A picture containing circle, text, map, globe

Description automatically generatedA black background with a black border and a black frame with a black border and a black background with a black border and a black border with a black border and a black border with a black border

Description automatically generated

**Cairo University**

**Faculty of Computers & Artificial Intelligence**

**Operation Research & Decision Support Dept.**

# Ta2menak

The Graduation Project Submitted to

The Faculty of Computers and Artificial Intelligence,

Cairo University

In Partial Fulfillment of the Requirements

for thebachelor’s degree

In

**Operations Research and Decision Support**

*Under Supervision of:*

**Ihab Elkhodary**

**CAIRO UNIVERSITY**

**JULY/2024**

A black background with a black border and a black frame with a black border and a black background with a black border and a black border with a black border and a black border with a black border

Description automatically generatedA picture containing circle, text, map, globe

Description automatically generated

## Ta2menak -تأمينك

**Dina Ahmed Mohamed Abdelaal**

**Habiba Abdelkareem Ahmed Ali**

**Mayar Ahmed Ibrahim Farag**

**Mona Abdelaziz Ali Abdelaziz**

*Supervised By*

**Ihab** **Elkhodary**

**CAIRO UNIVERSITY JULY 2024**

#### **ABSTRACT**

The project is about enhancing car insurance risk assessment and claim prediction using machine learning.

Basically, with the goal being able to categorize customers with respect to their risk profile and predict the claim likelihood over a 6-month duration.

Key Highlights

Risk Assessment Module: Developed a machine learning module that would look into the customer's age, location, and driving history to give a risk score for each.

Classification Based on Risk Score: The classification of customers based on the score into low, medium, and high risk.

Claim Prediction Model: Utilized machine learning to predict the likelihood of claims within six months with data-driven insights from historical claim data.

Issues Addressed

Data Limitations: Difficulty in sourcing data from Egypt, limited data that affect model accuracy, and poor data quality, which therefore requires cleaning.

Model Limitations: There could be inaccuracy in models developed.

Computational Resources: Demanding high computing power to train and deploy complex models.

Future Work Recommendations

Enhanced Data Collection: In collaboration with the telematics companies, data should be collected in real-time on driving patterns with a lot more emphasis on variables like speed and braking.

Leverage Advanced Features: Use weather data, including snowfall, for enabling risk assessment. Risk scores can thus be suitably adjusted.

Deployment and Integration: Develop a dashboard (web-based) where insurance agents may enter data concerning their customers to get results related to the immediate prediction of the risk and claims.

#### **DECLARATION**

We hereby declare that our dissertation is entirely our work and genuine / original. We understand that in case of discovery of any PLAGIARISM at any stage, our group will be assigned an F (FAIL) grade, and it may result in withdrawal of our bachelor’s degree.

Group members:

##### **Name Signature**

Dina Ahmed Mohamed Abdelaal \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Habiba Abdelkareem Ahmed Ali \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Mona Abdelaziz Ali Abdelaziz \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Mayar Ahmed Ibrahim Farag \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

#### **PLAIGRISM CERTIFICATE**

This is to certify that the project entitled “**-----------------------------------**”, which is being submitted here with for the award of the “**Bachelor of Computer and Artificial Intelligence Degree” in “Operations Research and Decision Support”**. This is the result of the original work by **Student 1** and **Student 2** under my supervision and guidance. The work embodied in this project has not been done earlier for the basis of award of any degree or compatible certificate or similar tile of this for any other diploma/examining body or university to the best of my knowledge and belief.

##### **Turnitin Originality Report**

Processed on:

ID:

Word Count:

Similarity Index

Similarity by Source

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Internet Sources: |  |  |  |  |  |  |
| Publications: |  |  |  |  |  |  |
| Student Papers: |  |  |  |  |  |  |
| Date: |  |  |  |  |  | Supervisor Name: Ihab Elkhodary |

#### **ACKNOWLEDGMENT**

we must express my gratitude for the guidance and invaluable support that our supervisor DR.Ehab-Elkhodary extended to us in the course of undertaking our graduation project. we would, therefore, like to appreciate their expertise and advice that has played a significant role in seeing this work to success. I genuinely appreciate his persistence and patience in helping me achieve my academic goals.

#### **TABLE OF CONTENTS**

Contents

[Ta2menak 1](#_Toc171277804)

[Ta2menak -تأمينك 2](#_Toc171277805)

[Chapter 1 11](#_Toc171277806)

[Introduction 11](#_Toc171277807)

[1.1 Car Insurance Types. 11](#_Toc171277808)

[1.1.1 Compulsory car insurance: 11](#_Toc171277809)

[1.1.2 Comprehensive (supplemental) car insurance: 11](#_Toc171277810)

[1.2) Problem statement 12](#_Toc171277811)

[1.2.1 High Fraud Rates: 12](#_Toc171277812)

[1.2.2 Inadequate Risk Assessment: 12](#_Toc171277813)

[1.2.3 Limited Data Availability: 12](#_Toc171277814)

[1.2.4 Inefficient Claims Processing: 13](#_Toc171277815)

[1.2.5 Low Insurance Penetration: 13](#_Toc171277816)

[1.3) Problem clarification 13](#_Toc171277817)

[If the number of clients that will claim premium insurance in 6 months exceeds their profit, car insurance companies can face several problems, including: 13](#_Toc171277818)

[If the type of insurance provided to clients is not suitable for their needs, several problems can arise: 14](#_Toc171277819)

[Car insurance companies can face several problems if they set premiums that are not optimal: 15](#_Toc171277820)

[1.3.4 If an insurance company makes a wrong decision to insure car someone, they can face several problems, including: 16](#_Toc171277821)

[1.4) Project Scope 17](#_Toc171277822)

[1.4.1 Risk Assessment Module: 17](#_Toc171277823)

[1.4.2 Risk Score Classification: 18](#_Toc171277824)

[1.4.3 Claim Prediction Model: 18](#_Toc171277825)

[1.5) Beneficiaries of this project 18](#_Toc171277826)

[1.5.1. Insurance companies: 18](#_Toc171277827)

[1.5.2. The Society: 18](#_Toc171277828)

[1.6) project timeframe 19](#_Toc171277829)

[Chapter 2 20](#_Toc171277830)

[2.1) Objectives of our model 20](#_Toc171277831)

[2.2) Expected output 20](#_Toc171277832)

[2.3) Activities 20](#_Toc171277833)

[2.3.1 Determine the project idea and objectives: 20](#_Toc171277834)

[2.3.2 Data collection: 20](#_Toc171277835)

[2.3.3 Data cleaning and processing: 21](#_Toc171277836)

[2.4.2 filling blank cells: 24](#_Toc171277837)

[2.5) Data Analysis 26](#_Toc171277838)

[Chapter 4 31](#_Toc171277839)

[4.1. The difference between the 3 Major Methods of Machine Learning. 31](#_Toc171277840)

[4.1.1 Main Reason for Choosing Supervised Learning: 32](#_Toc171277841)

[4.2) What are predictive analytics, and how does it work? 32](#_Toc171277842)

[4.3) The most widely used predictive models are. 33](#_Toc171277843)

[4.3.1 Decision trees: 33](#_Toc171277844)

[4.3.2 Regression (linear and logistic): 33](#_Toc171277845)

[4.3.3 Neural networks: 33](#_Toc171277846)

[Advantages of logistic regression: 34](#_Toc171277847)

[Disadvantages of logistic regression: 34](#_Toc171277848)

[Logistic Regression Equation and Assumptions: 35](#_Toc171277849)

[Key properties of the logistic regression equation: 36](#_Toc171277850)

[Key assumptions for implementing logistic regression: 36](#_Toc171277851)

[Types of logistic regression with examples: 37](#_Toc171277852)

[4.4.7) Summary of the points for choosing logistic regression: 38](#_Toc171277853)

[4.5) Boosting and Bagging Algorithms 38](#_Toc171277854)

[4.5.1 Bagging (Bootstrap Aggregating) 38](#_Toc171277855)

[4.5.3 Differences Between Bagging and Boosting 40](#_Toc171277856)

