Major Project
On
Statistical Analysis of Vehicle Insurance Policy

Submitted to the Amity University Uttar Pradesh In partial fulfillment of requirements for the award of the Degree of

MASTER OF STATISTICS

By

Arpit Saxena
Enrollment No: A4479222033

Under the Supervision of:

Supervisor

Dr. Bavita Singh

Department of Statistics
Amity Institute of Applied Science

Amity Institute of Applied Sciences, Amity University Uttar Pradesh, Sector 125, Noida – 201303 (India)

AMITY UNIVERSITY

UTTAR PRADESH

AMITY INSTITUTE OF APPLIED SCIENCES

Synopsis of Major Project:

Title: Statistical Analysis of Vehicle Insurance Policy

Name of Guide: Dr. Bavita Singh

Programme: - M.Stat		Year/Semester: - 4th Semester	
S.No.	**Enrollment No.**	**Name**	**Signature**
1	**A4479222033**	**Arpit Saxena**	

Summary:- Different Machine Learning Models has been fitted in order to identify which model best showcases whether a customer having health insurance is also interested in purchasing vehicle insurance.

Schedule of work completion:- 2nd January 2024 – 10th May 2024

Signature of Student

Signature of Guide

Signature of Programme Leader

Approval by Board of Faculty

Member	Signature	Remark (Approved / Not Approved)

DECLARATION

I, Arpit Saxena, student of Master Of Statistics hereby declare that the Major Project project titled “Statistical Analysis of Vehicle Insurance Policy” which is submitted by me to Department of Statistics, Amity Institute of Applied Sciences, Amity University, Uttar Pradesh, Noida, in partial fulfillment of requirement for the award of the degree of Master Of Statistics, has not been previously formed the basis for the award of any degree, diploma or other similar title or recognition.

Noida

Date: 10 May 2024

Arpit Saxena

CERTIFICATE

On the basis of declaration submitted by Arpit Saxena ,student of Master Of Statistics, I hereby certify that the Major Project titled "Statistical Analysis of Vehicle Insurance Policy" which is submitted to Department of Statistics, Amity Institute of Applied Sciences, Amity University, Uttar Pradesh, Noida, in partial fulfillment of requirement for the award of the degree of Master Of Statistics, is an original contribution with existing knowledge and faithful record of work carried out by him under my guidance and supervision.

To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Noida

Date: 10 May 2024

Dr. Bavita Singh

Department of Statistics
Amity Institute of Applied Sciences
Amity University, Uttar Pradesh, Noida

ACKNOWLEDGEMENT

I, Arpit Saxena, would like to express my deep sense of gratitude towards my faculty guide Dr. Bavita Singh, for guiding me from the inception till the completion of the project. The experience of working under my faculty guide has been a value addition to the learning during my course of Master of Statistics. I would like to express my heartfelt gratitude for Dr. Bavita Singh with whose guidance and valuable suggestions I was able to maximize on the learning curve during the completion of my project. Her timely responses to any issues that came along, and her promptness helped me to successfully complete the project.

I am thankful to her that she provided me an opportunity to work under her guidance and sharing her valuable experience.

Also, I am thankful to Amity Institute of Applied Science (AIAS) and Department of Statistics for providing me with such an opportunity to gain analytical knowledge and skills, imparted as a part of curriculum.

Arpit Saxena

ABSTRACT

This study shows the comparison between different Machine Learning Models to determine if a person with health insurance will prefer automobile insurance. In the insurance sector, Accurate prediction models are used for risk assessment, policy pricing and decision-making due to complexity and volume of insurance data that is expanding. Using Health Insurance dataset, the study shows the prediction abilities of Decision trees, Logistic Regression and Random Forest Models. The dataset consists of numerous demographic and categorical variables for which missing values are being handled and then categorical variables are converted into numerical values during pre-processing stage. After pre-processing stage, the model is being used for training using decision trees, logistic regression and random forest model for which the model’s accuracy is being estimated. For improving model’s accuracy, hyper tuning is done to find the best estimators for the model which are being applied to increase its accuracy. Then the classification report is being made to determine the precision score, recall score and f1 score of the model. Confusion matrix has also been made for the models. The project's results will tell which model is more accurate and predict who would purchase vehicle insurance for the business, among other things. The comparison analysis will show the benefits and drawbacks of each model, empowering insurers to choose the algorithm that best suits their unique requirements.

CONTENTS

I. Introduction

The practice of insurance companies which offers additional insurance products to its existing customers is referred to as Health Insurance cross-selling which plays major role in the growth of a business and profitability of insurance companies. The aim of this research is to develop a predictive model which can help us to determine if a person is interested in purchasing vehicle insurance. Various factors like demographics, socio-economic factors and demographics have been analysed to identify patterns which can help for prediction of interest of a customer to purchase the vehicle insurance.

This research holds an important purpose in strengthening the marketing strategies of health insurance companies. Insurers can identify the interest of their customers more accurately by predicting the interest of vehicle insurance which can further help in improving customer acquisition and retention rates which will result in increasing business revenue.

Also, the key findings of this study hold valuable resource especially for health insurance providers to increase their customers by analysing customer behaviour patterns which will determine when the customer will be interested in purchasing the vehicle insurance. With the help of this deep understanding of this behaviour, they can make data-driven strategies in order to adapt product offerings, enhance marketing strategies and create customized services that helps in meeting the needs and preferences of their customers which also helps in boosting satisfaction of customer as well as build strong client relationships which will further contributes towards the growth and sustainability of the insurance industry.

Key Components:

Data Analysis and Integration: In order to obtain important insights regarding risk patterns, claim frequencies, and severity distributions, the vast amounts of historical data needed to be analysed and then combine different data sources. Key risk indicators will be uncovered and reliable predictive models will be created with the help of data-driven methodology.

Risk segmentation and profiling: The policyholders will be divided into homogeneous risk groups according to pertinent characteristics including past claims history, demographics and lifestyle with the help of sophisticated statistical methods. By this fine-grained segmentation, it will be easier for creating customised pricing schemes and risk profiles.

Pricing optimisation: Pricing models, which reflects the risk involved in each group of policyholders appropriately, will be created through the framework for risk segmentation as a foundation. These models will be used to take into account variables which included anticipated claim amounts, expense considerations and the chance of loss so that competitive and profitable premium levels can be determined.

Regulatory Compliance: Our project will include regulatory requirements so that the established pricing models are compliant with pertinent laws, rules and standards.

It is necessary to analyse and update pricing models continuously due to the ongoing shift in the insurance markets. A structure which can routinely assess model performance, monitor the new risk trends and can incorporate fresh data will be set so that the dependability and precision of the pricing models can be improved.

Therefore, the project's results will be considered highly advantageous to the insurance firms and clients. Through enhancement of risk assessment accuracy and pricing fairness, the insurers can achieve better alignment between premiums and risk exposure which would further increase financial stability and profitability. Additionally, in order to improve affordability and satisfaction, the policyholder will gain from more specialised insurance products with rates which can take into account their unique risk profiles.

With the growing availability of data and development in analytics, the risk assessment and pricing procedures which are used by the insurance business have the potential to be completely transformed. We seek to lay groundwork for insurers for optimisation of pricing strategies and promotion of just and open insurance markets by combining data-driven insights, compliance consideration and cutting-edge modelling approaches.

A. Overview of Insurance and the techniques of data science in insurance industry:

The insurance sector is very crucial for offering both consumers and corporations financial security and also for risk management. It often covers industries which includes life, property, health, casualty and other insurances. Large volumes of information regarding claims, policyholders, market trends, risk factors and external variables are gathered by insurance companies. In order to improve their operations and decision-making process, substantial number of opportunities exist for insurers which results in expanding availability of this data and development in data science.

Data Science is referred to the practice of mining data to gather information and insights using scientific algorithms, techniques and tools which is important in the insurance sector due to having potential to solve many major problems which often produce beneficial results that is why data science is important for many reasons.

Risk Assessment: The risk analysis is often useful to the insurance industry since it can help to analyse large amounts of data, highlight pertinent risk indicators, and further useful for development of predictive models which allow for more accurate risk assessment which improves pricing plans, profitability and underwriting judgements.

Fraud detection: Insurance fraud is a major concern as it causes loss to insurers financially. The data science approaches such as pattern recognition, anomaly detection and predictive modelling can be useful for spot fraudulent claims and enhancement of fraud detection and prevention systems.

Customer segmentation and personalization: Data Science plays a huge role for insurers by segmenting clientele into several groups by taking account traits like demographics, behaviours or risk profiles which is helpful for insurers since it can offer targeted marketing techniques, adjusted pricing and personalised insurance policies for increment of customer satisfaction and retention by knowing consumer categories.

B. Types of Insurance

i. Life Insurance:

A contract formed between an insurance company and a policy holder on a condition that the insurer agrees to pay money to one or more beneficiaries in exchange of the premium which is paid by the insured person before he dies. The companies having the best life insurance contains good financial strength, high customer satisfaction, many policy types and low customer complaints.

Types of Life Insurance: There are many types of life insurances such as Term Life Insurance, Whole Life Insurance, Universal Life Insurance, Variable Life Insurance and many more having unique characteristics, advantages and premium arrangements in each type.

Factors Affecting Life Insurance Premiums and Costs: There are many factors which can affect the cost of life insurance premiums such as age, health status, smoking status, gender, occupation and lifestyle, family medical history, etc.

Components of Life Insurance: There are two components of Life Insurance which are death benefit and premium while permanent or whole life insurance also has a component namely cash value.

Death Benefit: It is termed as the amount of money which is guaranteed by the insurance company to the beneficiaries which can be identified in the policy when the insured person dies. The insured person might be parent while beneficiaries might be their children. On the basis of the beneficiaries estimated future needs, the insured person chooses the desired death benefit amount. The insurance company determines if there is any insurable interest and if the proposed insured person will qualify for the coverage based on the underwriting requirements of the company which is related to health, age and many more.

Premium: These are the money which is paid by the policyholder for the insurance. If the policyholder pays the required premiums, which are estimated by how likely the insurer will have to pay policy's death benefit on the basis of the life expectancy of the insured person, then the insurer should pay the death benefit when the insured person dies. There are various factors which can influence the life expectancy such as the insured person's age, health, gender, medical history, occupation and lifestyle, and many more.

Cash Value: There are purposes of the cash value of permanent life insurance. It is referred to as the savings account which is used by the policyholder during the insured person's life. Depending upon the use of the money, some policies consist of the restrictions on withdrawals.

ii. Health Insurance:

It is defined as a type of insurance which covers either whole or some part of the risk including medical expenses. Through estimation of overall risk of health risk, an insurer can plan and develop a finance structure such as monthly premium in order to provide money which can paid for the health care benefits by which risk is shared among many individuals which is specified in the insurance agreement. It is also referred to a coverage which provides benefits through payments to an injury or sickness which also includes coverage of losses from medical expense, accidents, disability and dismemberment.

However, the obligations of an insured person can occur in several forms such as:

- **Premium:** It is the amount that the policyholder should pay for their health coverage to the health plan so that their expenses can be paid in case of any medical situation or emergency. According to healthcare law, the premium is estimated by taking some factors into consideration which includes age, location, use of tobacco, enrolment of an individual or a family and also the type of plan, insurer choses. Tax credit is paid by the government in order to cover part of the premium for the persons who purchases private insurance through insurance marketplace under the Affordable Care Act.
- **Deductible:** Before the insurer pay its share for their health care, the insured should pay an amount which is out of pocket is termed as deductible. For instance, if a policy holder has to pay $5000 deductible per year before the health insurer pays for their health care then it might take several doctor's visit before the insured person reaches the deductible and payment is proceeded by the insurance company. Moreover, co-pays for doctor's visits against the deductible are not applied to many policies.
- **Co-payment:** Before the health insurer pays for a particular visit, the insured person has to pay an out-of-pocket amount which is termed as co-payment.
- **Coinsurance:** It is referred to as a percentage of the total cost which an insurance person should also pay instead of just paying a fixed amount of co-payment.
- **Exclusions:** Some billed items such as use-and-throw, taxes, etc. are the ones which gets excluded from admissible claim. So, all the services are not covered since the insurer are expected to pay for the non-covered services on their own.
- **Coverage Limits:** There is a limit that some policies related to health insurance pay for the health care up to a certain amount which are coverage limits. The insured person should pay any charge which arises after the health plan reaches the benefit maximum since it will stop further payment and the policy-holder will have to pay for the remaining costs.
- **Out-of-pocket maximum:** The insured person reaches the out-of-pocket maximum when their payment obligation ends and health insurance pays all other covered costs. Out-of-pocket maximum is generally limited to some specific benefits or can be applied to all the coverages which are provided during a specific benefit year.

iii. Non-Life Insurance

It is often known as the general insurance which is also a category of insurance which protects people, companies as well as organisations from many risks and losses other than the ones which are related to health or lives. Non-life Insurance helps in covering many assets, liabilities and property which are useful for offering protection against disease or death. It has many important features which are as follows:

- **Protection Types:** Risks such as theft, accident, liability claims, natural disasters, property damage, and other events which occurs in the life of an insured are being protected by the non-life insurance. Some types of non-life insurance are home insurance, liability insurance, auto insurance, travel insurance, marine insurance and commercial property insurance.
- **Premiums:** The insurance provider is paid the required premium by the policyholder for the purpose of keeping their non-life insurance active. There are only a few factors which affect premium costs such as risk factors, claim history of the policyholder, kind of coverage, insured sum and deductible option.
- **Policy Limits:** There is a limit known as coverage limit which highlights the maximum sum that will be covered by the insurance company for the insured person. If the amount reaches the coverage limits, then the insurance company will stop covering for further payments and the insured person has to pay for the remaining losses. Depending on the type of coverage and terms of the policy, the policy for the coverage limits might change but it is advisable for the insured person to carefully evaluate the limits for the future.
- **Deductibles:** This is the sum which is not covered by the insurance company and the policyholder has to pay the amount on his own and are included in non-life insurance in which reduced deductibles leads to higher premiums while bigger deductibles lead to reduced premiums.
- **Claim Procedure:** In the claiming procedure, the policyholders have to submit a claim to their insurance provider in covered loss. They must report the loss, provide the required documents and information and collaborate with the insurance provider for evaluation of the severity of the damage and then establish amount which has to be claimed. There are particular steps and deadlines which insurance companies follows when they process claims.

- **Benefits:** There are several benefits in a non-life insurance policy which includes the financial help which is provided at any medical emergency, compensation is paid by the third party in case of damages on property or life, etc. It can cover damages of the residential property such as fire, natural calamities, riots as well as burglary. It can cover insurance coverage of senior citizens as well as children along with issues such as accidents, loss of any documents and baggage, etc in a foreign land. It also benefits the businesses with policies such as shopkeepers' insurance, property and marine insurance, benefits insurance, etc.

C. IRDAI

IRDAI stands for Insurance Regulatory and Development Authority of India which is a regulatory agency that is responsible for monitoring and controlling the Indian insurance market which was mainly created in accordance with the Insurance Regulatory and Development Authority Act of 1999 for defending the interests of policyholders and encouraging of the expansion and development of the nation's insurance sector. IRDAI's responsibilities can b described in the insurance industry:

Control and Regulation: IRDAI is termed as the principal regulating body for the insurance industry in India. The regulations, standards and the policies that are developed and enforced by IRDAI governs the conduct and operations of insurance companies and other market participants. The authority ensures adherence to these rules for the purpose of keeping the insurance market transparent, equitable and stable.

Registration and licensing: The registering and licensing of insurance companies, agents, surveyors, brokers and other businesses which are involved in the insurance industry are the responsibilities of IRDAI. It also plays a key role in establishment of eligibility standards, focus on the insurance businesses' solvency and soundness, and ensures that only accredited and qualified parties carry out operations which are related to insurance.

Protection of Policyholder Interests: The IRDAI is essential for protecting the interests of the Policyholder. It develops standards for disclosure norms, ethical behaviour and consumer protection measures for the purpose of ensuring insurance products being available, open and advantageous to policyholders. Complaints are processed by IRDAI from the policyholders and forum is also offered to resolve them using integrated grievance management system.

Market Development: The expansion and development of the Indian insurance market as well as the rivalry, industry innovation and product variety has been promoted by the IRDAI and safeguarding the interests of policyholders. IRDAI works for stabilising the market and also monitor market trends so that an appropriate action can be taken which also encourages healthy competition and long-term growth in the insurance industry.

Financial Regulation: The financial operations are controlled by IRDAI to safeguard the solvency and financial stability of insurance businesses. It also established investment rules, capital requirements and risk management for insurers for keeping the insurance sector financially stable. It is also useful in evaluating the insurance industry's financial standing, performing audits and keeping track of their adherence to accounting rules.

Market Conduct: IRDAI focuses on the behaviour of insurance companies, intermediaries and agents for ensuring the ethical practises and equitable treatment of policyholders. It also develops standards for sales practises, regulations and code of conduct for eliminating the fraud and unfair practices in the insurance market. If the organisation fails to follow these rules, then investigation, inspections and sanctions are being conducted by IRDAI against that organisation.

For promotion of financial inclusion, IRDAI encourages insurers for providing affordable and accessible insurance products to the underprivileged segments of the society. It encourages micro-insurance projects and also implements the government-backed insurance systems for social welfare which also makes easier for expansion of insurance services in rural and isolated areas.

International partnership: IRDAI is active in participating in international cooperation partnership with other regulatory bodies and organisations for sharing expertise, advance cross-border regulatory harmonisation, and stay current on worldwide insurance practices. It also takes part in international forums, seminars and projects for supporting the expansion and growth of the insurance sector worldwide.

D. Pricing – Health Insurance Product

Estimation of premium costs is important for the consumer of a health insurance plan which would incur in return for coverage. The adequacy of the premiums is tested in order to cover the risks and projected expenditures for delivering services related to healthcare by taking number of variables into account for the pricing procedure. There are several key factors which determine the cost of health insurance plan that are as follows:

Claim Experience and Historical Data: The claims experience and historical data are analysed by insurers so that the cost patterns and trends can be comprehended. The claim sums, claim occurrence rates and the cost which is distributed across various demographics and the medical conditions are being examined for this process. In order to determine predicted costs, and establish suitable premium rates, the historical data has been useful for it.

Medical Cost Inflation: The Health Insurance costs depends on the cost of medical services, treatments and prescriptions medications. Insurers takes estimated rates of inflation in medical costs into account to ensure premiums are high so that the anticipated costs are paid during the policy period. Healthcare market dynamics, economic variables and governmental regulations are taken into account when future cost trends have to be projected.

Regulatory Requirements: In order to ensure fairness and consumer protection, the health insurance pricing is subject to regulatory oversight. The insurance regulators might also impose restrictions or guidelines on profit margins, rate filling processes and premium rates. While pricing their products, insurers should comply with these regulations.

Competitive Market Analysis: When premium rates are set, then the competitive landscape and market dynamics are taken into account by the insurers who also evaluate the pricing strategies of competitors, market demand, and consumer preferences. The insurers also balance the competitive pricing with adequate coverage and profitability.

For building price models and forecast future claim costs such as credibility theory, loss reserving and statistical modelling, actuarial methods are being used which also assist insurers to make adjustments for future uncertainty along with projection of the anticipated claims experience on the basis of the historical data.

II. Literature Review

This study has been made using existing research about the prediction of interest of a customer purchasing vehicle insurance with the help of machine learning models in the insurance industry. By identifying the interests of a person for vehicle insurance and gain a deep understanding of the pattern of the customer’s purchase, insurers can enhance their marketing strategies and enhance their business revenue. The insurance market has been highly competitive so accurate prediction has become crucial for the insurers for the growth of the industry so with the arrival of predictive analytics and machine learning models, insurers now have the opportunity to make data-driven decisions with the predictive analytics which helps in enhancing their business. Furthermore, with the help of machine learning algorithms, the insurance company has now got the opportunity to analyse large health insurance customer datasets and identify more patterns and insights which may not be possible through traditional methods.

A. Predictive Modelling Techniques

Many studies have explored the different predictive modelling techniques and determine its effectiveness to identify its potential in prediction. An Algorithm for gradient boosting has been made into use to predict the customers who wanted to purchase vehicle insurance based on their historical health claims and demographics. The results showed that the ensemble methods performs better than traditional Decision Tree models in terms of accuracy so making use of these ensemble methods helps us in improving the accuracy of the model for prediction. The models demonstrated the prediction by identifying potential customers who are interested in purchasing vehicle insurance by taking customer attributes and historical data into consideration. Similarly, ensemble methods are being applied to Random Forest Model and Logistic Model as well to improve their performance in prediction. The comparison drawn between the models helps in highlighting the importance of feature engineering and model selection in prediction more accurately.

B. Imbalance Data Handling

Imbalance data is a very common challenge in any analysis where some instances of a class or classes outweigh the other classes. When the distribution of classes in a dataset is highly skewed on class have more instances than the other class then the dataset is said to be imbalanced. Due to lack of adequate training samples, it becomes difficult for the classifier to identify patterns and predict unbiased outcomes for the minority class. This problem is resolved using various techniques such as oversampling which is used to increase the number of instances in the minority class. Different evaluation metrics are also used to check the class imbalance. Their findings help in handling imbalanced data to obtain unbiased predictions.

C. Customer Segmentation

Customer Segmentation is a way to enhance the effectiveness of the models in which customers are segmented based on their characteristics and behaviours. Using unsupervised clustering techniques to separate customers with similar profiles can give better predictions of the models with more accuracy by identifying distinct customer segments which enables more targeted strategies and then they built predictive models separately for each cluster which leads to more accurate predictions by capturing distinct purchasing behaviours with different customer segments.

D. Customer Satisfaction

Customer Satisfaction is a critical factor in the analysis since satisfied customers would more likely to purchase insurance later and also recommend others leading to more growth of the company. In the study, sentiment analysis can be used to consider customer reviews and feedback to identify customer satisfaction levels. Machine Learning Techniques were used further to identify the patterns in customer behaviour and their satisfaction levels.

The usage of prediction in health insurance has been advanced significantly through various machine learning techniques. Researchers have highlighted challenges like imbalanced data, customer segmentation and satisfaction to improve the accuracy and effectiveness of the models which results in growth and profitability of the health insurance companies.

The study is based on the existing knowledge to contribute more significantly in prediction by not only highlighting the challenges faced during prediction but to also provide valuable insights and practical solutions to the challenges to overcome it which will lead to enhancing the accuracy of the model which will further improve the growth of the insurance company.

E. Performance Evaluation Matrix

i. Accuracy:

It refers to the most basic and intuitive metric which represents the correctly predicted instances out of the total. It is not suitable for the imbalanced datasets since it can lead to misleading results so it is suitable mainly for balanced datasets.

ii. **Precision:**

It is used for measuring the proportion of correctly predicted positive instances which are predicted as positive. It is mainly concerned with the accuracy of positive predictions and is mainly useful when the cost of false positives is high. It is calculated as true positives divided by the sum of true positives and false positives.

iii. **Recall:**

It is used for calculating the proportion of correctly predicted positive instances out of the actual positive instances. It measures the model's ability to identify all the positive instances and is valuable when the cost of false negatives is high. Recall is calculated as true positives divided by the sum of true positives and false negatives.

iv. **F1 Score:**

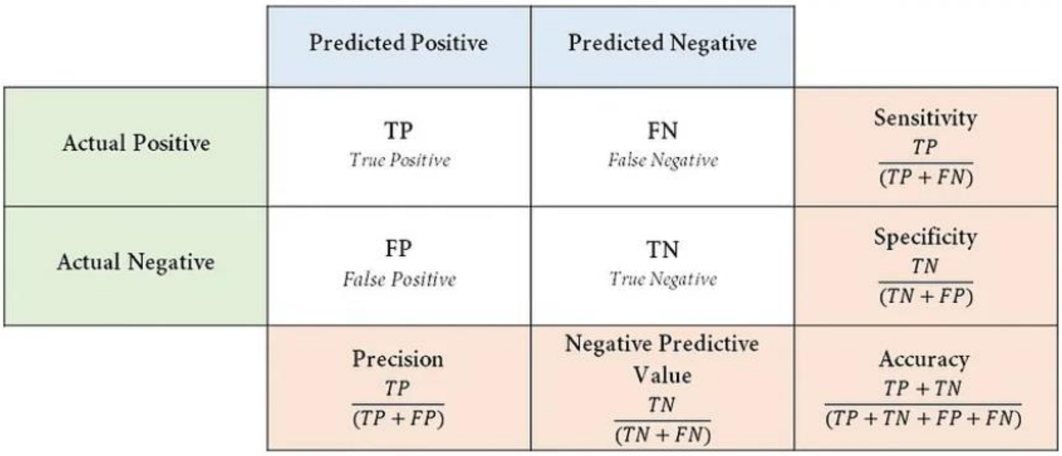
The F1 score is the harmonic mean of precision and recall. It provides a balanced measure that combines both metrics and is useful when there is an imbalance between precision and recall. The F1 score considers both false positives and false negatives and is calculated as $\frac{2 * (\text{precision} * \text{recall})}{(\text{precision} + \text{recall})}$.

v. **Specificity (True Negative Rate):**

Specificity calculates the proportion of correctly predicted negative instances out of the actual negative instances. It focuses on the accuracy of negative predictions and is useful when the cost of false negatives is high. Specificity is calculated as true negatives divided by the sum of true negatives and false positives.

vi. **Confusion Matrix:**

The general idea is that it counts the number of times the classifier identifies the positive class as negative and vice versa. In this matrix the row represents the actual class and the column represents the predicted one. This metric is used to compare the number of predictions for each class that are incorrect and those that are correct. Confusion matrix is shown below:



	Predicted Positive	Predicted Negative	
Actual Positive	TP *True Positive*	FN *False Negative*	Sensitivity $\frac{TP}{(TP+FN)}$
Actual Negative	FP *False Positive*	TN *True Negative*	Specificity $\frac{TN}{(TN+FP)}$
	Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

Figure 1: Confusion Matrix

vii. Machine Learning Models

a. Logistic Regression:

Logistic Regression is termed as a statistical technique which is used for estimation of likelihood that an event will occur while Regression analysis which is termed as a class of statistical techniques that is used to represent the relationship between a dependent and one or more independent variable which is what this type of study is.

In this project, the dependent variable is a categorical variable having only two possible values namely ‘yes’ or ‘no’, ‘true’ or ‘false’, or ‘male’ or ‘female’. Both continuous or categorical variables can be used as independent variable. The connection between dependent and independent variable is modelled using logistic function which has a S-shaped curve and is also a sigmoid function. The probability is shown in a curve in the given figure below. The Maximum Likelihood Estimation (MLE) technique is useful for estimation of logistic regression model and the values of the model parameters which maximise the likelihood of the observed data are found with the help of MLE approach.

After estimation of the logistic model, it is useful for forecasting the likelihood that additional data points will contain the dependent variable. With the usage of anticipated probability, a person can take a decision of whether a particular event is likely to happen or not.

