Project Summary

Batch details	PGPDSE - Online June22
Team members	1. Anoosh Kumar 2. Nidhi Jaiswal 3. Pranjal Jalota 4. Rushika Bokde 5. Sylesh JL
Domain of Project	Insurance
Proposed project title	Car Insurance Claim Prediction
Group Number	Group-03
Team Leader	Nidhi Jaiswal
Mentor Name	Ms. Vibha Santhanam

Date: 17th Dec 2022

Vibha Santhanam

Signature of the Mentor

Nidhi Jaiswal

Signature of the Team Leader

Table of Contents

OVERVIEW

We are implementing machine learning model in predicting whether a policy holder's likelihood in claiming the insurance in next 6 months. Here we use variety of classification algorithms techniques like Logistic Regression, Naïve Bayes, K-Nearest Neighbours, Decision Tree or Random Forest to classify a particular policy ID is a "claim" or "not a claim". The idea is to build a machine learning solution which is diversified, hence this project will have more diversity in combining models and various ensemble techniques for an accurate claim prediction.

Business Problem Goals (Problem Statement)

1. **Business Problem Understanding**

An insurance policy is an agreement between an insurance provider company and policyholder, wherein the company is liable in providing the guaranteed compensation for a specific loss or damage when a certain amount of premium is paid by the policyholder for taking the insurance with that company.

For an instance, if someone pays a premium of Rs. 6000/- every year for car insurance premium with a cover of Rs. 190,000/-. Unfortunately, in the event of an unexpected accident and the car is damaged, In such a case the insurance company will bear the cost for the damage happened up to Rs. 190,000.

Considering that the company charges a premium of only Rs. 3000/- per annum, the concept of probability plays an important role here. In other words, there might be thousands of customers paying a premium of Rs. 6000 every year just to cover the cost of the policy.

2. **Business Objective :**

Our objective is to help the Insurance company to understand the behavior of customers and predict a future trait if a car policy holder will claim his insurance or not based on the insurance data provided.. With the above objective we aim to develop a prediction model that helps the insurance company to understand the car policyholders behavior and classify if the policy holders are likely to claim insurance.

3. **Approach**
    - Understanding the Dataset
    - Data pre-processing
    - Handling Data types
    - Scaling & Transformation of Data wherever required
    - Dealing with the null values
    - Handling of Outliers with help of EDA techniques
    - Visualizations
    - Building classification Model
    - Evaluate & Deployment of the Model

4. Conclusions

After Exploration and Analyzation, we can build a machine Learning model to correctly identify whether a policyholders, given certain attributes, has a high probability to claim his/her insurance. We can use the model to identify certain traits of future that could have the potential to recognize the policyholders.

About Dataset :

The Dataset contains information on policyholders having the attributes like policy tenure, age of the car, age of the car owner, the population density of the city, make and model of the car, power, engine type, etc, and the target variable indicating whether the policyholder files a claim in the next 6 months or not.

Variable	Description
policy_id	Unique identifier of the policyholder
policy_tenure	Time period of the policy
age_of_car	Normalized age of the car in years
age_of_policyholder	Normalized age of policyholder in years
area_cluster	Area cluster of the policyholder
population density	Population density of the city (Policyholder City)
make	Encoded Manufacturer/company of the car
segment	Segment of the car (A/ B1/ B2/ C1/ C2)
model	Encoded name of the car
fuel_type	Type of fuel used by the car
max_torque	Maximum Torque generated by the car (Nm@rpm)
max_power	Maximum Power generated by the car (bhp@rpm)
engine_type	Type of engine used in the car
airbags	Number of airbags installed in the car
is_esc	Boolean flag indicating whether Electronic Stability Control (ESC) is present in the car or not.
is_adjustable_steering	Boolean flag indicating whether the steering wheel of the car is adjustable or not.
is_tpms	Boolean flag indicating whether Tyre Pressure Monitoring System (TPMS) is present in the car or not.
is_parking_sensors	Boolean flag indicating whether parking sensors are present in the car or not.
is_parking_camera	Boolean flag indicating whether the parking camera is present in the car or not.
rear_brakes_type	Type of brakes used in the rear of the car
displacement	Engine displacement of the car (cc)
cylinder	Number of cylinders present in the engine of the car
transmission_type	Transmission type of the car
gear_box	Number of gears in the car