[4.6) Why Choose XGBoost Over Other Techniques? 40](#_Toc171277857)

[4.6.1 Performance 40](#_Toc171277858)

[4.6.2 Speed and Efficiency 40](#_Toc171277859)

[4.6.3 Flexibility 40](#_Toc171277860)

[4.6.4 Scalability 40](#_Toc171277861)

[4.6.5 Advanced Features 41](#_Toc171277862)

[4.6.6 Community and Ecosystem 41](#_Toc171277863)

[4.6.7 Implementation Details 41](#_Toc171277864)

[4.6.8 Evaluation Metrics 41](#_Toc171277865)

[Risk Assessment and Premium Calculation 42](#_Toc171277866)

[4.11.2 Fraud Detection 43](#_Toc171277867)

[4.11.3 Customer Retention and Marketing 43](#_Toc171277868)

[4.14.1 Handling Missing Values 44](#_Toc171277869)

[4.14.2 Encoding Categorical Variables 44](#_Toc171277870)

[4.14.3 Scaling and Normalization 44](#_Toc171277871)

[Model Evaluation and Interpretation 44](#_Toc171277872)

[4.14.5 Evaluation Metrics 44](#_Toc171277873)

[4.14.6 Feature Importance 44](#_Toc171277874)

[4.14.7 Visualization 45](#_Toc171277875)

[Chapter 5 46](#_Toc171277876)

[5.2.1 Data Preparation and Exploration 50](#_Toc171277877)

[5.2.3 Data Preprocessing 50](#_Toc171277878)

[5.2.4 Hyperparameter Optimization 50](#_Toc171277879)

[5.2.5 Model Evaluation 50](#_Toc171277880)

[Chapter 6 53](#_Toc171277881)

[6.1.1 Overview of the Project 53](#_Toc171277882)

[Features: 53](#_Toc171277883)

[6.1.3 Limitations: 53](#_Toc171277884)

[6.2.1 Enhanced Data Collection: 54](#_Toc171277885)

[6.2.3 Deployment and Integration: 54](#_Toc171277886)

[6.2.4 Continuous Monitoring and Updating: 55](#_Toc171277887)

#### **LIST OF FIGURES**

[Figure 1 ( (car-age col before filling blank) 24](#_Toc171277981)

[Figure 2 ( (car age col after filling blank) ) 24](#_Toc171277982)

[Figure 3 (income col before filling blank) 24](#_Toc171277983)

[Figure 4 (income col after filling blank) 24](#_Toc171277984)

[Figure 5 ( (Home\_val) col before filling blank) 25](#_Toc171277985)

[Figure 6 ( (income col after filling blank) 25](#_Toc171277986)

[Figure 7 (claim flag plot ) 26](#_Toc171277987)

[Figure 8 (education plot) 26](#_Toc171277988)

[Figure 9 (occupation plot) 27](#_Toc171277989)

[Figure 10 (car type plot) 28](#_Toc171277990)

[Figure 11 (urbanicity plot) 29](#_Toc171277991)

[Figure 12 (ROC plot) 29](#_Toc171277992)

[Figure 13 (confusion matrix) 30](#_Toc171277993)

[Figure 14 (logistic regression activation function) 35](#_Toc171277994)

[Figure 15 (logistic regression equation) 35](#_Toc171277995)

[Figure 16 (xgboost objective function) 41](#_Toc171277996)

[Figure 17(Data Types in Dataset) 46](#_Toc171277997)

[Figure 18 (blank cells in cleaned data) 46](#_Toc171277998)

[Figure 19 (Attributes names in dataset) 46](#_Toc171277999)

[Figure 20(target column before undersampling) 47](#_Toc171278000)

[Figure 21 (target column after undersampling with miority) 47](#_Toc171278001)

[Figure 22 (target column before undersampling) 47](#_Toc171278002)

[Figure 23 (target column after sampling with maority) 47](#_Toc171278003)

[Figure 24 (confusion matrix in majority undersampling) 47](#_Toc171278004)

[Figure 25 (confusion matrix in minority undersampling)) 48](#_Toc171278005)

[Figure 26(accuracy in majority undersampling) 48](#_Toc171278006)

[Figure 27(accuracy in minority undersampling) 48](#_Toc171278007)

[Figure 28 (roc curve of logistic regression) 48](#_Toc171278008)

[Figure 29 (AUC score of logistic regression) 49](#_Toc171278009)

[Figure 30 (PCA for logistic regression) 49](#_Toc171278010)

[Figure 31 (accuracy of logistic regression) 49](#_Toc171278011)

[Figure 32(Initial Data Inspection) 50](#_Toc171278012)

[Figure 33(Model Performance) 51](#_Toc171278013)

[Figure 34 (PCA Feature) 51](#_Toc171278014)

[Figure 35 (Feature importance) 52](#_Toc171278015)

[Figure 36 (Best cross valdition score ) 52](#_Toc171278016)

[Figure 37 (XGBoost Accuracy) 52](#_Toc171278017)

# Chapter 1

## Introduction

The changing landscape within the automobile insurance sector demands that adaptation and innovation occur urgently. Companies must now understand that, slowly but surely, vehicles are plying the roads, and they need to be at the forefront of developments. This means the understanding and mitigation of risks generally associated with automobile insurance, juxtaposed with advances in technology and data-driven insights at a proactive level. A majority of the insurance space has not proven to be profitable for more than a decade, with underwriting loss trends recorded by high combined ratios across the industry. For most of the previous reasons, like environmental concerns and social inflations and supply chain distributions, Egypt's insurance industry is one of the oldest in the region, and it's viewed as very competitive more than the other emerging markets.

The first if you don't know that there are several types of insurances for cars

## Car Insurance Types.

### 1.1.1 Compulsory car insurance:

Which must be obtained as a necessary document for any car to have permission to be used and driven in public and must be paid while obtaining the car license.

### 1.1.2 Comprehensive (supplemental) car insurance:

There are three types of supplemental car insurance policies.

**1.1.2.1 Insurance policy for commercial purposes**

Designed for cars that have commercial purposes such as taxes, buses, rented cars, and some types of private cars that have commercial purposes such as agricultural tractors, ambulances fire trucks, etc.

**1.1.2.2 Insurance policy for license plates**

license plates designed for newly imported cars or cars that are still in the auto show which means they are still in possession of those who manufacture or trade in cars or the maintenance centers that repair them.

**1.1.2.3 Insurance policy for private cars**

It covers any car which is used for personal purposes except race cars.

Policy for private cars has many types of policies of supplementary car insurance.

**1.1.2.3.1 Liability coverage**

Pays for property damage and/ or injury to another person caused by an accident for which you are responsible.

This coverage is divided into two parts:

* Property damage.
* bodily injury.

**1.1.2.3.2 third party, fire, and theft (TPFT)**

It means the personnel is covered for three specific areas.

* damage to someone else's vehicle.
* if the vehicle accidentally catches fire, and it’s also insured for theft or attempted theft of the vehicle.
* It’s also covering third-party claims which means that in case you caused damage to another vehicle due to your driving.

**1.1.2.3.3 collision coverage**

It covers the person in the case that a car is damaged in an accident with another vehicle or object such as a fence or tree.

**1.1.2.3.4 Comprehensive coverage**

This policy usually includes everything in the previous policies, as well as adding extra coverage for person’s vehicles, persons, and any other passengers.

## Problem statement

High costs of repairs, car parts, and other related factors have forced insurance companies to adjust their rates. much of this due to environmental, social inflation, and chain distribution.

Car insurance companies in Egypt face a range of challenges that hinder their growth and profitability. These challenges include:

### 1.2.1 High Fraud Rates:

Fraudulent claims pose a significant financial burden on car insurance companies in Egypt. Fraudsters often manipulate accidents or inflate repair costs to obtain illegitimate payouts. This rampant fraud erodes the financial stability of insurance companies and drives up premiums for policyholders.

### 1.2.2 Inadequate Risk Assessment:

Traditional risk assessment methods for car insurance often rely heavily on demographic factors, such as age, gender, and driving experience. While these factors play a role in determining risk, they may not capture the full picture, particularly in Egypt's complex traffic environment.

