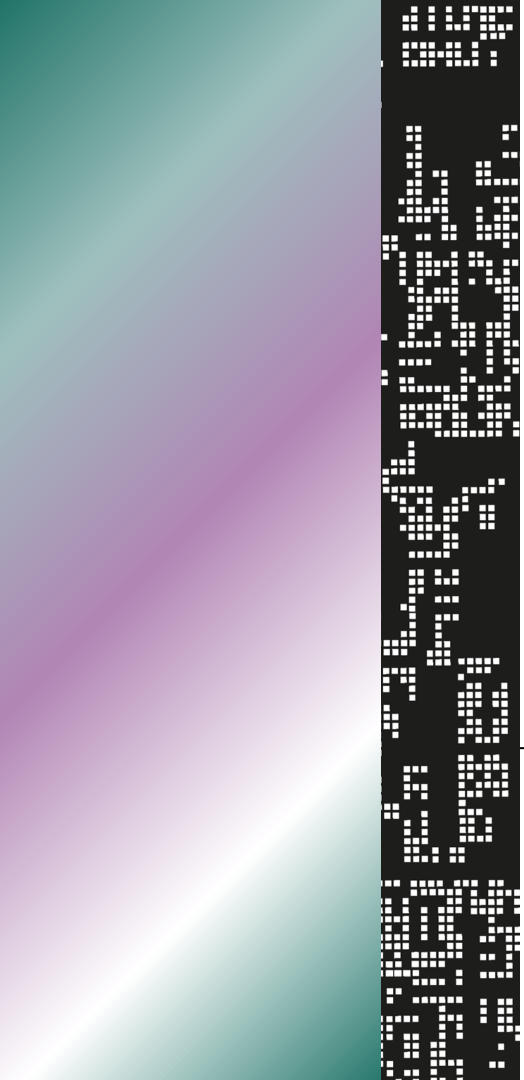


Predicting Vehicle Insurance Cross-Sell for Health Insurance Customers

Data Science Project

Group Members

01	Hoàng Hà Đăng	BI12-077
02	Nguyễn Anh Quân	BI12-365
03	Phạm Xuân Trung	BI12-458

Table of contents

Business Understanding	Define the problem and objectives of the project to address specific business needs.
Data Collection & Understanding	Gather relevant data and assess its quality and suitability for analysis.
Data Preparation	Clean, transform, and preprocess the data to make it suitable for modeling.
Model Building	Develop predictive or analytical models to extract insights or make predictions.
Model Evaluation & Selection	Assess model performance and select the best-performing model.
Deployment & Conclusion	Summarize findings, difficulties and future work with the business problem.

01 Business Understanding

A. What is cross-selling?

- A sale technique

- The company offers additional products or services to customers who have already purchased

- Increase sales and revenue

CROSS-SELLING
before
after

B. The problem

- Our client, an insurance provider, offers health insurance policies

- The company wants to expand their offerings to include vehicle insurance

=> Develop a solution to cross-sell car insurance for revenue & costumer happiness boosting using health insurance data

