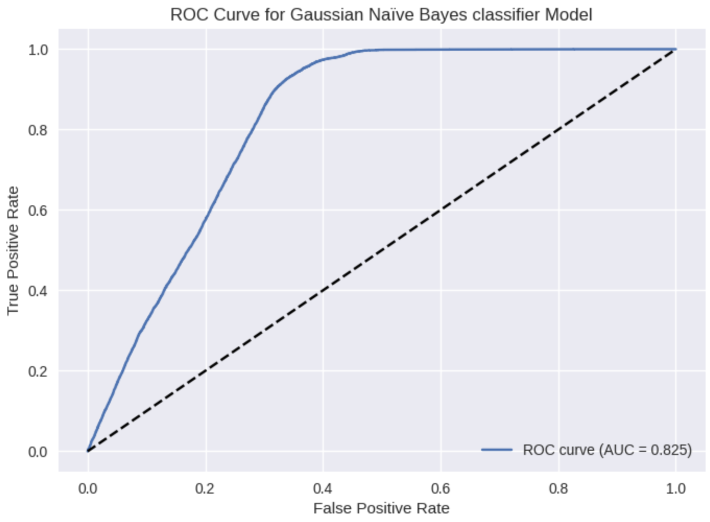


Figure 23 AUC figure to evaluate the Gaussian Naïve Bayes model

Procedure	Score
cross-validate with precision	70.5%
cross-validate with recall	97.7%
cross-validate with AUC score	82.5%

Table 16 The procedures used to evaluate the model and their scores

2. Model comparison and summarization

	Logistic Regression	Random Forest	XGB Classifier	Gradient Boosting Classifier	Gaussian Naïve Bayes
Precision	24.8%	27.2%	28.7%	27.8%	24.8%
Recall	97.6%	93.7%	90.1%	92.8%	97.6%
AUC_score	78.4%	79.6%	79.6%	79.8%	78.4%
Precision with Kfold	70.5%	92.1%	75.3%	73.6%	70.5%
Recall with Kfold	97.6%	99.9%	94.2%	92.8%	97.7%
AUC_score with Kfold	82%	85.4%	87.3%	85.7%	82.5%

Table 17 Machine learning algorithms comparison result

All algorithms used to measure the accuracy of the model produced logical and impressive results. Table 17 analyzes the evaluation scores for different algorithms, including Logistic Regression, Random Forest, XGB Classifier, Gradient Boosting Classifier, and Gaussian Naïve Bayes. As shown in the table, XGBClassifier and Random Forest outperformed the other models on all metrics. The random forest model achieved a recall score of 94% and a precision score of 73.4% with Kfold, significantly higher than the other models. XGBClassifier achieved a precision score of 28.7 on the test set. The remaining models performed poorly on all evaluation metrics.

All models made very good predictions, meeting the business case requirements. However, for this problem, we should focus on two evaluation metrics: recall and AUC score. A high recall score indicates that the models can reach out to the maximum number of potential car insurance customers. A high AUC score indicates that the models can accurately distinguish between potential customers and those who are unlikely to buy insurance. Finally, the precision score is not a reliable measure of model effectiveness in this problem because the test set contains too many customers who are not interested in buying insurance. To address this issue, we need to collect more data on potential car insurance customers. This will improve the model's learning efficiency and allow for more objective evaluations.

V. CONCLUSION

Cross-selling is a technique of selling or providing additional services to existing customers. It is an effective way for companies to multiply their profits. However, reaching target customers quickly and cost-effectively is a challenge for many companies, especially in the insurance sector, where companies want to sell as many insurance packages as possible to customers.

This project proposes a machine learning solution to classify customers and find potential customers. It also helps businesses identify important factors to easily identify potential customers and improve services to attract more customers.

Several supervised machine learning models, including XGB Classifier, Random Forest, Gaussian Naïve Bayes, Gradient Boosting Classifier, and Logistic Regression, were used. The model was trained and evaluated with various metrics, including precision, recall, and auc_score. Random Forest and XGB Classifier performed best on this dataset.

However, a limitation is that the dataset used was imbalanced, resulting in a low precision score due to poor handling of the imbalance. Future work will involve collecting larger and more diverse datasets to address the data imbalance and improve model training, resulting in more accurate predictions of potential customers.

VI. REFERENCES

Su Hyun AN, Seong Hee YEO, Minsoo KANG, "A Study on a car Insurance purchase Prediction Using Two-Class Logistic Regression and Two-Class Boosted Decision Tree", *Korean Journal of Artificial Intelligence, 2021.*

Kamakura, W. A, "Cross-selling: Offering the right product to the right customer at the right time". *Journal of Relationship Marketing, 2008.*

Marryville University, "Predictive Analytics in Insurance: Types, Tools, and the Future". *2020.*

Kamakura, W. A., Kossar, B. S., & Wedel, M. "Identifying innovators for the cross-selling of new products", *Management Science, 2004.*

Sotiris Kotsiantis. et al, "Handling imbalanced datasets: A review", *GESTS International Transactions on Computer Science and Engineering, Vol.30, 2006.*