The logistic function is given by: F(x)= $\frac{1}{e^{-(\beta_0+\beta_1 x).}}$. The graph of sigmoid function is given below:

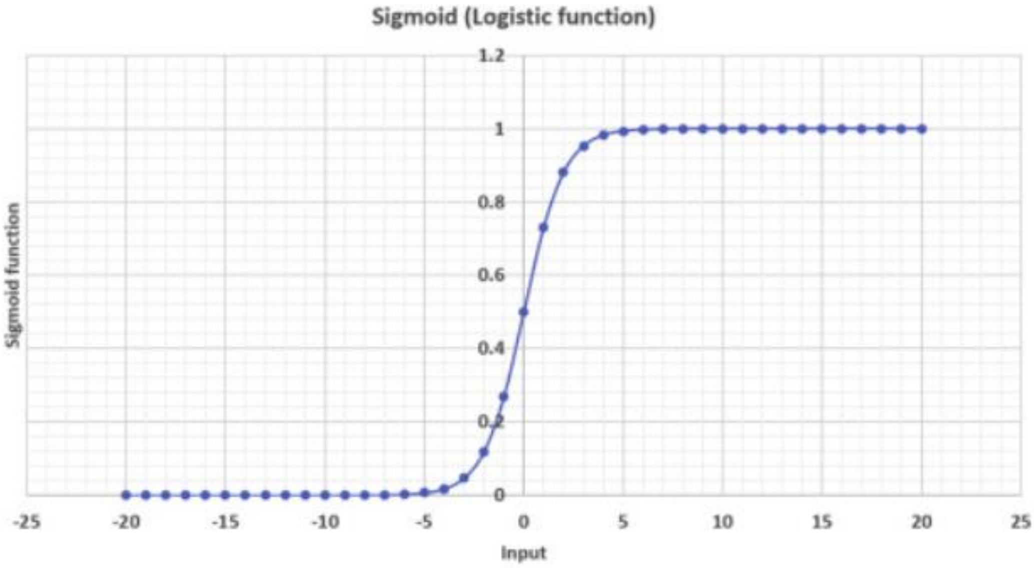


Figure 2: Sigmoid Function in Logistic Regression

b. Decision Tree Model

A supervised learning algorithm that is non-parametric and is used for both regression and classification tasks is termed as a decision tree which has a hierarchical tree structure consisting of an internal node, root nodes, branches and leaf nodes.

In the figure given below, it begins with a root node which do not contain any incoming branches but its outgoing branches feed into internal nodes which is also known as decision nodes. On the basis of the features available, both types of nodes are used for conducting the evaluations so that homogeneous subsets can be formed which are often denoted by leaf nodes or terminal nodes. All the possible outcomes within the dataset are represented by the leaf nodes.

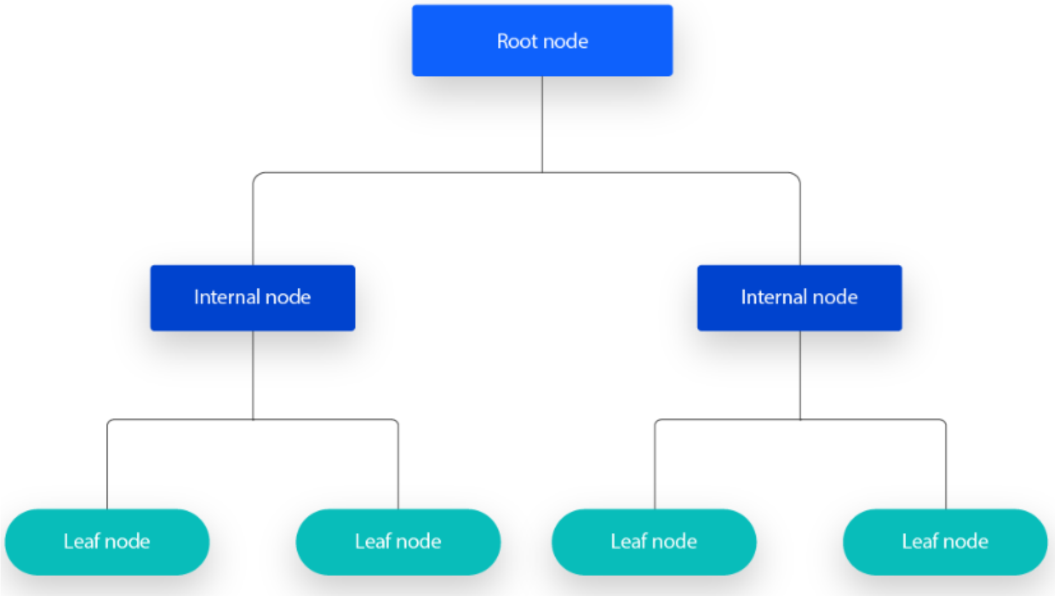


Figure 3: Decision Tree

Since Decision tree classifiers have the capacity to handle numerical and categorical data as well as recording the intricate decision rules along with its interpretability, they are considered as a common machine learning technique which are used in applications of health insurance. How special decision trees are and how ideal they are for health insurance will be discovered in this article.

Decision trees provides a simple and understandable illustration of the decision-making process. The decision tree structure has been useful for simplicity of all parties involved such as policyholders and insurance experts in order to comprehend the variables which affect the outcomes of health insurance. Decision rules in the decision tree provide the trust and transparency which are simple to understand and communicate.

The datasets on health insurance includes both numerical and categorical variables. Decision trees can handle both types of data without the need for specific feature engineering and encoding. The algorithm naturally accommodates them by dividing data into groups according to the different values of the categorical variables. In order to partition the data for numerical variables, decision trees can be used to determine the most informative thresholds.

Decision trees captures the non-linear relationships between features and the goal variable. Numerous variables which frequently interacts in complicated ways have an impact on health insurance outcomes. Due to their capacity to capture and model such interactions, decision trees are often well suitable for spotting complex patterns and hidden links in the data.

Decision trees provided a measure of feature importance which helps in quantifying the relative value of various factors in forecasting outcomes that is related to health insurance. The important variables can be identified which affect decisions about claim submission, risk assessment, policy renewal, premium calculation and fraud detection that can be made with the help of this information. The relevance of a feature can help in aiding insurers in setting the priorities and also efficiently allocation of resources.

However, health insurance datasets contain some missing values which can affect the performance of some algorithms. So, selection of most useful features and routes which is based on the data at hand, decision trees can effectively handle the missing data. This trait is especially useful when we have to work with insufficient or imperfectly gathered health insurance data.

In order to enhance the decision tree's prediction ability, ensemble techniques can be used for the same such as Random Forest and Gradient Boosting. On combining the recommendations of several trees, overfitting is decreased in ensembles, resilience is improved, and predictions are produced which are more accurate. This method is helpful for health insurance programmes which strive for improved robustness and accuracy in outcome prediction.

Decision tree methods can handle large datasets with many of attributes and instances effectively especially optimised version such as Classification and Regression Trees (CART) technique. For health insurance initiatives which need processing large amounts of data from policyholders, medical records, claims and demographic data, scalability is essential.

c. Random Forest Model

Random Forest is referred to as the composition of decision tree classifiers. An ensemble learning algorithm which combines multiple decision trees to make predictions are called as Random Forest Classifier which is widely used in many domains including health insurance projects. With the help of distinct subset of the data, each tree in a Random Forest ensemble is trained. Bagging which is also called as bootstrap aggregating is the term which is mainly used to describe this procedure.

Each tree in a decision tree learns a random sample of the training data and employs a random selection of features for splitting so that overfitting can be decreased and generalisations can be increased.

Random Forest adds another level of randomization by selection of subset of features for each tree at random which makes the tree more coherent and make sures that various feature sets are taken into account. So, in this way, a great variety of feature interactions are captured by Random Forest as well as overall prediction power has also been boosted.

Majority Voting (classification) or average (regression) of the individual tree predictions has been used for the final forecast. Each tree separately provides a prediction in the Random Forest during prediction and therefore, the final predictions are more accurate and stable due to this voting system.

Random Forest is robust to overfitting since the ensemble of trees decrease the influence of a single noisy or biased tree. To reduce the possibility or overfitting and enhance the performance of generalisation, predictions from various trees are combined.

Random Forest can properly handle the high-dimensional data with several attributes. For decision-making, it can automatically determine the most instructive features as random subsets of features at each tree are taken into account. Projects which involve health insurance industry requires large and complicated datasets with numerous attributes.

Random Forest shows the measure of feature importance with the help of proportional relevance of each feature in formulation of predictions. It is possible to highlight the variables which have more impact on health insurance results such as policy renewal, risk assessment, premium calculation and claim filing with the aid of this data. Feature importance can influence the choice of features, data pre-processing and decision-making procedures.

Health insurance datasets contains the class imbalance or underrepresentation of some classes. During training of data, minority classes are given more weight to effectively manage the skewed data in Random Forest which, in turn, is able to identify trends and predict outcomes for both minority and majority classes with more accuracy.

Random Forest can offer an objective assessment of the performance of the model without a separate validation set. Each tree in the ensemble is trained with the help of separate bootstrap sample with some cases (i.e., out of bag samples) excluded. Without the requirement for extra data splitting, the out of bag samples for evaluation are utilised which enables effective model evaluation.

d. Gradient Boosting

Gradient Boosting is a potent machine learning approach which is frequently used for utilisation of health insurance projects for categorization and predictive modelling applications. It is defined as an ensemble technique which combines number of decision trees that are poor learners so that a powerful prediction model can be produced.

Gradient Boosting is used for constructing the predictive model by adding weak learners successively to the ensemble, which are often decision trees. Each weak learner is trained on a subset of the data with an emphasis on the cases where had high residuals from prior learners or those which were misclassified. Decision Trees are employed as weak learners due to their capacity in capturing intricate correlations between data and the target variable.

In order to ensemble's overall prediction error, the Gradient Boosting develops weak learners iteratively. The algorithm often modifies the weights of the training cases so that the misclassified or high residual instances can be prioritised after each iteration in which a new weak learner is added to the ensemble. This method continues until and unless a stopping requirement is satisfied or a predetermined number of iterations have been completed.

Gradient descent has been used by gradient boosting for optimisation of the ensemble. The ensemble predictions are adjusted in a way which minimises loss after computation of the gradient (also called as slope) of the loss function with respect to them. It is used in enhancing the ability of the ensemble for capturing the complicated patterns and make predictions precisely by updating the predictions of the ensemble iteratively.

Gradient Boosting introduces the hyperparameter referred to as learning rate which shows the contribution of each weak learner to the ensemble. When the algorithm iterates more slowly, then it indicates a lower learning rate which can reduce overfitting and increase generalisation. However, many convergence repetitions have been made necessary for a slower learning rate.

To quantify feature importance, Gradient Boosting offers a way which shows the importance of one feature as compared to the other when predictions are produced. The relevance of the feature is determined by taking into account the contribution of each feature for lowering the loss function throughout the training phase. Then, decision-making procedures for selection of the feature and health insurance results are facilitated with the use of this information.

In health insurance dataset, Class Imbalance or the under representation of some classes are considered as a common problem which quietly occurs. By giving instances which belongs to the minority class more weights during training, gradient boosting can handle unbalanced data by showing their significance and minimise the bias towards their majority class.

In order to limit complexity of the model and also avoid overfitting, gradient boosting offer regularisation methods which are frequently used by imposing restrictions on the maximum depth of the decision trees, capping the number of leaves and addition of penalties to the loss function for complicated models.

Gradient Boosting consists of a number of hyperparameters which can be adjusted for enhancing the performance of the model which comprises of the maximum decision tree depth, learning rate, subsampling rate and regularisation parameters as well as the number of iterations (weak learners). Hyperparameters must be carefully optimised by finding the ideal configuration for the current health insurance dilemma.

e. XGBoost Model

XGBoost Model, which is also referred to as the extreme gradient boosting, is an optimised and effective version of gradient boosting library for the purpose of scalable training in a machine learning model. XGBoost model is employed in projects related to health insurance for problems which involved predictive modelling and classification. XGBoost model is considered as a preferred option in data-intensive applications due to its outstanding scalability and performance. The Gradient Boosting Technique is known as the foundation of XGBoost which often combines a number of weak learners to produce a powerful predictive model. Gradient Descent is used for training weak learners iteratively and reduction of the overall prediction error. It also offers regularisation strategies for limiting the complexity of the model and also to avoid overfitting. These techniques often include limiting the depth of tree, penalisation of complex models and subsampling the data.

In order to increase the scalability and efficiency of the model, Numerous optimisation techniques are implemented which consists of approximate greedy techniques for splitting points, compressed data storage and construction of parallel trees. XGBoost models make use of these improvements for the purpose of processing massive datasets will millions of features and instances which further makes it appropriate for health insurance projects which requires handling comprehensive and intricate data.

Just like conventional decision tree methods, XGBoost models also builds decision trees level-wise. It often employs a more algorithm which works effectively to determine the best splits for each tree node. It considers a variety of split candidates, assessment of their quality using a special loss function, and then the split point is chosen which will result in the highest loss reduction.

Class disparity is often showed by datasets on health insurance with some classes which are underrepresented. XGBoost models offers methods to deal with imbalanced data which involves allocation of weights to various classes during training. XGBoost model can also prioritise the minority class by giving greater weights to the underrepresented classes and improvement of performance of the prediction.

In XGBoost Model, the importance of each feature which is in relation to other features when predications are produced is indicated by the rankings of the feature importance based on how often a characteristic has been employed for dividing the data among all of the trees of the ensembles and the assessment of its significance. Choice of features, risk assessment and better decision-making have been made possible using this information, that assists in identification of the critical variables which affects health insurance outcomes.

XGBoost consists of a number of hyperparameters which can be adjusted for enhancing the performance of the model which comprises of the maximum decision tree depth, subsampling rate and regularisation parameters as well as the number of iterations (boosting rounds). Hyperparameters must be carefully adjusted by finding the ideal setup which maximises the performance of the model for the particular health insurance challenge.

For rating the effectiveness of the XGBoost model, there are some evaluation criteria such as accuracy, recall, precision, F1-score and AUC-ROC which is are under the receiver operating characteristic curve. It is essential to use different test set for testing purpose in order to gauge the generalisation of the model and also to avoid overfitting.

For a reliable and effective solution for health insurance projects, XGBoost make use of the strength of the gradient boosting and includes a number of optimisations. Due to its ability to handle large-scale datasets as well as the imbalanced data, it is considered as a useful tool for the tasks which is related to claim prediction, policy renewal forecasts, risk assessment, fraud detection and other aspects of health insurance. It also helps in delivering insights of feature importance.

III. Methodology

This study adopted quantitative research since it uses a deductive approach to identify patterns in human existence by separating the social realm into measurable elements referred to as variables which can be quantified numerically.

A. Quantitative Methodology

This method was being applied to focus on investigation of the answers to the questions such as how much, how many, and to what extent. In this study, data is extracted from an open-source database which focus on customer behaviour which can be quantified and interpreted to gather insights. Quantitative data is referred to as the data which is in the form of counts having numerical data on each dataset. This data was statistically analysed to examine the conclusive results of vehicle insurance prediction. Quantitative data was statistically analysed to determine if the person with health insurance is interested in purchasing vehicle insurance.

B. Data Collection and Processing

For this research, we have gathered and combined data sources, including claim details, policyholder information, historical claim data, external data (such as public records, social media) and any fraud signs that may be present. Then pre-processing is done on the data to ensure its reliability, accuracy and analytical suitability.

I. Machine Learning Lifecycle

The Machine Learning lifecycle has various stages which provides a key structure in handling data and build a suitable machine learning model. The specific techniques and algorithms used within each stage may vary from dataset to dataset depending on the nature of the health insurance dataset, its variables and mainly on its objective which means that the stages may not be always sequential. It often involves iterating through stages at multiple times, model refining, and improving its performance based on new insights and feedback during the process.

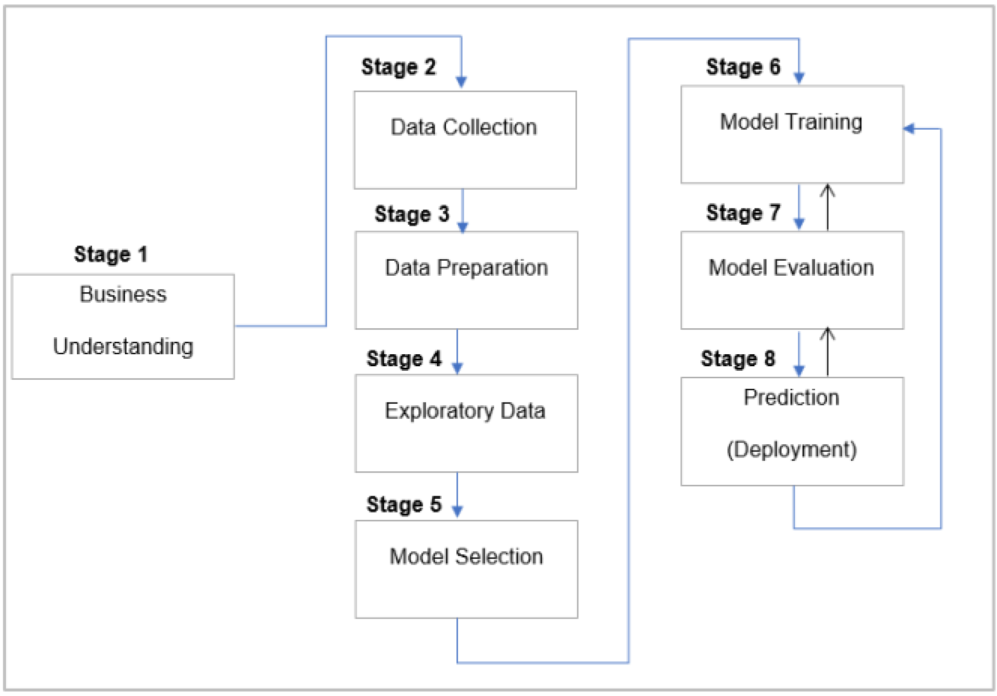


Figure 4: Machine Learning Lifecycle

II. Health Insurance Data Source

A large insurance company database contains the health insurance data which contains the information which is related to insurance policies, claims, members and other relevant variables.

III. Understanding Health Insurance Dataset

Leading insurance provider, our client, is starting a fascinating project to make use of their vast customer data and broaden their product offers. They have a solid base in the provision of health insurance; therefore, they are eager to delve into the world of auto insurance and find new clients who could be interested in buying a policy.

Building a strong predictive model that can accurately anticipate whether policyholders who have previously engaged with their health insurance offers will also display interest in their automobile insurance plans requires drawing on their rich dataset from prior years.

Insurance plans are essential for protecting people and organisations from unforeseen dangers and monetary losses. By providing auto insurance, companies hope to extend their protection to their clients' priceless possessions and give them assurance and thorough protection. In order to do this, we will use cutting-edge statistical methods and machine learning algorithms to analyse a variety of client attributes, demographic data, and historical insurance engagement data. We will create a predictive model capable of precisely identifying those clients who are likely to express interest in buying vehicle insurance by revealing hidden patterns and important indicators within the dataset.

By utilising this predictive model, we are able to maximise their marketing initiatives, streamline their customer acquisition techniques, and especially adapt their product offers to the requirements and preferences of their current clientele. This individualised strategy will not only increase customer happiness but also support our client's company's long-term expansion and success.

The dataset contains the following information:

Column	**Description**	**Data type**
id	Customer’s Unique ID	Integer
Gender	Customer’s Gender	Object
Age	Customer’s Age	Integer
Driving_License	0: Have Driving License 1: Do not Driving License	Integer
Region_Code	Customer’s Unique Code by region	Float
Previously_Insured	1: Have Vehicle Insurance 0: Do not have Vehicle Insurance	Integer
Vehicle_Age	Vehicle’s Age	Object
Vehicle_Damage	1: Vehicle got damaged in the past 0: Vehicle does not get damaged in the past	Object

Annual_Premium	Premium that Customer has to pay in the year	Float
Policy_Sales_Channel	Channel of reaching to the Customer via Anonymized Code	Float
Vintage	Number of days for which Customer has been associated with the Company	Integer
Response	1: Interested 0: Not Interested	integer

Table 1: Health Insurance Dataset- Features, Description and Data Type

IV. Data Collection Method

Gathering and combining pertinent data sources, including as claim details, policyholder information, historical claims data, external data (such as public records, social media), and any fraud signs that may be present, preparing and pre-processing the data to ensure its accuracy, reliability, and analytical suitability.

V. Google Colab and Python Programming

The dataset was imported into Google Colab after extraction and used to train the model. The extracted dataset was observed by selecting the first 10 rows to ensure that fields and features were imported successfully. Table given below represents he dataset before it was cleaned.

id	Gender	Age	Driving_License	Region_Code	Previously_Insured
1	Male	44	1	28	0
2	Male	76	1	3	0
3	Male	47	1	28	0
4	Male	21	1	11	1
5	Female	29	1	41	1
6	Female	24	1	33	0
7	Male	23	1	11	0
8	Female	56	1	28	0
9	Female	24	1	3	1
10	Female	32	1	6	1
11	Female	47	1	35	0
12	Female	24	1	50	1
13	Female	41	1	15	1
14	Male	76	1	28	0
15	Male	71	1	28	1
16	Male	37	1	6	0

17	Female	25	1	45	0
18	Female	25	1	35	1
19	Male	42	1	28	0
20	Female	60	1	33	0

Vehicle Age	Vehicle Damage	Annual Premium	Policy Sales Channel	Vintage	Respons e
> 2 Years	Yes	40454	26	217	1
1-2 Year	No	33536	26	183	0
> 2 Years	Yes	38294	26	27	1
< 1 Year	No	28619	152	203	0
< 1 Year	No	27496	152	39	0
< 1 Year	Yes	2630	160	176	0
< 1 Year	Yes	23367	152	249	0
1-2 Year	Yes	32031	26	72	1
< 1 Year	No	27619	152	28	0
< 1 Year	No	28771	152	80	0
1-2 Year	Yes	47576	124	46	1
< 1 Year	No	48699	152	289	0
1-2 Year	No	31409	14	221	0
1-2 Year	Yes	36770	13	15	0
1-2 Year	No	46818	30	58	0
1-2 Year	Yes	2630	156	147	1
< 1 Year	Yes	26218	160	256	0
< 1 Year	No	46622	152	299	0
1-2 Year	Yes	33667	124	158	0
1-2 Year	Yes	32363	124	102	1

Table 2: Dataset Before Cleaning

The dataset shape refers to the dimensions of the dataset, i.e. the number of rows (instances) and columns (features).

Table given below displays the shape of training and testing dataset respectively.

	Training Dataset Shape	Testing Dataset Shape
Number of Rows	127037	127037
Number of Columns	11	11

Table 3: Shape of the Training and Testing Data

VI. Health Insurance Data Processing

Data Pre-Processing is a major step in the data analysis and machine learning which includes cleaning, transforming and preparing raw data in order to make it suitable for the analysis and modelling. During data cleaning, all the duplicate values were removed and since there were no missing values so it didn’t require any imputation for it in python. The dataset was then split into training and testing datasets for evaluation of model's performance. After processing, all the categorical values are converted into numerical values which can be displayed as follows:

Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age
1	0.369231	1	0.538462	0	3
1	0.861538	1	0.057692	0	2
1	0.415385	1	0.538462	0	3
1	0.015385	1	0.211538	1	1
0	0.138462	1	0.788462	1	1
0	0.061538	1	0.634615	0	1
1	0.046154	1	0.211538	0	1
0	0.553846	1	0.538462	0	2
0	0.061538	1	0.057692	1	1
0	0.184615	1	0.115385	1	1

Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Vintage	Response
1	0.070366	0.154321	0.716263	1
0	0.057496	0.154321	0.598616	0
1	0.066347	0.154321	0.058824	1
0	0.048348	0.932099	0.66782	0
0	0.046259	0.932099	0.100346	0
1	0	0.981481	0.574394	0
1	0.038578	0.932099	0.82699	0
1	0.054696	0.154321	0.214533	1
0	0.046488	0.932099	0.062284	0
0	0.048631	0.932099	0.242215	0

Table 4: Dataset After Cleaning

C. Exploratory Data Analysis

After performing the pre-processing of the data, exploratory data analysis has to be performed to gain an understanding of the data. During EDA, the dataset has to be explored by the researcher which is extracted from the insurance company database which involved assessing factors like dataset’s size, count of column, and variable data types. It also includes descriptive analysis, correlations, data visualizations and examining distributions. EDA is considered as the most important step in the data analysis which involved exploring the data and also summarizing its main characteristics in order to find insights and understand its trends, patterns and potential issues in a clear manner.

1. **Descriptive Statistics**

   Exploratory Data Analysis has some important components which are descriptive statistics and distribution which are useful for understanding the main characteristics of dataset and distribution of the values across variables.

   Table given below displays the descriptive statistics of the numerical variables of the dataset.

	id	**Age**	**Driving License**	**Region Code**	**Previously Insured**
count	381109.00	381109.00	381109.00	381109.00	381109.00
mean	190555.00	38.82	1.00	26.39	0.46
std	110016.84	15.51	0.05	13.23	0.50
min	1.00	20.00	0.00	0.00	0.00
25%	95278.00	25.00	1.00	15.00	0.00
60%	190555.00	36.00	1.00	28.00	0.00
75%	285832.00	49.00	1.00	35.00	1.00
max	381109.00	85.00	1.00	52.00	1.00

	Annual Premium	**Policy Sales Channel**	**Vintage**	**Response**
count	381109.00	381109.00	381109.00	381109.00
mean	30564.39	112.03	154.35	0.12
std	17213.16	54.20	83.67	0.33
min	2630.00	1.00	10.00	0.00
25%	24405.00	29.00	82.00	0.00
60%	31669.00	133.00	154.00	0.00
75%	39400.00	152.00	227.00	0.00
max	540165.00	163.00	299.00	1.00

Table 5: Summary Statistics of Numerical Variables

Table given below displays the summary of the categorical variables:

	Gender	Vehicle_Age	Vehicle_Damage
count	381109	381109	381109
unique	2	3	2
top	Male	1-2 Year	Yes
freq	206089	200316	192413

Table 6: Summary Statistics of Categorical Variables

2. **Correlation**

The statistical relationship between two or more variables is referred to as correlation which measures the direction and strength of the linear relationship between two or more variables which means that it shows how change in one variable will affect another variable. The Statistical measures are made to determine the linear relationship between two variables and also to determine if the two variables are correlated when they both movie in the same direction.

The following figure displays the Pearson correlation of 11 features on how data were correlated in this study.