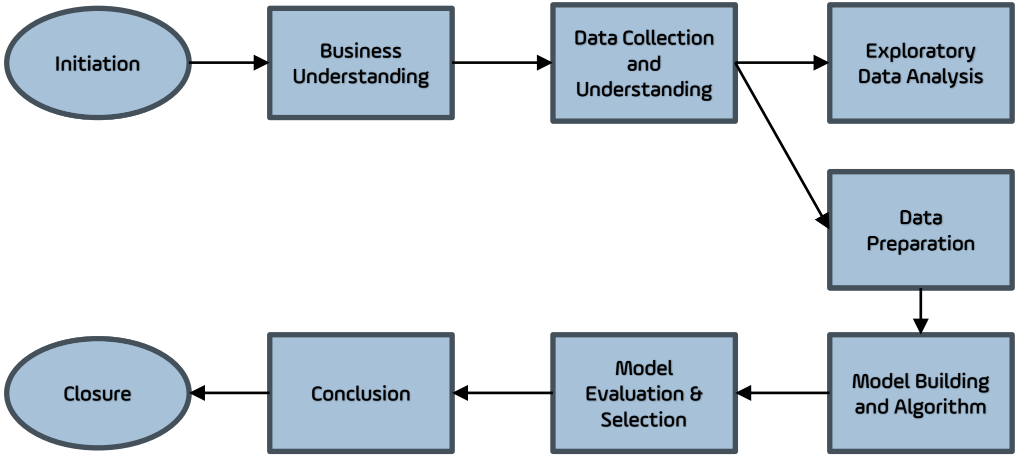
C. Non-data science approach

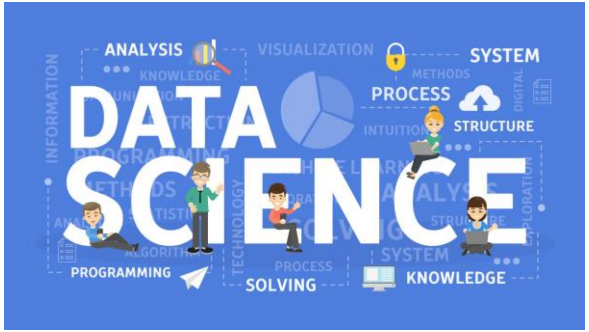
- Traditional Advertising
- Telemarketing, in-person sales meetings
- Refferal program

=> Cost great amount of time, money and labor

D. Data science approach

- Agile, modern, and cost-effective solution
- Using Python and necessary libraries


Initiation
Business Understanding
Data Collection and Understanding
Exploratory Data Analysis
Data Preparation
Model Building and Algorithm
Model Evaluation & Selection
Conclusion
Closure



E. Objective

- Build a prediction model to classify the response and find potential customers

- Handle issues related to the dataset

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

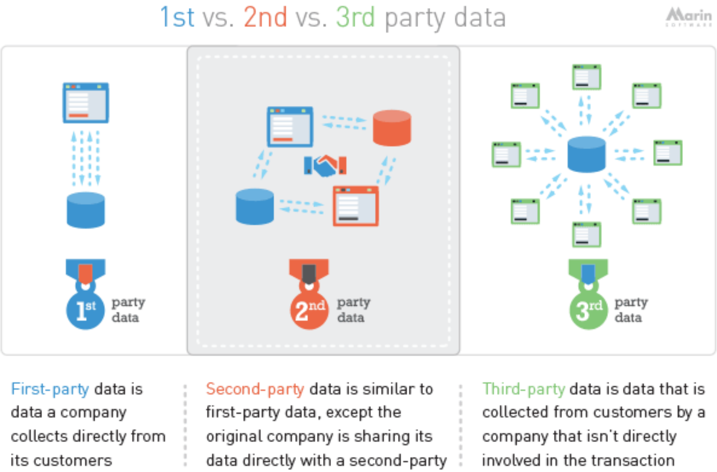
02 Data Collection & Understanding

A. Data collection

- The data contains costumer information

- Professional data collection may have a huge impact on the quality of expected outputs

- In this scenario,first party data and third-party sources (credit bureaus, government agencies, and public records)



1st vs. 2nd vs. 3rd party data

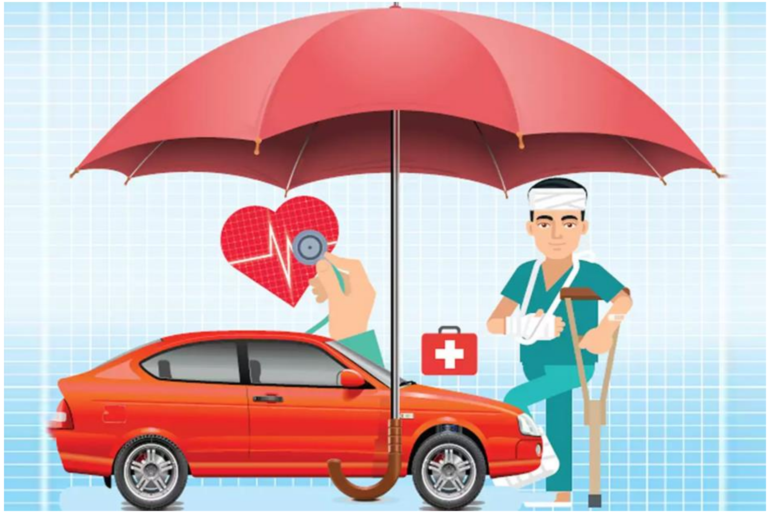
First-party data is data a company collects directly from its customers

Second-party data is similar to first-party data, except the original company is sharing its data directly with a second-party

Third-party data is data that is collected from customers by a company that isn't directly involved in the transaction

B. The Data set

- “Health Insurance Cross Sell Prediction” from Kaggle
- Health insurance owners data to forecast the personal interest in vehicle insurance
- 381109 records and 12 features