steering_type	Type of the power steering present in the car
turning_radius	The space a vehicle needs to make a certain turn (Meters)
length	Length of the car (Millimetre)
width	Width of the car (Millimetre)
height	Height of the car (Millimetre)
gross_weight	The maximum allowable weight of the fully-loaded car, including passengers, cargo and equipment (Kg)
is_front_fog_lights	Boolean flag indicating whether front fog lights are available in the car or not.
is_rear_window_wiper	Boolean flag indicating whether the rear window wiper is available in the car or not.
is_rear_window_washer	Boolean flag indicating whether the rear window washer is available in the car or not.
is_rear_window_defogger	Boolean flag indicating whether rear window defogger is available in the car or not.
is_brake_assist	Boolean flag indicating whether the brake assistance feature is available in the car or not.
is_power_door_lock	Boolean flag indicating whether a power door lock is available in the car or not.
is_central_locking	Boolean flag indicating whether the central locking feature is available in the car or not.
is_power_steering	Boolean flag indicating whether power steering is available in the car or not.
is_driver_seat_height_adjustable	Boolean flag indicating whether the height of the driver seat is adjustable or not.
is_day_night_rear_view_mirror	Boolean flag indicating whether day & night rearview mirror is present in the car or not.
is_ecw	Boolean flag indicating whether Engine Check Warning (ECW) is available in the car or not.
is_speed_alert	Boolean flag indicating whether the speed alert system is available in the car or not.
ncap_rating	Safety rating given by NCAP (out of 5)
is_claim	Outcome: Boolean flag indicating whether the policyholder file a claim in the next 6 months or not.

TOPIC SURVEY

1. **Problem understanding :**

The insurance company wants to understand if a policy holder will claim his/her policy in next 6 months based on the data survey to take some business decisions.

The dataset provided contains the 44 variables out of which 43 are independent variables that specify the details of the car such as age of car, model, torque , power , the sensors the car has equipped with , ncap rating of the car for safety, dimensions of the car etc , along with this the details of the policyholder such as age of the policyholder is also shared.

The one remaining column is Dependent variable (is_claim) which will help in fetching whether a particular policy will be claimed or not based on the various given independent variables' conditions

2. **Current solution to the problem :**

Due to the rapid modernisation, a surge in the usage of number of cars is seen in the last two decades compelling the insurance companies in utilizing most advanced data techniques in order to predict whether a policyholder opts for the claim or not. In this quest companies started using the machine learning algorithms in their business, Major categories in where the companies currently using ML models include behavior of driver monitoring system and in depth market analysis on Data of insurance claims based on consumer previous behavior, Example of insurance analytics is the model that identifies the severity of claim and total amount of funds to be released for various scenarios of accident using Linear regression or multiple linear regression models.

3. **Proposed solution to the problem (Abstract) :**

Our ultimate goal of this project is building a classification model. A classification model helps to make some inference from the input values given for training. It will predict the class labels of new data. Here we mainly focus on those independent variables that describe only the various prospects of the car which include cars's build , shape , mileage, age, safety rating, Engine power, torque it produces, availability of various sensors etc.

We are implementing complex machine learning models in predicting whether a policy holder's likelihood in claiming the insurance in next 6 months. Here we use variety of classification algorithms techniques like Logistic Regression, Naïve Bayes, K-Nearest Neighbours, Decision Tree or Random Forest to classify a particular policy ID is a "claim" or "not a claim". The idea is to build a machine learning solution which is diversified, hence this project will have more diversity in combining models and various ensemble techniques for an accurate claim prediction.