### 1.2.3 Limited Data Availability:

The lack of comprehensive and standardized accident data hinders insurance companies' ability to accurately assess and price risk. This data scarcity can lead to misaligned premiums, either overcharging low-risk drivers or undercharging high-risk drivers.

### 1.2.4 Inefficient Claims Processing:

The current claims processing system in Egypt is often manual, slow, and prone to errors. This inefficiency delays payouts to policyholders and increases administrative costs for insurance companies.

### 1.2.5 Low Insurance Penetration:

Despite the growing number of vehicles on Egyptian roads, insurance penetration remains low. This is partly due to a lack of awareness about the benefits of car insurance and the affordability of premiums.

## Problem clarification

A lot of problems can face car insurance companies if they assess their risk in a wrong way such as.

### If the number of clients that will claim premium insurance in 6 months exceeds their profit, car insurance companies can face several problems, including:

**1.3.1.1 Financial instability:**

If claims exceed premiums, insurance companies will have to dip into their reserves to cover the payouts. This can deplete their reserves and make them vulnerable to financial instability. In extreme cases, it could even lead to insolvency.

**1.3.1.2 Increased premiums:**

To recoup their losses, insurance companies may be forced to raise premiums for all policyholders, even those who haven't filed claims. This can make car insurance unaffordable for many people.

**1.3.1.3 Damage to reputation:**

A high number of claims can damage the reputation of an insurance company, making it difficult to attract new customers and retain existing ones.

**1. 3.1.4 Regulatory scrutiny:**

Insurance companies are regulated by state governments, and excessive claims could attract the attention of regulators. Regulators may investigate the company's underwriting practices or impose additional requirements, which could increase costs and reduce profitability.

these immediate problems, an excessive number of claims can also have long-term consequences for car insurance companies, such as:

* Reduced investment in innovation:

Insurance companies may be forced to cut back on investments in new products, services, and technologies to conserve capital. This can make them less competitive and less able to meet the needs of their customers.

* Exit from the market:

In extreme cases, insurance companies may be forced to exit the market altogether if they are unable to manage their claims costs. This can reduce competition and make it more difficult for consumers to find affordable insurance.

### If the type of insurance provided to clients is not suitable for their needs, several problems can arise:

**1.3.2.1 For clients:**

* **Uncovered expenses:**

If clients have insufficient coverage for their actual needs, they may be left financially responsible for a large portion of repair costs or medical bills in case of an accident.

* **Limited protection:**

Clients may lack adequate protection against specific risks they face, leaving them vulnerable to financial losses.

* **False sense of security:**

Clients might believe they have comprehensive coverage, leading to disappointment and frustration when claims are denied or fall short of expectations.

* **Loss of trust:**

Clients may lose trust in the insurance company and its agents, potentially switching to other providers.

**1.3.2.2 For insurance companies:**

* **Increased claims:**

Unsuitable coverage can lead to claims exceeding expectations, causing financial strain, and impacting profitability.

* **Fraudulent claims:**

Clients with inadequate coverage might be tempted to exaggerate or falsify claims to access benefits they originally didn’t purchase.

* **Regulatory issues:**

Insurance companies might face regulatory scrutiny and penalties for misrepresenting or selling inadequate coverage.

* **Damaged reputation:**

Negative experiences from clients can damage the company's reputation, hindering customer acquisition and retention.

* **Increased costs:**

Companies might need to invest in additional resources and training to ensure proper risk assessment and product suitability.

**1.3.2.3 Additional problems:**

* **Market instability:**

If a significant portion of clients have unsuitable coverage, it can create market instability, impacting pricing and overall insurance accessibility.

* **Public perception:**

Negative publicity from dissatisfied clients can negatively impact public perception of the insurance industry.

### Car insurance companies can face several problems if they set premiums that are not optimal:

**1.3.3.1 For clients:**

* **Unaffordability:**

Premiums that are too high can make car insurance unaffordable for many people, leading to a decline in insurance penetration. This can create a situation where only high-risk drivers remain insured, further pushing up premiums and creating a vicious cycle.

* **Insufficient coverage:**

If premiums are too low, insurance companies may not generate enough revenue to cover claims. This could lead to financial instability for the company, potentially affecting future payouts to policyholders.

* **Unfair Pricing:**

Premiums that are not based on individual risk profiles can lead to unfair pricing, where low-risk drivers subsidize the costs of high-risk drivers. This can discourage safe driving behavior and create resentment among policyholders.

* **Market instability:**

Inaccurate pricing can create instability in the insurance market, leading to unpredictable fluctuations in premiums and making it difficult for consumers to budget for car insurance.

**1.3.3.2 For insurance companies:**

* **Financial losses:**

If premiums are too low, insurance companies may incur significant financial losses from claims exceeding their revenue. This can threaten their solvency and lead to increased regulatory scrutiny.

* **Adverse selection:**

Low premiums can attract high-risk drivers, disproportionately increasing claims and further jeopardizing financial stability.

* **Loss of competitiveness:**

Overpriced insurance can drive customers to competitors offering more competitive premiums, leading to market share loss.

* **Damaged reputation:**

Unfair pricing practices and inadequate payouts can damage the company's reputation, making it difficult to attract and retain customers.

* **Increased regulatory scrutiny:**

Insurance regulators closely monitor pricing practices to ensure fairness and financial stability. Inaccurate or unfair pricing can lead to fines, penalties, and additional reporting requirements.

**1.3.3.3 Additional problems:**

* **Reduced innovation:**

Financial instability can hinder investments in new technologies and innovative products, making the company less competitive in the long run.

* **Limited market access:**

Companies with unsustainable pricing models may find it difficult to enter new markets or expand their existing customer base.

* **Reduced consumer trust:**

Unfair pricing and financial instability can erode consumer trust in the insurance industry.

### 1.3.4 If an insurance company makes a wrong decision to insure car someone, they can face several problems, including:

* + - 1. **Financial problems:**
* **Increased claims:**

Insuring high-risk drivers can lead to a significant increase in claims, exceeding the premiums collected. This can cause financial losses and threaten the company's solvency.

* **Reserve depletion:**

Insurance companies need to maintain adequate reserves to cover future claims. If they deplete their reserves due to excessive payouts, they may become financially unstable and unable to meet their obligations.

* **Higher reinsurance costs:**

Reinsurance companies help insurance companies mitigate risk by sharing a portion of their claims burden. However, ensuring high-risk individuals can lead to higher reinsurance costs, further impacting the company's finances.

**1.3.4.2 Legal issues:**

* **Fraudulent claims:**

High-risk drivers might be more likely to submit fraudulent claims, leading to legal entanglements and financial losses.

Regulatory penalties: Insurance companies are subject to strict regulations and must comply with fair pricing and underwriting practices. Wrongly ensuring high-risk individuals can lead to regulatory scrutiny, fines, and even penalties.

* **Litigation:**

If an insurance company denies a claim based on a wrong decision to insure someone, they may face legal challenges and potential lawsuits.

**1.3.4.3 Reputational damage:**

* **Negative publicity:**

If an insurance company is known for insuring high-risk drivers and paying out excessive claims, it can damage its reputation and make it difficult to attract new customers.

* **Loss of trust:**

Policyholders who discover that the company ensures high-risk individuals may lose trust and switch to competitors.

* **Difficulty Attracting Safe Drivers:**

A reputation for insuring high-risk drivers can make it difficult to attract safe drivers, leading to an adverse selection problem and further financial instability.

**1.3.4.4 Other problems:**

* **Increased workload:**

Ensuring high-risk individuals requires more resources to manage claims, investigate fraud, and handle legal issues, leading to higher administrative costs.

* **Market instability:**

Wrongly ensuring high-risk individuals can contribute to market instability, making it difficult for insurance companies to accurately price their products and assess their financial risks.

* **Reduced innovation:**

Financial strain from excessive claims can hinder the company's ability to invest in new technologies and innovative products, making it less competitive in the long run.