C. Data description

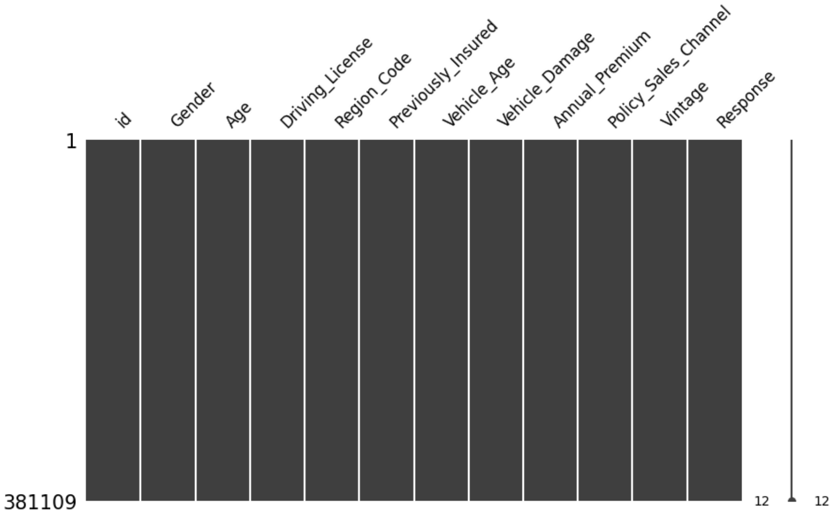
Variable	Definition	Data Type
Id	Unique ID for the customer	Qualitative
Gender	Gender of the customer	Qualitative
Age	Age of the customer	Quantitative
Driving_License	0 : Customer does not have DL, 1 : Customer already has DL	Qualitative
Region_Code	Unique code for the region of the customer	Qualitative
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

C. Data description

Variable	Definition	Data Type
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

D. Exploratory Data Analysis

- The data contains no missing values or duplicated values



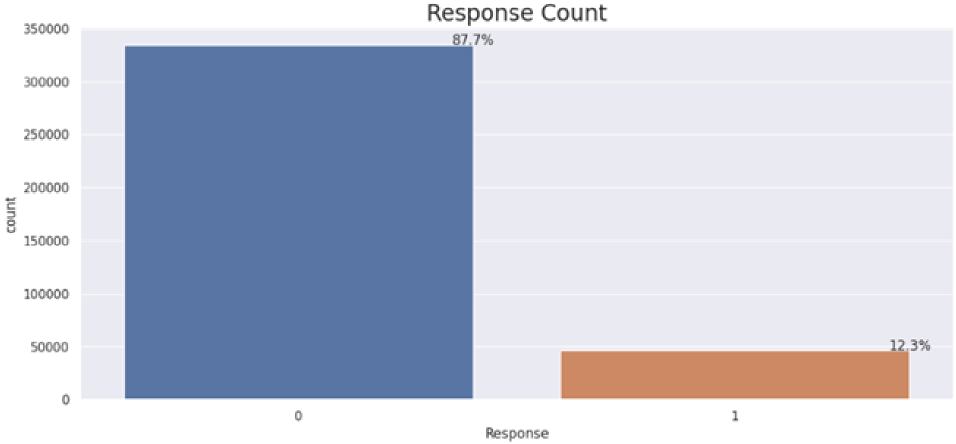
D. Exploratory Data Analysis

Respone

- 87.7% of customers are not interested in purchasing car insurance

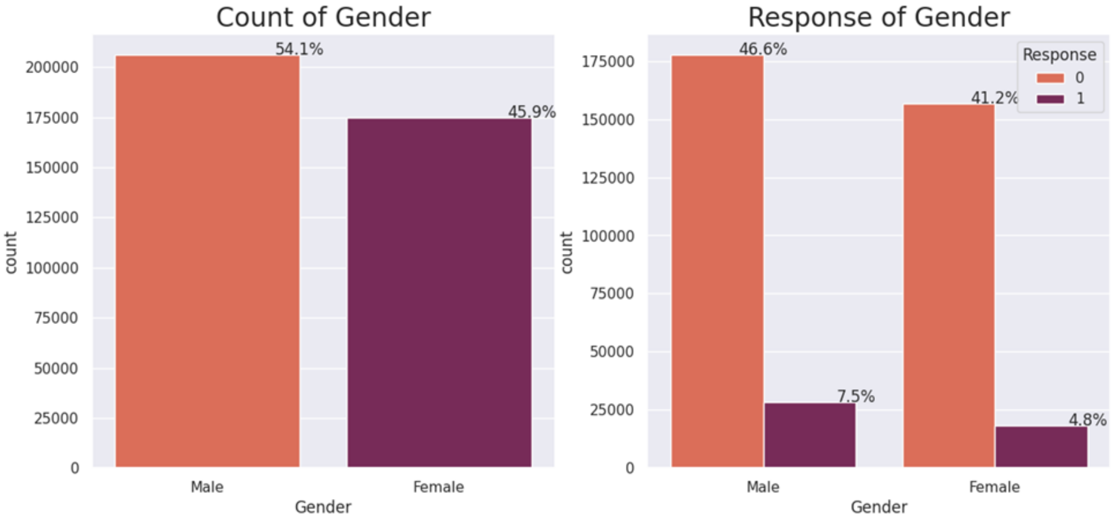
- Large ratio between the major class (not interested) and the minor class (interested)

