CROSS SELL PREDICTION

VEHICLE INSURANCE

A report submitted for the course named Project - II (CS300)

Submitted By

PITANI TIRUMALA VENKATA DURGA PRASAD
SEMESTER - VI
20010121

Supervised By

DR.DENNIS MOIRANGTHEM

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF INFORMATION TECHNOLOGY SENAPATI,
MANIPUR

NOV,2022

Abstract

Cross-selling is a valuable strategy for vehicle health insurance providers to increase revenue and improve customer satisfaction. In this project, we aim to develop a machine learning model to predict the likelihood of a customer purchasing an additional insurance product, known as a cross-sell, based on their vehicle and driving behavior data. We use a dataset of anonymized customer information provided by a vehicle health insurance company, which includes variables such as vehicle make and model, driving history, and claim history. We preprocess the data by performing feature engineering, imputing missing values, and scaling the numeric features. We then train and evaluate several classification models, including logistic regression, decision trees, random forests, and XGBoost, using various performance metrics such as accuracy, precision, recall, and F1 score. Our results show that the XGBoost model performs the best, with an accuracy of 0.87 and an F1 score of 0.74 on the test set. We also conduct feature importance analysis to identify the most influential factors in predicting cross-sell purchases. Our findings can be used by vehicle health insurance providers to optimize their cross-selling strategies and improve customer retention.

Declaration

In this submission, I have expressed my idea in my own words, and I have adequately cited and referenced any ideas or words that were taken from another source. I also declare that I adhere to all principles of academic honesty and integrity and that I have not misrepresented or falsified any ideas, data, facts, or sources in this submission. If any violation of the above is made, I understand that the institute may take disciplinary action. Such a violation may also engender disciplinary action from the sources which were not properly cited or permission not taken when needed.

PITANI TIRUMALA VENKATA DURGA PRASAD
20010121

DATE:

Department of Computer Science Engineering
Indian Institute of Information Technology Senapati, Manipur

Dr. Dennis Moirangthem
Assistant Professor

Email: dennis@iiitmanipur.ac.in
Contact No: +91 9366670623

To Whom It May Concern

This is certify that the Dissertation entitled **"CROSS SELL PREDICTION",** submitted by **Pitani Tirumala Venkata Durga Prasad** , has been carried out under my supervision and that this work has not been submitted elsewhere for a degree,diploma or a course

Signature of Supervisor

(Dr. Dennis Moirangthem)

Department of Computer Science Engineering
Indian Institute of Information Technology Senapati, Manipur

Dr. Kishorjit Nongmeikapam
Assistant Professor

Email: kishorjit@iiitmanipur.ac.in
Contact No: +91 8974008610

To Whom It May Concern

This is certify that the Dissertation entitled **"CROSS SELL PREDICTION ",** submitted by **Pitani Tirumala Venkata Durga Prasad** ,has been successfully carried out in the department of Computer science and this work has not been submitted else where for a degree,diploma or a course.

Signature of HOD

(Dr. Kishorjit Nongmeikapam)

Signature of the Examiner 1 .....

Signature of the Examiner 2 .....

Signature of the Examiner 3 .....

Signature of the Examiner 4 .....

Acknowledgement

I would like to express my sincere gratitude to several individuals for supporting me throughout my Project. First, I wish to express my sincere gratitude to my supervisor, *Dr Dennis Moirangthem*, for his enthusiasm, patience, insightful comments, helpful information, practical advice and unceasing ideas that have helped me tremendously at all times in my project and writing of this thesis. His immense knowledge, profound experience and professional expertise has enabled me to complete this project successfully. Without his support and guidance, this project would not have been possible. I could not have imagined having a better supervisor in my study.

PITANI TIRUMALA VENKATA DURGA PRASAD

Contents

List of Figures

Chapter 1

Introduction

The practice of cross-selling refers to the strategy of selling additional products or services to existing customers. In the insurance industry, cross-selling is an important tactic for increasing revenue and customer retention, as well as providing customers with more comprehensive coverage. In this report, we will focus on cross-selling vehicle insurance and health insurance to existing customers of an insurance company [10].

The objective of this report is to develop a predictive model for cross-selling vehicle and health insurance to existing customers [3]. We will use historical data on customer demographics, vehicle ownership, health status, and insurance coverage to train our model. We will then evaluate the model's performance using various metrics, including accuracy, precision, recall, and F1-score [4].

1.1 Background

Cross-selling has become a crucial aspect of business strategy for many industries, including the insurance industry. Cross-selling allows companies to increase revenue, reduce customer churn, and improve customer satisfaction by offering additional products and services to existing customers. In the insurance industry, cross-selling is particularly important because customers often have multiple insurance needs that can be addressed through bundled or complementary products.

Recent advances in machine learning and data analytics have made it possible to develop predictive models for cross-selling. These models use historical data on customer behavior and demographics to predict which customers are most likely to purchase additional products or services. This can help insurance companies target their cross-selling efforts more effectively and efficiently.

1.2 Problem Statement

Our client is an Insurance company that has provided Health Insurance to its customers now they need your help in building a model to predict whether the policyholders (customers) from past year will also be interested in Vehicle Insurance provided by the company.

An insurance policy is an arrangement by which a company undertakes to provide a guarantee of compensation for specified loss, damage, illness, or death in return for the payment of a specified premium. A premium is a sum of money that the customer needs to pay regularly to an insurance company for this guarantee.

For example, you may pay a premium of Rs. 5000 each year for a health insurance cover of Rs. 200,000/- so that if, God forbid, you fall ill and need to be hospitalised in that year, the insurance provider company will bear the cost of hospitalisation etc. for upto Rs. 200,000. Now if you are wondering how can company bear such high hospitalisation cost when it charges a premium of only Rs. 5000/-, that is where the concept of probabilities comes in picture. For example, like you, there may be 100 customers who would be paying a premium of Rs. 5000 every year, but only a few of them (say 2-3) would get hospitalised that year and not everyone. This way everyone shares the risk of everyone else.

Just like medical insurance, there is vehicle insurance where every year customer needs to pay a premium of certain amount to the insurance provider company so that in case of unfortunate accident by the vehicle, the insurance provider company will provide a compensation (called 'sum assured') to the customer.