4. **References**

https://dalpozz.github.io/static/pdf/Claim_prediction.pdf

https://arxiv.org/pdf/2204.06109.pdf

https://www.jatit.org/volumes/Vol98No22/8Vol98No22.pdf

https://www.virtusa.com/perspectives/article/predictive-analytics-in-insurance-claims

https://www.genpact.com/insight/analytics-in-claims-how-to-gain-better-faster-consumable-insights-using-augmented-intelligence

https://thesai.org/Downloads/Volume12No3/Paper_54-Motor_Insurance_Claim_Status_Prediction.pdf

CRITICAL ASSESSMENT OF TOPIC SURVEY

The vital practice in an Insurance industry is to set the premium before the contract inception. To decide an accurate premium for the consecutive years in an insurance company, a precise and reliable estimate of the number of claims occurrences and the total claim amounts is humongously important. The claims history is the backbone for new insurance products which driven to the market expecting huge margins of attraction and profits. But not large volume of products are released because manual analysis are still performed in many insurance sector. Faster and accurate predictions are still an ordeal in this sector.

We are aiming to provide a prediction model with Enhanced accurate forecast of claims which can help insurers maneuveur with new release of products with multiple offerings to insurers and added benefits to customer as well

METHODOLOGY

Methodology indicates the routine for finding solutions to a specific problem. This is a cyclic process that undergoes a critic behavior guiding us to act accordingly.