=> The data set is imbalanced



D. Exploratory Data Analysis

Gender



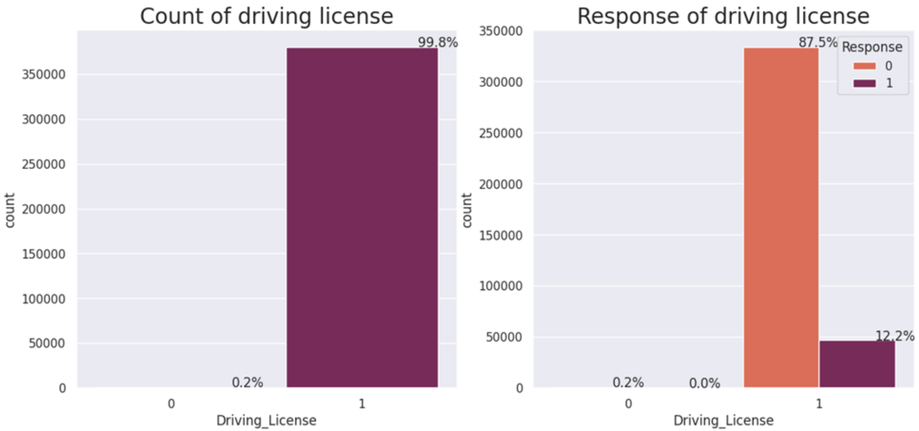
D. Exploratory Data Analysis

Driving license

- Everyone must obtain a driving license before purchasing vehicle insurance

- No effect on predicting

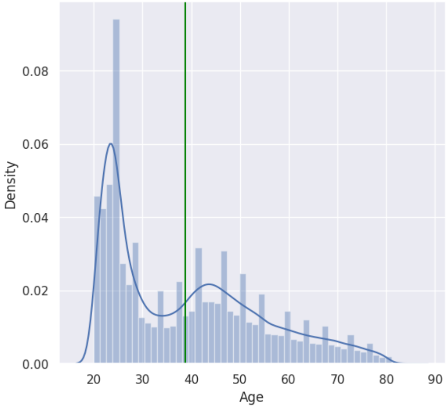
- Should be dropped

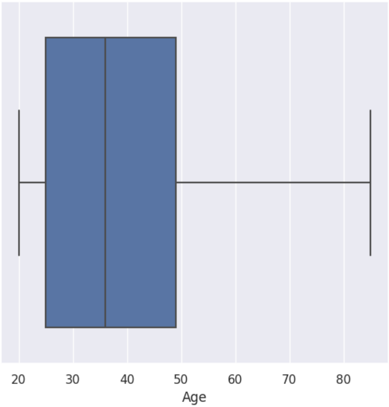


D. Exploratory Data Analysis

Age

- The population is relatively young
- Decline with age, fewest in 80+ range
- No outliers