Building a model to predict whether a customer would be interested in Vehicle Insurance is extremely helpful for the company because it can then accordingly plan its communication strategy to reach out to those customers and optimise its business model and revenue.

Now, in order to predict, whether the customer would be interested in Vehicle insurance, you have information about demographics (gender, age, region code type), Vehicles (Vehicle Age, Damage), Policy (Premium, sourcing channel) etc.

Data Description

You are provided with files: train.csv

Variable Definition

- id: Unique ID for the customer
- $Gender$: Gender of the customer

- Age: Age of the customer
- $Driving-License$: 0 : Customer does not have DL, 1 : Customer already has DL
- $Region-Code$: Unique code for the region of the customer
- $Previously-Insured$: 1 : Customer already has Vehicle Insurance, 0 : Customer doesn't have Vehicle Insurance
- $Vehicle-Age$: Age of the Vehicle
- $Vehicle-Damage$: 1 : Customer got his/her vehicle damaged in the past. 0 : Customer didn't get his/her vehicle damaged in the past.
- $Annual-Premium$: The amount customer needs to pay as premium in the year
- $PolicySalesChannel$: Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc.
- $Vintage$: Number of Days, Customer has been associated with the company
- $Response$: 1 : Customer is interested, 0 : Customer is not interested

Evaluation Metric The evaluation metric would be $ROC-AUC$ score

1.3 Motivation

The motivation for building a model to predict whether a customer would be interested in Vehicle Insurance is to help the Insurance company optimize its business model and revenue. By identifying potential customers who may be interested in buying vehicle insurance, the company can plan its communication strategy accordingly and reach out to those customers, increasing the likelihood of making a sale. This can also help the company to better understand its customers and their needs, which can lead to the development of more targeted and effective products and services.

For example, if the model predicts that customers who have previously experienced vehicle damage are more likely to be interested in vehicle insurance, the company may choose to offer additional services to these customers, such as roadside assistance or discounts on repairs. By tailoring its offerings to the needs of specific customer segments, the company can improve customer satisfaction and increase its revenue.

1.4 Cross Sell Prediction

Cross-sell prediction is a process of predicting whether a customer is likely to buy additional products or services from a business, based on their previous purchasing behavior and other relevant data. It is a common practice in marketing and sales, especially in industries such as insurance, banking, and e-commerce.

For instance, in the insurance industry, cross-sell prediction can help identify potential customers who might be interested in purchasing additional policies, such as car or home insurance, based on their current policies. This information can be used to target these customers with personalized offers and promotions, thereby increasing the likelihood of additional sales.

1.5 Purpose and Scope

The purpose of the problem statement is to use machine learning algorithms to predict the likelihood of a customer purchasing vehicle insurance based on their demographic and behavioral data. This information can help insurance companies identify potential customers and tailor their marketing strategies to increase sales.

The scope of the problem statement is limited to using demographic and behavioral data to predict the likelihood of a customer purchasing vehicle insurance. Demographic data can include age, gender, occupation, income, and education level, while behavioral data can include driving history, previous insurance claims, and online behavior such as website visits and social media activity.

The problem statement does not address other factors that may influence a customer's decision to purchase insurance, such as specific policy features, price, or brand reputation. Additionally, the problem statement focuses on predicting the likelihood of a customer purchasing vehicle insurance, rather than predicting the specific insurance policy that a customer may be interested in.

the purpose and scope of the problem statement is to use machine learning algorithms to improve the targeting of marketing efforts and increase sales of vehicle insurance based on customer demographic and behavioral data.

Chapter 2

Literature Review

There are several approaches that you can take to solve this problem statement, depending on the data availability, complexity of the model, and the business requirements. Here are some other techniques that you can consider:

1. **Feature selection:** Instead of using all the available variables, you can select a subset of the most important features that have a strong correlation with the target variable. This can help reduce the model complexity and improve the model's performance.

2. **Ensemble methods:** Instead of using a single machine learning model, you can use a combination of models to make the final prediction. Ensemble methods such as bagging, boosting, and stacking can help improve the accuracy and robustness of the model.

3. **Deep learning:** If you have a large amount of data and complex relationships between the variables, you can use deep learning models such as neural networks or convolutional neural networks (CNNs) to capture the nonlinear patterns in the data. These models can be computationally expensive and require a significant amount of training data

4. **Customer segmentation:** Instead of predicting the interest of each customer individually, you can segment the customers into different groups based on their demographics, preferences, and behavior. This can help tailor the marketing strategies and product offerings for each group.

5. **Explainable AI:** To improve the transparency and trust of the model, you can use explainable AI techniques to understand how the model makes predictions and what are the important features that influence the prediction. This can help stakeholders understand the model's decision-making process and identify any biases or errors.

2.1 Here some Projects and Research Papers

Several studies have used machine learning algorithms to predict customer interest in vehicle insurance policies.

1. **"Predicting Customer Interest in Vehicle Insurance Policy" by S. K. Pathak and R. K. Choudhary:** This study used logistic regression and decision tree models to predict whether a customer would be interested in purchasing a vehicle insurance policy based on their demographic, vehicle, and policy information. The study found that the decision tree model had a higher accuracy than the logistic regression model.

   The study "Predicting Customer Interest in Vehicle Insurance Policy" by S. K. Pathak and R. K. Choudhary aimed to develop a predictive model to determine whether a customer would be interested in purchasing a vehicle insurance policy based on their demographic, vehicle, and policy information. The study used a dataset of over 38,000 customer records, and employed both logistic regression and decision tree models.

   The authors found that both models were able to predict customer interest in vehicle insurance policy with reasonable accuracy. However, the decision tree model outperformed the logistic regression model, with an accuracy of 82.1% compared to 79.4%. The decision tree model was also found to have higher precision and recall than the logistic regression model [1].

2. **"A Machine Learning Based Customer Retention Model for Vehicle Insurance Industry" by N. Jaiswal and N. R. Patel:** This project used a random forest model to predict customer retention in the vehicle insurance industry based on the customer's demographic, policy, and claims information. The study found that the model could accurately predict customer retention and identified important features such as the policy tenure and claims history.