This includes below mentioned processes-

1.Business Understanding

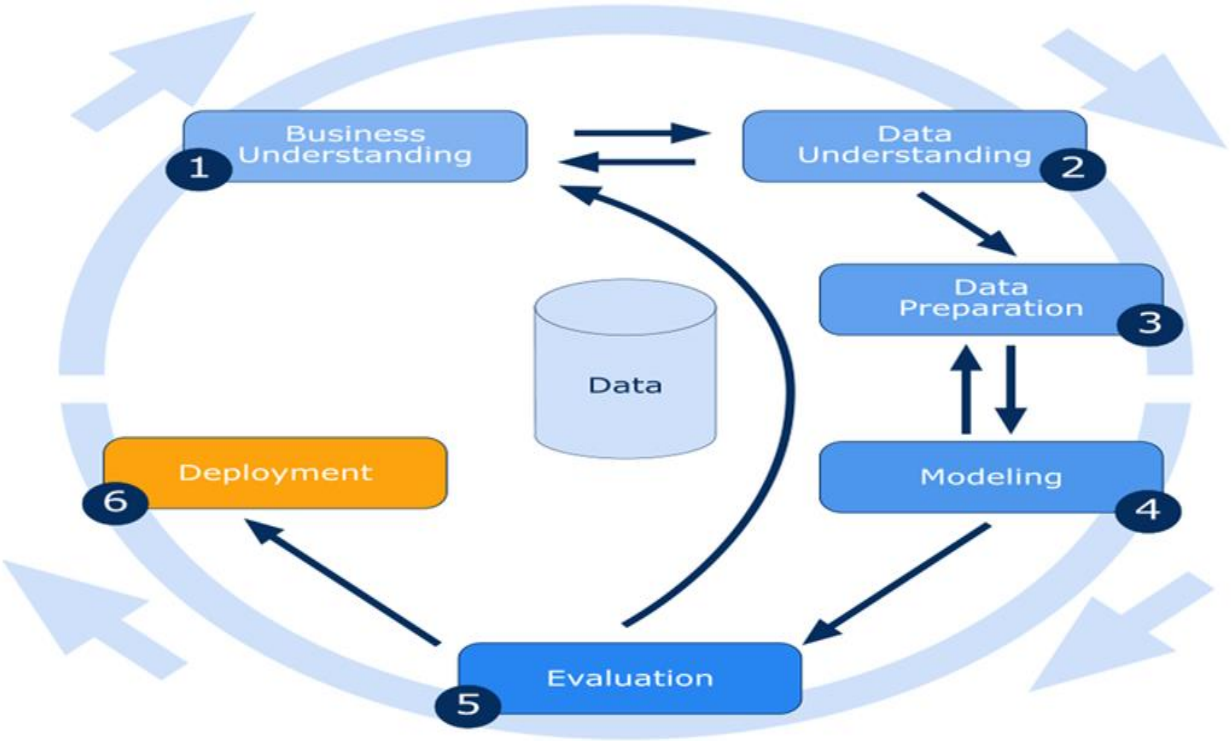
2.Data Understanding

3.Data Preparation

4.Modeling

5.Evaluation

6.Deployment



Business understanding:

We will specify the key variables that are to serve as the model targets and whose related metrics are used determine the success of the project. Then we identify the relevant data sources that the business has access to or needs to obtain.

There are two main tasks addressed in this stage:

- Define objectives: Work with dataset and online sources to understand and identify the business problems. Formulate questions that define the business goals that the data science techniques can target.
- Identify data sources: Find the relevant data that helps you answer the questions that define the objectives of the project.

Data Understanding:

Data understanding involves accessing the data and exploring it using tables and graphics.

This determines the quality of the data and describe the results of these steps in the project documentation.

Data Preparation:

This step is concerned with transforming the raw data that was collected into a form that can be used in modeling.

On a predictive modeling project, such as classification or regression, raw data typically cannot be used directly.

This is because of reasons such as:

- Machine learning algorithms require data to be numbers.
- Some machine learning algorithms impose requirements on the data.
- Statistical noise and errors in the data may need to be corrected.
- Complex nonlinear relationships may be teased out of the data.

As such, the raw data must be pre-processed prior to being used to fit and evaluate a machine learning model. This step in a predictive modeling project is referred to as "**data preparation**".

We can define data preparation as the transformation of raw data into a form that is more suitable for modeling.

This is highly specific to data, to the goals of our project, and to the algorithms that will be used to model data.

Nevertheless, there are common or standard tasks that we may use or explore during the data preparation step in a machine learning project.

These tasks include:

- **Data Cleaning**: Identifying and correcting mistakes or errors in the data.
- **Feature Selection**: Identifying those input variables that are most relevant to the task.
- **Data Transforms**: Changing the scale or distribution of variables.
- **Feature Engineering**: Deriving new variables from available data.
- **Dimensionality Reduction**: Creating compact projections of the data.

Modeling:

Data modeling is the process of analyzing and defining the dataset with business perspective, as well as the relationships between those bits of data

Model used:

We will be using **Binary Classification Model** to analyze our Dataset as we are trying to predict if a person will claim his car insurance or not.

Classification Predictive Modeling

A classification problem in machine learning is one in which a class label is anticipated for a specific example of input data.

Problems with categorization include the following:

- Give an example and indicate whether it is spam or not.
- Identify a handwritten character as one of the recognized characters.
- Determine whether to label the current user behavior as churn.

A training dataset with numerous examples of inputs and outputs is necessary for classification from a modeling standpoint.

A model will determine the optimal way to map samples of input data to certain class labels using the training dataset. The training dataset must therefore contain a large number of samples of each class label and be suitably representative of the problem.

When providing class labels to a modeling algorithm, string values like "spam" or "not spam" must first be converted to numeric values. Label encoding, which is frequently used, assigns a distinct integer to every class label, such as "spam" = 0, "no spam," = 1.

There are numerous varieties of algorithms for classification in modeling problems, including predictive modeling and classification.

Based on their output, classification predictive modeling algorithms are assessed. A common statistic for assessing a model's performance based on projected class labels is classification accuracy. Although not perfect, classification accuracy is a reasonable place to start for many classification jobs.

There are four different types of Classification Tasks in Machine Learning and they are following -

- Binary Classification
- Multi-Class Classification
- Multi-Label Classification

Imbalanced Classification

We will be using the binary classification for our analysis.