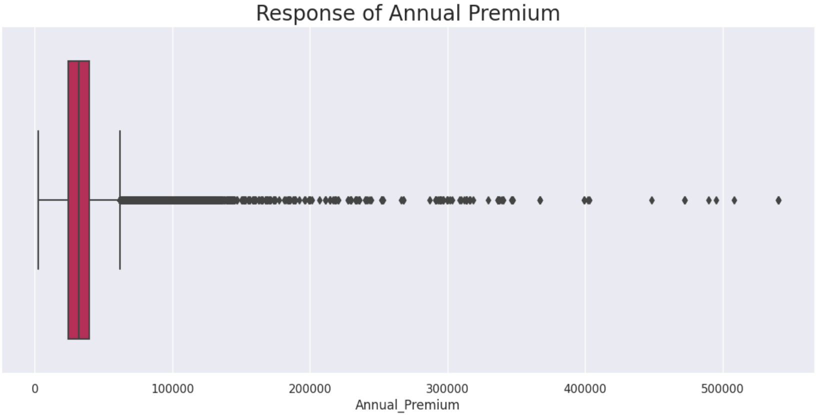




D. Exploratory Data Analysis

Annual Premium

- Lots of outliers in the variable



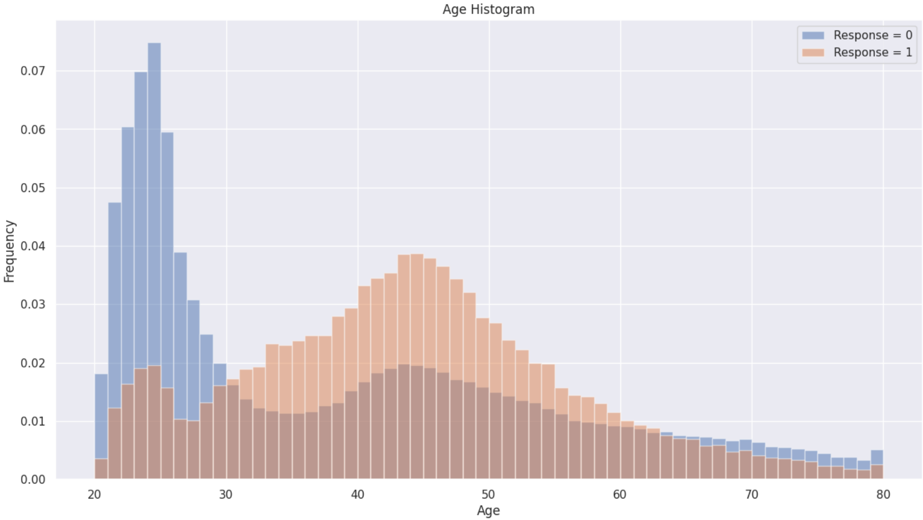
D. Exploratory Data Analysis

Age

- People aged 30 to 50 are more likely to purchase insurance

- These customers may own modern and luxury vehicles

=> Expensive repair and maintenance



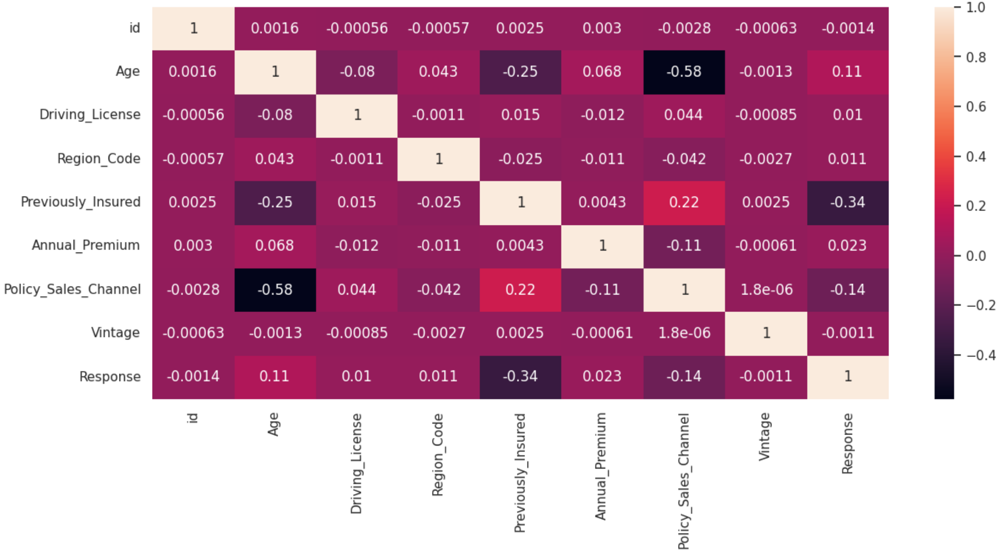
D. Exploratory Data Analysis

Correlation Matrix

- Only shows linear relationships between variables

- Does not identify strongly correlated pairs of variables

- Dropping features based on correlatio matrix analysis is not recommended



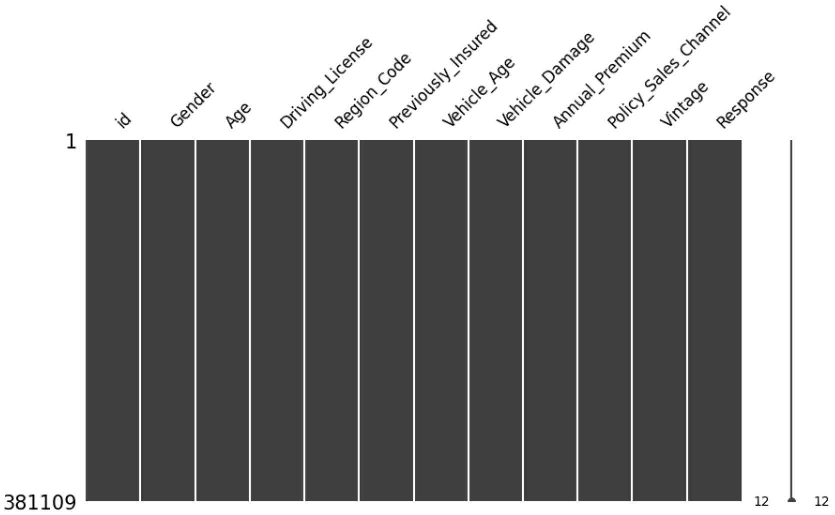
	id	Age	Driving_License	Region_Code	Previously_Insured	Annual_Premium	Policy_Sales_Channel	Vintage	Response
id	1	0.0016	-0.00056	-0.00057	0.0025	0.003	-0.0028	-0.00063	-0.0014
Age	0.0016	1	-0.08	0.043	-0.25	0.068	-0.58	-0.0013	0.11