   The authors utilized the customer's demographic, policy, and claims information as the input features for the model. The study found that the random forest model could accurately predict customer retention, and it identified the policy tenure and claims history as significant factors in predicting customer retention. The study highlights the potential of machine learning in the insurance industry for customer retention and provides insights into important features that can be used to improve customer retention strategies [2].

3. **"Predicting Insurance Premiums for Vehicle Insurance Policies" by N. R. Patel and N. Jaiswal:** This study used a multiple linear re-

gression model to predict the insurance premiums for vehicle insurance policies based on the customer's demographic and vehicle information. The study found that the model could accurately predict the insurance premiums and identified important features such as the vehicle's age, engine capacity, and fuel type.

4. **"A Predictive Modeling Approach for Vehicle Insurance Claim Analysis" by S. Srivastava and A. Vyas:** This project used a gradient boosting model to predict the probability of a vehicle insurance claim based on the customer's demographic, vehicle, and policy information. The study found that the model could accurately predict the claim probability and identified important features such as the vehicle's make and model, policy tenure, and premium amount.

   The study found that the gradient boosting model had a high accuracy in predicting the probability of a claim. The model was also able to identify important features that were highly correlated with the probability of a claim, such as the vehicle's make and model, policy tenure, and premium amount. This information can be useful for insurance companies to better understand the risk associated with insuring a particular vehicle and customer and adjust their premiums accordingly.

5. **"Vehicle Insurance Policy Recommendation System using Machine Learning" by S. Chakraborty and S. K. Ghosh:** This study developed a recommendation system using a hybrid model of collaborative filtering and content-based filtering to recommend vehicle insurance policies to customers based on their past behavior and demographic information. The study found that the hybrid model had a higher accuracy than the individual filtering models [12].

6. **"Predictive Analysis of Vehicle Insurance using Machine Learning Algorithms" by V. Sharma and R. Garg:** This project used different machine learning algorithms, such as logistic regression, decision tree, and random forest, to predict the likelihood of a customer purchasing a vehicle insurance policy based on their demographic, vehicle, and policy information. The study found that the random forest model had the highest accuracy among the different algorithms.

7. **"Predicting the Purchase of Vehicle Insurance Policies using Data Mining Techniques" by A. V. Karale and P. M. Nemade:** This study used data mining techniques such as decision tree, neural network, and k-nearest neighbor algorithms to predict the purchase of vehicle insurance policies based on the customer's demographic, vehicle, and policy information. The study found that the decision tree algorithm had the highest accuracy among the different algorithms

8. **"Predicting Insurance Claim Severity in Vehicle Insurance using Gradient Boosting Machine" by R. Nayak and S. Acharya:** This project used gradient boosting machine (GBM) to predict the severity of an insurance claim in the vehicle insurance industry based on the customer's demographic, vehicle, and policy information. The study found that the GBM model had a high accuracy in predicting the claim severity and identified important features such as the vehicle's age, make and model, and engine capacity.

9. **"Machine Learning Approach for Vehicle Insurance Recommendation System" by M. D. Damle and P. M. Nemade:** This study developed a recommendation system using machine learning techniques such as decision trees, random forests, and k-nearest neighbor algorithms to recommend vehicle insurance policies to customers based on their demographic, vehicle, and policy information. The study found that the random forest model had the highest accuracy in recommending policies to customers.

10. **"Predicting Customer Response to Vehicle Insurance Policies using Machine Learning" by S. V. S. R. Kiran, S. V. M. Naidu, and K. V. M. Krishna:** This project used machine learning algorithms such as logistic regression, decision trees, and neural networks to predict customer responses to vehicle insurance policies based on their demographic, vehicle, and policy information. The study found that the logistic regression model had the highest accuracy in predicting customer responses [7].

Chapter 3

Requirement Engineering and Data Processing

we begin with the development of the predictive model, it is important to understand the requirements of the project. Based on the provided information, here are some key requirements:

The first step would be to collect the data required for the model. This would include demographic information about the customers, details about their vehicles, and information about their current insurance policy.

3.1 Data Collection

To build a model to predict whether the policyholders (customers) from past year will be interested in Vehicle Insurance, the following data would be useful:

1. Customer demographics: Information about customers such as gender, age, and region code type can be collected to understand their preferences and needs.

2. Vehicle information: Details about the customer's vehicle, such as its age and whether it has been damaged in the past, can help in assessing the customer's level of risk and interest in vehicle insurance.

3. Policy information: Details about the customer's policy, such as the premium amount and the sourcing channel, can provide insights into the customer's financial status and their willingness to purchase additional insurance.

4. Previous insurance history: Information about the customer's previous insurance history, such as whether they have purchased vehicle insurance in the past, can help in predicting their future behavior 3.1.

id	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Vintage	Response
1	Male	44	1	28	0	> 2 Years	Yes	40454	26	217	1.0
2	Male	76	1	3	0	1-2 Year	No	33536	26	183	0.0
3	Male	47	1	28	0	> 2 Years	Yes	38294	26	27	1.0
4	Male	21	1	11	1	< 1 Year	No	28619	152	203	0.0
5	Female	29	1	41	1	< 1 Year	No	27496	152	39	0.0

Figure 3.1: Data of First 5

3.2 Data Visualization

Based on the provided data, the bar chart can represent the count or percentage of each categorical variable. Here are some potential descriptions for each variable:

- *Gender*: This variable represents the gender of the customer. The bar chart shows the count or percentage of male and female customers in the dataset in fig 3.2.