## **1.4) Project Scope**

In essence, our tool aids companies by automatically identifying, from their extensive portfolios, the customers most likely to benefit from expanding their insurance coverage. This application goes beyond mere insurance recommendations; it is crafted to offer you peace of mind. We aim to streamline the insurance shopping process, enabling you to concentrate on the journey ahead without concerns about insufficient coverage.

### 1.4.1 Risk Assessment Module:

Develop a risk assessment module that utilizes machine learning algorithms and data analytics to evaluate the risk associated with each car insurance customer.

Incorporate relevant data points such as occupation, number of children who can drive, motor points, geographical location, customer age, and other relevant factors.

### 1.4.2 Risk Score Classification:

Define thresholds for risk scores to categorize customers into low, medium, and high-risk categories.

### 1.4.3 Claim Prediction Model:

Create a predictive model that analyzes historical data to forecast whether a customer is likely to claim the insurance premium within the next 6 months or not to avoid loss.

Utilize machine learning techniques such as regression or classification algorithms to build the prediction model.

## 1.5) Beneficiaries of this project

The project to determine insurance eligibility and conditions for cars is a valuable tool for many parties. By carefully assessing risks and determining appropriate insurance terms, this project can benefit both insurance companies and customers alike.

### 1.5.1. Insurance companies:

**1.5.1.1 Improve the accuracy of risk assessment.**

Accurately determine the customer's likelihood of accepting insurance.

Use a predictive model that combines personal information with vehicle information, driving history, and vehicle use. Assess the client's likelihood of being involved in an accident.

* Focus on low-risk clients.
* Accept clients with low loss potential. Reduce the risk of financial losses.
* Increase the profits of insurance companies.
* Reduce financial losses.

**1.5.1.2 Increase the efficiency of the insurance process.**

* Reducing the time and effort spent on customer evaluation.
* Improve customer service.

**1.5.1.3 Expanding the scope of insurance business.**

* Attract new customers.

### 1.5.2. The Society:

**1.5.2.1 Improving financial stability:**

* Helping individuals deal with unexpected events.
* Improving the financial stability of individuals.

**1.5.2.2 Promoting economic stability:**

* Reducing the burden on the financial system.
* Improving economic stability.

## 1.6) project timeframe

* 2/10/2023: start to define the project idea.
* 5/10/2023: search for real data in real insurance companies and online
* 1/11/2023: problem formulation.
* 29/12/2023:data cleaning
* 29/1/2024: searching for the boosting model.
* 7/2/2024:data preparation.

# Chapter 2

## Objectives of our model

* To assess risk for car insurance customers to minimize it as much as possible.
* To predict if the customer will claim the insurance premium within 6 months or not.
* To accept or reject making new insurance policies for customers depending on their risk score.

## Expected output

* Show the risk score for each company customer after calculating it depending on some important features such as policyholder age, number of home kids of the policyholder, age of the car, and other features.
* Print an expected claim status for each company customer (1 if the insurance has already been done 0 otherwise)
* Deciding if the insurance company will make a new insurance policy for that customer or reject it depending on his/her risk score.

## Activities

Manually evaluating insurance eligibility and insurance terms for cars is a tedious and time-consuming process, which can lead to delays in customer service and evaluation errors.

### 2.3.1 Determine the project idea and objectives:

* Improving the accuracy of insurance eligibility assessment.
* Reducing the time and effort spent on customer evaluation.
* Improve customer service.
* Increase the efficiency of the insurance process.

### 2.3.2 Data collection:

Data sources:

* Customer personal data (name, age, gender, profession, ...).
* Information about the car (type of car, year of manufacture, car value, ...).
* Information about driving record (number of accidents, traffic violations, ...).
* Information about car use (annual mileage, car use (personal, commercial, ...)).

### 2.3.3 Data cleaning and processing:

**2.3.3.1 Removal of incomplete or contradictory data:**

Check for null values:

Delete data that contains empty values in basic fields (such as customer name, car type, year of manufacture).

Replace blank values in non-essential fields with default values (such as annual mileage).

Data validation:

Verify the validity of personal information (such as date of birth, and phone number).

Verify the validity of the vehicle information (such as chassis number, and year of manufacture).

Verify the validity of driving record information (such as the number of accidents, and traffic violations).

**2.3.3.2 Handling missing data:**

Use completion techniques:

Completing missing data using the most common mean, median, or value.

Using machine learning models to predict missing data.

Remove data that contains many missing values:

Delete data that contains more than 50% missing values.

Delete data that contains missing values in key fields.

**2.3.3.3 Convert data into a format suitable for analysis:**

Convert data to CSV or Excel format:

Easy analysis using data analysis tools like Pandas and NumPy.

Easy integration with insurance systems.

Standardization of data formats:

Standardize date and time formats.

Standardize units of measurement (such as distance, and currency).

Create new variables:

Create new variables by combining data (such as the customer's age when purchasing the car).

Create new variables through data analysis (eg safe driving record).

* 1. **Data we have used for our project**

So, the data we have used for our project we have found from Kaggle was most suitable for our objectives. It describes driver’s behavior, characteristics and his demographic characteristic.

|  |  |  |
| --- | --- | --- |
| **column name** | **Description** | **hypothetical effect** |
| ID | Identification |  |
| KIDSDRIV | number of driving children | children driving cars means a higher risk of collision |
| BIRTH | date of driver’s birth |  |
| AGE | age of driver | younger drivers tend to be riskier |
| HOMEKIDS | number of children at home | having children may reduce tendencies of driving dangerously |
| INCOME | income | less income may cause risk in paying insurance premiums |
| PARENT1 | single parent | single parents may be more tired and possibly less alert |
| HOME\_VAL | home value | higher home values may be correlated with income |
| MSTATUS | marital status | single and married people may have different driving behaviors |
| GENDER | Gender | urban legends says that women crash less than men do |
| EDUCATION | highest level of education | more educated people should be more careful when driving |
| OCUUPATION | seniority at work | different levels of profession may be correlate with different driving styles |
| TRAVTIME | distance to work | longer journeys are at higher risk of collisions than shorter ones |
| CAR\_USE | purpose of vehicle | private and commercial vehicles may be driven differentially |
| BLUEBOOK | value of vehicle | the higher price of the car the more careful its driver |
| TIF | time in-force | time the policy has been in force longer term customers |
| CAR\_TYPE | type of vehicle | unknown effect |
| RED\_CAR | whether or not the car color is red | another urban legend amongst insurers that red cars are riskier |
| OLDCLAIM | total claims in the past 5 years | previous behavior as an indicator of future likelihood |
| CLM\_FREQ | claim frequency in the past 5 years | previous behavior as an indicator of future likelihood |
| REVOKED | license revoked in the past 7 years | having license revoked is an indication of bad road behavior |
| MVR\_PTS | motor vehicle record points | if there are lots of traffic violations there may be higher risk of collisions |
| CLAIM\_AMT | claim amount | claim amount if a claim occurred |
| CAR\_AGE | age of car | newer cars tend to have more safety features |
| CLAIM\_FLAG | claim or no claim | true/false of whether a claim occurred |
| URBANCITY | urban or rural | urban environments may have more vehicle density than rural areas |

**2.4.1 dropping unnecessary data:**

2.4.1.1 we started to clean the data by dropping/removing commercial cars from car types as our project specialized for private cars due to complexity of assessing risk for commercial cars.

2.4.1.2 we dropped red car column as it’s not assessment risk technique in real life

2.4.1.3 we dropped birth column as age of the driver is the same meaning

### 2.4.2 filling blank cells:

2.4.2.1 filling blank cells in income column with average of income based on his/her occupation

A screenshot of a screen

Description automatically generatedA screenshot of a spreadsheet

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

Figure 4 (income col after filling blank)

Figure 1 ( (car-age col before filling blank)

Figure 2 ( (car age col after filling blank) )

Figure 3 (income col before filling blank)

2.4.2.2 filling blank cells in car age by mode



2.4.2.3 filling blank cells in home value with average based on demographic area

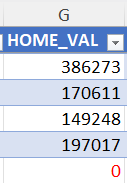
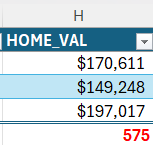


Figure 5 ( (Home\_val) col before filling blank)

Figure 6 ( (income col after filling blank)

## Data Analysis

A green and white bar graph

Description automatically generated

Figure 7 (claim flag plot )

this graph shows the number of claims whether it occurred or not .