	Gender	Age	Driving License	Region Code	Previously Insured	Vehicle Age
Gender	1	0.15	-0.018	0.0006	-0.082	0.16
Age	0.15	1	-0.08	0.043	-0.25	0.77
Driving License	-0.018	-0.08	1	-0.0011	0.015	-0.037
Region Code	0.0006	0.043	-0.0011	1	-0.025	0.044
Previously Insured	-0.082	-0.25	0.015	-0.025	1	-0.38
Vehicle Age	0.16	0.77	-0.037	0.044	-0.38	1
Vehicle Damage	0.092	0.27	-0.017	0.028	-0.82	0.4
Annual Premium	0.0037	0.068	-0.012	-0.011	0.0043	0.0042
Policy Sales Channel	-0.11	-0.58	0.044	-0.042	0.22	-0.55
Vintage	-0.0025	-0.0013	-0.00085	-0.0027	0.0025	-0.0019
Response	0.052	0.11	0.01	0.011	-0.34	0.22

	Vehicle Damage	Annual Premium	Policy Sales Channel	Vintage	Response
Gender	0.092	0.0037	-0.11	-0.0025	0.052
Age	0.27	0.068	-0.58	-0.0013	0.11
Driving License	-0.017	-0.012	0.044	-0.00085	0.01
Region Code	0.028	-0.011	-0.042	-0.0027	0.011

Previously Insured	-0.82	0.0043	0.22	0.0025	-0.34
Vehicle Age	0.4	0.042	-0.55	-0.0019	0.22
Vehicle Damage	1	0.0093	-0.22	-0.0021	0.35
Annual Premium	0.0093	1	-0.11	-0.00061	0.023
Policy Sales Channel	-0.22	-0.11	1	1.80E-06	-0.14
Vintage	-0.0021	-0.00061	1.80E-06	1	-0.0011
Response	0.35	0.023	-0.14	-0.0011	1

Table 7: Correlation between Features of Health Insurance Dataset

Vehicle Damage, Vehicle Age and Age are most positively correlated with Response which indicates that when Vehicle Damage, Vehicle Age and Age increases then Response also increases i.e., vehicle having more damage would result in more response for insurance claims, older vehicles have more responses due to increased wear and tear and maintenance issues and old people have different response rates due to driving habits and risk factors.

Previously Insured and Policy Sales Channel are most negatively correlated with Response which indicates that when Previously Insured and Policy Sales Channel increases then Response decreases i.e., Individuals who are already insured have lower response rates since they do not require additional services and certain sales channel have lower response rates due to inefficiencies and customer preferences.

D. Data Visualization

The graphical representation of data is a crucial step for understanding of the data with the help of charts, graphs, etc. where the data is presented in such a way that it affects the target variable. In this study, the data is displayed through different visualizations to get a clear picture of the data and gain insights from it and also it helps the researcher to ensure that data has been distributed correctly. It helps the researcher in exploration of the data, identification of outliers and patterns, and also interpret the key findings efficiently.

i. Age and Gender

The following figure displays the visualization of Age and Gender.

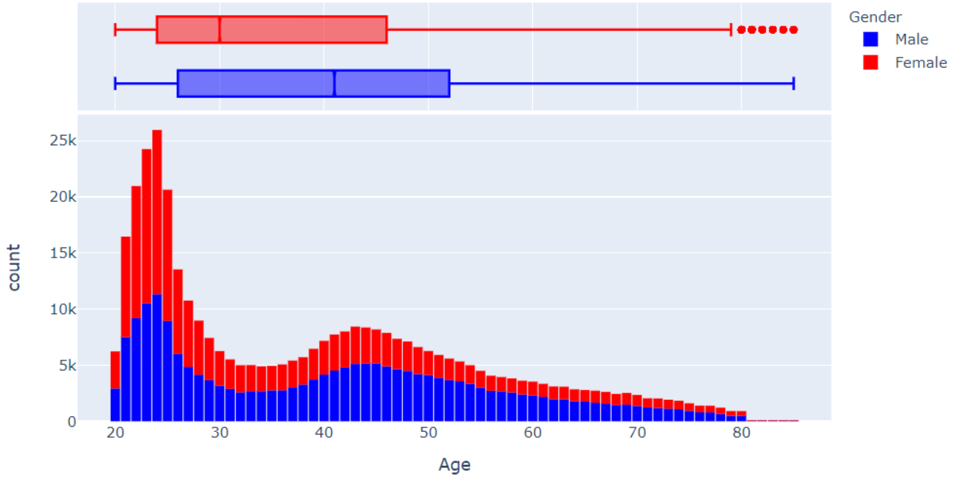


Figure 5: Age and Gender Distribution

It is considered that people of age between 20 and 30 have more policy holder than other ages and within that age group, females are than males.

ii. Vehicle Damage and Response:

The following figure displays the visualization of cross tabulation of vehicle damage and Response.

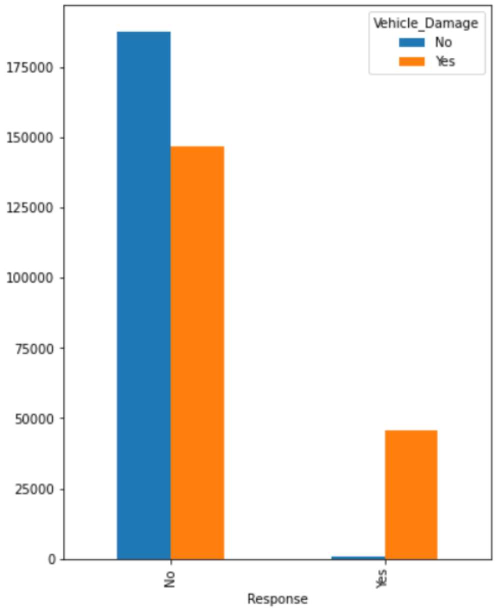


Figure 6: Vehicle Damage and Response

It is observed that the customers whose vehicle got damaged in the past are more likely to be interested in insurance and customers who are not interested in insurance do not get their vehicles damaged in the past.

iii. Age and Response:

The following visualization displays how age is distributed with response.

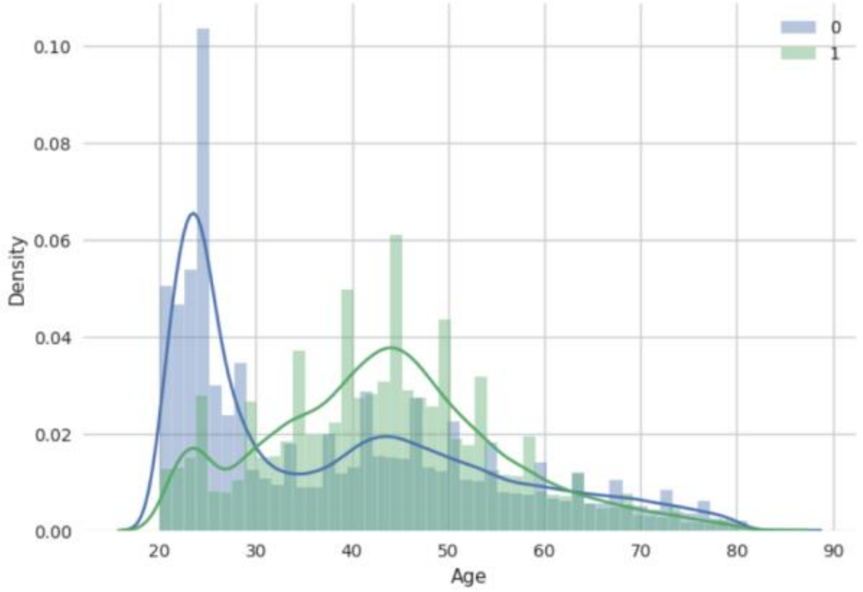


Figure 7: Age and Response

Young people whose age is less than 30 are not interested in purchasing vehicle insurance which could be due to lack of experience, not having expensive vehicles, etc. whereas people aged 30-60 are more interested in vehicle insurance and people with age more than 60 are less interested in vehicle insurance since they might not drive much at old age.

iv. Vehicle Age and Response:

The following visualization displays the distribution of vehicle age with response.

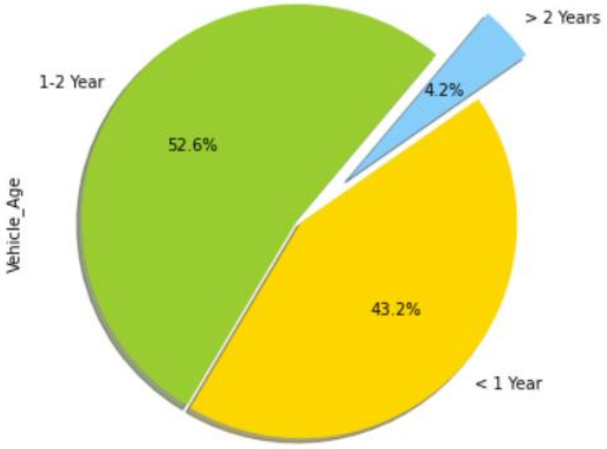


Figure 8: Vehicle Age and Response

It is observed that vehicle age between 1-2 years have more than half of the samples of data and less than 5% of the data has samples with vehicle age more than 2 years and 43% of the data has samples with vehicle age less than 1 year.

v. Annual Premium and Age

The following visualization displays the Distribution of Annual Premium with Age.

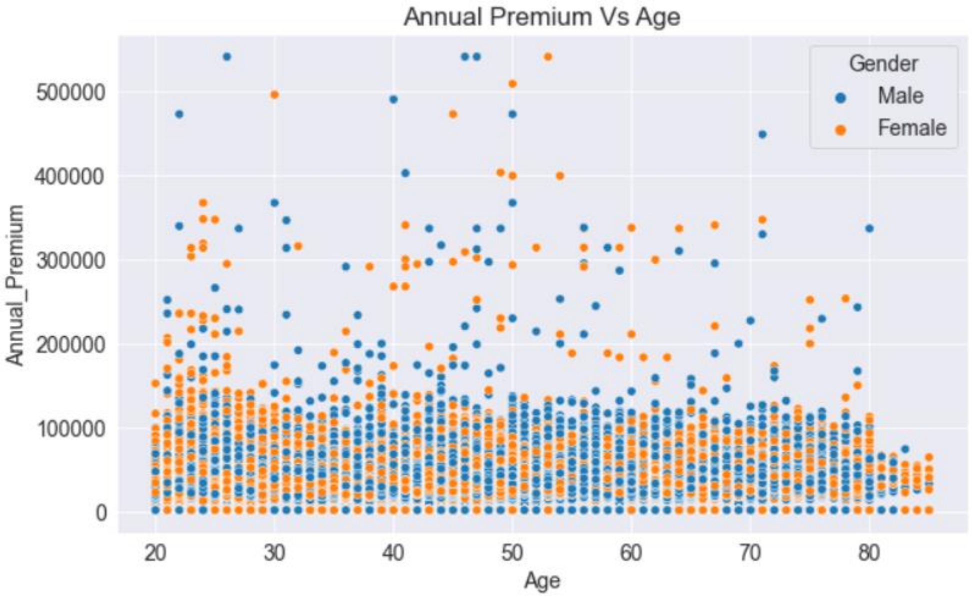


Figure 9: Annual Premium and Age

The annual premium for many of the policy holders is around 100K of all age.

E. Best Features in the Health Insurance Data

After exploring the health insurance data, the features of the data are scaled or normalized ensuring that they have a similar scale which can be useful for improving the performance of certain machine learning algorithms. It also involves encoding categorical variables into numerical format with the help of techniques using one-hot encoding and label encoding.

After Transforming the dataset, best features are selected out of all the features which is available in the data using SelectKBest selection feature.

The bar graph below shows a visual representation of the best features that needs to be considered for the analysis.

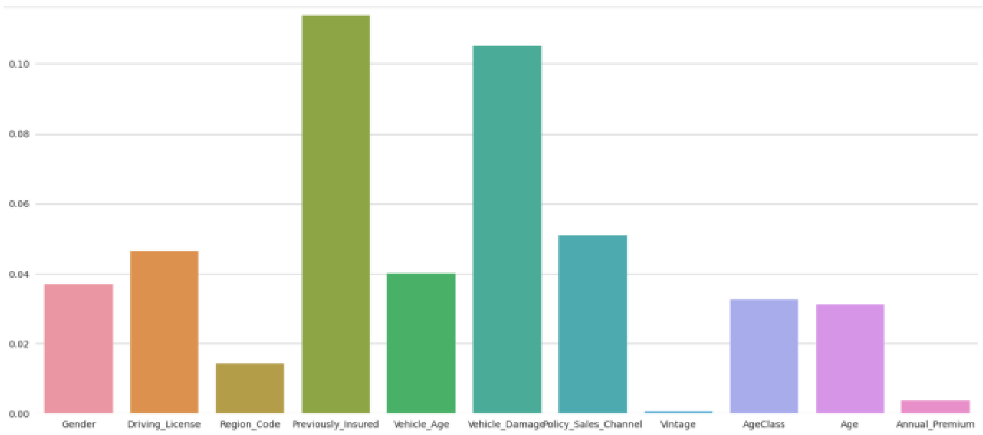


Figure 10: Best Features in the Health Insurance Dataset

The graph shows that "Premium Insured" is the most important feature while Vintage being the least as for that the value is least i.e., 0.0007.

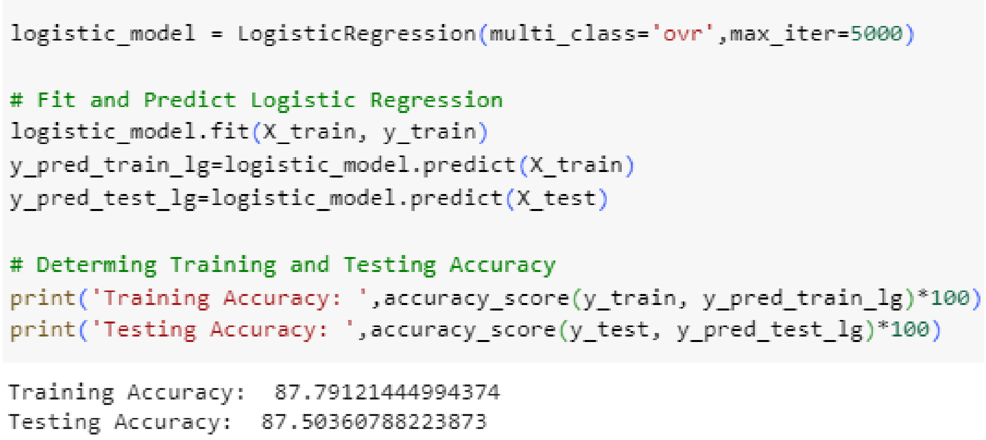
F. Model Selection

The process of selecting the most appropriate model for a data to perform a specific task is called Model Selection. It involves the evaluation of different models and comparison of their performances and selecting the one which best fits the data and also gives more accurate predictions.

The following chart displays which machine learning algorithm has been selected, trained and evaluated to build best predictive model.

G. Model Training and Model Evaluation Training

(a) Logistic Regression



```
logistic_model = LogisticRegression(multi_class='ovr',max_iter=5000)

# Fit and Predict Logistic Regression
logistic_model.fit(X_train, y_train)
y_pred_train_lg=logistic_model.predict(X_train)
y_pred_test_lg=logistic_model.predict(X_test)

# Determing Training and Testing Accuracy
print('Training Accuracy: ',accuracy_score(y_train, y_pred_train_lg)*100)
print('Testing Accuracy: ',accuracy_score(y_test, y_pred_test_lg)*100)
```

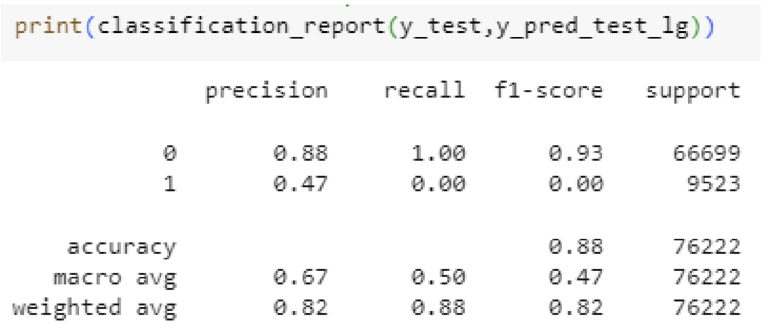
```
Training Accuracy:  87.79121444994374
Testing Accuracy:  87.50360788223873
```

Figure 11: Fitting of Logistic Regression Model

The logistic Regression model achieved an accuracy of 87.79% using training data which means that it made 87.79% correct predictions of the examples in the training data i.e., the model correctly predicted around 87 examples out of every 100 examples in the training data.

The model has achieved an accuracy of 87.50% when it has been tested on a testing data which has not been trained before so when it is applied to new and unseen data then accuracy drops slightly as compare to the training accuracy which has been obtained before.

Since the testing accuracy is slightly lower than the training accuracy, it indicates slighting overfitting of the model in the training data which means that the model in the training data has learned to capture noise or random fluctuations instead of generalizing well to unseen data on a separate testing dataset helps to assess its ability to generalize to new data and provides insights into potential overfitting issues.



```
print(classification_report(y_test,y_pred_test_lg))

              precision    recall  f1-score   support

           0       0.88      1.00      0.93     66699
           1       0.47      0.00      0.00      9523

    accuracy                           0.88     76222
   macro avg       0.67      0.50      0.47     76222
weighted avg       0.82      0.88      0.82     76222
```

Figure 12: Classification Report of Logistic Regression Model

Precision:

A precision highlights how many positive predictions are correct which are made by the model. The precision of class 0 is 88% that means that 88% of the instances which have been predicted by the model of class 0 are not interested in Vehicle Insurance whereas the precision of class 1 is 47% that means that 47% of the instances which have been predicted by the model of class 1 are interested in Vehicle Insurance.

Recall:

Recall, which is also termed as sensitivity, measures the correctly identified positive predictions that have been made by the model. The recall of class 0 is 1.00 which indicates that the model is able to correctly identify all the instances of customers that are not interested in Vehicle Insurance whereas for class 1, it is 0 which indicates the model is not able to correctly identify all the instances who are interested in Vehicle Insurance.

Support:

Support indicates the count of actual occurrences of each class in a specified dataset. The class consists of highest support value i.e., 66699 that means that the dataset is imbalanced towards class 0 which highlights that there are more instances of customers who are interested in purchasing Vehicle Insurance than those who are interested in purchasing Vehicle Insurance which is indicated by class 1 having support value 9523.

F1 Score:

The F1 Score of class 0 highlights the performance of the model in identification of instances of customers who are not interested in Vehicle Insurance which is, however, higher which indicates a good balance between precision and recall for class 0. This means it has identified instances of customers who are not interested in Vehicle Insurance effectively while also minimizing both false positives and false negatives. However, the F1 Score of class 1 is 0 which indicates no balance is present between precision and recall for class 1.

Accuracy:

The model attained an accuracy of 88% which indicates that 88% of the predictions which have been made by the model are correct but accuracy has a limitation which is its biasedness towards the majority class which is class 0 in this case. Since it cannot handle imbalanced data where one is dominated by the other class so it can lead to misleading conclusions when assessing the model's performance.

Hence, the Logistic Regression Model performs well in identifying the instances of customers who are not interested in Vehicle Insurance but does not perform well in identifying instances of customers who are interested in Vehicle Insurance.

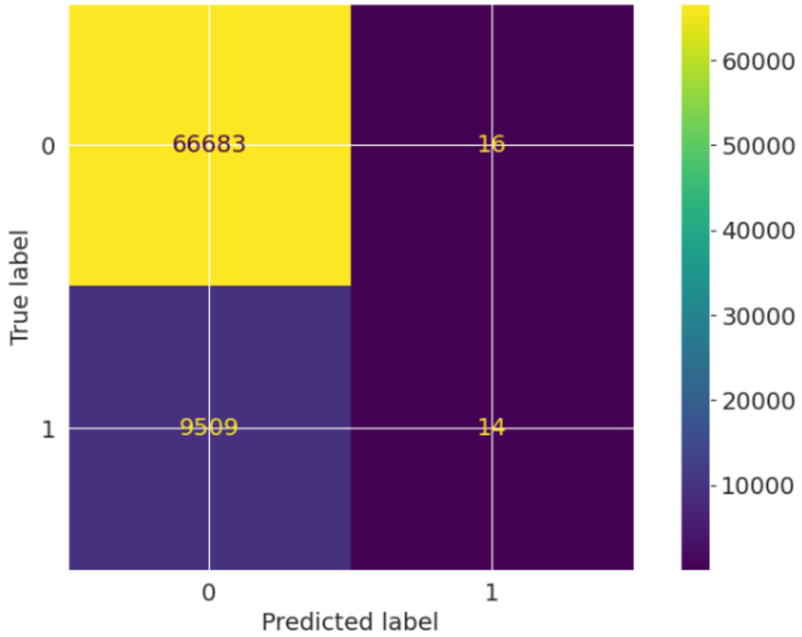
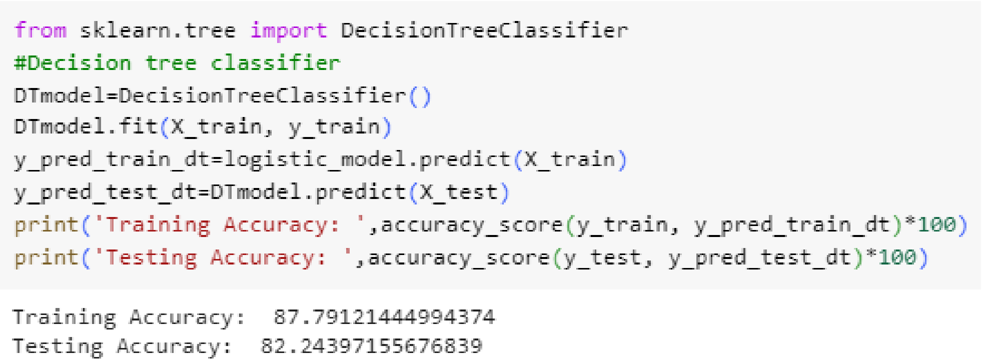


Figure 13: Confusion Matrix of Logistic Regression Model

The confusion Matrix shows that 66683 instances are correctly identified instances of label 0 and 14 instances are correctly identified instances of label 1. Also, 9509 instances are predicted as label 0 but are actually label 1 and 16 instances are actually label 0 but are predicted as label 1.

(b) Decision Tree Model



```
from sklearn.tree import DecisionTreeClassifier
#Decision tree classifier
DTmodel=DecisionTreeClassifier()
DTmodel.fit(X_train, y_train)
y_pred_train_dt=logistic_model.predict(X_train)
y_pred_test_dt=DTmodel.predict(X_test)
print('Training Accuracy: ',accuracy_score(y_train, y_pred_train_dt)*100)
print('Testing Accuracy: ',accuracy_score(y_test, y_pred_test_dt)*100)
```

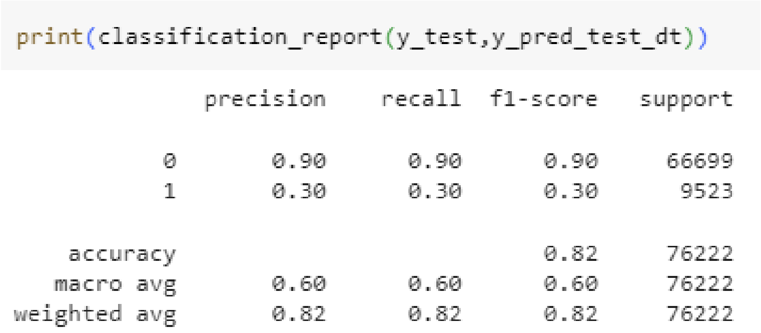
```
Training Accuracy:  87.79121444994374
Testing Accuracy:  82.24397155676839
```

Figure 14: Fitting of Decision Tree Model

The Decision Tree model achieved an accuracy of 87.79% using training data which means that it made 87.79% correct predictions of the examples in the training data i.e., the model correctly predicted around 87 examples out of every 100 examples in the training data.

The model has achieved an accuracy of 82.24% when it has been tested on a testing data which has not been trained before so when it is applied to new and unseen data then accuracy drops slightly as compare to the training accuracy which has been obtained before.

Since the testing accuracy is slightly lower than the training accuracy, it indicates slighting overfitting of the model in the training data which means that the model in the training data has learned to capture noise or random fluctuations instead of generalizing well to unseen data on a separate testing dataset helps to assess its ability to generalize to new data and provides insights into potential overfitting issues.



```
print(classification_report(y_test,y_pred_test_dt))

              precision    recall  f1-score   support

           0       0.90      0.90      0.90     66699
           1       0.30      0.30      0.30      9523

    accuracy                           0.82     76222
   macro avg       0.60      0.60      0.60     76222
weighted avg       0.82      0.82      0.82     76222
```

Figure 15: Classification Report of Decision Tree Model

Precision:

A precision highlights how many positive predictions are correct which are made by the model. The precision of class 0 is 82% that means that 82% of the instances which have been predicted by the model of class 0 are not interested in Vehicle Insurance whereas the precision of class 1 is 30% that means that 30% of the instances which have been predicted by the model of class 1 are interested in Vehicle Insurance.

Recall:

Recall, which is also termed as sensitivity, measures the correctly identified positive predictions that have been made by the model. The recall of class 0 is 0.90 which indicates that the model is able to correctly identify 90% of the instances of customers that are not interested in Vehicle Insurance whereas for class 1, it is 0.30 which indicates the model is able to correctly identify 30% of the instances who are interested in Vehicle Insurance which is comparatively less than the one achieved in class 0.

Support:

Support indicates the count of actual occurrences of each class in a specified dataset. The class consists of highest support value i.e., 66699 that means that the dataset is imbalanced towards class 0 which highlights that there are more instances of customers who are interested in purchasing Vehicle Insurance than those who are interested in purchasing Vehicle Insurance which is indicated by class 1 having support value 9523.

F1 Score:

The F1 Score of class 0 highlights the performance of the model in identification of instances of customers who are not interested in Vehicle Insurance which is, however, higher i.e., 0.90 which indicates a good balance between precision and recall for class 0. This means it has identified 90% instances of customers who are not interested in Vehicle Insurance effectively while also minimizing both false positives and false negatives. However, the F1 Score of class 1 is 0.30 which indicates a poor balance between precision and recall for class 1.

Accuracy:

The model attained an accuracy of 82% which indicates that 82% of the predictions which have been made by the model are correct but accuracy has a limitation which is its biasedness towards the majority class which is class 0 in this case. Since it cannot handle imbalanced data where one is dominated by the other class so it can lead to misleading conclusions when assessing the model's performance.

Hence, the Decision Tree Model performs well in identifying the instances of customers who are not interested in Vehicle Insurance than identifying instances of customers who are interested in Vehicle Insurance.