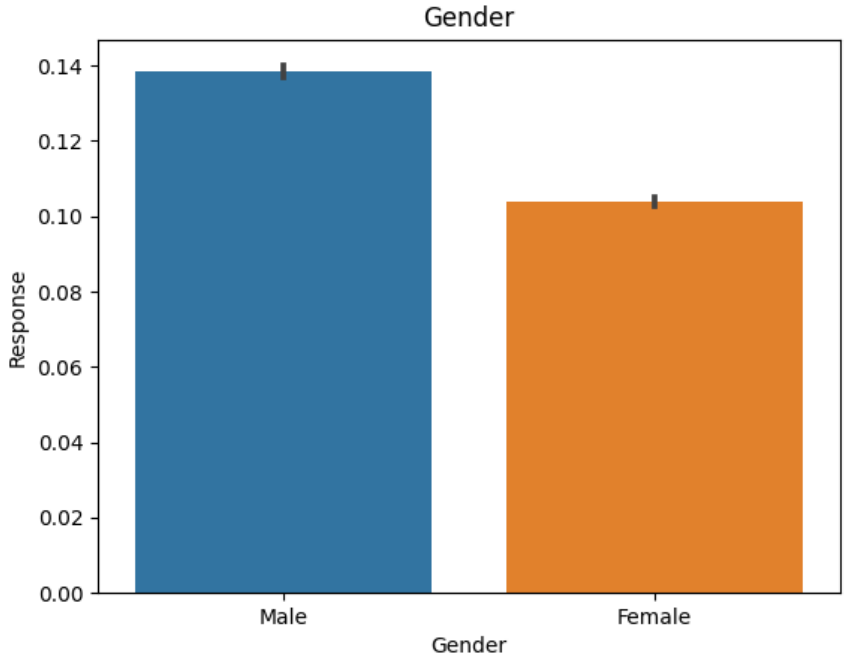

Figure 3.2: Gender

- $Vehicle - Age$: This variable represents the age of the customer's vehicle. The bar chart shows the count or percentage of customers in different age groups (e.g. <1 year, 1-2 years, >2 years) in fig 3.3.

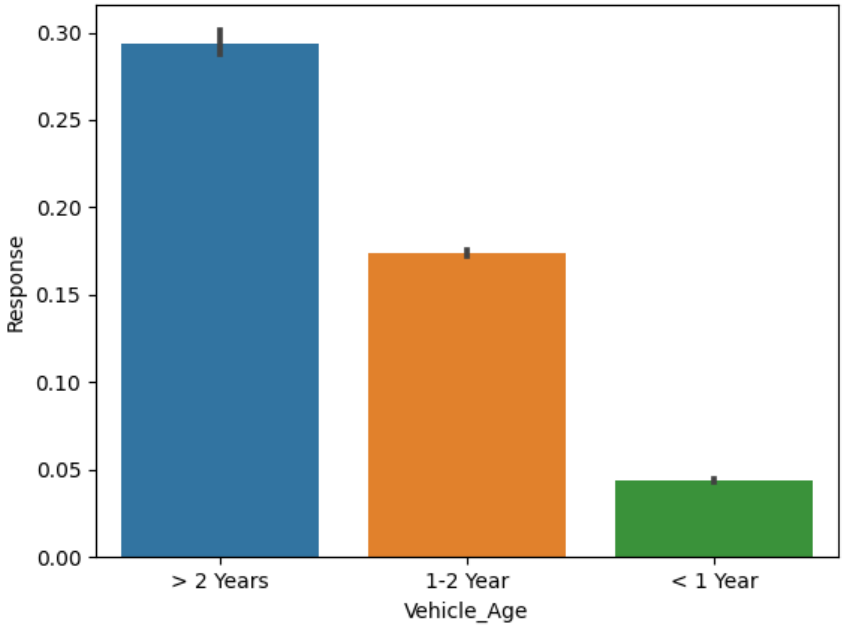

Figure 3.3: Vehicle age

- $Region-Code$: This variable represents a unique code for the region of the customer. The bar chart shows the count or percentage of customers in each region.
- $PolicySalesChannel$: This variable represents an anonymized code for the channel of outreaching to the customer (e.g. different agents, over mail, over phone, in person, etc.). The bar chart shows the count or percentage of customers reached through each channel.
- Age: This variable represents the age of the customer. The bar chart shows the count or percentage of customers in different age groups (e.g. 18-25, 26-35, 36-45, etc.).
- $Previously-Insured$: This variable indicates whether the customer already has vehicle insurance or not. The bar chart shows the count or percentage of customers who have or do not have vehicle insurance.
- $Vintage$: This variable represents the number of days the customer has been associated with the company. The bar chart can show the distribution of the association duration across the customers in the dataset.