1436 (22.05%) claimed and 5077 (77.95%) do not.

A graph of a number of students

Description automatically generated with medium confidence

Figure 8 (education plot)

this graph shows that the persons who have the degree of masters are more safe

Because in bachelor 1323 (26.06%) do not claimed but 309 (21.12%)claimed

In high school1927 (37.95%)do not claim but 776 (54.04%)claimed.

In masters 1285 (25.3%)do not claim but 261 (18.17%) claimed.

A graph showing the number of jobs in the united states

Description automatically generatedIn PHD 542 (10.67%)do not claim but 90 (6.2%)claimed.

Figure 9 (occupation plot)

this graph shows that the persons who are manager are safer.

In Professional there are 853 (16.8%) do not claim but 205 (14.27%) claimed.

In manager there are 835 (16.45%) do not claim but 119 (8.29%) claimed.

In clerical there are 928 (18.28%) do not claim but 382 (26.6%) claimed.

In Blue collar there are 408 (8.04%) do not claim but 158 (11%) claimed.

In Doctor there are 307 (6.05%) do not claim but 41 (2.85%) claimed.

In lawyer there are 850 (16.74%) do not claim but 188 (13.1%) claimed.

In Home maker there are 582 (11.46%) do not claim but 204 (14.21%) claimed.

In Student there are 314 (6.18%) do not claim but 139 (9.68%) claimed.

A graph of a number of cars

Description automatically generated

Figure 10 (car type plot)

This graph shows that the persons who drive minivan are more safe.

Because in minivan there are 1830 (36.04%) do not claim but 311 (21.7%) claimed.

In pickup there are 563 (11.09%) do not claim but 141 (9.82%) claimed.

In sports car there are 662 (13.04%) do not claim but 317(22.07%) claimed.

In SUV cars there are 1722 (33.92%) do not claim but 606(42.2%) claimed.

In van cars there are 300 (5.91%) do not claim but 61(4.25%) claimed.

A graph showing a number of different cities

Description automatically generated with medium confidence

Figure 11 (urbanicity plot)

This graph shows that the persons who live in rural areas are more safe.

Because in rural areas 1188 (23.4%) do not claim but 81 (5.6%) claimed.

In urban areas 3889 (76.6%) do not claim but 1355(94.3%) claimed.

A line graph with a line graph and a line graph

Description automatically generated with medium confidence

Figure 12 (ROC plot)

This graph shows that the standardization and visualization of the training data.

A chart of a number of colored squares

Description automatically generated with medium confidence

Figure 13 (confusion matrix)

First Row:

Low Risk (Actual Class):

Low Risk (Predicted Class): 19 customers were correctly classified as low risk.

Medium Risk: 0 customers were incorrectly classified as medium risk.

High Risk: 0 customers were incorrectly classified as high risk.

Second Row:

Medium Risk (Actual Class):

Low Risk: 0 customer was incorrectly classified as low risk.

Medium Risk: 11 customers were correctly classified as medium risk.

High Risk: 2 customers were incorrectly classified as high risk.

Third Row:

High Risk (Actual Class):

Low Risk: 0 customers were incorrectly classified as low risk.

Medium Risk: 0 customers were incorrectly classified as medium risk.

High Risk: 13 customers were correctly classified as high risk.

# Chapter 4

For our second objective which is predicting if the customer will claim the insurance premium within 6 months or not, we searched for many prediction algorithms and reached for the suitable algorithm and will explain in this chapter the steps we took to find it but before we talk about the steps we will talk about some concepts.

## 4.1. The difference between the 3 Major Methods of Machine Learning.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Supervised Learning | Unsupervised Learning | Reinforcement Learning |
| Data | Relies on labeled data, where each data point has a corresponding label or desired output. Imagine a dataset of emails with "spam" or "not spam" labels for each email | Deals with unlabeled data, where the data points lack predefined labels. The goal is to uncover hidden patterns within the data itself. Think of a dataset of customer purchases, where unsupervised learning can identify customer groups with similar buying habits. | Doesn't use labeled data in the traditional sense. The learning agent interacts with its environment, takes actions, and receives rewards or penalties for those actions. Through trial and error, the agent learns optimal behavior. Imagine an AI playing a game; it gets rewarded for winning and penalized for losing, gradually learning the best strategies. |
| Goal | Aims to build a model that can predict future outcomes based on the learned relationship between labeled inputs and outputs. This could be classifying emails as spam or predicting housing prices based on features like size and location. | Focuses on discovering hidden structures within the data. This might involve tasks like grouping customers into segments, identifying anomalies in sensor data, or summarizing document topics. | Seeks to train an agent to make optimal decisions within an environment by maximizing its reward. This is useful for robotics control, game playing, and dynamic resource allocation problems. |
| Learning Style | Learns from examples (labeled data) provided by a teacher (human programmer). | Learns by exploring the data on its own, identifying patterns and relationships without guidance. | Learns through trial and error by interacting with the environment and receiving feedback in the form of rewards or penalties. |

### 4.1.1 Main Reason for Choosing Supervised Learning:

Data Nature: The available data is labeled, meaning each data point has a corresponding label or desired outcome.

Goal: We aim to build a model that can predict future outcomes based on the learned relationship between labeled inputs and outputs.

Suitability of Supervised Learning Algorithms: Supervised learning algorithms are specifically designed to handle labeled data, making them well-suited for our needs.

So Supervised learning is the optimal choice for Labeled data availability, models are easier to understand, which can be important for explaining decisions. good for single outcomes, while unsupervised or reinforcement learning might be better for complex tasks.

## 4.2) What are predictive analytics, and how does it work?

A common misconception is that predictive analytics and machine learning are the same thing. This is not the case. (Where the two do overlap, however, is predictive modelling – but more on that later.)

At its core, predictive analytics encompasses a variety of statistical techniques (including machine learning, predictive modelling and data mining) and uses statistics (both historical and current) to estimate, or ‘predict’, future outcomes. These outcomes might be behaviors a customer is likely to exhibit or possible changes in the market, for example. Predictive analytics help us to understand possible future occurrences by analyzing the past.

Machine learning, on the other hand, is a subfield of computer science that, as per Arthur Samuel’s definition from 1959, gives ‘computers the ability to learn without being explicitly programmed’. Machine learning evolved from the study of pattern recognition and explores the notion that algorithms can learn from and make predictions on data. And, as they begin to become more ‘intelligent’, these algorithms can overcome program instructions to make highly accurate, data-driven decisions.

Predictive analytics is driven by predictive modelling. It’s more of an approach than a process. Predictive analytics and machine learning go hand-in-hand, as predictive models typically include a machine learning algorithm. These models can be trained over time to respond to new data or values, delivering the results the business needs. Predictive modelling largely overlaps with the field of machine learning.

There are two types of predictive models. They are Classification models, that predict class membership, and Regression models that predict a number. These models are then made up of algorithms. The algorithms perform data mining and statistical analysis, determining trends and patterns in data. Predictive analytics software solutions will have built-in algorithms that can be used to make predictive models. The algorithms are defined as ‘classifiers’, identifying which set of categories data belongs to.

## The most widely used predictive models are.

### 4.3.1 Decision trees:

Decision trees are a simple, but powerful form of multiple variable analysis. They are produced by algorithms that identify various ways of splitting data into branch-like segments. Decision trees partition data into subsets based on categories of input variables, helping you to understand someone’s path of decisions.

### 4.3.2 Regression (linear and logistic):

Regression is one of the most popular methods in statistics. Regression analysis estimates relationships among variables, finding key patterns in large and diverse data sets and how they relate to each other.

### 4.3.3 Neural networks:

Patterned after the operation of neurons in the human brain, neural network (also called artificial neural networks) are a variety of deep learning technologies. They’re typically used to solve complex pattern recognition problems – and are incredibly useful for analyzing large data sets. They are great at handling nonlinear relationships in data – and work well when certain variables are unknow.