Binary Classification

Those classification jobs with only two class labels are referred to as binary classification.

Examples comprise -

- Prediction of conversion (buy or not).
- Churn forecast (churn or not).
- Detection of spam email (spam or not).

Binary classification problems often require two classes, one representing the normal state and the other representing the aberrant state.

For instance, the normal condition is "not spam," while the abnormal state is "spam." Another illustration is when a task involving a medical test has a normal condition of "cancer not identified" and an abnormal state of "cancer detected."

Class label 0 is given to the class in the normal state, whereas class label 1 is given to the class in the abnormal condition.

A model that forecasts a Bernoulli probability distribution for each case is frequently used to represent a binary classification task.

The discrete probability distribution known as the Bernoulli distribution deals with the situation where an event has a binary result of either 0 or 1. In terms of classification, this indicates that the model forecasts the likelihood that an example would fall within class 1, or the abnormal state.

The following are well-known binary classification algorithms:

- Logistic Regression
- Support Vector Machines
- Simple Bayes
- Decision Trees

Evaluation-

Model evaluation is the process of using different evaluation metrics to understand a machine learning model's performance, as well as its strengths and weaknesses. Model evaluation is important to assess the efficacy of a model during initial research phases, and it also plays a role in model monitoring.

The most popular metrics for measuring classification performance include accuracy, precision, confusion matrix, log-loss, and AUC (area under the ROC curve).

- **Accuracy** measures how often the classifier makes the correct predictions, as it is the ratio between the number of correct predictions and the total number of predictions.
- **Precision** measures the proportion of predicted Positives that are truly Positive. Precision is a good choice of evaluation metrics when you want to be very sure of your prediction. For example, if you are building a system to predict whether to decrease the credit limit on a particular account, you want to be very sure about the prediction or it may result in customer dissatisfaction.
- The **confusion matrix** (or confusion table) shows a more detailed breakdown of correct and incorrect classifications for each class. Using a confusion matrix is useful when you want to understand the distinction between classes, particularly when the cost of misclassification might differ for the two classes, or you have a lot more test data on one class than the other. For example, the consequences of making a false positive or false negative in a cancer diagnosis are very different.

Time line chart (Weekly Plan)

Task	Timeline
Finalising the dataset	5th Dec to 7th Dec, 2022
Overview dataset and finalizing model	8th Dec to 14th Dec,2022
Synopsis Creation	15th Dec to 17th Dec,2022
Business Understanding	18th Dec to 31st Dec 2022
Data Understanding	1st Jan to 15th Jan 2023
Modeling	16th Jan to 10th Feb 2023
Evaluation	11th Feb to 20th Feb 2023
Final project report	21st Feb to 28th Feb

REFERENCES

The references can be blogs, articles or even social media news relevant to explain the importance of the projects.

https://www.ijitee.org/wp-content/uploads/papers/v8i6s4/F11180486S419.pdf

https://ijiset.com/vol8/v8s5/IJISET_V8_I05_19.pdf

https://www.iii.org/fact-statistic/facts-statistics-auto-insurance

https://www.drnishikantjha.com/papersCollection/A%20STUDY%20ON%20OPERATION%20AND%20CLAIM%20PROCEDURE%20OF%20MOTOR%20VEHICLE%20INSURANCE%20.pdf

Notes For Project Team

Sample Reference for Datasets (to be filled by team and mentor)

Original owner of data	IFTESHA NAJNIN
Data set information	The Dataset contains information on policyholders having the attributes like policy tenure, age of the car, age of the car owner, the population density of the city, make and model of the car, power, engine type, etc, and the target variable indicating whether the policyholder files a claim in the next 6 months or not.
Any past relevant articles using the dataset	https://www.kaggle.com/code/fredrue/car-claims-prediction-with-xgbclassifier
Link to web page	https://www.kaggle.com/datasets/ifteshanajnin/carinsuranceclaimprediction-classification/code?resource=download