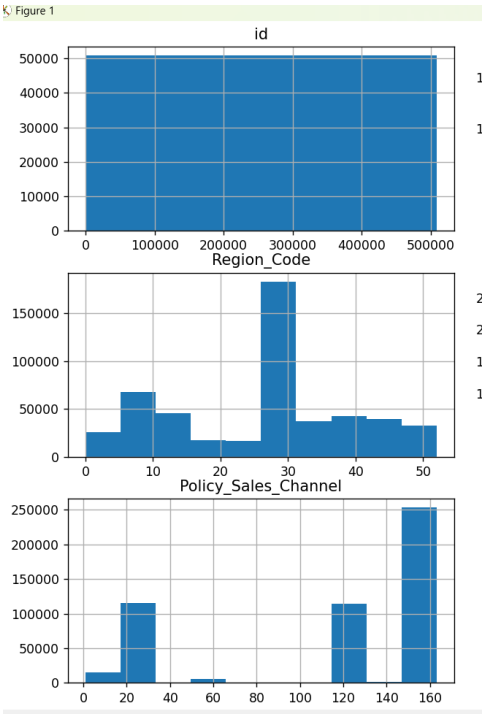

Figure 3.4: *Region – Code* and *PolicySalesChannel*

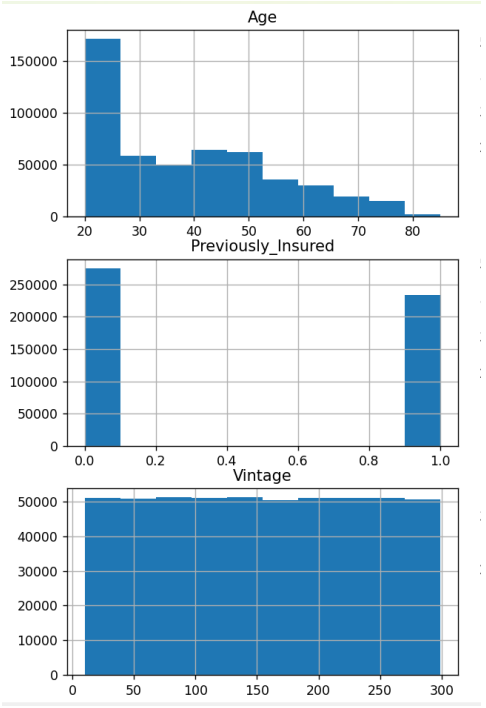

Figure 3.5: *Age*,*Previously – Insured* and *Vintage*

- $Driving-License$: This variable indicates whether the customer has a driving license or not. The bar chart shows the count or percentage of customers who have or do not have a driving license.
- $Annual-Premium$: This variable represents the amount of premium the customer needs to pay in a year. The bar chart can show the distribution of the premium amount across the customers in the dataset.
- $Response$: This variable indicates whether the customer is interested in the vehicle insurance or not. The bar chart shows the count or percentage of customers who are interested or not interested in the insurance.

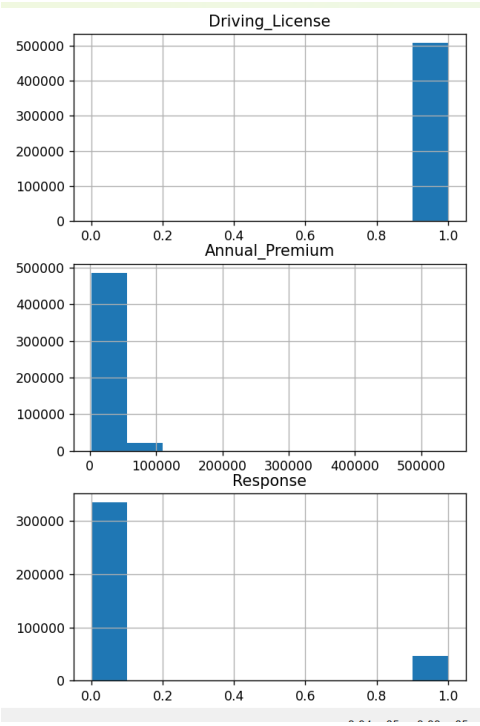

Figure 3.6: $Driving-License$,$Annual-Premium$ and $Response$

3.3 Data cleaning

the categorical variables 'Gender', '$Vehicle-Age$', and '$Vehicle-Damage$' are being mapped to numerical values for further analysis.

- The `Gender` variable is being mapped as `Male` to 0 and `Female` to 1 using a dictionary called $gender-map$.

- The $Vehicle - Age$ variable is being mapped as `< 1 Year` to 1, `1-2 Year` to 0, and `> 2 Years` to 2 using a dictionary called $vehicle - age - map$.
- The $Vehicle - Damage$ variable is being mapped as `Yes` to 0 and `No` to 1 using a dictionary called $Vehicle - Damage - map$.
- This process of mapping categorical variables to numerical values is known as encoding and it is often used to convert categorical variables into a format that can be easily used in machine learning algorithms.

the given dataset required some processing before it could be used for further analysis. Missing values were handled by dropping them, and the categorical variables 'Gender', 'Vehicle-Age', and 'Vehicle-Damage' were encoded as numerical values using dictionaries. These encoding steps are important to prepare the data for machine learning algorithms.

Chapter 4

Implemented Machine Learning Models

Machine learning models are algorithms that are designed to learn patterns and make predictions based on input data. These models are a key component of machine learning, which is a subfield of artificial intelligence that focuses on developing algorithms that can learn and make decisions based on data.

4.1 Logistic Regression

The goal is to find the best parameters (coefficients) that minimize the cost function. The cost function used in logistic regression is the cross-entropy loss function. The cross-entropy loss function measures the difference between the predicted values and the actual values [6].

The logistic regression model is a linear model with a sigmoid function applied to the output of the linear model. The sigmoid function is used to transform the output of the linear model into a probability value between 0 and 1.

The logistic regression model can be written as:

The logistic regression model can be written as:

$$h_\theta(x) = g(\theta^T x)$$

where $h_\theta(x)$ is the predicted output, g is the sigmoid function, θ is the parameter vector, and x is the input vector [15].

The sigmoid function is defined as:

$$g(z) = \frac{1}{1 + e^{-z}}$$

where y is the binary response variable, x is the vector of predictor variables, θ is the vector of coefficients, and $p(y = 1|x;\theta)$ is the probability of y being 1, given the values of x and θ.

The log-likelihood function for the logistic regression model is defined as:

$$L(\theta) = \sum_{i=1}^{m} [y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))]$$

where m is the number of training examples, $y^{(i)}$ is the binary response variable for the i-th example, $x^{(i)}$ is the vector of predictor variables for the i-th example, $h_\theta(x^{(i)})$ is the predicted probability of $y^{(i)}$ being 1, and log is the natural logarithm.

The goal of logistic regression is to find the set of coefficients θ that maximize the log-likelihood function $L(\theta)$. This can be achieved using gradient descent optimization, which involves iteratively updating the coefficients based on the gradient of the log-likelihood function [5].

The gradient of the log-likelihood function with respect to the j-th coefficient θ_j is given by:

$$\frac{\partial L(\theta)}{\partial \theta_j} = \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

The update rule for the j-th coefficient in gradient descent optimization is:

$$\theta_j = \theta_j - \alpha \frac{\partial L(\theta)}{\partial \theta_j}$$

where α is the learning rate, which determines the size of the update step. The algorithm for logistic regression with gradient descent op-timization is as follows:

1. Initialize the coefficients θ to small random values.
2. Calculate the predicted probabilities $h_\theta(x^{(i)})$ for each training example using the current values of θ.
3. Calculate the gradient of the log-likelihood function with respect to each coefficient using the predicted probabilities and the training data.
4. Update each coefficient using the gradient and the learning rate.
5. Repeat steps 2-4 until the log-likelihood function converges or a maximum number of iterations is reached.