**In our project we used logistic regression as base model, and we will explain it in detail and explain how we use it in our project.**

* 1. **What Is Logistic Regression?**

Logistic regression is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation. The model delivers a binary or dichotomous outcome limited to two possible outcomes: yes/no, 0/1, or true/false.

Logical regression analyzes the relationship between one or more independent variables and classifies data into discrete classes. It is extensively used in predictive modeling, where the model estimates the mathematical probability of whether an instance belongs to a specific category or not.

For example, 0 – represents a negative class; 1 – represents a positive class. Logistic regression is commonly used in binary classification problems where the outcome variable reveals either of the two categories (0 and 1).

### Advantages of logistic regression:

* + - 1. **Simplicity and Interpretability:**

Logistic regression is easy to understand and implement. The model coefficients provide insights into how each feature influences the outcome, making it a good choice for building interpretable models.

**4.4.1.2 Efficiency:**

Training a logistic regression model is computationally efficient, especially compared to more complex algorithms like neural networks. This makes it ideal for large datasets.

**4.4.1.3 Probabilistic Outputs:**

Logistic regression doesn't just predict a class label (e.g., spam/not spam), it also outputs the probability of belonging to each class. This provides valuable confidence scores for predictions.

* + - 1. **Works Well with Linearly Separable Data:**

When the data can be clearly separated into classes by a straight line, logistic regression often performs very well.

Flexible for Multi-class Classification: Logistic regression can be extended to handle more than two classes (multinomial logistic regression) with some adjustments.

### Disadvantages of logistic regression:

* + - 1. **Assumes Linearity:**

Logistic regression assumes a linear relationship between the independent variables and the outcome. If the underlying relationships are more complex, the model may not capture them accurately.

* + - 1. **Susceptible to Overfitting:**

With a high number of features or a small dataset, logistic regression can overfit the training data and perform poorly on unseen data. Regularization techniques can help mitigate this.

* + - 1. **Limited to Binary Classification by Default:**

The basic form of logistic regression is designed for binary classification (two classes). Multi-class problems require modifications.

* + - 1. **Not Ideal for High-Dimensional Data:**

When dealing with a large number of features, logistic regression may struggle, especially if there are issues like multicollinearity (correlated features).

* + - 1. **Performance Can be Outshined by More Complex Models:**

For certain tasks, especially with very complex data, other machine learning algorithms like decision trees or neural networks might achieve higher accuracy.

### Logistic Regression Equation and Assumptions:

Logistic regression uses a logistic function called a sigmoid function to map predictions and their probabilities. The sigmoid function refers to an S-shaped curve that converts any real value to a range between 0 and 1.

Moreover, if the output of the sigmoid function (estimated probability) is greater than a predefined threshold on the graph, the model predicts that the instance belongs to that class. If the estimated probability is less than the predefined threshold, the model predicts that the instance does not belong to the class.

For example, if the output of the sigmoid function is above 0.5, the output is considered as 1. On the other hand, if the output is less than 0.5, the output is classified as 0. Also, if the graph goes further to the negative end, the predicted value of y will be 0 and vice versa. In other words, if the output of the sigmoid function is 0.65, it implies that there are 65% chances of the event occurring: a coin toss, for example.

The sigmoid function is referred to as an activation function for logistic regression and is defined as:

A black and white math symbols

Description automatically generated with medium confidenceA math equation with numbers and plus

Description automatically generated

Figure 14 (logistic regression activation function)

where,

* e = base of natural logarithms
* value = numerical value one wishes to transform

The following equation represents logistic regression:

Figure 15 (logistic regression equation)

where,

* x = input value
* y = predicted output
* b0 = bias or intercept term
* b1 = coefficient for input (x)

### Key properties of the logistic regression equation:

* Logistic regression’s dependent variable obeys ‘Bernoulli distribution.
* Estimation/prediction is based on ‘maximum likelihood.
* Logistic regression does not evaluate the coefficient of determination (or R squared) as observed in linear regression. Instead, the model’s fitness is assessed through a concordance.

### Key assumptions for implementing logistic regression:

* + - 1. **The dependent/response variable is binary or dichotomous:**

The first assumption of logistic regression is that response variables can only take on two possible outcomes – pass/fail, male/female, and malignant/benign. This assumption can be checked by simply counting the unique outcomes of the dependent variable. If more than two possible outcomes surface, then one can consider that this assumption is violated.

* + - 1. **Little or no multicollinearity between the predictor/explanatory variables:**

This assumption implies that the predictor variables (or the independent variables) should be independent of each other. Multicollinearity relates to two or more highly correlated independent variables. Such variables do not provide unique information in the regression model and lead to wrongful interpretation. The assumption can be verified with the variance inflation factor (VIF), which determines the correlation strength between the independent variables in a regression model**.**

* + - 1. **Linear relationship of independent variables to log odds:**

Log odds refer to the ways of expressing probabilities. Log odds are different from probabilities. Odds refer to the ratio of success to failure, while probability refers to the ratio of success to everything that can occur.

For example, consider that you play twelve tennis games with your friend. Here, the odds of you winning are 5 to 7 (or 5/7), while the probability of you winning is 5 to 12 (as the total games played = 12).

* + - 1. **Prefers large sample size:**

Logistic regression analysis yields reliable, robust, and valid results when a larger sample size of the dataset is considered.

This assumption can be validated by taking into account a minimum of 10 cases considering the least frequent outcome for each estimator variable. Let’s consider a case where you have three predictor variables, and the probability of the least frequent outcome is 0.30. Here, the sample size would be (10\*3) / 0.30 = 100.

* + - 1. **Problem with extreme outliers:**

Another critical assumption of logistic regression is the requirement of no extreme outliers in the dataset.

This assumption can be verified by calculating Cook’s distance (Di) for each observation to identify influential data points that may negatively affect the regression model. In situations when outliers exist, one can implement the following solutions:

* Eliminate or remove the outliers
* Consider a value of mean or median instead of outliers, or
* Keep the outliers in the model but maintain a record of them while reporting the regression results
  + - 1. **Consider independent observations:**

This assumption states that the dataset observations should be independent of each other. The observations should not be related to each other or emerge from repeated measurements of the same individual type.

The assumption can be verified by plotting residuals against time, which signifies the order of observations. The plot helps in determining the presence or absence of a random pattern. If a random pattern is present or detected, this assumption may be considered violated.

### Types of logistic regression with examples:

* + - 1. **Binary logistic regression:**

Binary logistic regression predicts the relationship between the independent and binary dependent variables. Some examples of the output of this regression type may be, success/failure, 0/1, or true/false.

Examples:

* Deciding on whether or not to offer a loan to a bank customer: Outcome = yes or no.
* Evaluating the risk of cancer: Outcome = high or low.
* Predicting a team’s win in a football match: Outcome = yes or no.
  + - 1. **Multinomial logistic regression:**

A categorical dependent variable has two or more discrete outcomes in a multinomial regression type. This implies that this regression type has more than two possible outcomes.

Examples:

* Let’s say you want to predict the most popular transportation type for 2040. Here, transport type equates to the dependent variable, and the possible outcomes can be electric cars, electric trains, electric buses, and electric bikes.
* Predicting whether a student will join a college, vocational/trade school, or corporate industry.
* Estimating the type of food consumed by pets, the outcome may be wet food, dry food, or junk food.
  + - 1. **Ordinal logistic regression:**

Ordinal logistic regression applies when the dependent variable is in an ordered state (i.e., ordinal). The dependent variable (y) specifies an order with two or more categories or levels.

Examples: Dependent variables represent,

* Formal shirt size: Outcomes = XS/S/M/L/XL
* Survey answers: Outcomes = Agree/Disagree/Unsure
* Scores on a math test: Outcomes = Poor/Average/Good

### 4.4.7) Summary of the points for choosing logistic regression:

Easy to understand and implement: Simple concept and straightforward structure, a well-understood and widely used algorithm, and there are many resources available for learning about it and implementing it.

High computational efficiency: Fast processing of large datasets.