Driving_License	-0.00056	-0.08	1	-0.0011	0.015	-0.012	0.044	-0.00085	0.01
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

03 Data Preparation

A. Handling duplicated & missing values

- No duplicated and missing values

- Skip this part



B. Feature eliminating

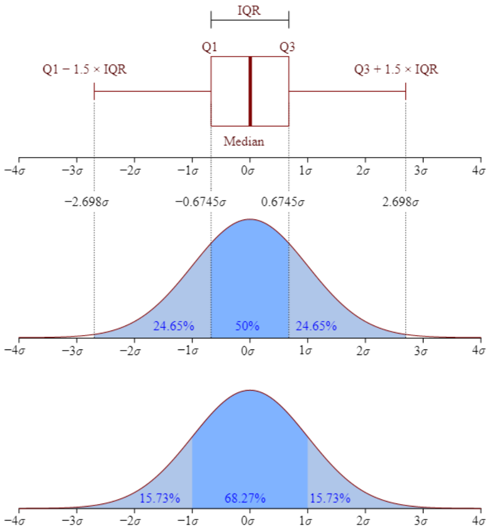
- ID
- Driving license

C. Outlier detection

- An outlier is a data point that differs significantly from other observations

- The interquartile range (IQR) is often used to find outliers in data

- Below Q1 – 1.5 IQR or above Q3 + 1.5 IQR



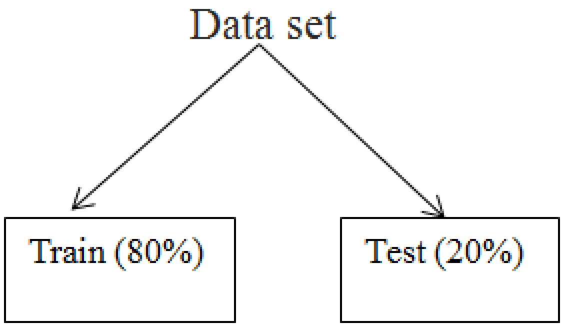
D. Categorical Value Encoder

- Convert categorical data into a numerical format

- Using Ordinal Encoder

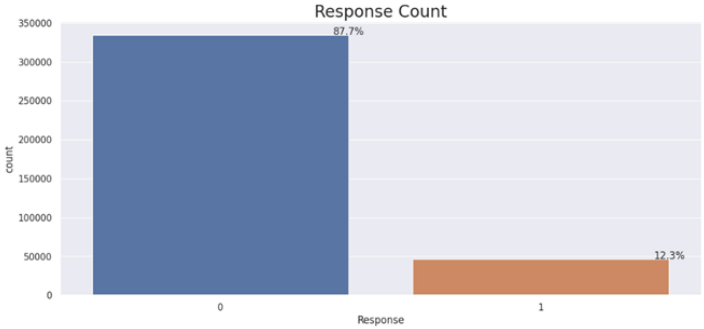
E. Train-Test Splitting

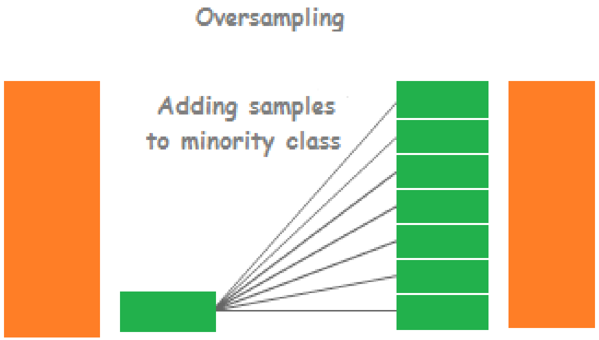
- The dataset is split into a training set (80%) and a test set (20%)