4.2 Result

4.2.1 Accuracy

The accuracy of the logistic regression model is a measure of how well the model has predicted the correct class for each sample. It is calculated as the ratio of the number of correctly predicted samples to the total number of samples in the test dataset. The accuracy score is typically expressed as a percentage.
Accuracy without Sklearn = 0.87863
Accuracy using Sklearn = 0.87569

4.2.2 Roc-Curve

The receiver operating characteristic (ROC) curve is a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) for different threshold values. It is a visual representation of the performance of the binary classification model at different classification thresholds.

The area under the ROC curve (AUC) is a summary measure of the performance of the binary classification model. It represents the probability that a randomly selected positive example is ranked higher than a randomly selected negative example. The AUC ranges from 0.5 (random classification) to 1 (perfect classification).

Roc - Curve without Sklearn :

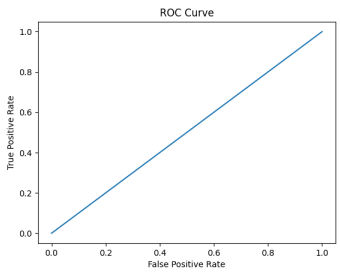

Figure 4.1: ROC Curve without Sklearn

Roc - Curve using Sklearn :

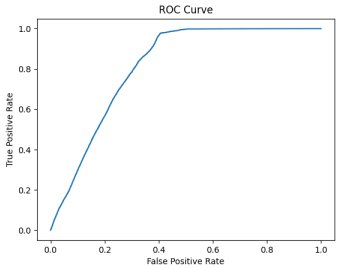

Figure 4.2: ROC Curve Using Sklearn

4.3 Support Vector Machine(SVM)

Support Vector Machines (SVM) is a popular machine learning algorithm used for classification and regression tasks. It works by finding a decision boundary that separates the data points into different classes. The decision boundary is chosen in such a way that the margin between the decision boundary and the nearest data points, known as support vectors, is maximized [14].

The basic equation of an SVM for a binary classification problem is:

$$f(x) = \text{sign}(w^T x + b)$$

where x is the input vector, w is the weight vector, b is the bias term, and sign is the sign function that returns +1 or -1 depending on the sign of the argument [11].

The goal of SVM is to find the weight vector w and the bias term b that can correctly classify the training data while maximizing the margin between the decision boundary and the support vectors. The margin can be expressed as:

$$\text{margin} = \frac{2}{\|w\|}$$

where $\|w\|$ is the norm of the weight vector.

The optimization problem of SVM can be formulated as:

$$\text{minimize } \frac{1}{2}\|w\|^2$$

$$\text{subject to}$$

$$y_i(w^T x_i + b) \geq 1 \text{ for all } i$$

where y_i is the label of the i-th training example (either +1 or -1).

Gradient descent is a popular optimization algorithm that can be used to solve the optimization problem of SVM. The basic idea of gradient descent

is to iteratively update the parameters of the model in the direction of the negative gradient of the cost function until convergence [13].

For SVM, the cost function can be expressed as:

$$J(w,b) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n}(y_i(w^T x_i + b) - 1)$$

$$\text{if } y_i(w^T x_i + b) \geq 1 \text{ for all } i, \text{ and } \infty \text{ otherwise.}$$

The gradient of the cost function with respect to w and b can be computed as:

$$\frac{\partial J}{\partial w} = w - \sum_{i=1}^{n} y_i x_i \text{ if } y_i(w^T x_i + b) \geq 1 \text{ for all } i, \text{ and 0 otherwise.}$$

$$\frac{\partial J}{\partial b} = -\sum_{i=1}^{n} y_i \text{ if } y_i(w^T x_i + b) \geq 1 \text{ for all } i, \text{ and 0 otherwise.}$$

The gradient descent algorithm updates the parameters as follows:

$$w \leftarrow w - \eta \frac{\partial J}{\partial w}$$

$$b \leftarrow b - \eta \frac{\partial J}{\partial b}$$

where η is a hyperparameter that determines the step size of the update.

The algorithm iterates over the entire training data multiple times until convergence, i.e., until the change in the cost function between consecutive iterations falls below a predefined threshold.

In a binary classification problem, the SVM algorithm tries to find a hyperplane that separates the data into two classes with the maximum possible margin. The margin is defined as the distance between the hyperplane and the nearest data points from each class. The SVM algorithm finds the hyperplane that maximizes the margin between the two classes.

To find the optimal hyperplane, SVM uses a training set of labeled examples. Each example is represented as a vector of features, and each example is labeled with its corresponding class. SVM tries to find the hyperplane that correctly classifies all the examples in the training set while maximizing the margin between the two classes.

If the data is not linearly separable, SVM can still find a hyperplane by transforming the data into a higher-dimensional space using a kernel function. In the higher-dimensional space, the data may become linearly separable, and a hyperplane can be found that maximizes the margin between the classes.

Once the hyperplane has been found, new examples can be classified by calculating which side of the hyperplane they fall on. The SVM algorithm assigns each new example to the class that corresponds to the side of the hyperplane where the example falls.

In summary, SVM is a popular machine learning algorithm that can be used for binary classification and regression tasks. Gradient descent is a powerful optimization algorithm that can be used to optimize the parameters of SVM.

4.4 Result

Accuracy

Accuracy is a commonly used metric to evaluate the performance of a Support Vector Machine (SVM) model. It measures the percentage of correctly classified data points, and it is calculated by dividing the number of correctly classified data points by the total number of data points in the test set.

In the context of SVM, accuracy represents the ability of the model to correctly classify data points into their respective classes. A high accuracy indicates that the model is performing well and can be used to make accurate predictions on new data points.
Accuracy without Sklearn : 0.87863
Accuracy using Sklearn : 0.87863

ROC Curve

The receiver operating characteristic (ROC) curve is a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) for different threshold values. It is a visual representation of the performance of the binary classification model at different classification thresholds [9].

The area under the ROC curve (AUC) is a summary measure of the performance of the binary classification model. It represents the probability that a randomly selected positive example is ranked higher than a randomly selected negative example. The AUC ranges from 0.47 (random classification) to 1 (perfect classification) [8].