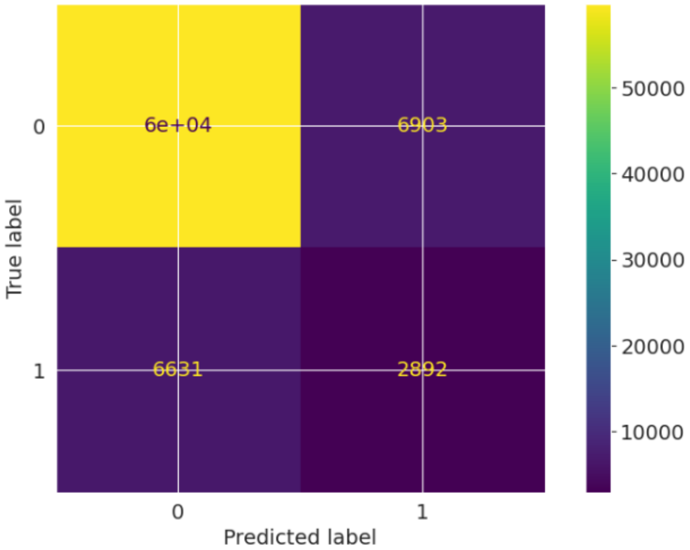
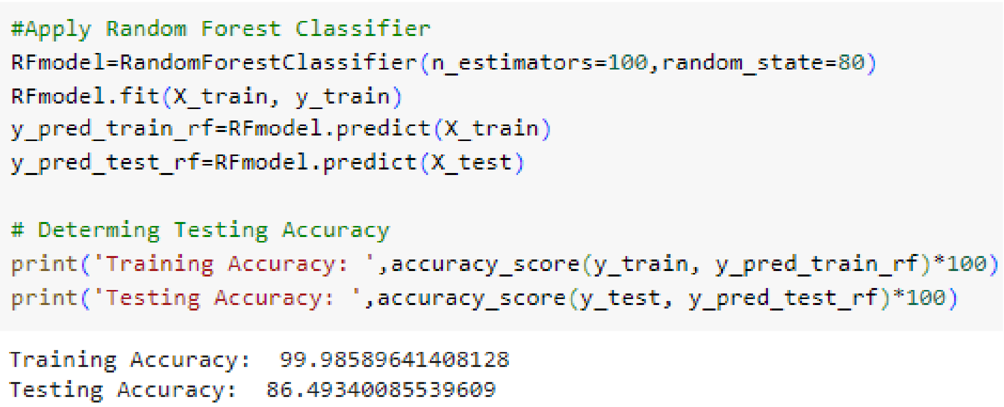


Figure 16: Confusion Matrix of Decision Tree Model

The confusion Matrix shows that 60000 instances are correctly identified instances of label 0 and 2892 instances are correctly identified instances of label 1. Also, 6631 instances are predicted as label 0 but are actually label 1 and 6903 instances are actually label 0 but are predicted as label 1.

(c) Random Forest Model



```
#Apply Random Forest Classifier
RFmodel=RandomForestClassifier(n_estimators=100,random_state=80)
RFmodel.fit(X_train, y_train)
y_pred_train_rf=RFmodel.predict(X_train)
y_pred_test_rf=RFmodel.predict(X_test)

# Determing Testing Accuracy
print('Training Accuracy: ',accuracy_score(y_train, y_pred_train_rf)*100)
print('Testing Accuracy: ',accuracy_score(y_test, y_pred_test_rf)*100)
```

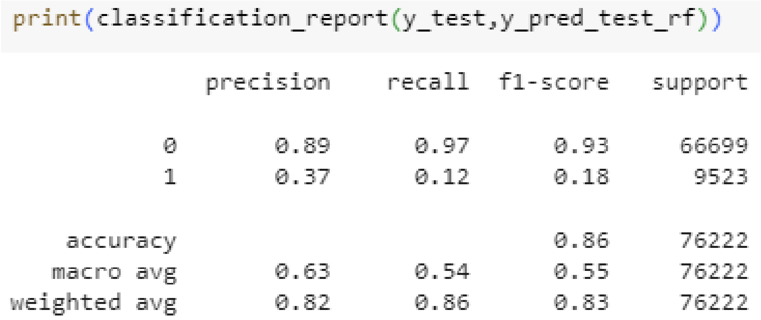
```
Training Accuracy:  99.98589641408128
Testing Accuracy:  86.49340085539609
```

Figure 17: Fitting of Random Forest Model

The Random Forest Model achieved an accuracy of 99.98% using training data which means that it made 99.98% correct predictions of the examples in the training data i.e., the model correctly predicted around 99 examples out of every 100 examples in the training data.

The model has achieved an accuracy of 86.49% when it has been tested on a testing data which has not been trained before so when it is applied to new and unseen data then accuracy drops slightly as compare to the training accuracy which has been obtained before.

Since the testing accuracy is slightly lower than the training accuracy, it indicates slighting overfitting of the model in the training data which means that the model in the training data has learned to capture noise or random fluctuations instead of generalizing well to unseen data on a separate testing dataset helps to assess its ability to generalize to new data and provides insights into potential overfitting issues.



```
print(classification_report(y_test,y_pred_test_rf))

              precision    recall  f1-score   support

           0       0.89      0.97      0.93     66699
           1       0.37      0.12      0.18      9523

    accuracy                           0.86     76222
   macro avg       0.63      0.54      0.55     76222
weighted avg       0.82      0.86      0.83     76222
```

Figure 18: Classification Report of Random Forest Model

Precision:

A precision highlights how many positive predictions are correct which are made by the model. The precision of class 0 is 89% that means that 89% of the instances which have been predicted by the model of class 0 are not interested in Vehicle Insurance whereas the precision of class 1 is 37% that means that 37% of the instances which have been predicted by the model of class 1 are interested in Vehicle Insurance.

Recall:

Recall, which is also termed as sensitivity, measures the correctly identified positive predictions that have been made by the model. The recall of class 0 is 0.97 which indicates that the model is able to correctly identify 97% of the instances of customers that are not interested in Vehicle Insurance whereas for class 1, it is 0.12 which indicates the model is able to correctly identify 12% of the instances who are interested in Vehicle Insurance which is comparatively less than the one achieved in class 0.

Support:

Support indicates the count of actual occurrences of each class in a specified dataset. The class consists of highest support value i.e., 66699 that means that the dataset is imbalanced towards class 0 which highlights that there are more instances of customers who are interested in purchasing Vehicle Insurance than those who are interested in purchasing Vehicle Insurance which is indicated by class 1 having support value 9523.

F1 Score:

The F1 Score of class 0 highlights the performance of the model in identification of instances of customers who are not interested in Vehicle Insurance which is, however, higher i.e., 0.93 which indicates a good balance between precision and recall for class 0. This means it has identified 93% instances of customers who are not interested in Vehicle Insurance effectively while also minimizing both false positives and false negatives. However, the F1 Score of class 1 is 0.18 which indicates a poor balance between precision and recall for class 1.

Accuracy:

The model attained an accuracy of 86% which indicates that 86% of the predictions which have been made by the model are correct but accuracy has a limitation which is its biasedness towards the majority class which is class 0 in this case. Since it cannot handle imbalanced data where one is dominated by the other class so it can lead to misleading conclusions when assessing the model's performance.

Hence, the Random Forest Model performs well in identifying the instances of customers who are not interested in Vehicle Insurance than identifying instances of customers who are interested in Vehicle Insurance.

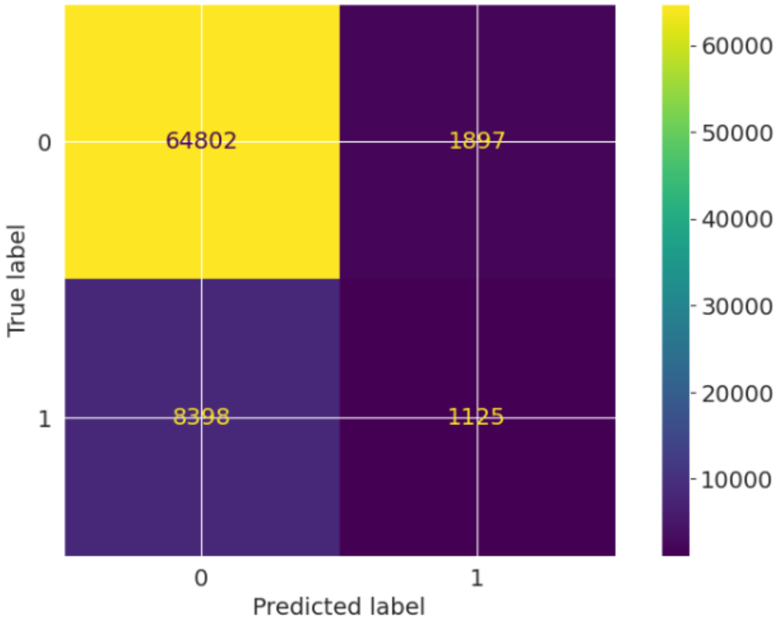
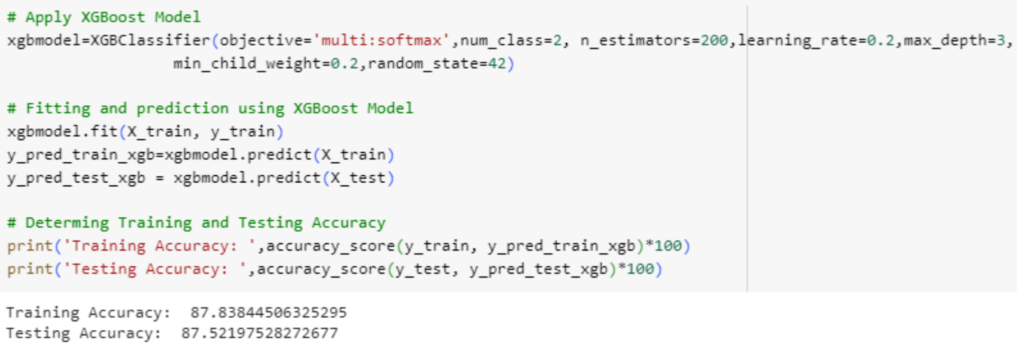


Figure 19: Confusion Matrix of Random Forest Model

The confusion Matrix shows that 64802 instances are correctly identified instances of label 0 and 1125 instances are correctly identified instances of label 1. Also, 8398 instances are predicted as label 0 but are actually label 1 and 1897 instances are actually label 0 but are predicted as label 1.

(d) XGBoost Model



```
# Apply XGBoost Model
xgbmodel=XGBClassifier(objective='multi:softmax',num_class=2, n_estimators=200,learning_rate=0.2,max_depth=3,
                    min_child_weight=0.2,random_state=42)

# Fitting and prediction using XGBoost Model
xgbmodel.fit(X_train, y_train)
y_pred_train_xgb=xgbmodel.predict(X_train)
y_pred_test_xgb = xgbmodel.predict(X_test)

# Determing Training and Testing Accuracy
print('Training Accuracy: ',accuracy_score(y_train, y_pred_train_xgb)*100)
print('Testing Accuracy: ',accuracy_score(y_test, y_pred_test_xgb)*100)
```

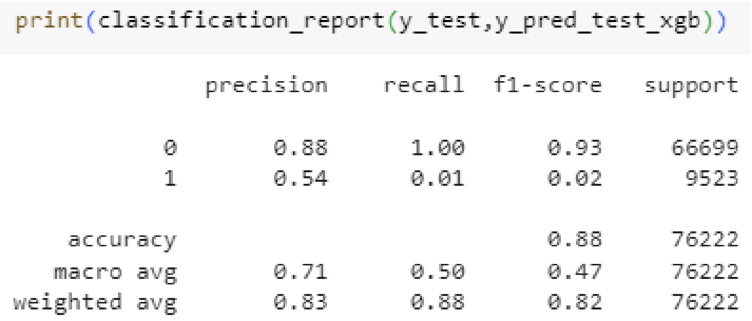
```
Training Accuracy:  87.83844506325295
Testing Accuracy:  87.52197528272677
```

Figure 20: Fitting of XGBoost Model

The XGBoost Model achieved an accuracy of 87.83% using training data which means that it made 87.83% correct predictions of the examples in the training data i.e., the model correctly predicted around 87 examples out of every 100 examples in the training data.

The model has achieved an accuracy of 87.52% when it has been tested on a testing data which has not been trained before so when it is applied to new and unseen data then accuracy drops slightly as compare to the training accuracy which has been obtained before.

Since the testing accuracy is slightly lower than the training accuracy, it indicates slighting overfitting of the model in the training data which means that the model in the training data has learned to capture noise or random fluctuations instead of generalizing well to unseen data on a separate testing dataset helps to assess its ability to generalize to new data and provides insights into potential overfitting issues.



```
print(classification_report(y_test,y_pred_test_xgb))

              precision    recall  f1-score   support

           0       0.88      1.00      0.93     66699
           1       0.54      0.01      0.02      9523

    accuracy                           0.88     76222
   macro avg       0.71      0.50      0.47     76222
weighted avg       0.83      0.88      0.82     76222
```

Figure 21: Classification Report of XGBoost Model

Precision:

A precision highlights how many positive predictions are correct which are made by the model. The precision of class 0 is 88% that means that 88% of the instances which have been predicted by the model of class 0 are not interested in Vehicle Insurance whereas the precision of class 1 is 54% that means that 54% of the instances which have been predicted by the model of class 1 are interested in Vehicle Insurance.

Recall:

Recall, which is also termed as sensitivity, measures the correctly identified positive predictions that have been made by the model. The recall of class 0 is 1.00 which indicates that the model is able to correctly identify 100% of the instances of customers that are not interested in Vehicle Insurance whereas for class 1, it is 0.01 which indicates the model is able to correctly identify 1% of the instances who are interested in Vehicle Insurance which is comparatively less than the one achieved in class 0.

Support:

Support indicates the count of actual occurrences of each class in a specified dataset. The class consists of highest support value i.e., 66699 that means that the dataset is imbalanced towards class 0 which highlights that there are more instances of customers who are interested in purchasing Vehicle Insurance than those who are interested in purchasing Vehicle Insurance which is indicated by class 1 having support value 9523.

F1 Score:

The F1 Score of class 0 highlights the performance of the model in identification of instances of customers who are not interested in Vehicle Insurance which is, however, higher i.e., 0.93 which indicates a good balance between precision and recall for class 0. This means it has identified 93% instances of customers who are not interested in Vehicle Insurance effectively while also minimizing both false positives and false negatives. However, the F1 Score of class 1 is 0.02 which indicates a poor balance between precision and recall for class 1.

Accuracy:

The model attained an accuracy of 88% which indicates that 88% of the predictions which have been made by the model are correct but accuracy has a limitation which is its biasedness towards the majority class which is class 0 in this case. Since it cannot handle imbalanced data where one is dominated by the other class so it can lead to misleading conclusions when assessing the model's performance.

Hence, the XGBoost Model performs well in identifying the instances of customers who are not interested in Vehicle Insurance than identifying instances of customers who are interested in Vehicle Insurance.

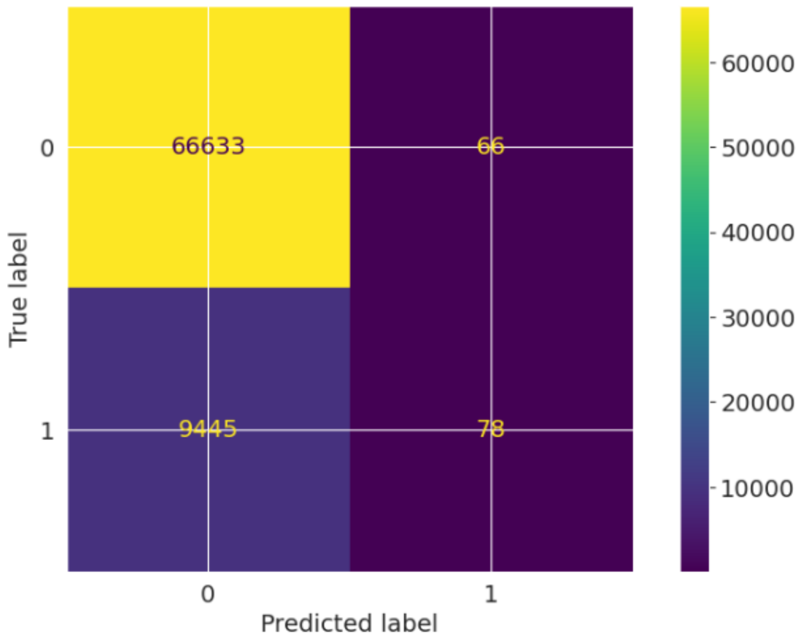


Figure 22: Confusion Matrix of XGBoost Model

The confusion Matrix shows that 66633 instances are correctly identified instances of label 0 and 78 instances are correctly identified instances of label 1. Also, 9445 instances are predicted as label 0 but are actually label 1 and 66 instances are actually label 0 but are predicted as label 1.

IV. Conclusion

XGBoost Model is the best model which has performed well in predicting the customers who are interested in Vehicle Insurance as compared to other models. Although Logistic Regression's accuracy is close to XGBoost Model's accuracy but XGBoost model can predict more instances of customers who are interested in Vehicle Insurance as compared to Logistic Regression.

V. References

[1] https://www.researchgate.net/publication/376958786_Analysing_Health_Insurance_Customer_Dataset_to_Determine_Cross-Selling_Potential

[2] https://www.analyticsvidhya.com/blog/2021/10/building-an-end-to-end-logistic-regression-model/

[3] https://www.ibm.com/topics/decision-trees

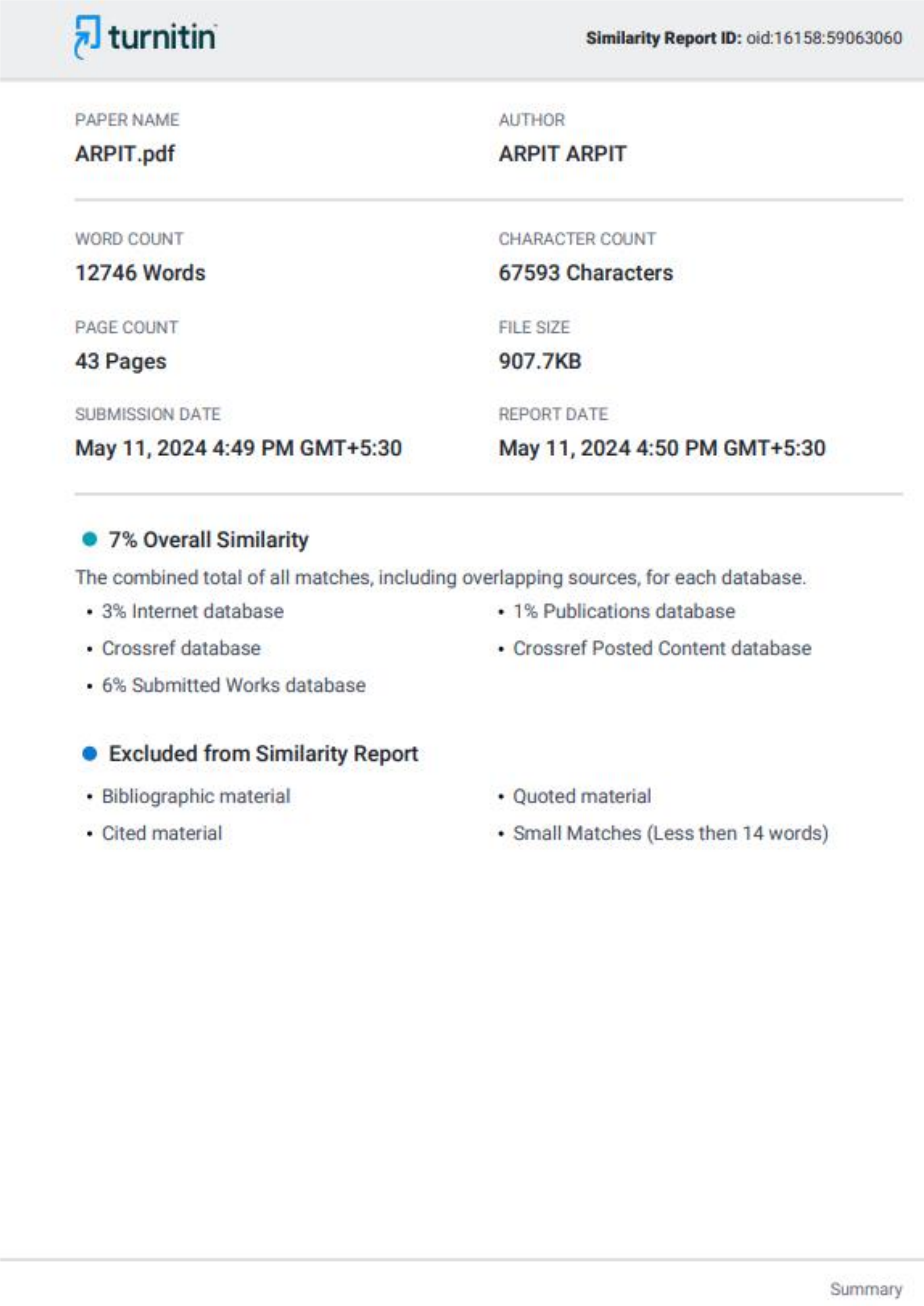
[4] **https://www.investopedia.com/terms/l/lifeinsurance.asp#toc-who-needs-life-insurance**

[5] **https://www.investopedia.com/terms/l/lifeinsurance.asp#toc-benefits-of-life-insurance**

[6] **https://medium.com/@danyal.wainstein1/understanding-the-confusion-matrix-b9bc45ba2679**

[7] **https://irdai.gov.in/**

Plagiarism Report



turnitin

Similarity Report ID: oid:16158:59063060

PAPER NAME

ARPIT.pdf

AUTHOR

ARPIT ARPIT

WORD COUNT

12746 Words

CHARACTER COUNT

67593 Characters

PAGE COUNT

43 Pages

FILE SIZE

907.7KB

SUBMISSION DATE

May 11, 2024 4:49 PM GMT+5:30

REPORT DATE

May 11, 2024 4:50 PM GMT+5:30

● **7% Overall Similarity**

The combined total of all matches, including overlapping sources, for each database.

- 3% Internet database
- Crossref database
- 6% Submitted Works database
- 1% Publications database
- Crossref Posted Content database

● **Excluded from Similarity Report**

- Bibliographic material
- Cited material
- Quoted material
- Small Matches (Less then 14 words)

Summary

Amity Institute of Applied Sciences

Weekly Progress Report (WPR)

For the week commencing: 02/01/2024-08/01/2024.

WPR: 1

Enrollment Number: A4479222033

Program: Master of Statistics (2022-2024)

Student Name: Arpit Saxena

Faculty Guide's Name: Dr. Bavita Singh

Project Title: Statistical Analysis of Vehicle Insurance Policy

Target For the Week: Research work on Insurance

Achievements: Research work on Insurance

Future Work Plans: Find Trend and patterns

Amity Institute of Applied Sciences

Weekly Progress Report (WPR)

For the week commencing: 09/01/2024-15/01/2024.

WPR: 2

Enrollment Number: A4479222033

Program: Master of Statistics (2022-2024)

Student Name: Arpit Saxena

Faculty Guide's Name: Dr. Bavita Singh

Project Title: Statistical Analysis of Vehicle Insurance Policy

Target For the Week: Find trends and patterns in dataset

Achievements: Find trends and patterns in dataset

Future Work Plans: Analyse the dataset

Amity Institute of Applied Sciences

Weekly Progress Report (WPR)

For the week commencing: 16/01/2024-22/01/2024.

WPR: 3

Enrollment Number: A4479222033

Program: Master of Statistics (2022-2024)

Student Name: Arpit Saxena

Faculty Guide's Name: Dr. Bavita Singh

Project Title: Statistical Analysis of Vehicle Insurance Policy

Target For the Week: Pre-processing of Data has to be done

Achievements: Pre-processing of Data is done

Future Work Plans: Perform Data Transformation

Amity Institute of Applied Sciences

Weekly Progress Report (WPR)

For the week commencing: 23/01/2024-29/01/2024.

WPR: 4

Enrollment Number: A4479222033

Program: Master of Statistics (2022-2024)

Student Name: Arpit Saxena

Faculty Guide's Name: Dr. Bavita Singh

Project Title: Statistical Analysis of Vehicle Insurance Policy

Target For the Week: Data Transformation

Achievements: Data Transformation

Future Work Plans: Research Work about Vehicle Insurance Policy

Amity Institute of Applied Sciences

Weekly Progress Report (WPR)

For the week commencing: 30/01/2024-05/02/2024.

WPR: 5

Enrollment Number: A4479222033

Program: Master of Statistics (2022-2024)

Student Name: Arpit Saxena

Faculty Guide's Name: Dr. Bavita Singh

Project Title: Statistical Analysis of Vehicle Insurance Policy

Target For the Week: Research Work about Vehicle Insurance Policy

Achievements: Research Work about Vehicle Insurance Policy

Future Work Plans: Splitting of Data into train and test data.

Amity Institute of Applied Sciences

Weekly Progress Report (WPR)

For the week commencing: 06/02/2024-12/02/2024.

WPR: 6

Enrollment Number: A4479222033

Program: Master of Statistics (2022-2024)

Student Name: Arpit Saxena

Faculty Guide's Name: Dr. Bavita Singh

Project Title: Statistical Analysis of Vehicle Insurance Policy

Target For the Week: Splitting of Data into train and test data

Achievements: Splitting of Data into train and test data

Work Plans: Handling Text and Categorical Attributes and Normalization.

Amity Institute of Applied Sciences

Weekly Progress Report (WPR)

For the week commencing: 13/02/2024-19/02/2024.

WPR: 7

Enrollment Number: A4479222033

Program: Master of Statistics (2022-2024)

Student Name: Arpit Saxena

Faculty Guide's Name: Dr. Bavita Singh

Project Title: Statistical Analysis of Vehicle Insurance Policy

Target For the Week: Handling Text and Categorical Attributes and Normalization.

Achievements: Handled Text and Categorical Attributes and Normalization.

Future Work Plans: Visualization of Data in Python to find some relationship between them.

Amity Institute of Applied Sciences

Weekly Progress Report (WPR)

For the week commencing: 20/02/2024-26/02/2024.

WPR: 8

Enrollment Number: A4479222033

Program: Master of Statistics (2022-2024)

Student Name: Arpit Saxena

Faculty Guide's Name: Dr. Bavita Singh

Project Title: Statistical Analysis of Vehicle Insurance Policy

Target For the Week: Visualization of Data in Python to find some relationship between them

Achievements: Visualized data in Python and found relationship between them.

Future Work Plans: Research Work on Different types of Models.

Amity Institute of Applied Sciences

Weekly Progress Report (WPR)

For the week commencing: 27/02/2024-04/03/2024.

WPR: 9

Enrollment Number: A4479222033

Program: Master of Statistics (2022-2024)

Student Name: Arpit Saxena

Faculty Guide's Name: Dr. Bavita Singh

Project Title: Statistical Analysis of Vehicle Insurance Policy

Target For the Week: Research Work on Different types of Models.

Achievements: Research Work on Different types of Models.

Future Work Plans: Random Forest Classifier is used.

Amity Institute of Applied Sciences

Weekly Progress Report (WPR)

For the week commencing: 05/03/2024-11/03/2024.