Probability outputs: Determining the probability of an observation belonging to a specific class.

Effective with linear data: Excellent performance with data that has a linear relationship between variables.

Flexible for multi-class classification: Can be modified for multi-class classification tasks, Logistic regression is a discriminative model, which means it learns a decision boundary that separates the data into two classes.

## Boosting and Bagging Algorithms

Boosting and Bagging are two popular ensemble learning techniques used in machine learning to improve the accuracy and robustness of models. Both methods aim to create a strong classifier by combining multiple weak classifiers, but they do so in different ways.

### 4.5.1 Bagging (Bootstrap Aggregating)

**4.5.1.1 Overview:** Bagging is a method that generates multiple versions of a predictor and uses these to get an aggregated predictor. This approach reduces variance and helps to avoid overfitting.

**4.5.1.2 Steps:**

1. **Data Sampling:** Create multiple subsets of the original dataset by sampling with replacement (bootstrap sampling).
2. **Training:** Train a base model (e.g., decision tree) on each subset independently.
3. **Aggregation:** Combine the predictions from all the base models by averaging (for regression) or voting (for classification).

**4.5.1.3 Advantages:**

* Reduces variance and helps to prevent overfitting.
* Simple and easy to implement.
* Can be parallelized since each model is trained independently.

**4.5.1.4 Common Algorithms:**

* Random Forest (an extension of bagging using decision trees).

**4.5.2 Boosting**

**4.5.2.1 Overview:** Boosting focuses on training weak models sequentially, with each new model attempting to correct the errors of the previous ones. This approach reduces both bias and variance.

**4.5.2.2 Steps:**

1. **Initialize:** Start with an initial model and assign equal weights to all data points.
2. **Sequential Training:** Train a series of models, each one focusing more on the data points that were misclassified by the previous models.
3. **Weight Adjustment:** Adjust the weights of the data points based on the errors. Misclassified points get higher weights.
4. **Final Model:** Combine all the weak models through a weighted sum (for regression) or weighted voting (for classification).

**4.5.2.3 Advantages:**

* Reduces both bias and variance.
* Can handle complex relationships by focusing on difficult cases.
* Often provides better accuracy than bagging.

**4.5.2.4 Common Algorithms:**

* AdaBoost (Adaptive Boosting)
* Gradient Boosting Machines (GBM)
* XGBoost (Extreme Gradient Boosting)

### 4.5.3 Differences Between Bagging and Boosting

A screenshot of a computer

Description automatically generated

## Why Choose XGBoost Over Other Techniques?

XGBoost (Extreme Gradient Boosting) is a popular machine learning algorithm known for its speed and performance. Here are several reasons why you might choose XGBoost over other techniques:

### 4.6.1 Performance

Accuracy: XGBoost often provides higher accuracy compared to other algorithms due to its optimization and regularization techniques.

Handling Overfitting: It includes regularization parameters (L1 and L2) that help to prevent overfitting, making it more robust.

### 4.6.2 Speed and Efficiency

Parallel Processing: XGBoost supports parallel and distributed computing, which makes it faster to train large datasets.

Out-of-Core Computation: It can process data that doesn't fit into memory, using efficient disk-based methods.

### 4.6.3 Flexibility

Custom Objective Functions: XGBoost allows the use of custom objective functions, which makes it highly flexible for different types of tasks.

Handling Missing Values: It has a built-in mechanism to handle missing values, making data preprocessing easier.

### 4.6.4 Scalability

Distributed Computing: XGBoost can be run on a cluster of machines, making it suitable for very large datasets.

Cross-Platform: It supports multiple languages (Python, R, Julia, C++, Java, Scala) and platforms, making it versatile and easy to integrate into various pipelines.

### 4.6.5 Advanced Features

Tree Pruning: XGBoost uses a more sophisticated approach to tree pruning (e.g., max\_depth, gamma parameters) that improves performance.

Weighted Quantile Sketch: This technique allows XGBoost to handle weighted data better, providing more accurate results.

### 4.6.6 Community and Ecosystem

Active Community: XGBoost has a large and active community, which means there is plenty of support, tutorials, and resources available.

Integrations: It integrates well with other machine learning libraries and frameworks like scikit-learn, Spark, and Hadoop.

### 4.6.7 Implementation Details

Regularization: XGBoost uses regularization to avoid overfitting and improve generalization.

Handling Imbalanced Datasets: It has parameters like scale\_pos\_weight that help in handling imbalanced datasets effectively.

### 4.6.8 Evaluation Metrics

Evaluation Metrics: XGBoost supports various evaluation metrics and provides tools to track the performance of the model during training.

* 1. **Disadvantages of XGBOOST**

Despite its strengths, XGBOOST has some disadvantages:

Complexity: Implementing and tuning XGBOOST can be complex, requiring expertise to achieve optimal performance.

Computational Resources: While efficient, XGBOOST can still be resource-intensive, especially with large and complex datasets.

Interpretability: Models created with XGBOOST can be difficult to interpret compared to simpler algorithms, making it harder to understand the influence of specific behaviors.

* 1. **XGBOOST Equation and Assumptions**

XGBOOST builds an ensemble of trees using the following objective function:



Figure 16 (xgboost objective function)

Where:

* 𝑙 l is the loss function that measures the difference between the predicted value 𝑦^𝑖 and the actual value y(i)
* ​Ω is a regularization term to control the complexity of the model.
* θ represents the model parameters.

Key assumptions include:

The data should be in a suitable format, typically numerical or easily convertible to numerical.

There should be no missing values or they should be handled appropriately.

The model assumes a certain level of independence between the features, such as different behavioral factors being independently observed.

* 1. **Key Properties of the XGBOOST Equation**
* **Additive Nature:** XGBOOST builds the model in an additive manner, where each new tree corrects the errors of the previous trees.
* **Gradient Descent:** It uses gradient descent to minimize the loss function, adjusting the model parameters iteratively.
* **Regularization:** The regularization term Ω(fk)\Omega(f\_k)Ω(fk​) helps in controlling overfitting by penalizing complex models.
* **Second-order Approximation:** XGBOOST uses second-order approximation (Taylor expansion) for the loss function, which improves the optimization process.
  1. **Key Assumptions for Implementing XGBOOST**
* **Independence of Features:**

It is assumed that the behavioral features are not highly correlated, though XGBOOST can handle some degree of multicollinearity.

* **Quality of Data:**

The data should be preprocessed to handle missing values, outliers, and irrelevant features.

* **Parameter Tuning:**

Proper tuning of hyperparameters like learning rate, max depth, and the number of estimators is crucial for optimal performance.

* **Data Size:**

XGBOOST is designed to handle large datasets, but it requires sufficient computational resources.

* 1. **Real-World Applications of XGBOOST in Car Insurance**

XGBOOST has several real-world applications in the car insurance industry, especially when it comes to predicting claims based on human behaviors. Here are some key applications:

### Risk Assessment and Premium Calculation

* **Risk Stratification:** Insurers can use XGBOOST to stratify policyholders into different risk categories based on their driving behavior, demographics, and historical claim data. This stratification helps in determining the premium rates more accurately.
* **Personalized Premiums:** By analyzing the driving habits of individuals, insurers can offer personalized premiums. For example, safer drivers can be rewarded with lower premiums, whereas high-risk drivers may be charged more.

### 4.11.2 Fraud Detection

* **Identifying Fraudulent Claims:** XGBOOST can be trained to detect patterns indicative of fraudulent claims. By analyzing anomalies in the data, insurers can flag suspicious claims for further investigation.
* **Predictive Analytics:** Utilizing historical data, XGBOOST can predict the likelihood of a claim being fraudulent based on the policyholder's behavior and claim history.