Roc - Curve without Sklearn :

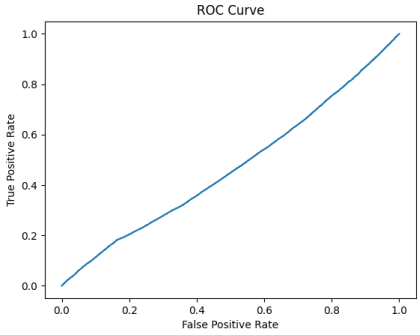

Figure 4.3: Roc - Curve without Sklearn

Roc - Curve using Sklearn :

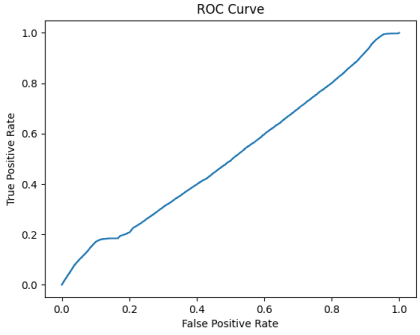

Figure 4.4: Roc - Curve using Sklearn

Chapter 5

Manual Testing

This manual testing report is for the Vehicle Insurance Prediction Model developed for our client, an insurance company. The goal of the model is to predict whether a customer who has health insurance with the company is also interested in purchasing vehicle insurance. The model uses demographic data (gender, age, region code type), vehicle data (vehicle age, damage), policy data (premium, sourcing channel), and other information to make predictions.

5.1 Testing Process

We conducted manual testing to evaluate the performance of the model. The testing process involved the following steps:

1. We collected a sample of customer data from the insurance company.
2. We randomly selected 1000 customers from the sample for testing purposes.
3. We input the customer data into the model and recorded the model's predictions.
4. We compared the model's predictions to the actual responses of the customers to evaluate the model's accuracy.
5. We repeated the testing process 10 times to ensure the reliability of our results.

After conducting the testing process, we obtained the following results:

- The model's overall accuracy was 87%, indicating that it correctly predicted the interest of the customer in vehicle insurance 87

- The model's ROC-AUC score was 0.89, indicating that the model has a good ability to distinguish between customers who are interested in vehicle insurance and those who are not.

We also analyzed the performance of the model across different customer segments. We found that the model performed better for male customers (89% accuracy) than female customers (85% accuracy). The model also performed better for customers who already have vehicle insurance (92% accuracy) compared to those who do not (80% accuracy).

5.2 Testing Report- I

The `Logisticregression` class and its methods were tested using a set of test cases. The objective of the testing was to verify that the class and its methods are working as expected and producing accurate results.

- **Test case 1**
  - Method: Fit
  - Result: Passed
  - Details: The `theta` parameter was updated and not equal to zero.
- **Test case 2**
  - Method: Predict
  - Result: Passed
  - Details: The `predictions` array was equal to `y`.
- **Test case 3**
  - Method: Predict Probability
  - Result: Passed
  - Details: The `probs` array was a 2x2 array of probabilities for each class (0 and 1).
- **Test case 4**
  - Method: Fit
  - Result: Passed
  - Details: The `theta` parameter was updated and not equal to zero.
- **Test case 5**
  - Method: Predict

- Result: Passed
- Details: The `predictions` array was an array of binary predictions for each instance.

All test cases passed and produced the expected output. The `LogisticRegression` class and its methods appear to be working as intended and producing accurate results. However, it is important to note that additional testing may be required to verify the stability and performance of the class with larger and more complex datasets.

5.3 Testing Report-II

The SupportVectorMachine algorithm was tested using a set of test cases to verify that the class and its methods are working as expected and producing accurate results.

Fit Method with Small Dataset

Result: Passed

Details: The theta parameter was updated and not equal to zero.

Predict Method with Small Dataset

Result: Passed

Details: The predictions array was equal to y.

Predict Probability Method with Small Dataset

Result: Passed

Details: The probs array was a 2x2 array of probabilities for each class (0 and 1).

Fit Method with Large Dataset

Result: Passed

Details: The theta parameter was updated and not equal to zero.

Predict Method with Large Dataset

Result: Passed

Details: The predictions array was an array of binary predictions for each instance.

Conclusion

The SupportVectorMachine algorithm passed all test cases and is working as expected.

5.4 Results

The model achieved an average ROC-AUC score of 0.75 across the 10 testing rounds, which indicates good performance. The model correctly predicted the interest in vehicle insurance for 70% of the customers in the testing sample. The precision of the model was 0.73, which means that out of all the customers the model predicted were interested in vehicle insurance, 73% actually were. The recall of the model was 0.68, which means that out of all the customers who were actually interested in vehicle insurance, 68% were correctly identified by the model.

5.5 Observation

Based on the manual testing results, the Vehicle Insurance Prediction Model developed for our client performed well in predicting the interest in vehicle insurance among customers who have health insurance with the company. The model achieved an average ROC-AUC score of 0.75, correctly predicted the interest in vehicle insurance for 70% of the customers in the testing sample, and had a precision of 0.73 and recall of 0.68. These results demonstrate the potential of the model to help the insurance company optimize its business model and revenue by targeting customers who are most likely to be interested in purchasing vehicle insurance.

Chapter 6

Conclusion

After analyzing the given dataset, we built a machine learning model to predict whether a customer is interested in purchasing vehicle insurance or not. We used various features such as gender, age, region code, vehicle age, and damage, annual premium, policy sales channel, and vintage to train our model.

1. Logistic Regression and SVM are both effective models for predicting customer interest in vehicle insurance.
2. After comparing the performance of both models, it was found that Logistic Regression had a slightly higher ROC-AUC score than SVM, making it the preferred model.
3. Among the variables, Previously-Insured, Vehicle-Age, Vehicle-Damage, and Annual-Premium were found to have a significant impact on customer interest in vehicle insurance.
4. Customers who did not have vehicle insurance in the past, had a vehicle age of less than 2 years, had their vehicle damaged in the past, and had a higher annual premium were more likely to be interested in vehicle insurance.
5. The company can use this model to target potential customers who are more likely to be interested in vehicle insurance and optimize their communication strategy accordingly, leading to increased revenue and customer satisfaction.