WPR: 10

Enrollment Number: A4479222033

Program: Master of Statistics (2022-2024)

Student Name: Arpit Saxena

Faculty Guide's Name: Dr. Bavita Singh

Project Title: Statistical Analysis of Vehicle Insurance Policy

Target For the Week: Random Forest Classifier is used.

Achievements: Random Forest Classifier is used.

Future Work Plans: Research about XGBoost Model.

Amity Institute of Applied Sciences

Weekly Progress Report (WPR)

For the week commencing: 12/03/2024-18/03/2024.

WPR: 11

Enrollment Number: A4479222033

Program: Master of Statistics (2022-2024)

Student Name: Arpit Saxena

Faculty Guide’s Name: Dr. Bavita Singh

Project Title: Statistical Analysis of Vehicle Insurance Policy

Target For the Week: Research about XGBoost Model.

Achievements: Research about XGBoost Model.

Future Work Plans: To make use of XGBoost Model

Amity Institute of Applied Sciences

Weekly Progress Report (WPR)

For the week commencing: 19/03/2024-25/03/2024.

WPR: 12

Enrollment Number: A4479222033

Program: Master of Statistics (2022-2024)

Student Name: Arpit Saxena

Faculty Guide's Name: Dr. Bavita Singh

Project Title: Statistical Analysis of Vehicle Insurance Policy

Target For the Week: To make use of XGBoost Model

Achievements: To make use of XGBoost Model

Future Work Plans: Hypertuning of the XGBoost Model

Amity Institute of Applied Sciences

Weekly Progress Report (WPR)

For the week commencing: 26/03/2024-01/04/2024.

WPR: 13

Enrollment Number: A4479222033

Program: Master of Statistics (2022-2024)

Student Name: Arpit Saxena

Faculty Guide's Name: Dr. Bavita Singh

Project Title: Statistical Analysis of Vehicle Insurance Policy

Target For the Week: Hypertuning of the XGBoost Model

Achievements: Hypertuning of the XGBoost Model

Future Work Plans: To improve accuracy of the model

Amity Institute of Applied Sciences

Weekly Progress Report (WPR)

For the week commencing: 02/04/2024-08/04/2024.

WPR: 14

Enrollment Number: A4479222033

Program: Master of Statistics (2022-2024)

Student Name: Arpit Saxena

Faculty Guide's Name: Dr. Bavita Singh

Project Title: Statistical Analysis of Vehicle Insurance Policy

Target For the Week: To improve accuracy of the model

Achievements: Model is improved.

Future Work Plans: Decision Tree Model will be used.

Amity Institute of Applied Sciences

Weekly Progress Report (WPR)

For the week commencing: 09/04/2024-15/04/2024.

WPR: 15

Enrollment Number: A4479222033

Program: Master of Statistics (2022-2024)

Student Name: Arpit Saxena

Faculty Guide's Name: Dr. Bavita Singh

Project Title: Statistical Analysis of Vehicle Insurance Policy

Target For the Week: Decision Tree Model will be used.

Achievements: Decision Tree Model has been fitted and its accuracy has also been estimated.

Future Work Plans: Classification Report and Confusion Matrix of the Decision Tree Model.

Amity Institute of Applied Sciences

Weekly Progress Report (WPR)

For the week commencing: 16/04/2024-22/04/2024.

WPR: 16

Enrollment Number: A4479222033

Program: Master of Statistics (2022-2024)

Student Name: Arpit Saxena

Faculty Guide's Name: Dr. Bavita Singh

Project Title: Statistical Analysis of Vehicle Insurance Policy

Target For the Week: Classification Report and Confusion Matrix of the Decision Tree Model.

Achievements: Classification Report and Confusion Matrix of the Decision Tree Model has been achieved.

Future Work Plans: Fitting of Logistic Regression Model

Amity Institute of Applied Sciences

Weekly Progress Report (WPR)

For the week commencing: 23/04/2024-29/04/2024.

WPR: 17

Enrollment Number: A4479222033

Program: Master of Statistics (2022-2024)

Student Name: Arpit Saxena

Faculty Guide's Name: Dr. Bavita Singh

Project Title: Statistical Analysis of Vehicle Insurance Policy

Target For the Week: Fitting of Logistic Regression Model

Achievements: Logistic Regression Model has been fitted along with their Classification report and confusion matrix.

Future Work Plans: Comparison between different models

Amity Institute of Applied Sciences

Weekly Progress Report (WPR)

For the week commencing: 30/04/2024-06/05/2024.

WPR: 18

Enrollment Number: A4479222033

Program: Master of Statistics (2022-2024)

Student Name: Arpit Saxena

Faculty Guide's Name: Dr. Bavita Singh

Project Title: Statistical Analysis of Vehicle Insurance Policy

Target For the Week: Comparison between different models

Achievements: Comparison between different models is done

Future Work Plans: To conclude the project

Amity Institute of Applied Sciences

Weekly Progress Report (WPR)

For the week commencing: 07/05/2024-13/05/2024.

WPR: 19

Enrollment Number: A4479222033

Program: Master of Statistics (2022-2024)

Student Name: Arpit Saxena

Faculty Guide's Name: Dr. Bavita Singh

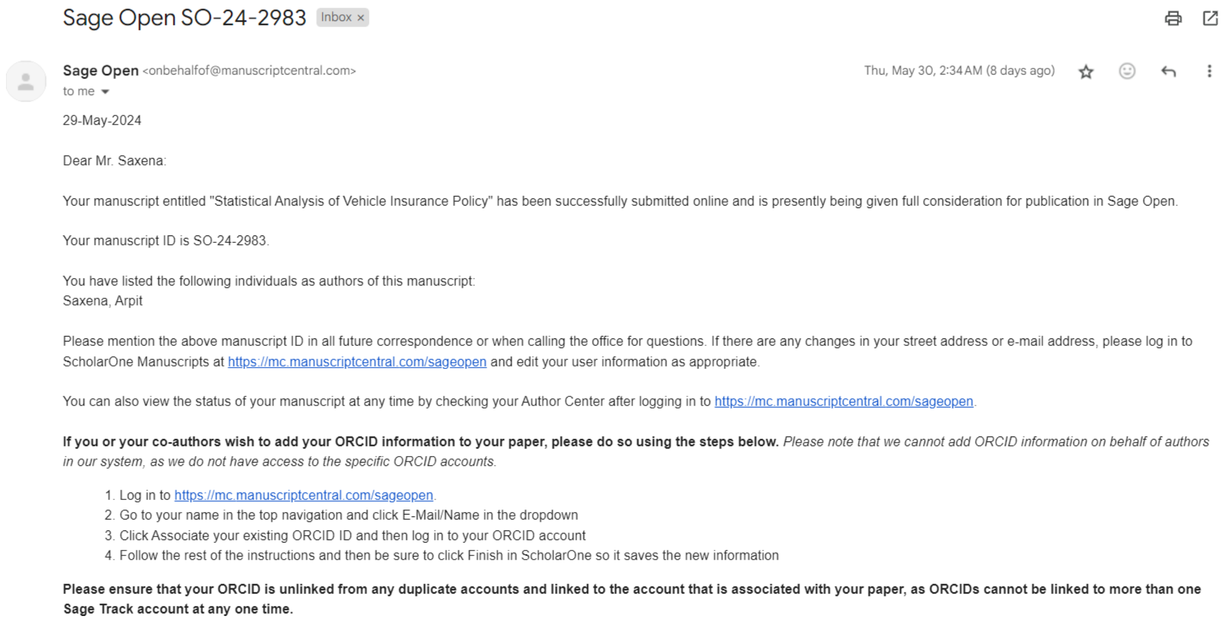
Project Title: Statistical Analysis of Vehicle Insurance Policy

Target For the Week: To conclude the project

Achievements: Concluded the Project

Future Work Plans: Submission of the Project

Communicated Paper in Sage Publications



Sage Open SO-24-2983

Sage Open <onbehalfof@manuscriptcentral.com>

Thu, May 30, 2:34 AM (8 days ago)

29-May-2024

Dear Mr. Saxena:

Your manuscript entitled "Statistical Analysis of Vehicle Insurance Policy" has been successfully submitted online and is presently being given full consideration for publication in Sage Open.

Your manuscript ID is SO-24-2983.

You have listed the following individuals as authors of this manuscript:
Saxena, Arpit

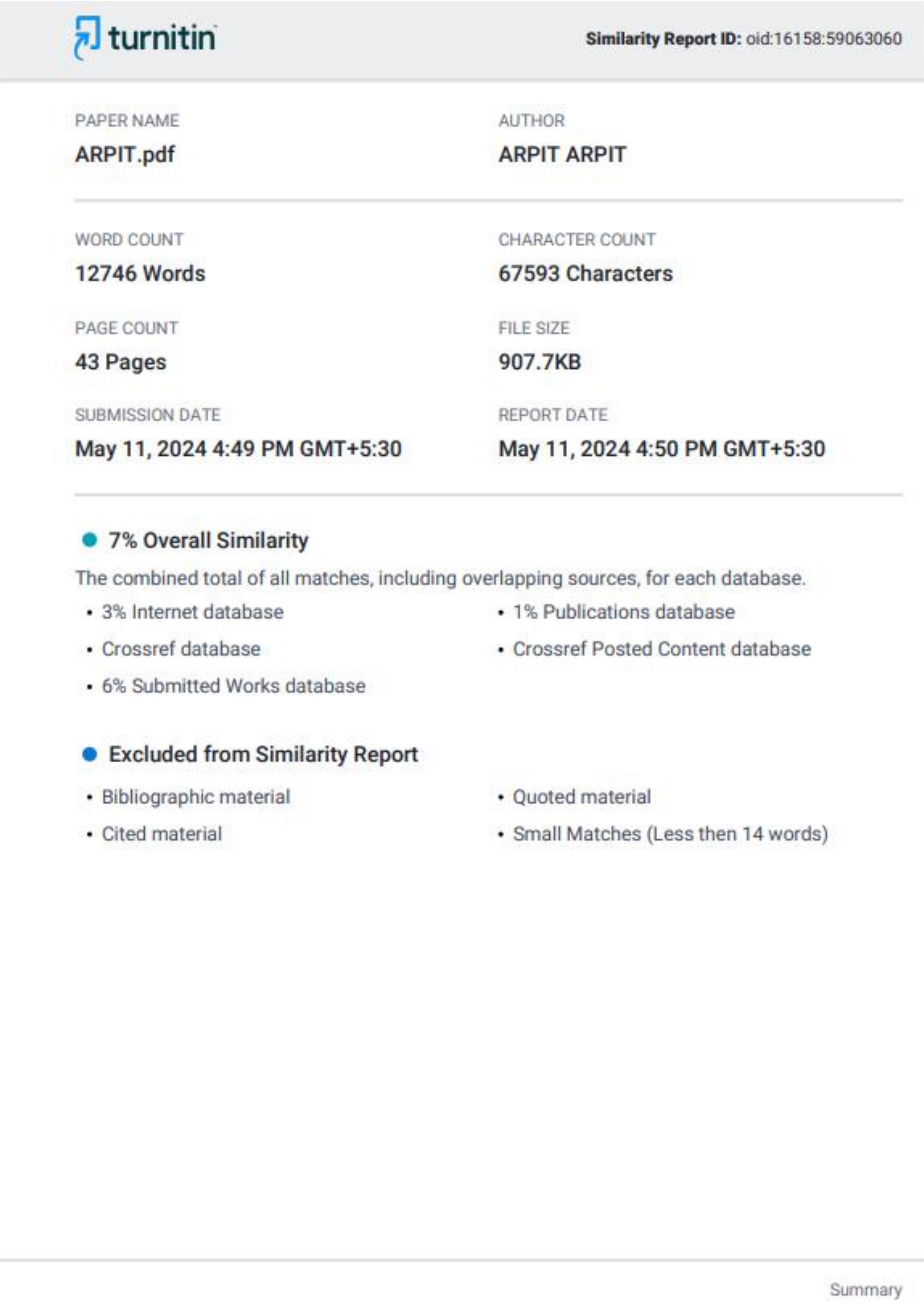
Please mention the above manuscript ID in all future correspondence or when calling the office for questions. If there are any changes in your street address or e-mail address, please log in to ScholarOne Manuscripts at https://mc.manuscriptcentral.com/sageopen and edit your user information as appropriate.

You can also view the status of your manuscript at any time by checking your Author Center after logging in to https://mc.manuscriptcentral.com/sageopen.

If you or your co-authors wish to add your ORCID information to your paper, please do so using the steps below. *Please note that we cannot add ORCID information on behalf of authors in our system, as we do not have access to the specific ORCID accounts.*

1. Log in to https://mc.manuscriptcentral.com/sageopen.
2. Go to your name in the top navigation and click E-Mail/Name in the dropdown
3. Click Associate your existing ORCID ID and then log in to your ORCID account
4. Follow the rest of the instructions and then be sure to click Finish in ScholarOne so it saves the new information

Please ensure that your ORCID is unlinked from any duplicate accounts and linked to the account that is associated with your paper, as ORCIDs cannot be linked to more than one Sage Track account at any one time.



turnitin

Similarity Report ID: oid:16158:59063060

PAPER NAME **ARPIT.pdf**	AUTHOR **ARPIT ARPIT**
WORD COUNT **12746 Words**	CHARACTER COUNT **67593 Characters**
PAGE COUNT **43 Pages**	FILE SIZE **907.7KB**
SUBMISSION DATE **May 11, 2024 4:49 PM GMT+5:30**	REPORT DATE **May 11, 2024 4:50 PM GMT+5:30**

● **7% Overall Similarity**

The combined total of all matches, including overlapping sources, for each database.

- 3% Internet database
- Crossref database
- 6% Submitted Works database
- 1% Publications database
- Crossref Posted Content database

● **Excluded from Similarity Report**

- Bibliographic material
- Cited material
- Quoted material
- Small Matches (Less then 14 words)

Summary

Abstract

This study shows the comparison between different Machine Learning Models to determine if a person with health insurance will prefer automobile insurance. In the insurance sector, Accurate prediction models are used for risk assessment, policy pricing and decision-making due to complexity and volume of insurance data that is expanding. Using Health Insurance dataset, the study shows the prediction abilities of Decision trees, Logistic Regression and Random Forest Models. The dataset consists of numerous demographic and categorical variables for which missing values are being handled and then categorical variables are converted into numerical values during pre-processing stage. After pre-processing stage, the model is being used for training using decision trees, logistic regression and random forest model for which the model's accuracy is being estimated. For improving model's accuracy, hyper tuning is done to find the best estimators for the model which are being applied to increase its accuracy. Then the classification report is being made to determine the precision score, recall score and f1 score of the model. Confusion matrix has also been made for the models. The project's results will tell which model is more accurate and predict who would purchase vehicle insurance for the business, among other things. The comparison analysis will show the benefits and drawbacks of each model, empowering insurers to choose the algorithm that best suits their unique requirements.

I. INTRODUCTION

The practice of insurance companies which offers additional insurance products to its existing customers is referred to as Health Insurance cross-selling which plays major role in the growth of a business and profitability of insurance companies. The aim of this research is to develop a predictive model which can help us to determine if a person is interested in purchasing vehicle insurance. Various factors like demographics, socio-economic factors and demographics have been analysed to identify patterns which can help for prediction of interest of a customer to purchase the vehicle insurance.

This research holds an important purpose in strengthening the marketing strategies of health insurance companies. Insurers can identify the interest of their customers more accurately by predicting the interest of vehicle insurance which can further help in improving customer acquisition and retention rates which will result in increasing business revenue.

Also, the key findings of this study hold valuable resource especially for health insurance providers to increase their customers by analysing customer behaviour patterns which will determine when the customer will be interested in purchasing the vehicle insurance. With the help of this deep understanding of this behaviour, they can make data-driven strategies in order to adapt product offerings, enhance marketing strategies and create customized services that helps in meeting the needs and preferences of their customers which also helps in boosting satisfaction of customer as well as build strong client relationships which will further contributes towards the growth and sustainability of the insurance industry.

Key Components:

Data Analysis and Integration: In order to obtain important insights regarding risk patterns, claim frequencies, and severity distributions, the vast amounts of historical data needed to be analysed and then combine different data sources. Key risk indicators will be uncovered and reliable predictive models will be created with the help of data-driven methodology.

Risk segmentation and profiling: The policyholders will be divided into homogeneous risk groups according to pertinent characteristics including past claims history, demographics and lifestyle with the help of sophisticated statistical methods. By this fine-grained segmentation, it will be easier for creating customised pricing schemes and risk profiles.

Pricing optimisation: Pricing models, which reflects the risk involved in each group of policyholders appropriately, will be created through the framework for risk segmentation as a foundation. These models will be used to take into account variables which included anticipated claim amounts, expense considerations and the chance of loss so that competitive and profitable premium levels can be determined.

Regulatory Compliance: Our project will include regulatory requirements so that the established pricing models are compliant with pertinent laws, rules and standards.

It is necessary to analyse and update pricing models continuously due to the ongoing shift in the insurance markets. A structure which can routinely assess model performance, monitor the new risk trends and can incorporate fresh data will be set so that the dependability and precision of the pricing models can be improved.

Therefore, the project's results will be considered highly advantageous to the insurance firms and clients. Through enhancement of risk assessment accuracy and pricing fairness, the insurers can achieve better alignment between premiums and risk exposure which would further increase financial stability and profitability. Additionally, in order to improve affordability and satisfaction, the policyholder will gain from more specialised insurance products with rates which can take into account their unique risk profiles.

With the growing availability of data and development in analytics, the risk assessment and pricing procedures which are used by the insurance business have the potential to be completely transformed. We seek to lay groundwork for insurers for optimisation of pricing strategies and promotion of just and open insurance markets by combining data-driven insights, compliance consideration and cutting-edge modelling approaches.

A. Overview of Insurance and the techniques of data science in insurance industry:

The insurance sector is very crucial for offering both consumers and corporations financial security and also for risk management. It often covers industries which includes life, property, health, casualty and other insurances. Large volumes of information regarding claims, policyholders, market trends, risk factors and external variables are gathered by insurance companies. In order to improve their operations and decision-making process, substantial number of opportunities exist for insurers which results in expanding availability of this data and development in data science.

Data Science is referred to the practice of mining data to gather information and insights using scientific algorithms, techniques and tools which is important in the insurance sector due to having potential to solve many major problems which often produce beneficial results that is why data science is important for many reasons.

Risk Assessment: The risk analysis is often useful to the insurance industry since it can help to analyse large amounts of data, highlight pertinent risk indicators, and further useful for development of predictive models which allow for more accurate risk assessment which improves pricing plans, profitability and underwriting judgements.

Fraud detection: Insurance fraud is a major concern as it causes loss to insurers financially. The data science approaches such as pattern recognition, anomaly detection and predictive modelling can be useful for spot fraudulent claims and enhancement of fraud detection and prevention systems.

Customer segmentation and personalization: Data Science plays a huge role for insurers by segmenting clientele into several groups by taking account traits like demographics, behaviours or risk profiles which is helpful for insurers since it can offer targeted marketing techniques, adjusted pricing and personalised insurance policies for increment of customer satisfaction and retention by knowing consumer categories.

B. Types of Insurance

i. Life Insurance:

A contract formed between an insurance company and a policy holder on a condition that the insurer agrees to pay money to one or more beneficiaries in exchange of the premium which is paid by the insured person before he dies. The companies having the best life insurance contains good financial strength, high customer satisfaction, many policy types and low customer complaints.

Types of Life Insurance: There are many types of life insurances such as Term Life Insurance, Whole Life Insurance, Universal Life Insurance, Variable Life Insurance and many more having unique characteristics, advantages and premium arrangements in each type.

Factors Affecting Life Insurance Premiums and Costs: There are many factors which can affect the cost of life insurance premiums such as age, health status, smoking status, gender, occupation and lifestyle, family medical history, etc.

Components of Life Insurance: There are two components of Life Insurance which are death benefit and premium while permanent or whole life insurance also has a component namely cash value.

Death Benefit: It is termed as the amount of money which is guaranteed by the insurance company to the beneficiaries which can be identified in the policy when the insured person dies. The insured person might be parent while beneficiaries might be their children. On the basis of the beneficiaries estimated future needs, the insured person chooses the desired death benefit amount. The insurance company determines if there is any insurable interest and if the proposed insured person will qualify for the coverage based on the underwriting requirements of the company which is related to health, age and many more.

Premium: These are the money which is paid by the policyholder for the insurance. If the policyholder pays the required premiums, which are estimated by how likely the insurer will have to pay policy's death benefit on the basis of the life expectancy of the insured person, then the insurer should pay the death benefit when the insured person dies. There are various factors which can influence the life expectancy such as the insured person's age, health, gender, medical history, occupation and lifestyle, and many more.

Cash Value: There are purposes of the cash value of permanent life insurance. It is referred to as the savings account which is used by the policyholder during the insured person's life. Depending upon the use of the money, some policies consist of the restrictions on withdrawals.

ii. Health Insurance:

It is defined as a type of insurance which covers either whole or some part of the risk including medical expenses. Through estimation of overall risk of health risk, an insurer can plan and develop a finance structure such as monthly premium in order to provide money which can paid for the health care benefits by which risk is shared among many individuals which is specified in the insurance agreement. It is also referred to a coverage which provides benefits through payments to an injury or sickness which also includes coverage of losses from medical expense, accidents, disability and dismemberment.

However, the obligations of an insured person can occur in several forms such as:

- **Premium:** It is the amount that the policyholder should pay for their health coverage to the health plan so that their expenses can be paid in case of any medical situation or emergency. According to healthcare law, the premium is estimated by taking some factors into consideration which includes age, location, use of tobacco, enrolment of an individual or a family and also the type of plan, insurer choses. Tax credit is paid by the government in order to cover part of the premium for the persons who purchases private insurance through insurance marketplace under the Affordable Care Act.
- **Deductible:** Before the insurer pay its share for their health care, the insured should pay an amount which is out of pocket is termed as deductible. For instance, if a policy holder has to pay $5000 deductible per year before the health insurer pays for their health care then it might take several doctor's visit before the insured person reaches the deductible and payment is proceeded by the insurance company. Moreover, co-pays for doctor's visits against the deductible are not applied to many policies.
- **Co-payment:** Before the health insurer pays for a particular visit, the insured person has to pay an out-of-pocket amount which is termed as co-payment.
- **Coinsurance:** It is referred to as a percentage of the total cost which an insurance person should also pay instead of just paying a fixed amount of co-payment.
- **Exclusions:** Some billed items such as use-and-throw, taxes, etc. are the ones which gets excluded from admissible claim. So, all the services are not covered since the insurer are expected to pay for the non-covered services on their own.
- **Coverage Limits:** There is a limit that some policies related to health insurance pay for the health care up to a certain amount which are coverage limits. The insured person should pay any charge which arises after the health plan reaches the benefit maximum since it will stop further payment and the policy-holder will have to pay for the remaining costs.

- **Out-of-pocket maximum:** The insured person reaches the out-of-pocket maximum when their payment obligation ends and health insurance pays all other covered costs. Out-ofpocket maximum is generally limited to some specific benefits or can be applied to all the coverages which are provided during a specific benefit year.

iii. Non-Life Insurance

It is often known as the general insurance which is also a category of insurance which protects people, companies as well as organisations from many risks and losses other than the ones which are related to health or lives. Non-life Insurance helps in covering many assets, liabilities and property which are useful for offering protection against disease or death. It has many important features which are as follows:

- **Protection Types:** Risks such as theft, accident, liability claims, natural disasters, property damage, and other events which occurs in the life of an insured are being protected by the non-life insurance. Some types of non-life insurance are home insurance, liability insurance, auto insurance, travel insurance, marine insurance and commercial property insurance.
- **Premiums:** The insurance provider is paid the required premium by the policyholder for the purpose of keeping their non-life insurance active. There are only a few factors which affect premium costs such as risk factors, claim history of the policyholder, kind of coverage, insured sum and deductible option.
- **Policy Limits:** There is a limit known as coverage limit which highlights the maximum sum that will be covered by the insurance company for the insured person. If the amount reaches the coverage limits, then the insurance company will stop covering for further payments and the insured person has to pay for the remaining losses. Depending on the type of coverage and terms of the policy, the policy for the coverage limits might change but it is advisable for the insured person to carefully evaluate the limits for the future.
- **Deductibles:** This is the sum which is not covered by the insurance company and the policyholder has to pay the amount on his own and are included in non-life insurance in which reduced deductibles leads to higher premiums while bigger deductibles lead to reduced premiums.
- **Claim Procedure:** In the claiming procedure, the policyholders have to submit a claim to their insurance provider in covered loss. They must report the loss, provide the required documents and information and collaborate with the insurance provider for evaluation of the severity of

the damage and then establish amount which has to be claimed. There are particular steps and deadlines which insurance companies follows when they process claims.

- **Benefits:** There are several benefits in a non-life insurance policy which includes the financial help which is provided at any medical emergency, compensation is paid by the third party in case of damages on property or life, etc. It can cover damages of the residential property such as fire, natural calamities, riots as well as burglary. It can cover insurance coverage of senior citizens as well as children along with issues such as accidents, loss of any documents and baggage, etc in a foreign land. It also benefits the businesses with policies such as shopkeepers' insurance, property and marine insurance, benefits insurance, etc.

C. IRDAI

IRDAI stands for Insurance Regulatory and Development Authority of India which is a regulatory agency that is responsible for monitoring and controlling the Indian insurance market which was mainly created in accordance with the Insurance Regulatory and Development Authority Act of 1999 for defending the interests of policyholders and encouraging of the expansion and development of the nation's insurance sector. IRDAI's responsibilities can b described in the insurance industry:

Control and Regulation: IRDAI is termed as the principal regulating body for the insurance industry in India. The regulations, standards and the policies that are developed and enforced by IRDAI governs the conduct and operations of insurance companies and other market participants. The authority ensures adherence to these rules for the purpose of keeping the insurance market transparent, equitable and stable.

Registration and licensing: The registering and licensing of insurance companies, agents, surveyors, brokers and other businesses which are involved in the insurance industry are the responsibilities of IRDAI. It also plays a key role in establishment of eligibility standards, focus on the insurance businesses' solvency and soundness, and ensures that only accredited and qualified parties carry out operations which are related to insurance.

Protection of Policyholder Interests: The IRDAI is essential for protecting the interests of the Policyholder. It develops standards for disclosure norms, ethical behaviour and consumer protection measures for the purpose of ensuring insurance products being available, open and advantageous to policyholders. Complaints are processed by IRDAI from the policyholders and forum is also offered to resolve them using integrated grievance management system.

Market Development: The expansion and development of the Indian insurance market as well as the rivalry, industry innovation and product variety has been promoted by the IRDAI and safeguarding the interests of policyholders. IRDAI works for stabilising the market and also monitor market trends so that an appropriate action can be taken which also encourages healthy competition and long-term growth in the insurance industry.

Financial Regulation: The financial operations are controlled by IRDAI to safeguard the solvency and financial stability of insurance businesses. It also established investment rules, capital requirements and risk management for insurers for keeping the insurance sector financially stable. It is also useful in evaluating the insurance industry's financial standing, performing audits and keeping track of their adherence to accounting rules.