F. Imbalance data handling

- Random Oversampling using Random Over Sampler method
- Randomly duplicating instances of the minority class until a more balanced distribution is achieved
- New training set with an equal number of examples from both classes

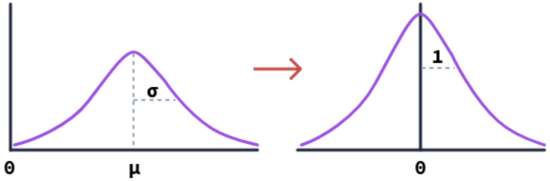




G. Standardization

- Transform the data so that it has a mean of 0 and a standard deviation of 1

- Prevent features with larger magnitudes from dominating the learning process



Standardization:

$$z = \frac{x-\mu}{\sigma}$$

with mean:

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i)$$

and standard deviation:

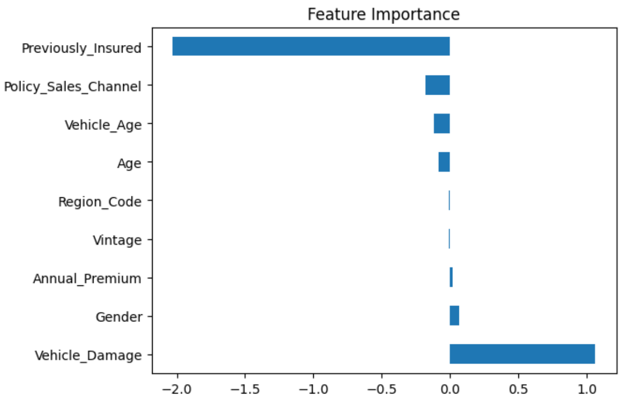
$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

04 Model Building

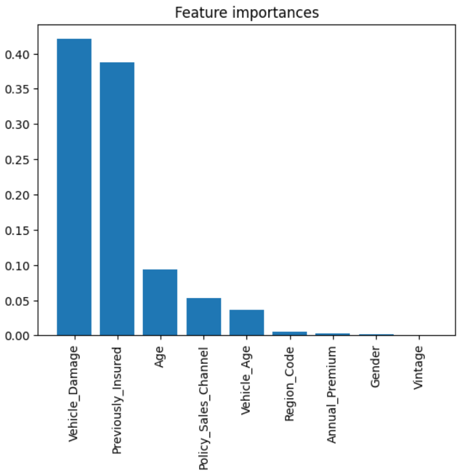
A. Model training

- Logistic Regression
- Random Forest (50 trees)
- XGB Classifier
- Gradient Boosting Classifier
- Gaussian Naive Bayes