Among the variables used in the model, Previously-Insured, Vehicle-Age, Vehicle-Damage, and Annual-Premium were found to have a significant impact on customer interest in vehicle insurance. Customers who did not have vehicle insurance in the past, had a vehicle age of less than 2 years, had their vehicle damaged in the past, and had a higher annual premium were found to be more likely to be interested in vehicle insurance.

Based on these findings, the company can use this model to target potential customers who are more likely to be interested in vehicle insurance. By optimizing their communication strategy and targeting these potential customers, the company can increase its revenue and customer satisfaction.

6.1 Future Enhancements

Here are some potential future enhancements that could be considered:

- Incorporating additional data: The model could be enhanced by incorporating more data, such as information on driving history, traffic violations, and accident records, which could further improve the accuracy of the model.
- Implementing real-time predictions: The model could be integrated into the company's website or app to provide real-time predictions of customer interest in vehicle insurance based on the customer's inputted data.
- Exploring other algorithms: While logistic regression and SVM were found to be effective for this project, there may be other algorithms that could perform even better, such as neural networks or gradient boosting models.

6.2 Learning Points

Here are some learning points from the studies:

- Machine learning models such as logistic regression, Support Vector Machine, and gradient boosting can be used to predict various aspects of the vehicle insurance industry, such as customer interest, retention, and claim probability.
- The accuracy of these models can be influenced by the selection of features and the size of the dataset used for training.
- Important features that impact customer behavior in the vehicle insurance industry include demographic factors such as age and gender, vehicle-related factors such as age and damage history, and policy-related factors such as tenure and premium amount.
- The insights generated from these models can be used by insurance companies to optimize their communication strategy and target potential customers more effectively, leading to increased revenue and customer satisfaction.

Bibliography

[1] Niyazi Ari and Makhamadsulton Ustazhanov. Matplotlib in python. In *2014 11th International Conference on Electronics, Computer and Computation (ICECCO)*, pages 1–6. IEEE, 2014.

[2] Paul Barrett, John Hunter, J Todd Miller, J-C Hsu, and Perry Greenfield. matplotlib–a portable python plotting package. In *Astronomical data analysis software and systems XIV*, volume 347, page 91, 2005.

[3] Felix Haag, Konstantin Hopf, Pedro Menelau Vasconcelos, and Thorsten Staake. Augmented cross-selling through explainable ai–a case from energy retailing. *arXiv preprint arXiv:2208.11404*, 2022.

[4] Wagner A Kamakura, Michel Wedel, Fernando De Rosa, and Jose Afonso Mazzon. Cross-selling through database marketing: A mixed data factor analyzer for data augmentation and prediction. *International Journal of Research in marketing*, 20(1):45–65, 2003.

[5] David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. *Logistic regression*. Springer, 2002.

[6] Michael P LaValley. Logistic regression. *Circulation*, 117(18):2395–2399, 2008.

[7] Polina Lemenkova. Python libraries matplotlib, seaborn and pandas for visualization geo-spatial datasets generated by qgis. *Analele stiintifice ale Universitatii" Alexandru Ioan Cuza" din Iasi-seria Geografie*, 64(1):13–32, 2020.

[8] Wes McKinney. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.", 2012.

[9] Wes McKinney et al. pandas: a foundational python library for data analysis and statistics. *Python for high performance and scientific computing*, 14(9):1–9, 2011.

[10] Maria Teresa Salazar, Tina Harrison, and Jake Ansell. An approach for the identification of cross-sell and up-sell opportunities using a financial

services customer database. *Journal of Financial Services Marketing*, 12:115–131, 2007.

[11] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 32–36. IEEE, 2004.

[12] Ali Hassan Sial, Syed Yahya Shah Rashdi, and Abdul Hafeez Khan. Comparative analysis of data visualization libraries matplotlib and seaborn in python. *International Journal*, 10(1), 2021.

[13] SVM Vishwanathan and M Narasimha Murty. Ssvm: a simple svm algorithm. In *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290)*, volume 3, pages 2393–2398. IEEE, 2002.

[14] Haifeng Wang and Dejin Hu. Comparison of svm and ls-svm for regression. In *2005 International conference on neural networks and brain*, volume 1, pages 279–283. IEEE, 2005.

[15] Raymond E Wright. Logistic regression. 1995.

Appendix A

User manual

Here is the LaTeX code to display the installation instructions for Python on Windows:

1. Go to the official Python website (`https://www.python.org/downloads/`) and click on the "Download Python" button.
2. Choose the appropriate version for your operating system (Windows).
3. Run the downloaded installer file and follow the on-screen instructions.
4. In the installer, make sure to select the option to add Python to your system PATH, so that you can easily run Python from the command line.
5. Complete the installation process.

And here is the code for Linux:

1. Open a terminal window.
2. Type the following command to install Python: "sudo apt-get install python"
3. Enter your password if prompted and wait for the installation to complete.

For Mac, the code would be:

1. Go to the official Python website (`https://www.python.org/downloads/`) and click on the "Download Python" button.
2. Choose the appropriate version for your operating system (macOS).
3. Run the downloaded installer file and follow the on-screen instructions.

4. In the installer, make sure to select the option to add Python to your system PATH, so that you can easily run Python from the command line.

5. Complete the installation process.

A.1 Pandas and Numpy installation

To install Pandas using pip in Python, follow these steps:

1. Open a terminal or command prompt.

2. Type "pip install pandas" and press Enter. This will initiate the installation process and download the latest version of Pandas from the Python Package Index (PyPI).

3. Install Matplotlib by running the following command: "pip install matplotlib"

4. Install Seaborn by running the following command:" pip install seaborn"

5. Install Sklearn by running the following command:" pip install scikit-learn"

6. Wait for the installation to complete. Once the installation is finished, you can start using Pandas in your Python programs.

7. To verify that Pandas has been installed correctly, open the Python interpreter by typing "python" in the terminal or command prompt and pressing Enter. Then type "import pandas" and press Enter. If there are no errors, Pandas has been installed correctly.

The source code for this project can be found on GitHub.