Market Conduct: IRDAI focuses on the behaviour of insurance companies, intermediaries and agents for ensuring the ethical practises and equitable treatment of policyholders. It also develops standards for sales practises, regulations and code of conduct for eliminating the fraud and unfair practices in the insurance market. If the organisation fails to follow these rules, then investigation, inspections and sanctions are being conducted by IRDAI against that organisation.

For promotion of financial inclusion, IRDAI encourages insurers for providing affordable and accessible insurance products to the underprivileged segments of the society. It encourages micro-insurance projects and also implements the government-backed insurance systems for social welfare which also makes easier for expansion of insurance services in rural and isolated areas.

International partnership: IRDAI is active in participating in international cooperation partnership with other regulatory bodies and organisations for sharing expertise, advance cross-border regulatory harmonisation, and stay current on worldwide insurance practices. It also takes part in international forums, seminars and projects for supporting the expansion and growth of the insurance sector worldwide.

D. Pricing – Health Insurance Product

Estimation of premium costs is important for the consumer of a health insurance plan which would incur in return for coverage. The adequacy of the premiums is tested in order to cover the risks and projected expenditures for delivering services related to healthcare by taking number of variables into account for the pricing procedure. There are several key factors which determine the cost of health insurance plan that are as follows:

Claim Experience and Historical Data: The claims experience and historical data are analysed by insurers so that the cost patterns and trends can be comprehended. The claim sums, claim occurrence rates and the cost which is distributed across various demographics and the medical conditions are being examined for this process. In order to determine predicted costs, and establish suitable premium rates, the historical data has been useful for it.

Medical Cost Inflation: The Health Insurance costs depends on the cost of medical services, treatments and prescriptions medications. Insurers takes estimated rates of inflation in medical costs into account to ensure premiums are high so that the anticipated costs are paid during the policy period. Healthcare market dynamics, economic variables and governmental regulations are taken into account when future cost trends have to be projected.

Regulatory Requirements: In order to ensure fairness and consumer protection, the health insurance pricing is subject to regulatory oversight. The insurance regulators might also impose restrictions or guidelines on profit margins, rate filling processes and premium rates. While pricing their products, insurers should comply with these regulations.

Competitive Market Analysis: When premium rates are set, then the competitive landscape and market dynamics are taken into account by the insurers who also evaluate the pricing strategies of competitors, market demand, and consumer preferences. The insurers also balance the competitive pricing with adequate coverage and profitability.

For building price models and forecast future claim costs such as credibility theory, loss reserving and statistical modelling, actuarial methods are being used which also assist insurers to make adjustments for future uncertainty along with projection of the anticipated claims experience on the basis of the historical data.

II. Literature Review

This study has been made using existing research about the prediction of interest of a customer purchasing vehicle insurance with the help of machine learning models in the insurance industry. By identifying the interests of a person for vehicle insurance and gain a deep understanding of the pattern of the customer’s purchase, insurers can enhance their marketing strategies and enhance their business revenue. The insurance market has been highly competitive so accurate prediction has become crucial for the insurers for the growth of the industry so with the arrival of predictive analytics and machine learning models, insurers now have the opportunity to make data-driven decisions with the predictive analytics which helps in enhancing their business. Furthermore, with the help of machine learning algorithms, the insurance company has now got the opportunity to analyse large health insurance customer datasets and identify more patterns and insights which may not be possible through traditional methods.

A. Predictive Modelling Techniques

Many studies have explored the different predictive modelling techniques and determine its effectiveness to identify its potential in prediction. An Algorithm for gradient boosting has been made into use to predict the customers who wanted to purchase vehicle insurance based on their historical health claims and demographics. The results showed that the ensemble methods performs better than traditional Decision Tree models in terms of accuracy so making use of these ensemble methods helps us in improving the accuracy of the model for prediction. The models demonstrated the prediction by identifying potential customers who are interested in purchasing vehicle insurance by taking customer attributes and historical data into consideration. Similarly, ensemble methods are being applied to Random Forest Model and Logistic Model as well to improve their performance in prediction. The comparison drawn between the models helps in highlighting the importance of feature engineering and model selection in prediction more accurately.

B. Imbalance Data Handling

Imbalance data is a very common challenge in any analysis where some instances of a class or classes outweigh the other classes. When the distribution of classes in a dataset is highly skewed on class have more instances than the other class then the dataset is said to be imbalanced. Due to lack of adequate training samples, it becomes difficult for the classifier to identify patterns and predict unbiased outcomes for the minority class. This problem is resolved using various techniques such as oversampling which is used to increase the number of instances in the minority class. Different evaluation metrics are also used to check the class imbalance. Their findings help in handling imbalanced data to obtain unbiased predictions.

C. Customer Segmentation

Customer Segmentation is a way to enhance the effectiveness of the models in which customers are segmented based on their characteristics and behaviours. Using unsupervised clustering techniques to separate customers with similar profiles can give better predictions of the models with more accuracy by identifying distinct customer segments which enables more targeted strategies and then they built predictive models separately for each cluster which leads to more accurate predictions by capturing distinct purchasing behaviours with different customer segments.

D. Customer Satisfaction

Customer Satisfaction is a critical factor in the analysis since satisfied customers would more likely to purchase insurance later and also recommend others leading to more growth of the company. In the study, sentiment analysis can be used to consider customer reviews and feedback to identify customer satisfaction levels. Machine Learning Techniques were used further to identify the patterns in customer behaviour and their satisfaction levels.

The usage of prediction in health insurance has been advanced significantly through various machine learning techniques. Researchers have highlighted challenges like imbalanced data, customer segmentation and satisfaction to improve the accuracy and effectiveness of the models which results in growth and profitability of the health insurance companies.

The study is based on the existing knowledge to contribute more significantly in prediction by not only highlighting the challenges faced during prediction but to also provide valuable insights and practical solutions to the challenges to overcome it which will lead to enhancing the accuracy of the model which will further improve the growth of the insurance company.

E. Performance Evaluation Matrix

i. Accuracy:

It refers to the most basic and intuitive metric which represents the correctly predicted instances out of the total. It is not suitable for the imbalanced datasets since it can lead to misleading results so it is suitable mainly for balanced datasets.

ii. Precision:

It is used for measuring the proportion of correctly predicted positive instances which are predicted as positive. It is mainly concerned with the accuracy of positive predictions and is mainly useful when the cost of false positives is high. It is calculated as true positives divided by the sum of true positives and false positives.

iii. Recall:

It is used for calculating the proportion of correctly predicted positive instances out of the actual positive instances. It measures the model's ability to identify all the positive instances and is valuable when the cost of false negatives is high. Recall is calculated as true positives divided by the sum of true positives and false negatives.

iv. F1 Score:

The F1 score is the harmonic mean of precision and recall. It provides a balanced measure that combines both metrics and is useful when there is an imbalance between precision and recall. The F1 score considers both false positives and false negatives and is calculated as

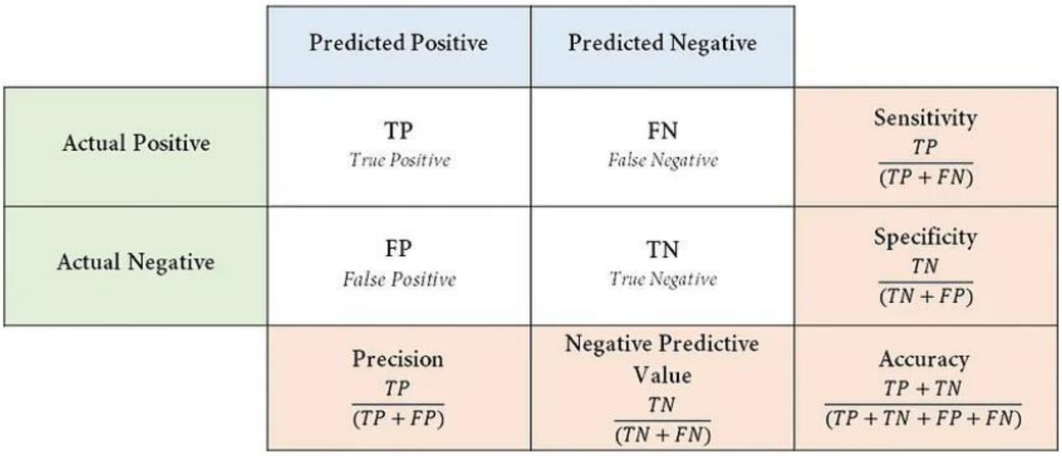
$$\frac{2 * (\text{precision} * \text{recall})}{(\text{precision} + \text{recall})}.$$

v. Specificity (True Negative Rate):

Specificity calculates the proportion of correctly predicted negative instances out of the actual negative instances. It focuses on the accuracy of negative predictions and is useful when the cost of false negatives is high. Specificity is calculated as true negatives divided by the sum of true negatives and false positives.

vi. Confusion Matrix:

The general idea is that it counts the number of times the classifier identifies the positive class as negative and vice versa. In this matrix the row represents the actual class and the column represents the predicted one. This metric is used to compare the number of predictions for each class that are incorrect and those that are correct. Confusion matrix is shown below:



	Predicted Positive	Predicted Negative	
Actual Positive	TP *True Positive*	FN *False Negative*	Sensitivity $\frac{TP}{(TP + FN)}$
Actual Negative	FP *False Positive*	TN *True Negative*	Specificity $\frac{TN}{(TN + FP)}$
	Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

Figure 1: Confusion Matrix

vii. Machine Learning Models

a. Logistic Regression:

Logistic Regression is termed as a statistical technique which is used for estimation of likelihood that an event will occur while Regression analysis which is termed as a class of statistical techniques that is used to represent the relationship between a dependent and one or more independent variable which is what this type of study is.

In this project, the dependent variable is a categorical variable having only two possible values namely ‘yes’ or ‘no’, ‘true’ or ‘false’, or ‘male’ or ‘female’. Both continuous or categorical variables can be used as independent variable. The connection between dependent and independent variable is modelled using logistic function which has a Sshaped curve and is also a sigmoid function. The probability is shown in a curve in the given figure below. The Maximum Likelihood Estimation (MLE) technique is useful for estimation of logistic regression model and the values of the model parameters which maximise the likelihood of the observed data are found with the help of MLE approach. After estimation of the logistic model, it is useful for forecasting the likelihood that additional data points will contain the dependent variable. With the usage of anticipated probability, a person can take a decision of whether a particular event is likely to happen or not.

The graph of sigmoid function is given below:

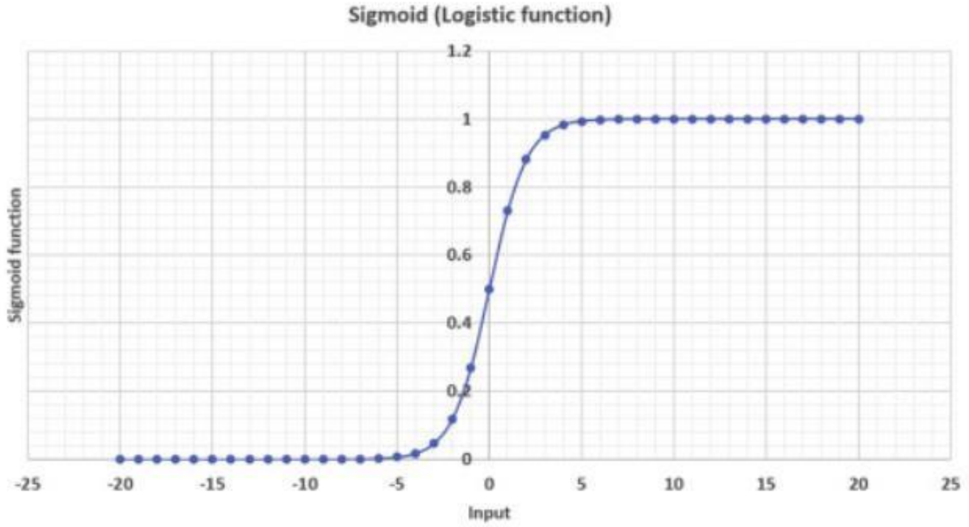


Figure 2: Sigmoid Function in Logistic Regression

b. Decision Tree Model

A supervised learning algorithm that is non-parametric and is used for both regression and classification tasks is termed as a decision tree which has a hierarchical tree structure consisting of an internal node, root nodes, branches and leaf nodes.

In the figure given below, it begins with a root node which do not contain any incoming branches but its outgoing branches feed into internal nodes which is also known as decision nodes. On the basis of the features available, both types of nodes are used for conducting the evaluations so that homogeneous subsets can be formed which are often denoted by leaf nodes or terminal nodes. All the possible outcomes within the dataset are represented by the leaf nodes.

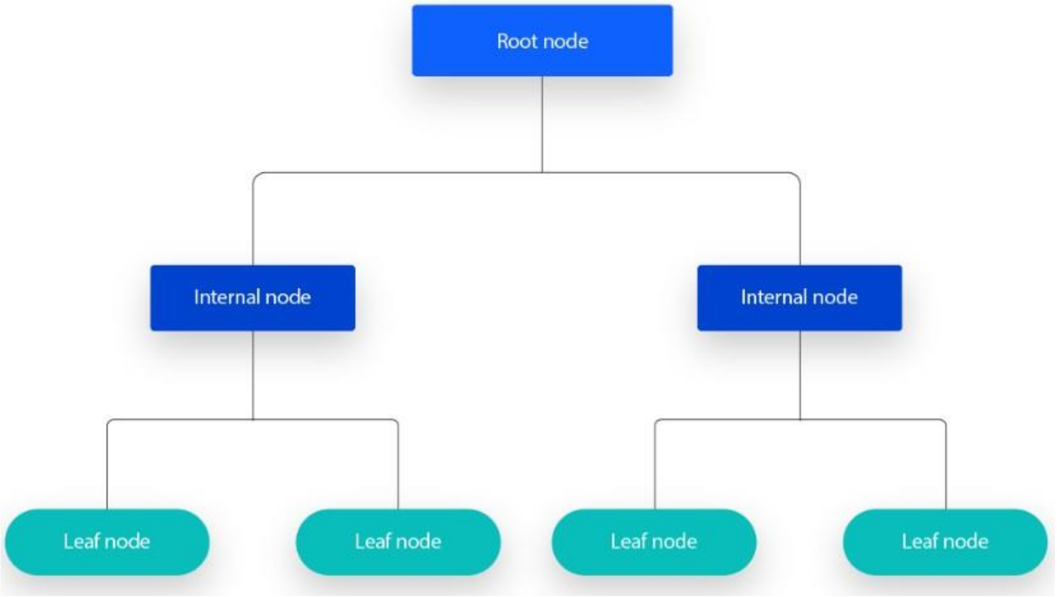


Figure 3: Decision Tree

Since Decision tree classifiers have the capacity to handle numerical and categorical data as well as recording the intricate decision rules along with its interpretability, they are considered as a common machine learning technique which are used in applications of health insurance. How special decision trees are and how ideal they are for health insurance will be discovered in this article.

Decision trees provides a simple and understandable illustration of the decision-making process. The decision tree structure has been useful for simplicity of all parties involved such as policyholders and insurance experts in order to comprehend the variables which affect the outcomes of health insurance. Decision rules in the decision tree provide the trust and transparency which are simple to understand and communicate.

The datasets on health insurance includes both numerical and categorical variables. Decision trees can handle both types of data without the need for specific feature engineering and encoding. The algorithm naturally accommodates them by dividing data into groups according to the different values of the categorical variables. In order to partition the data for numerical variables, decision trees can be used to determine the most informative thresholds.

Decision trees captures the non-linear relationships between features and the goal variable. Numerous variables which frequently interacts in complicated ways have an impact on health insurance outcomes. Due to their capacity to capture and model such interactions, decision trees are often well suitable for spotting complex patterns and hidden links in the data.

Decision trees provided a measure of feature importance which helps in quantifying the relative value of various factors in forecasting outcomes that is related to health insurance. The important variables can be identified which affect decisions about claim submission, risk assessment, policy renewal, premium calculation and fraud detection that can be made with the help of this information. The relevance of a feature can help in aiding insurers in setting the priorities and also efficiently allocation of resources.

However, health insurance datasets contain some missing values which can affect the performance of some algorithms. So, selection of most useful features and routes which is based on the data at hand, decision trees can effectively handle the missing data. This trait is especially useful when we have to work with insufficient or imperfectly gathered health insurance data.

In order to enhance the decision tree's prediction ability, ensemble techniques can be used for the same such as Random Forest and Gradient Boosting. On combining the recommendations of several trees, overfitting is decreased in ensembles, resilience is improved, and predictions are produced which are more accurate. This method is helpful for health insurance programmes which strive for improved robustness and accuracy in outcome prediction.

Decision tree methods can handle large datasets with many of attributes and instances effectively especially optimised version such as Classification and Regression Trees (CART) technique. For health insurance initiatives which need processing large amounts of data from policyholders, medical records, claims and demographic data, scalability is essential.

c. Random Forest Model

Random Forest is referred to as the composition of decision tree classifiers. An ensemble learning algorithm which combines multiple decision trees to make predictions are called as Random Forest Classifier which is widely used in many domains including health insurance projects. With the help of distinct subset of the data, each tree in a Random Forest ensemble is trained. Bagging which is also called as bootstrap aggregating is the term which is mainly used to describe this procedure. Each tree in a decision tree learns a random sample of the training data and employs a random selection of features for splitting so that overfitting can be decreased and generalisations can be increased. Random Forest adds another level of randomization by selection of subset of features for each tree at random which makes the tree more coherent and make sures that various feature sets are taken into account. So, in this way, a great variety of feature interactions are captured by Random Forest as well as overall prediction power has also been boosted. Majority Voting (classification) or average (regression) of the individual tree predictions has been used for the final forecast. Each tree separately provides a prediction in the Random Forest during prediction and therefore, the final predictions are more accurate and stable due to this voting system.

Random Forest is robust to overfitting since the ensemble of trees decrease the influence of a single noisy or biased tree. To reduce the possibility or overfitting and enhance the performance of generalisation, predictions from various trees are combined.

Random Forest can properly handle the high-dimensional data with several attributes. For decision-making, it can automatically determine the most instructive features as random subsets of features at each tree are taken into account. Projects which involve health insurance industry requires large and complicated datasets with numerous attributes. Random Forest shows the measure of feature importance with the help of proportional relevance of each feature in formulation of predictions. It is possible to highlight the variables which have more impact on health insurance results such as policy renewal, risk assessment, premium calculation and claim filing with the aid of this data. Feature importance can influence the choice of features, data pre-processing and decision-making procedures.

Health insurance datasets contains the class imbalance or underrepresentation of some classes. During training of data, minority classes are given more weight to effectively manage the skewed data in Random Forest which, in turn, is able to identify trends and predict outcomes for both minority and majority classes with more accuracy.

Random Forest can offer an objective assessment of the performance of the model without a separate validation set. Each tree in the ensemble is trained with the help of separate bootstrap sample with some cases (i.e., out of bag samples) excluded. Without the requirement for extra data splitting, the out of bag samples for evaluation are utilised which enables effective model evaluation.

d. Gradient Boosting

Gradient Boosting is a potent machine learning approach which is frequently used for utilisation of health insurance projects for categorization and predictive modelling applications. It is defined as an ensemble technique which combines number of decision trees that are poor learners so that a powerful prediction model can be produced. Gradient Boosting is used for constructing the predictive model by adding weak learners successively to the ensemble, which are often decision trees. Each weak learner is trained on a subset of the data with an emphasis on the cases where had high residuals from prior learners or those which were misclassified. Decision Trees are employed as weak learners due to their capacity in capturing intricate correlations between data and the target variable.

In order to ensemble's overall prediction error, the Gradient Boosting develops weak learners iteratively. The algorithm often modifies the weights of the training cases so that the misclassified or high residual instances can be prioritised after each iteration in which a new weak learner is added to the ensemble. This method continues until and unless a stopping requirement is satisfied or a predetermined number of iterations have been completed.

Gradient descent has been used by gradient boosting for optimisation of the ensemble. The ensemble predictions are adjusted in a way which minimises loss after computation of the gradient (also called as slope) of the loss function with respect to them. It is used in enhancing the ability of the ensemble for capturing the complicated patterns and make predictions precisely by updating the predictions of the ensemble iteratively.

Gradient Boosting introduces the hyperparameter referred to as learning rate which shows the contribution of each weak learner to the ensemble. When the algorithm iterates more slowly, then it indicates a lower learning rate which can reduce overfitting and increase generalisation. However, many convergence repetitions have been made necessary for a slower learning rate.

To quantify feature importance, Gradient Boosting offers a way which shows the importance of one feature as compared to the other when predictions are produced. The relevance of the feature is determined by taking into account the contribution of each feature for lowering the loss function throughout the training phase. Then, decisionmaking procedures for selection of the feature and health insurance results are facilitated with the use of this information.

In health insurance dataset, Class Imbalance or the under representation of some classes are considered as a common problem which quietly occurs. By giving instances which belongs to the minority class more weights during training, gradient boosting can handle unbalanced data by showing their significance and minimise the bias towards their majority class.

In order to limit complexity of the model and also avoid overfitting, gradient boosting offer regularisation methods which are frequently used by imposing restrictions on the maximum depth of the decision trees, capping the number of leaves and addition of penalties to the loss function for complicated models.

Gradient Boosting consists of a number of hyperparameters which can be adjusted for enhancing the performance of the model which comprises of the maximum decision tree depth, learning rate, subsampling rate and regularisation parameters as well as the number of iterations (weak learners). Hyperparameters must be carefully optimised by finding the ideal configuration for the current health insurance dilemma.

e. XGBoost Model

XGBoost Model, which is also referred to as the extreme gradient boosting, is an optimised and effective version of gradient boosting library for the purpose of scalable training in a machine learning model. XGBoost model is employed in projects related to health insurance for problems which involved predictive modelling and classification. XGBoost model is considered as a preferred option in data-intensive applications due to its outstanding scalability and performance. The Gradient Boosting Technique is known as the foundation of XGBoost which often combines a number of weak learners to produce a powerful predictive model. Gradient Descent is used for training weak learners iteratively and reduction of the overall prediction error. It also offers regularisation strategies for limiting the complexity of the model and also to avoid overfitting. These techniques often include limiting the depth of tree, penalisation of complex models and subsampling the data.

In order to increase the scalability and efficiency of the model, Numerous optimisation techniques are implemented which consists of approximate greedy techniques for splitting points, compressed data storage and construction of parallel trees. XGBoost models make use of these improvements for the purpose of processing massive datasets will millions of features and instances which further makes it appropriate for health insurance projects which requires handling comprehensive and intricate data.

Just like conventional decision tree methods, XGBoost models also builds decision trees level-wise. It often employs a more algorithm which works effectively to determine the best splits for each tree node. It considers a variety of split candidates, assessment of their quality using a special loss function, and then the split point is chosen which will result in the highest loss reduction.

Class disparity is often showed by datasets on health insurance with some classes which are underrepresented. XGBoost models offers methods to deal with imbalanced data which involves allocation of weights to various classes during training. XGBoost model can also prioritise the minority class by giving greater weights to the underrepresented classes and improvement of performance of the prediction.

In XGBoost Model, the importance of each feature which is in relation to other features when predications are produced is indicated by the rankings of the feature importance based on how often a characteristic has been employed for dividing the data among all of the trees of the ensembles and the assessment of its significance. Choice of features, risk assessment and better decision-making have been made possible using this information, that assists in identification of the critical variables which affects health insurance outcomes.

XGBoost consists of a number of hyperparameters which can be adjusted for enhancing the performance of the model which comprises of the maximum decision tree depth, subsampling rate and regularisation parameters as well as the number of iterations (boosting rounds). Hyperparameters must be carefully adjusted by finding the ideal setup which maximises the performance of the model for the particular health insurance challenge.

For rating the effectiveness of the XGBoost model, there are some evaluation criteria such as accuracy, recall, precision, F1-score and AUC-ROC which is are under the receiver operating characteristic curve. It is essential to use different test set for testing purpose in order to gauge the generalisation of the model and also to avoid overfitting.

For a reliable and effective solution for health insurance projects, XGBoost make use of the strength of the gradient boosting and includes a number of optimisations. Due to its ability to handle large-scale datasets as well as the imbalanced data, it is considered as a useful tool for the tasks which is related to claim prediction, policy renewal forecasts, risk assessment, fraud detection and other aspects of health insurance. It also helps in delivering insights of feature importance.

III. Methodology

This study adopted quantitative research since it uses a deductive approach to identify patterns in human existence by separating the social realm into measurable elements referred to as variables which can be quantified numerically.

A. Quantitative Methodology

This method was being applied to focus on investigation of the answers to the questions such as how much, how many, and to what extent. In this study, data is extracted from an opensource database which focus on customer behaviour which can be quantified and interpreted to gather insights. Quantitative data is referred to as the data which is in the form of counts having numerical data on each dataset. This data was statistically analysed to examine the conclusive results of vehicle insurance prediction. Quantitative data was statistically analysed to determine if the person with health insurance is interested in purchasing vehicle insurance.

B. Data Collection and Processing

For this research, we have gathered and combined data sources, including claim details, policyholder information, historical claim data, external data (such as public records, social media) and any fraud signs that may be present. Then pre-processing is done on the data to ensure its reliability, accuracy and analytical suitability.

I. Machine Learning Lifecycle

The Machine Learning lifecycle has various stages which provides a key structure in handling data and build a suitable machine learning model. The specific techniques and algorithms used within each stage may vary from dataset to dataset depending on the nature of the health insurance dataset, its variables and mainly on its objective which means that the stages may not be always sequential. It often involves iterating through stages at multiple times, model refining, and improving its performance based on new insights and feedback during the process.