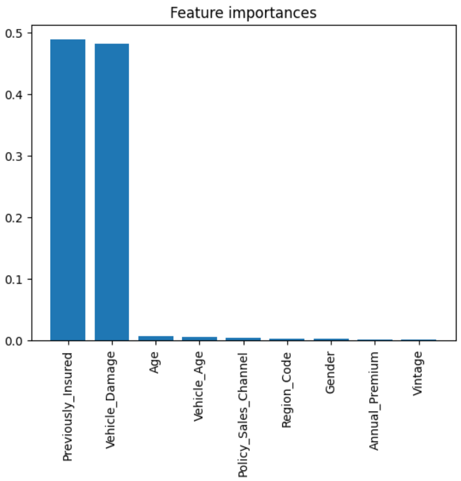
B. Feature importance



Logistic Regression

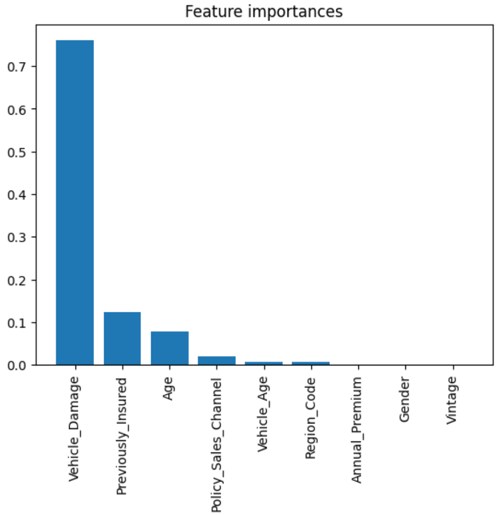


Random Forest

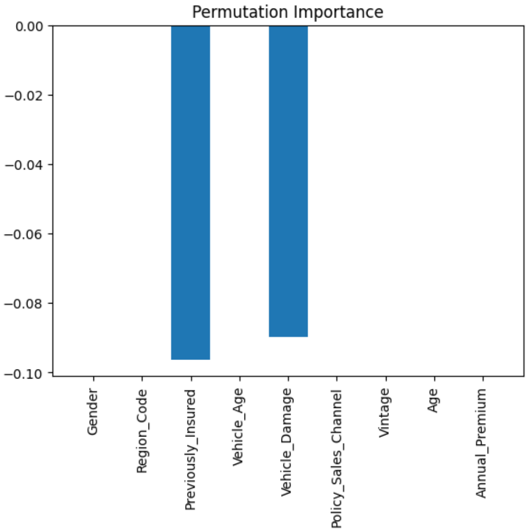


XGB Classifier

B. Feature importance



Gradient Boosting Classifier

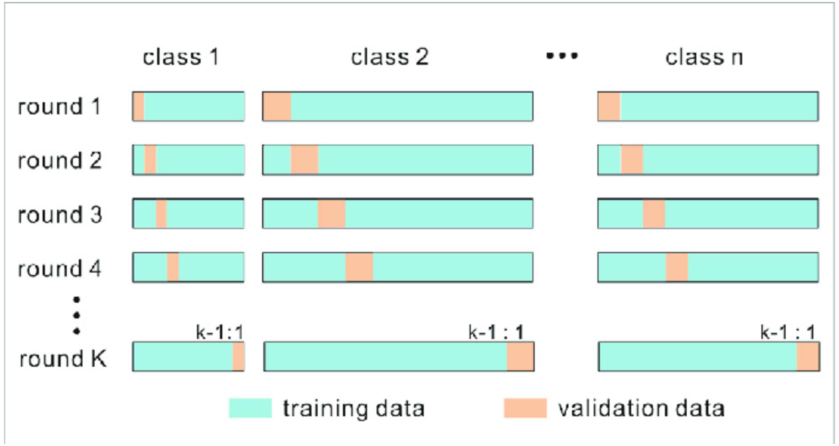


Gaussian Naive Bayes

05 Model Evaluation & Selection

A. Model Evaluation

K-fold Cross-validation



- Stratified K-fold cross validation

- K = 10

A. Model Evaluation

	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%	73.4%	75.3%	73.6%	70.5%
Recall with Kfold	97.6%	94%	94.2%	92.8%	97.7%
AUC_score with Kfold	82%	85.4%	87.3%	85.7%	82.5%

B. Model Selection

- The scores of XGB Classifier and Random Forest are better than others
- XGB Classifier will be selected to be deployed

B. Model Selection

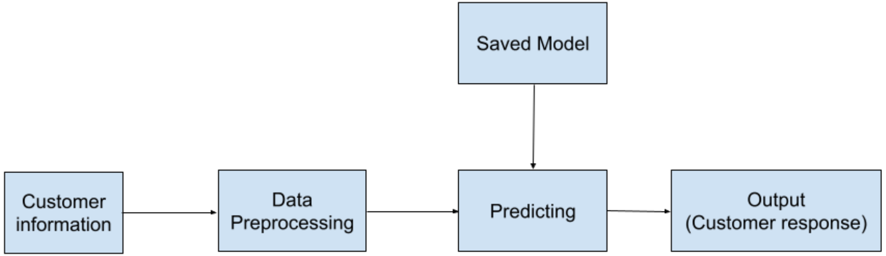
- The scores of XGB Classifier and Random Forest are better than others

- XGB Classifier will be selected to be deployed

C. Model Deployment

- Simple deployment using pickle to store the model

- The users enter the information, the function preprocesses the data and predicts using the saved model and return the output



06 Conclusion

A. Solution

- Machine learning models to classify and predict the potential customers
- Handling imbalanced data, using proper metrics

B. Limitations & Future Works

- Imbalanced dataset limitations affecting precision scores
- Simple deployment

B. Limitations & Future Works

- Collecting larger, more diverse data set
- Build a web or professional GUI for model deployment

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***

Thanks!

Do you have any questions?