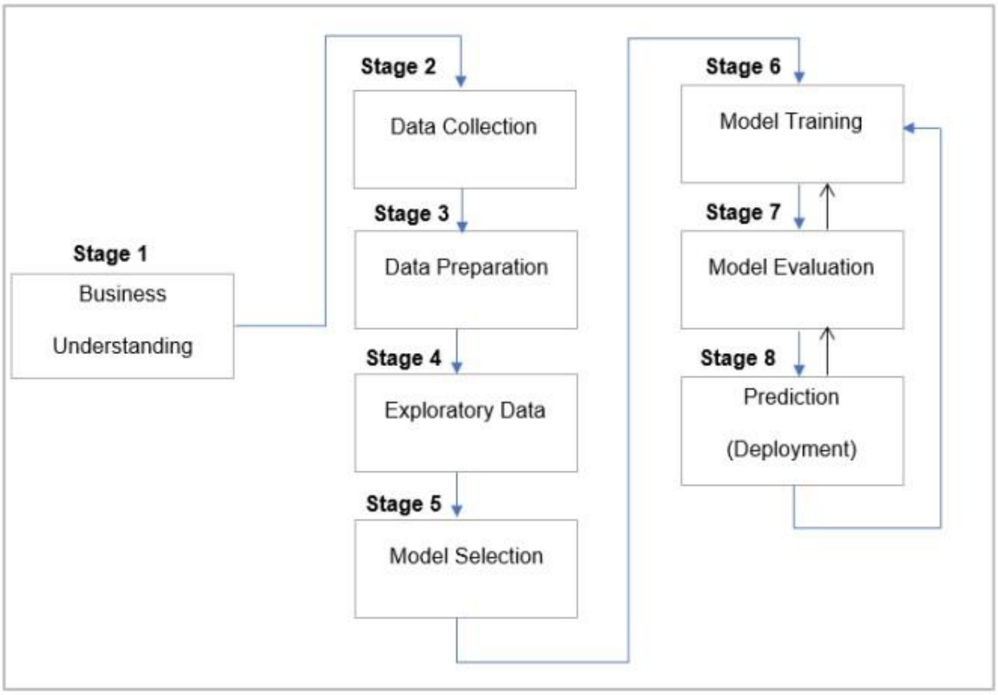


Figure 4: Machine Learning Lifecycle

II. Health Insurance Data Source

A large insurance company database contains the health insurance data which contains the information which is related to insurance policies, claims, members and other relevant variables.

III. Understanding Health Insurance Dataset

Leading insurance provider, our client, is starting a fascinating project to make use of their vast customer data and broaden their product offers. They have a solid base in the provision of health insurance; therefore, they are eager to delve into the world of auto insurance and find new clients who could be interested in buying a policy.

Building a strong predictive model that can accurately anticipate whether policyholders who have previously engaged with their health insurance offers will also display interest in their automobile insurance plans requires drawing on their rich dataset from prior years.

Insurance plans are essential for protecting people and organisations from unforeseen dangers and monetary losses. By providing auto insurance, companies hope to extend their protection to their clients' priceless possessions and give them assurance and thorough protection. In order to do this, we will use cutting-edge statistical methods and machine learning algorithms to analyse a variety of client attributes, demographic data, and historical insurance engagement data. We will create a predictive model capable of precisely identifying those clients who are likely to express interest in buying vehicle insurance by revealing hidden patterns and important indicators within the dataset.

By utilising this predictive model, we are able to maximise their marketing initiatives, streamline their customer acquisition techniques, and especially adapt their product offers to the requirements and preferences of their current clientele. This individualised strategy will not only increase customer happiness but also support our client's company's long-term expansion and success.

The dataset contains the following information:

Column	**Description**	**Data type**
id	Customer’s Unique ID	Integer
Gender	Customer’s Gender	Object
Age	Customer’s Age	Integer
Driving_License	0: Have Driving License 1: Do not Driving License	Integer
Region_Code	Customer’s Unique Code by region	Float
Previously_Insured	1: Have Vehicle Insurance 0: Do not have Vehicle Insurance	Integer
Vehicle_Age	Vehicle’s Age	Object
Vehicle_Damage	1: Vehicle got damaged in the past 0: Vehicle does not get damaged in the past	Object
Annual_Premium	Premium that Customer has to pay in the year	Float
Policy_Sales_Channel	Channel of reaching to the Customer via Anonymized Code	Float
Vintage	Number of days for which Customer has been associated with the Company	Integer
Response	1: Interested 0: Not Interested	integer

Table 1: Health Insurance Dataset- Features, Description and Data Type

IV. Data Collection Method

Gathering and combining pertinent data sources, including as claim details, policyholder information, historical claims data, external data (such as public records, social media), and any fraud signs that may be present, preparing and pre-processing the data to ensure its accuracy, reliability, and analytical suitability.

V. Google Colab and Python Programming

The dataset was imported into Google Colab after extraction and used to train the model. The extracted dataset was observed by selecting the first 10 rows to ensure that fields and features were imported successfully. Table given below represents he dataset before it was cleaned.

id	Gender	Age	Driving_License	Region_Code	Previously_Insured
1	Male	44	1	28	0
2	Male	76	1	3	0
3	Male	47	1	28	0
4	Male	21	1	11	1
5	Female	29	1	41	1
6	Female	24	1	33	0
7	Male	23	1	11	0
8	Female	56	1	28	0
9	Female	24	1	3	1
10	Female	32	1	6	1
11	Female	47	1	35	0
12	Female	24	1	50	1
13	Female	41	1	15	1
14	Male	76	1	28	0
15	Male	71	1	28	1
16	Male	37	1	6	0
17	Female	25	1	45	0
18	Female	25	1	35	1
19	Male	42	1	28	0
20	Female	60	1	33	0

Vehicle Age	Vehicle Damage	Annual Premium	Policy Sales Channel	Vintage	Response
> 2 Years	Yes	40454	26	217	1
1-2 Year	No	33536	26	183	0
> 2 Years	Yes	38294	26	27	1
< 1 Year	No	28619	152	203	0
< 1 Year	No	27496	152	39	0
< 1 Year	Yes	2630	160	176	0
< 1 Year	Yes	23367	152	249	0
1-2 Year	Yes	32031	26	72	1
< 1 Year	No	27619	152	28	0
< 1 Year	No	28771	152	80	0
1-2 Year	Yes	47576	124	46	1
< 1 Year	No	48699	152	289	0
1-2 Year	No	31409	14	221	0
1-2 Year	Yes	36770	13	15	0

1-2 Year	No	46818	30	58	0
1-2 Year	Yes	2630	156	147	1
< 1 Year	Yes	26218	160	256	0
< 1 Year	No	46622	152	299	0
1-2 Year	Yes	33667	124	158	0
1-2 Year	Yes	32363	124	102	1

Table 2: Dataset Before Cleaning

The dataset shape refers to the dimensions of the dataset, i.e. the number of rows (instances) and columns (features).

Table given below displays the shape of training and testing dataset respectively.

	Training Dataset Shape	Testing Dataset Shape
Number of Rows	127037	127037
Number of Columns	11	11

Table 3: Shape of the Training and Testing Data

VI. Health Insurance Data Processing

Data Pre-Processing is a major step in the data analysis and machine learning which includes cleaning, transforming and preparing raw data in order to make it suitable for the analysis and modelling. During data cleaning, all the duplicate values were removed and since there were no missing values so it didn't require any imputation for it in python. The dataset was then split into training and testing datasets for evaluation of model's performance. After processing, all the categorical values are converted into numerical values which can be displayed as follows:

Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age
1	0.369231	1	0.538462	0	3
1	0.861538	1	0.057692	0	2
1	0.415385	1	0.538462	0	3
1	0.015385	1	0.211538	1	1
0	0.138462	1	0.788462	1	1
0	0.061538	1	0.634615	0	1
1	0.046154	1	0.211538	0	1
0	0.553846	1	0.538462	0	2
0	0.061538	1	0.057692	1	1
0	0.184615	1	0.115385	1	1

Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Vintage	Response
1	0.070366	0.154321	0.716263	1
0	0.057496	0.154321	0.598616	0
1	0.066347	0.154321	0.058824	1
0	0.048348	0.932099	0.66782	0
0	0.046259	0.932099	0.100346	0
1	0	0.981481	0.574394	0
1	0.038578	0.932099	0.82699	0
1	0.054696	0.154321	0.214533	1
0	0.046488	0.932099	0.062284	0
0	0.048631	0.932099	0.242215	0

Table 4: Dataset After Cleaning

C. **Exploratory Data Analysis**

After performing the pre-processing of the data, exploratory data analysis has to be performed to gain an understanding of the data. During EDA, the dataset has to be explored by the researcher which is extracted from the insurance company database which involved assessing factors like dataset's size, count of column, and variable data types. It also includes descriptive analysis, correlations, data visualizations and examining distributions. EDA is considered as the most important step in the data analysis which involved exploring the data and also summarizing its main characteristics in order to find insights and understand its trends, patterns and potential issues in a clear manner.

1. Descriptive Statistics

Exploratory Data Analysis has some important components which are descriptive statistics and distribution which are useful for understanding the main characteristics of dataset and distribution of the values across variables.

Table given below displays the descriptive statistics of the numerical variables of the dataset.

	id	Age	Driving License	Region Code	Previously Insured
count	381109.00	381109.00	381109.00	381109.00	381109.00
mean	190555.00	38.82	1.00	26.39	0.46
std	110016.84	15.51	0.05	13.23	0.50
min	1.00	20.00	0.00	0.00	0.00

25%	95278.00	25.00	1.00	15.00	0.00
60%	190555.00	36.00	1.00	28.00	0.00
75%	285832.00	49.00	1.00	35.00	1.00
max	381109.00	85.00	1.00	52.00	1.00

	Annual Premium	Policy Sales Channel	Vintage	Response
count	381109.00	381109.00	381109.00	381109.00
mean	30564.39	112.03	154.35	0.12
std	17213.16	54.20	83.67	0.33
min	2630.00	1.00	10.00	0.00
25%	24405.00	29.00	82.00	0.00
60%	31669.00	133.00	154.00	0.00
75%	39400.00	152.00	227.00	0.00
max	540165.00	163.00	299.00	1.00

Table 5: Summary Statistics of Numerical Variables

Table given below displays the summary of the categorical variables:

	Gender	Vehicle_Age	Vehicle_Damage
count	381109	381109	381109
unique	2	3	2
top	Male	1-2 Year	Yes
freq	206089	200316	192413

Table 6: Summary Statistics of Categorical Variables

2. Correlation

The statistical relationship between two or more variables is referred to as correlation which measures the direction and strength of the linear relationship between two or more variables which means that it shows how change in one variable will affect another variable. The Statistical measures are made to determine the linear relationship between two variables and also to determine if the two variables are correlated when they both movie in the same direction.

The following figure displays the Pearson correlation of 11 features on how data were correlated in this study.

	Gender	Age	Driving License	Region Code	Previously Insured	Vehicle Age	Vehicle Damage	Annual Premium	Policy Sales Channel	Vintage	Response
Gender	1	0.15	-0.018	0.0006	-0.082	0.16	0.092	0.0037	-0.11	-0.0025	0.052
Age	0.15	1	-0.08	0.043	-0.25	0.77	0.27	0.068	-0.58	-0.0013	0.11
Driving License	-0.018	-0.08	1	-0.0011	0.015	-0.037	-0.017	-0.012	0.044	-0.00085	0.01
Region Code	0.0006	0.043	-0.0011	1	-0.025	0.044	0.028	-0.011	-0.042	-0.0027	0.011
Previously Insured	-0.082	-0.25	0.015	-0.025	1	-0.38	-0.82	0.0043	0.22	0.0025	-0.34
Vehicle Age	0.16	0.77	-0.037	0.044	-0.38	1	0.4	0.042	-0.55	-0.0019	0.22
Vehicle Damage	0.092	0.27	-0.017	0.028	-0.82	0.4	1	0.0093	-0.22	-0.0021	0.35
Annual Premium	0.0037	0.068	-0.012	-0.011	0.0043	0.0042	0.0093	1	-0.11	-0.00061	0.023
Policy Sales Channel	-0.11	-0.58	0.044	-0.042	0.22	-0.55	-0.22	-0.11	1	1.80E-06	-0.14
Vintage	-0.0025	-0.0013	-0.00085	-0.0027	0.0025	-0.0019	-0.0021	-0.00061	1.80E-06	1	-0.0011
Response	0.052	0.11	0.01	0.011	-0.34	0.22	0.35	0.023	-0.14	-0.0011	1

Table 7: Correlation between Features of Health Insurance Dataset

Vehicle Damage, Vehicle Age and Age are most positively correlated with Response which indicates that when Vehicle Damage, Vehicle Age and Age increases then Response also increases i.e., vehicle having more damage would result in more response for insurance claims, older vehicles have more responses due to increased wear and tear and maintenance issues and old people have different response rates due to driving habits and risk factors.

Previously Insured and Policy Sales Channel are most negatively correlated with Response which indicates that when Previously Insured and Policy Sales Channel increases then Response decreases i.e., Individuals who are already insured have lower response rates since they do not require additional services and certain sales channel have lower response rates due to inefficiencies and customer preferences.

3. Data Visualization

The graphical representation of data is a crucial step for understanding of the data with the help of charts, graphs, etc. where the data is presented in such a way that it affects the target variable. In this study, the data is displayed through different visualizations to get a clear picture of the data and gain insights from it and also it helps the researcher to ensure that data has been distributed correctly. It helps the researcher in exploration of the data, identification of outliers and patterns, and also interpret the key findings efficiently.

(a) Age and Gender

The following figure displays the visualization of Age and Gender.

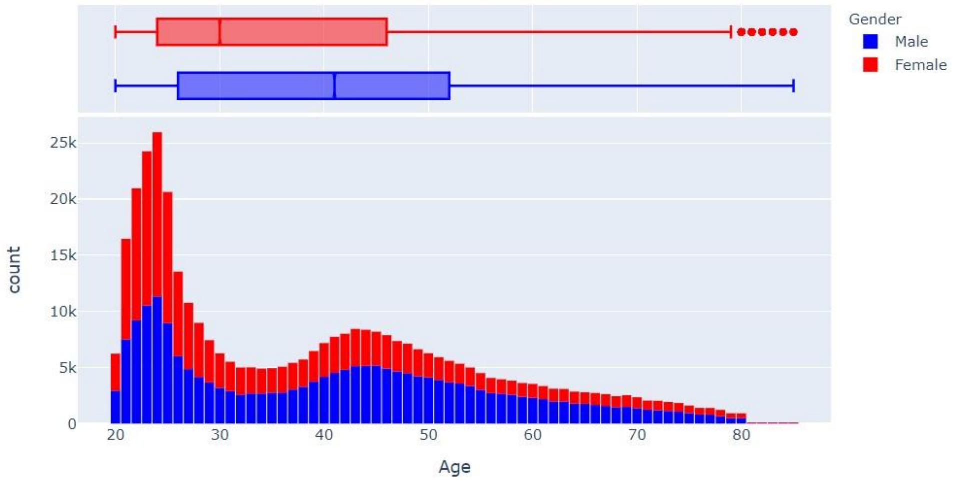


Figure 5: Age and Gender Distribution

It is considered that people of age between 20 and 30 have more policy holder than other ages and within that age group, females are than males.

(b) Vehicle Damage and Response:

The following figure displays the visualization of cross tabulation of vehicle damage and Response.

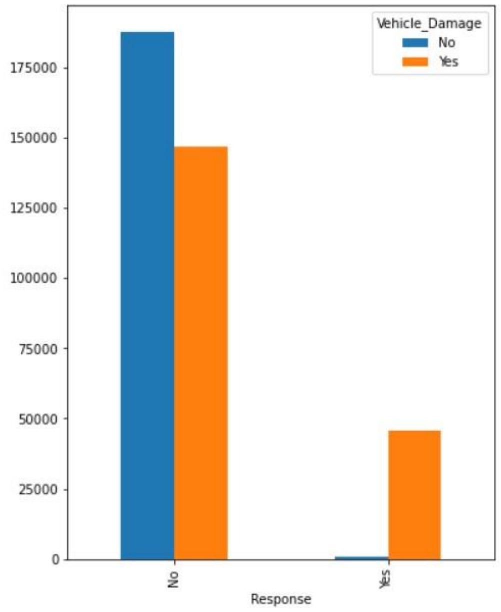


Figure 5: Vehicle Damage and Response

It is observed that the customers whose vehicle got damaged in the past are more likely to be interested in insurance and customers who are not interested in insurance do not get their vehicles damaged in the past.

(c) Age and Response:

The following visualization displays how age is distributed with response.

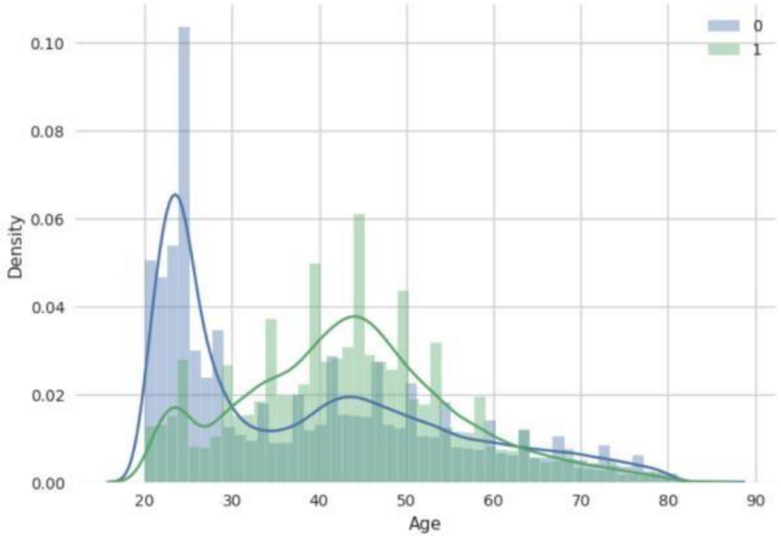


Figure 6: Age and Response

Young people whose age is less than 30 are not interested in purchasing vehicle insurance which could be due to lack of experience, not having expensive vehicles, etc. whereas people aged 30-60 are more interested in vehicle insurance and people with age more than 60 are less interested in vehicle insurance since they might not drive much at old age.

(d) Vehicle Age and Response:

The following visualization displays the distribution of vehicle age with response.

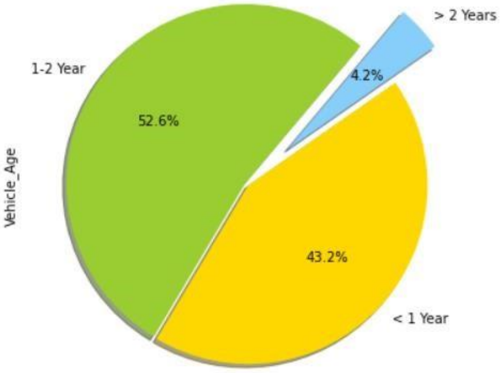


Figure 7: Vehicle Age and Response

It is observed that vehicle age between 1-2 years have more than half of the samples of data and less than 5% of the data has samples with vehicle age more than 2 years and 43% of the data has samples with vehicle age less than 1 year.

(e) Annual Premium and Age

The following visualization displays the Distribution of Annual Premium with Age.

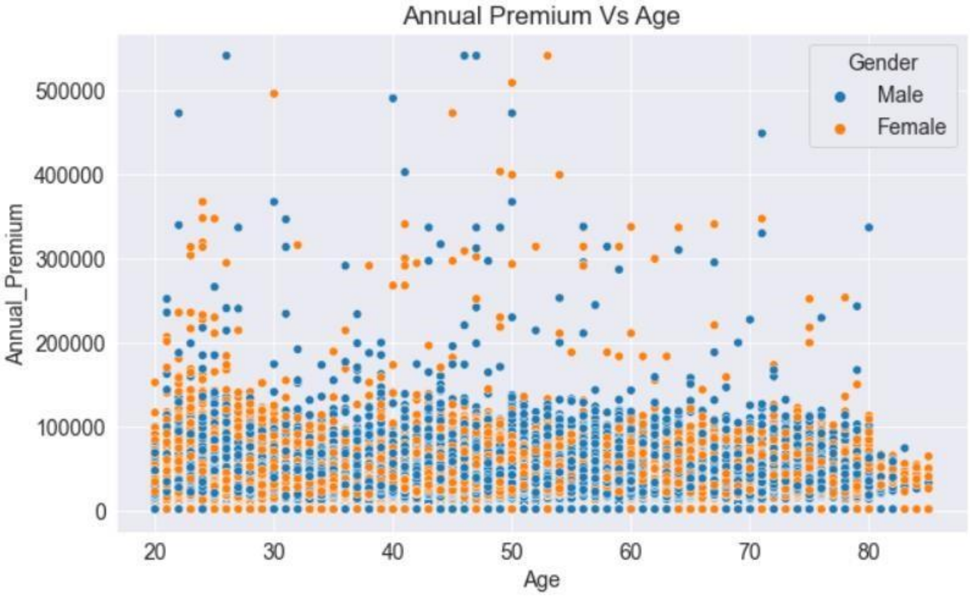


Figure 8: Annual Premium and Age

The annual premium for many of the policy holders is around 100K of all age.

4. Data Pre-processing

After exploring the health insurance data, the features of the data are scaled or normalized ensuring that they have a similar scale which can be useful for improving the performance of certain machine learning algorithms. It also involves encoding categorical variables into numerical format with the help of techniques using one-hot encoding and label encoding. In this study, the marital status field in data have categorical values such as “single”, “married”, or “divorced’. Using One-Hot Encoding, these categories are transformed into numerical values which can be used further for the analysis or model building process. With the help of sklearn’s Ordinal Encoder or Label Encoder, it is crucial when categorical variables are required to be encoded in a certain order.

After Transforming the dataset, best features are selected out of all the features which is available in the data using SelectKBest selection feature.

The bar graph below shows a visual representation of the best features that needs to be considered for the analysis.

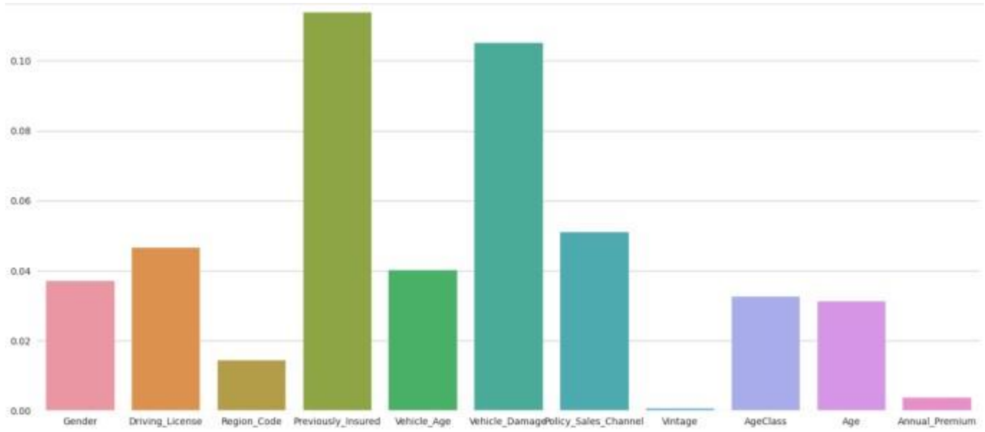


Figure 9: Best Features in the Health Insurance Dataset

The graph shows that “Premium Insured” is the most important feature while Vintage being the least as for that the value is least i.e., 0.0007.

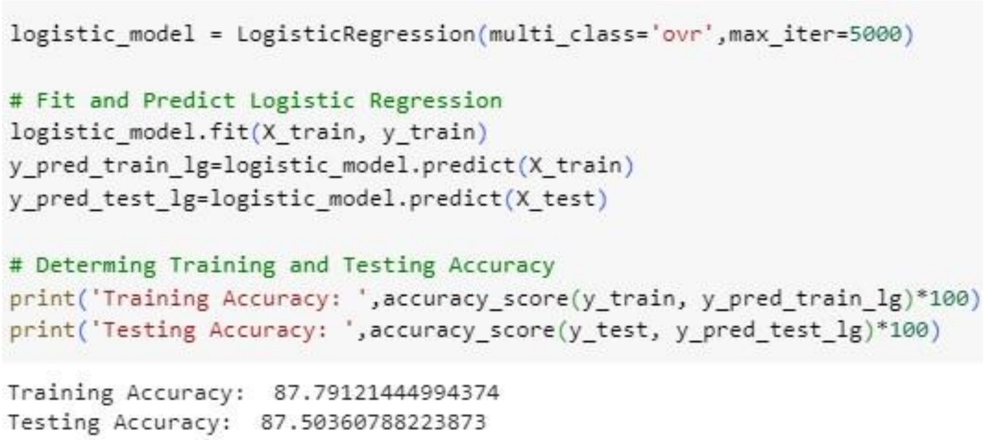
5. Model Selection

The process of selecting the most appropriate model for a data to perform a specific task is called Model Selection. It involves the evaluation of different models and comparison of their performances and selecting the one which best fits the data and also gives more accurate predictions.

The following chart displays which machine learning algorithm has been selected, trained and evaluated to build best predictive model.

6. Model Training and Model Evaluation Training

(a) Logistic Regression



```
logistic_model = LogisticRegression(multi_class='ovr',max_iter=5000)

# Fit and Predict Logistic Regression
logistic_model.fit(X_train, y_train)
y_pred_train_lg=logistic_model.predict(X_train)
y_pred_test_lg=logistic_model.predict(X_test)

# Determing Training and Testing Accuracy
print('Training Accuracy: ',accuracy_score(y_train, y_pred_train_lg)*100)
print('Testing Accuracy: ',accuracy_score(y_test, y_pred_test_lg)*100)
```

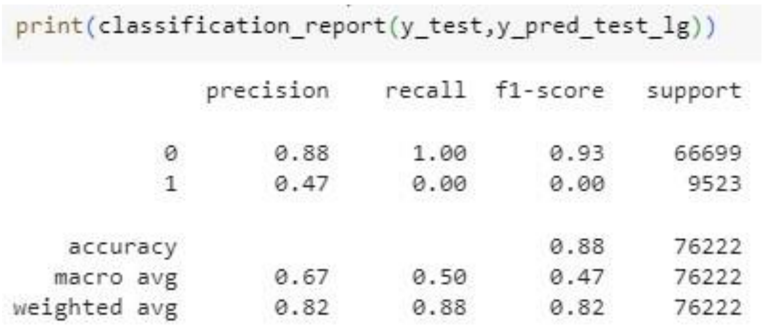
```
Training Accuracy:  87.79121444994374
Testing Accuracy:  87.50360788223873
```

Figure 10: Fitting of Logistic Regression Model

The logistic Regression model achieved an accuracy of 87.79% using training data which means that it made 87.79% correct predictions of the examples in the training data i.e., the model correctly predicted around 87 examples out of every 100 examples in the training data.

The model has achieved an accuracy of 87.50% when it has been tested on a testing data which has not been trained before so when it is applied to new and unseen data then accuracy drops slightly as compare to the training accuracy which has been obtained before.

Since the testing accuracy is slightly lower than the training accuracy, it indicates slighting overfitting of the model in the training data which means that the model in the training data has learned to capture noise or random fluctuations instead of generalizing well to unseen data on a separate testing dataset helps to assess its ability to generalize to new data and provides insights into potential overfitting issues.



```
print(classification_report(y_test,y_pred_test_lg))
```

```
              precision    recall  f1-score   support

           0       0.88      1.00      0.93     66699
           1       0.47      0.00      0.00      9523

    accuracy                           0.88     76222
   macro avg       0.67      0.50      0.47     76222
weighted avg       0.82      0.88      0.82     76222
```

Figure 11: Classification Report of Logistic Regression Model

Precision:

A precision highlights how many positive predictions are correct which are made by the model. The precision of class 0 is 88% that means that 88% of the instances which have been predicted by the model of class 0 are not interested in Vehicle Insurance whereas the precision of class 1 is 47% that means that 47% of the instances which have been predicted by the model of class 1 are interested in Vehicle Insurance.

Recall:

Recall, which is also termed as sensitivity, measures the correctly identified positive predictions that have been made by the model. The recall of class 0 is 1.00 which indicates that the model is able to correctly identify all the instances of customers that are not interested in Vehicle Insurance whereas for class 1, it is 0 which indicates the model is not able to correctly identify all the instances who are interested in Vehicle Insurance.

Support:

Support indicates the count of actual occurrences of each class in a specified dataset. The class consists of highest support value i.e., 66699 that means that the dataset is imbalanced towards class 0 which highlights that there are more instances of

5 customers who are interested in purchasing Vehicle Insurance than those who are interested in purchasing Vehicle Insurance which is indicated by class 1 having support value 9523.

F1 Score:

The F1 Score of class 0 highlights the performance of the model in identification of instances of customers who are not interested in Vehicle Insurance which is, however, higher which indicates a good balance between precision and recall for class 0. This means it has identified instances of customers who are not interested in Vehicle Insurance effectively while also minimizing both false positives and false negatives. However, the F1 Score of class 1 is 0 which indicates no balance is present between precision and recall for class 1.

Accuracy:

The model attained an accuracy of 88% which indicates that 88% of the predictions which have been made by the model are correct but accuracy has a limitation which is its biasedness towards the majority class which is class 0 in this case. Since it cannot handle imbalanced data where one is dominated by the other class so it can lead to misleading conclusions when assessing the model's performance.

Hence, the Logistic Regression Model performs well in identifying the instances of customers who are not interested in Vehicle Insurance but does not perform well in identifying instances of customers who are interested in Vehicle Insurance.

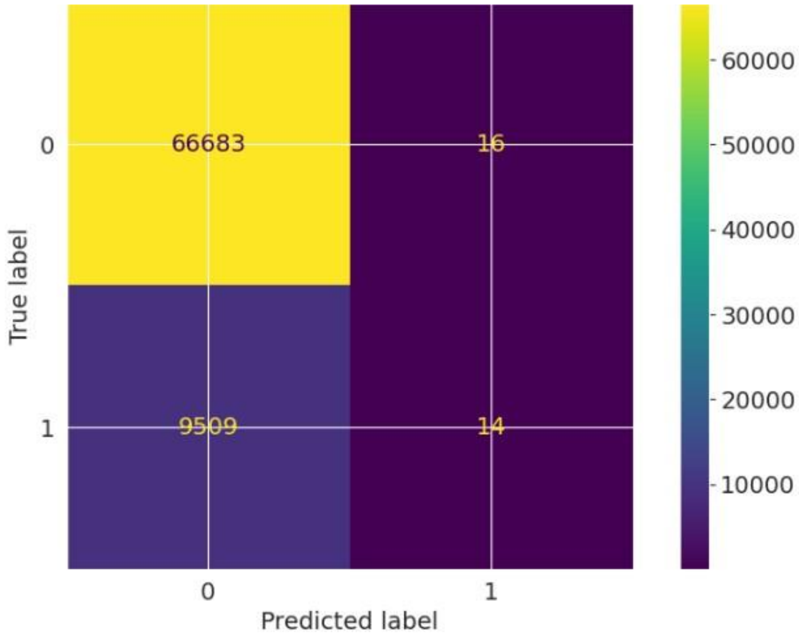
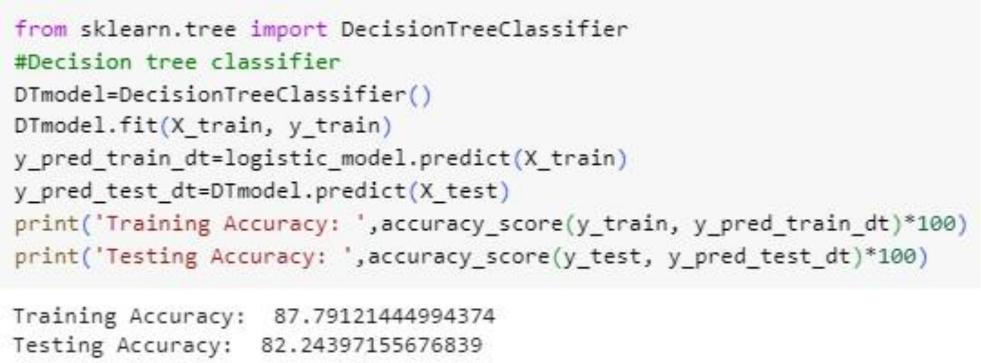


Figure 12: Confusion Matrix of Logistic Regression Model

The confusion Matrix shows that 66683 instances are correctly identified instances of label 0 and 14 instances are correctly identified instances of label 1. Also, 9509 instances are predicted as label 0 but are actually label 1 and 16 instances are actually label 0 but are predicted as label 1.

(b) Decision Tree Model



```
from sklearn.tree import DecisionTreeClassifier
#Decision tree classifier
DTmodel=DecisionTreeClassifier()
DTmodel.fit(X_train, y_train)
y_pred_train_dt=logistic_model.predict(X_train)
y_pred_test_dt=DTmodel.predict(X_test)
print('Training Accuracy: ',accuracy_score(y_train, y_pred_train_dt)*100)
print('Testing Accuracy: ',accuracy_score(y_test, y_pred_test_dt)*100)
```

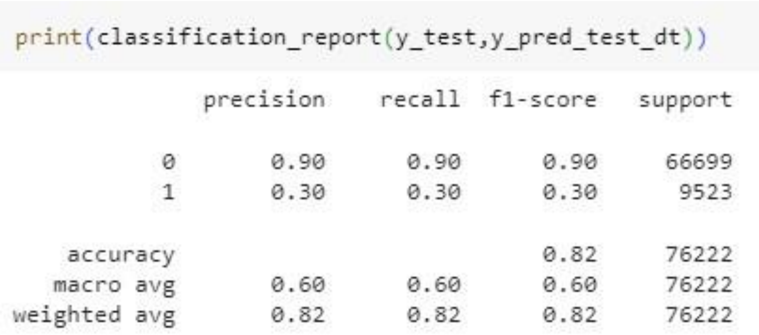
```
Training Accuracy:  87.79121444994374
Testing Accuracy:  82.24397155676839
```

Figure 13: Fitting of Decision Tree Model

The Decision Tree model achieved an accuracy of 87.79% using training data which means that it made 87.79% correct predictions of the examples in the training data i.e., the model correctly predicted around 87 examples out of every 100 examples in the training data.

The model has achieved an accuracy of 82.24% when it has been tested on a testing data which has not been trained before so when it is applied to new and unseen data then accuracy drops slightly as compare to the training accuracy which has been obtained before.

Since the testing accuracy is slightly lower than the training accuracy, it indicates slighting overfitting of the model in the training data which means that the model in the training data has learned to capture noise or random fluctuations instead of generalizing well to unseen data on a separate testing dataset helps to assess its ability to generalize to new data and provides insights into potential overfitting issues.



```
print(classification_report(y_test,y_pred_test_dt))
```

```
              precision    recall  f1-score   support

           0       0.90      0.90      0.90     66699
           1       0.30      0.30      0.30      9523

    accuracy                           0.82     76222
   macro avg       0.60      0.60      0.60     76222
weighted avg       0.82      0.82      0.82     76222
```

Figure 14: Classification Report of Decision Tree Model

Precision:

A precision highlights how many positive predictions are correct which are made by the model. The precision of class 0 is 82% that means that 82% of the instances which have been predicted by the model of class 0 are not interested in Vehicle Insurance whereas the precision of class 1 is 30% that means that 30% of the instances which have been predicted by the model of class 1 are interested in Vehicle Insurance.

Recall:

Recall, which is also termed as sensitivity, measures the correctly identified positive predictions that have been made by the model. The recall of class 0 is 0.90 which indicates that the model is able to correctly identify 90% of the instances of customers that are not interested in Vehicle Insurance whereas for class 1, it is 0.30 which indicates the model is able to correctly identify 30% of the instances who are interested in Vehicle Insurance which is comparatively less than the one achieved in class 0.

Support:

Support indicates the count of actual occurrences of each class in a specified dataset. The class consists of highest support value i.e., 66699 that means that the dataset is imbalanced towards class 0 which highlights that there are more instances of customers who are interested in purchasing Vehicle Insurance than those who are interested in purchasing Vehicle Insurance which is indicated by class 1 having support value 9523.

F1 Score:

The F1 Score of class 0 highlights the performance of the model in identification of instances of customers who are not interested in Vehicle Insurance which is, however, higher i.e., 0.90 which indicates a good balance between precision and recall for class 0. This means it has identified 90% instances of customers who are not interested in Vehicle Insurance effectively while also minimizing both false positives and false negatives. However, the F1 Score of class 1 is 0.30 which indicates a poor balance between precision and recall for class 1.

Accuracy:

The model attained an accuracy of 82% which indicates that 82% of the predictions which have been made by the model are correct but accuracy has a limitation which is its biasedness towards the majority class which is class 0 in this case. Since it cannot handle imbalanced data where one is dominated by the other class so it can lead to misleading conclusions when assessing the model's performance.

Hence, the Decision Tree Model performs well in identifying the instances of customers who are not interested in Vehicle Insurance than identifying instances of customers who are interested in Vehicle Insurance.

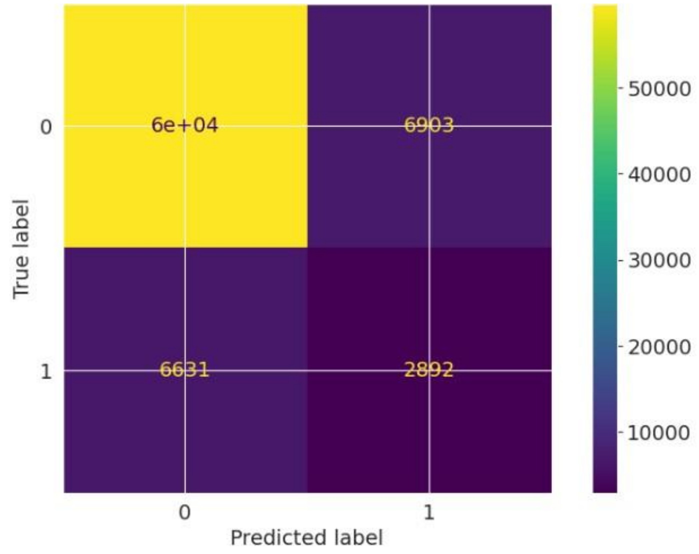
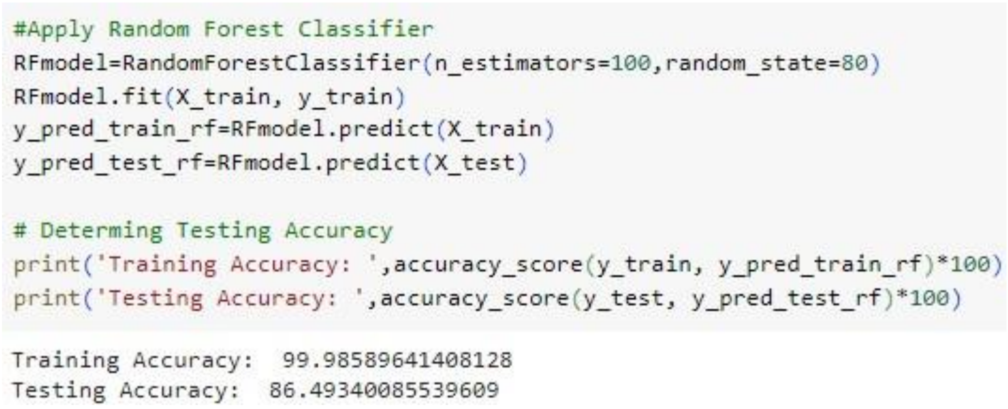


Figure 15: Confusion Matrix of Decision Tree Model

The confusion Matrix shows that 60000 instances are correctly identified instances of label 0 and 2892 instances are correctly identified instances of label 1. Also, 6631 instances are predicted as label 0 but are actually label 1 and 6903 instances are actually label 0 but are predicted as label 1.

(c) Random Forest Model



```
#Apply Random Forest Classifier
RFmodel=RandomForestClassifier(n_estimators=100,random_state=80)
RFmodel.fit(X_train, y_train)
y_pred_train_rf=RFmodel.predict(X_train)
y_pred_test_rf=RFmodel.predict(X_test)

# Determing Testing Accuracy
print('Training Accuracy: ',accuracy_score(y_train, y_pred_train_rf)*100)
print('Testing Accuracy: ',accuracy_score(y_test, y_pred_test_rf)*100)
```

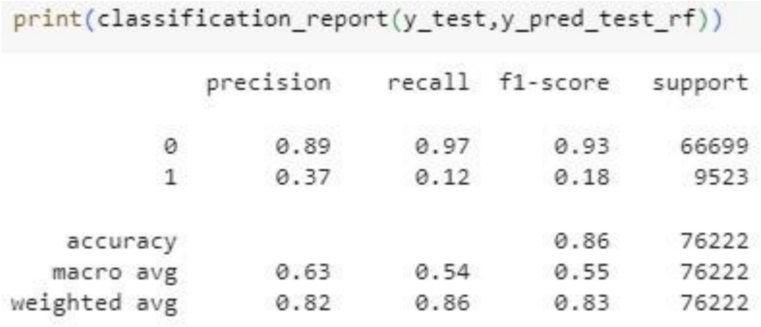
```
Training Accuracy:  99.98589641408128
Testing Accuracy:  86.49340085539609
```

Figure 16: Fitting of Random Forest Model

The Random Forest Model achieved an accuracy of 99.98% using training data which means that it made 99.98% correct predictions of the examples in the training data i.e., the model correctly predicted around 99 examples out of every 100 examples in the training data.

The model has achieved an accuracy of 86.49% when it has been tested on a testing data which has not been trained before so when it is applied to new and unseen data then accuracy drops slightly as compare to the training accuracy which has been obtained before.

Since the testing accuracy is slightly lower than the training accuracy, it indicates slighting overfitting of the model in the training data which means that the model in the training data has learned to capture noise or random fluctuations instead of generalizing well to unseen data on a separate testing dataset helps to assess its ability to generalize to new data and provides insights into potential overfitting issues.



```
print(classification_report(y_test,y_pred_test_rf))

              precision    recall  f1-score   support

           0       0.89      0.97      0.93     66699
           1       0.37      0.12      0.18      9523

    accuracy                           0.86     76222
   macro avg       0.63      0.54      0.55     76222
weighted avg       0.82      0.86      0.83     76222
```

Figure 17: Classification Report of Random Forest Model

Precision:

A precision highlights how many positive predictions are correct which are made by the model. The precision of class 0 is 89% that means that 89% of the instances which have been predicted by the model of class 0 are not interested in Vehicle Insurance whereas the precision of class 1 is 37% that means that 37% of the instances which have been predicted by the model of class 1 are interested in Vehicle Insurance.

Recall:

Recall, which is also termed as sensitivity, measures the correctly identified positive predictions that have been made by the model. The recall of class 0 is 0.97 which indicates that the model is able to correctly identify 97% of the instances of customers that are not interested in Vehicle Insurance whereas for class 1, it is 0.12 which indicates the model is able to correctly identify 12% of the instances who are interested in Vehicle Insurance which is comparatively less than the one achieved in class 0.

Support:

Support indicates the count of actual occurrences of each class in a specified dataset. The class consists of highest support value i.e., 66699 that means that the dataset is imbalanced towards class 0 which highlights that there are more instances of customers who are interested in purchasing Vehicle Insurance than those who are interested in purchasing Vehicle Insurance which is indicated by class 1 having support value 9523.

F1 Score:

The F1 Score of class 0 highlights the performance of the model in identification of instances of customers who are not interested in Vehicle Insurance which is, however, higher i.e., 0.93 which indicates a good balance between precision and recall for class 0. This means it has identified 93% instances of customers who are not interested in Vehicle Insurance effectively while also minimizing both false positives and false negatives. However, the F1 Score of class 1 is 0.18 which indicates a poor balance between precision and recall for class 1.

Accuracy:

The model attained an accuracy of 86% which indicates that 86% of the predictions which have been made by the model are correct but accuracy has a limitation which is its biasedness towards the majority class which is class 0 in this case. Since it cannot handle imbalanced data where one is dominated by the other class so it can lead to misleading conclusions when assessing the model's performance.

Hence, the Random Forest Model performs well in identifying the instances of customers who are not interested in Vehicle Insurance than identifying instances of customers who are interested in Vehicle Insurance.

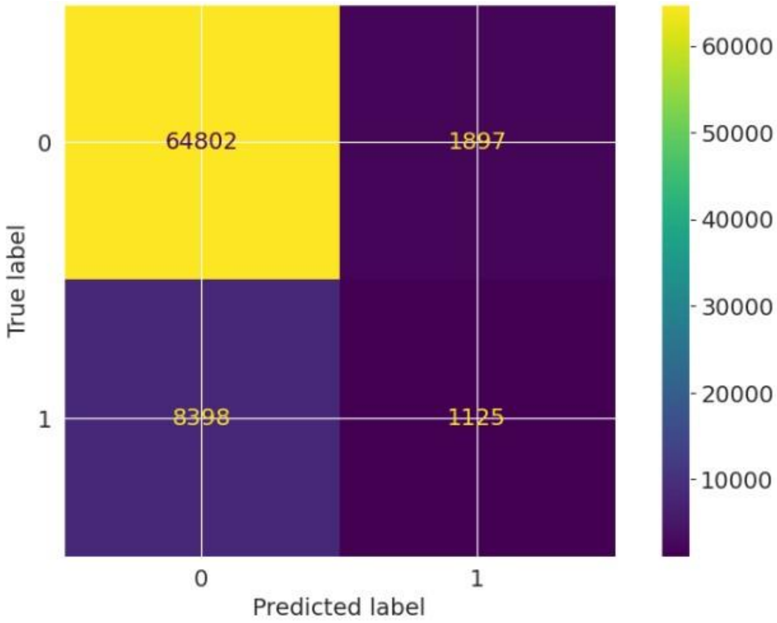
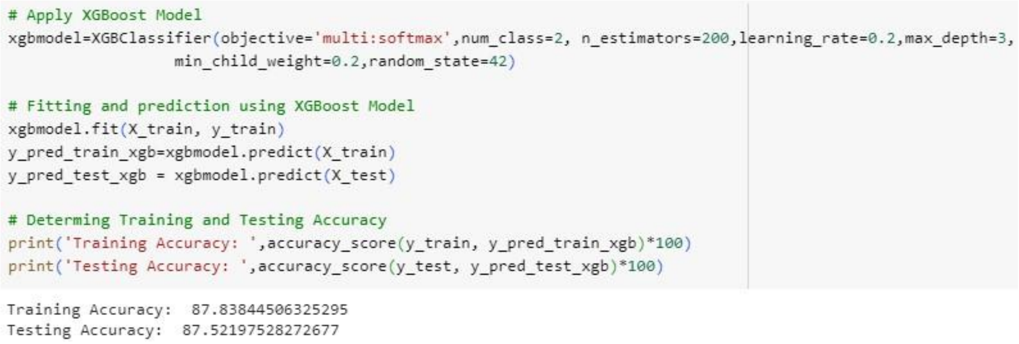


Figure 18: Confusion Matrix of Random Forest Model

The confusion Matrix shows that 64802 instances are correctly identified instances of label 0 and 1125 instances are correctly identified instances of label 1. Also, 8398 instances are predicted as label 0 but are actually label 1 and 1897 instances are actually label 0 but are predicted as label 1.

(d) XGBoost Model



```
# Apply XGBoost Model
xgbmodel=XGBClassifier(objective='multi:softmax',num_class=2, n_estimators=200,learning_rate=0.2,max_depth=3,
                  min_child_weight=0.2,random_state=42)

# Fitting and prediction using XGBoost Model
xgbmodel.fit(X_train, y_train)
y_pred_train_xgb=xgbmodel.predict(X_train)
y_pred_test_xgb = xgbmodel.predict(X_test)

# Determing Training and Testing Accuracy
print('Training Accuracy: ',accuracy_score(y_train, y_pred_train_xgb)*100)
print('Testing Accuracy: ',accuracy_score(y_test, y_pred_test_xgb)*100)
```

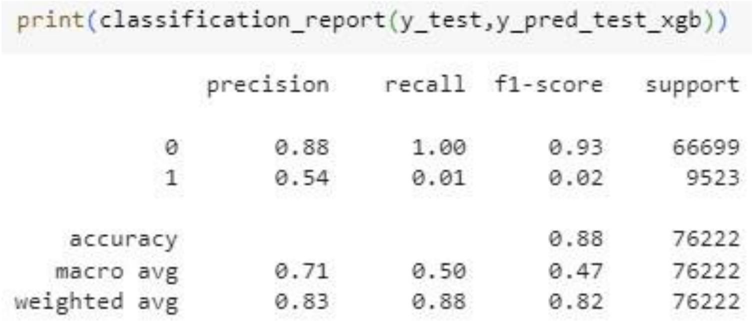
```
Training Accuracy:  87.83844506325295
Testing Accuracy:  87.52197528272677
```

Figure 19: Fitting of XGBoost Model

The XGBoost Model achieved an accuracy of 87.83% using training data which means that it made 87.83% correct predictions of the examples in the training data i.e., the model correctly predicted around 87 examples out of every 100 examples in the training data.

The model has achieved an accuracy of 87.52% when it has been tested on a testing data which has not been trained before so when it is applied to new and unseen data then accuracy drops slightly as compare to the training accuracy which has been obtained before.

Since the testing accuracy is slightly lower than the training accuracy, it indicates slighting overfitting of the model in the training data which means that the model in the training data has learned to capture noise or random fluctuations instead of generalizing well to unseen data on a separate testing dataset helps to assess its ability to generalize to new data and provides insights into potential overfitting issues.



```
print(classification_report(y_test,y_pred_test_xgb))

              precision    recall  f1-score   support

           0       0.88      1.00      0.93     66699
           1       0.54      0.01      0.02      9523

    accuracy                           0.88     76222
   macro avg       0.71      0.50      0.47     76222
weighted avg       0.83      0.88      0.82     76222
```

Figure 20: Classification Report of XGBoost Model

Precision:

A precision highlights how many positive predictions are correct which are made by the model. The precision of class 0 is 88% that means that 88% of the instances which have been predicted by the model of class 0 are not interested in Vehicle Insurance whereas the precision of class 1 is 54% that means that 54% of the instances which have been predicted by the model of class 1 are interested in Vehicle Insurance.

Recall:

Recall, which is also termed as sensitivity, measures the correctly identified positive predictions that have been made by the model. The recall of class 0 is 1.00 which indicates that the model is able to correctly identify 100% of the instances of customers that are not interested in Vehicle Insurance whereas for class 1, it is 0.01 which indicates the model is able to correctly identify 1% of the instances who are interested in Vehicle Insurance which is comparatively less than the one achieved in class 0.

Support:

Support indicates the count of actual occurrences of each class in a specified dataset. The class consists of highest support value i.e., 66699 that means that the dataset is imbalanced towards class 0 which highlights that there are more instances of customers who are interested in purchasing Vehicle Insurance than those who are interested in purchasing Vehicle Insurance which is indicated by class 1 having support value 9523.

F1 Score:

The F1 Score of class 0 highlights the performance of the model in identification of instances of customers who are not interested in Vehicle Insurance which is, however, higher i.e., 0.93 which indicates a good balance between precision and recall for class 0. This means it has identified 93% instances of customers who are not interested in Vehicle Insurance effectively while also minimizing both false positives and false negatives. However, the F1 Score of class 1 is 0.02 which indicates a poor balance between precision and recall for class 1.

Accuracy:

The model attained an accuracy of 88% which indicates that 88% of the predictions which have been made by the model are correct but accuracy has a limitation which is its biasedness towards the majority class which is class 0 in this case. Since it cannot handle imbalanced data where one is dominated by the other class so it can lead to misleading conclusions when assessing the model's performance.

Hence, the XGBoost Model performs well in identifying the instances of customers who are not interested in Vehicle Insurance than identifying instances of customers who are interested in Vehicle Insurance.

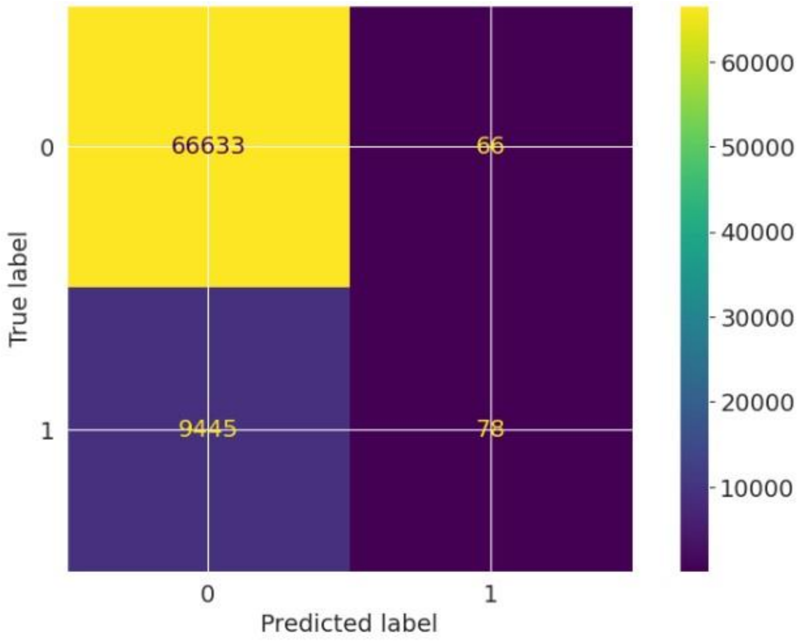


Figure 21: Confusion Matrix of XGBoost Model

The confusion Matrix shows that 66633 instances are correctly identified instances of label 0 and 78 instances are correctly identified instances of label 1. Also, 9445 instances are predicted as label 0 but are actually label 1 and 66 instances are actually label 0 but are predicted as label 1.

7. Conclusion

XGBoost Model is the best model which has performed well in predicting the customers who are interested in Vehicle Insurance as compared to other models. Although Logistic Regression's accuracy is close to XGBoost Model's accuracy but XGBoost model can predict more instances of customers who are interested in Vehicle Insurance as compared to Logistic Regression.

8. References

[1] https://www.researchgate.net/publication/376958786_Analysing_Health_Insurance_Customer_Dataset_to_Determine_Cross-Selling_Potential

[2] https://www.analyticsvidhya.com/blog/2021/10/building-an-end-to-end-logisticregression-model/

[3] https://www.ibm.com/topics/decision-trees

[4] https://www.investopedia.com/terms/l/lifeinsurance.asp#toc-who-needs-lifeinsurance

[5] https://www.investopedia.com/terms/l/lifeinsurance.asp#toc-benefits-of-lifeinsurance

[6] https://medium.com/@danyal.wainstein1/understanding-the-confusion-matrixb9bc45ba2679

[7] https://irdai.gov.in/