### 4.11.3 Customer Retention and Marketing

* **Churn Prediction:** XGBOOST can help in identifying policyholders who are likely to cancel their policies. By understanding the reasons behind customer churn, insurers can take proactive measures to retain customers.
* **Targeted Marketing:** Insurers can use XGBOOST to segment their customer base and tailor marketing strategies accordingly. This targeted approach ensures that marketing efforts are more effective and relevant to the customers.
  1. **Feature Engineering for Car Insurance Prediction**

Feature engineering plays a crucial role in the effectiveness of XGBOOST models. For car insurance prediction, the following features can be engineered from the data:

* 1. **Behavioral Features**
* **Driving Patterns:** Features such as average speed, frequency of sudden stops, and acceleration can be derived from telematics data.
* **Mileage:** Total distance driven over a period can be a significant predictor of risk.
* **Time of Day:** Driving behavior during different times of the day (e.g., rush hour vs. off-peak hours) can provide insights into risk levels.
* **Demographic Features**
* **Age and Gender:** Certain age groups or genders may exhibit different risk profiles.
* **Location:** Geographic location can influence risk due to varying traffic conditions and accident rates.
* **Vehicle Type:** The make and model of the vehicle can affect the likelihood of claims.
* **Historical Data**
* **Claim History:** Past claims can be a strong indicator of future claims. Features like the number of previous claims and the time since the last claim are useful.
* **Policy Information:** Details such as policy duration, coverage type, and deductible amount can be important predictors.
  1. **Data Preprocessing for XGBOOST**

Effective data preprocessing is essential for building robust XGBOOST models. Here are key steps involved in preprocessing data for car insurance prediction:

### 4.14.1 Handling Missing Values

* **Imputation:** Replace missing values with statistical measures like mean, median, or mode, or use more advanced techniques like K-Nearest Neighbors (KNN) imputation.
* **Removal:** In some cases, rows or columns with a high percentage of missing values may be removed to maintain data quality.

### 4.14.2 Encoding Categorical Variables

* **Label Encoding:** Convert categorical variables into numeric values using label encoding.
* **Target Encoding:** Use target encoding for categorical variables to capture the relationship between the category and the target variable (e.g., claim flag).

### 4.14.3 Scaling and Normalization

* **Feature Scaling:** Apply techniques like StandardScaler or MinMaxScaler to scale numerical features, ensuring they are on a similar scale and improving model convergence.

### Model Evaluation and Interpretation

Evaluating and interpreting XGBOOST models is crucial for understanding their performance and making informed decisions.

### 4.14.5 Evaluation Metrics

* **Accuracy:** Measures the overall correctness of the model's predictions.
* **ROC-AUC:** Evaluates the model's ability to distinguish between classes, useful for imbalanced datasets.
* **F1 Score:** Balances precision and recall, providing a single metric for classification performance.

### 4.14.6 Feature Importance

* **Gain:** Measures the contribution of each feature to the model based on the improvement in accuracy brought by that feature.
* **Weight:** Counts the number of times a feature is used in all the trees.
* **Cover:** Measures the relative number of observations affected by a feature.

### 4.14.7 Visualization

* **Feature Importance Plot:** Visualize the importance of different features using a bar plot.
* **SHAP Values:** Use SHAP (SHapley Additive exPlanations) values to interpret the impact of each feature on the model's predictions.
  1. **Why Choose XGBoost as an Enhancement with Logistic Regression**

Higher Accuracy: More accurate predictions, especially on complex datasets.

Faster Training: Handles large datasets efficiently for real-time applications.

Prevents Overfitting: Regularization helps the model generalize well to new data.

Feature Insights: Reveals which factors most influence predictions.

Adapts to Non-Linearity: Works well even when data doesn't have a linear relationship.

with Logistic Regression:

Boosts Performance: Significantly improves prediction accuracy compared to Logistic Regression alone.

Faster Learning: Speeds up the training process on large datasets.

Better Generalization: Prevents overfitting for better performance on unseen data.

Complements Feature Understanding: Provides insights beyond the linear relationships captured by Logistic Regression.

# Chapter 5

**5.1) Methodology for Building a Car Insurance Model using logistic regression**

* importing libraries and uploading datasets
* describe the data and visualization

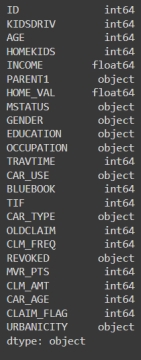


Figure 17(Data Types in Dataset)

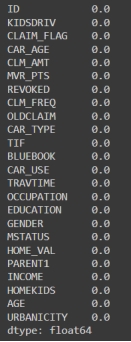


Figure 18 (blank cells in cleaned data)

A black background with white text

Description automatically generated

Figure 19 (Attributes names in dataset)

* scaling and encoding.
* dropping target columns.
* target column was imbalanced, so we used undersampling which is a technique used to address imbalanced datasets. An imbalanced dataset has a significant difference in the number of data points between different classes. Imagine you're training a model to detect fraudulent credit card transactions. Most transactions are legitimate (the majority class), and only a small portion are fraudulent (the minority class).

Undersampling tackles this imbalance by reducing the number of data points in the majority class. This makes the minority class more comparable in size and allows the model to focus on learning its patterns. There are different ways to undersample data, including randomly removing data points or using more sophisticated methods to select which ones to remove.

* So, we tried first to undersample the data by increasing the number of data points in minority class

A graph showing a number of blue and orange colored squares

Description automatically generated

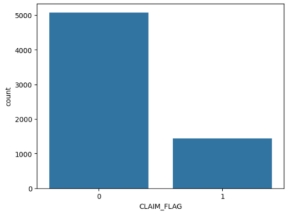


Figure 20(target column before undersampling)

Figure 21 (target column after undersampling with miority)

A blue rectangular bar graph

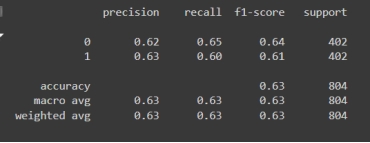
Description automatically generated with medium confidenceA graph showing a number of blue and orange colored squares

Description automatically generated

Figure 22 (target column before undersampling)

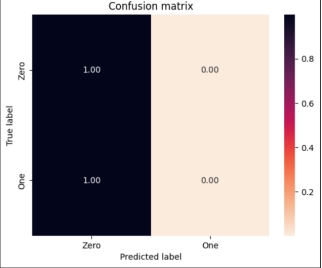
Figure 23 (target column after sampling with maority)

* Then we tried to undersample the data by decreasing the number of data points in majority class

 A diagram of a number of colored squares

Description automatically generated with medium confidence

Figure 24 (confusion matrix in majority undersampling)



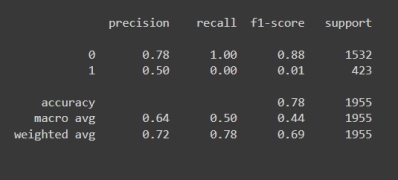


Figure 25 (confusion matrix in minority undersampling))

A black and white text

Description automatically generated

A screenshot of a computer

Description automatically generated

Figure 26(accuracy in majority undersampling)

Figure 27(accuracy in minority undersampling)

So, from this analysis we have concluded that majority undersampling is better for the model and the data set

* Train a Logistic Regression model using the selected features from the resampled data.
* Evaluate the models using accuracy score, confusion matrix, and classification report.
* Plot the ROC curve and calculate the AUC score to assess the model's A graph with a line

  Description automatically generatedperformance.

Figure 28 (roc curve of logistic regression)

A close up of a word

Description automatically generated

Figure 29 (AUC score of logistic regression)

* Principal Component Analysis (PCA):

Apply PCA to the standardized and encoded training data to visualize the variance explained by each principal component.

* Create a bar plot to show the variance explained by each PCA feature

A graph with numbers and a bar

Description automatically generated

Figure 30 (PCA for logistic regression)

* Encoding Categorical Variables:

Use LabelEncoder to encode categorical variables in the resampled data before fitting the final model.

* Logistic Regression Model:

Fit the LogisticRegression model on the resampled training data.

* Evaluate the model's performance on the resampled test data.
* print the accuracy

A screenshot of a computer

Description automatically generated

Figure 31 (accuracy of logistic regression)

**5.2 Methodology for Building a Car Insurance Model using XGBOOST**

### 5.2.1 Data Preparation and Exploration

1. **Upload and Load Data**:
   * Import necessary libraries and upload the dataset using google.colab.
   * Load the dataset into a Pandas DataFrame.
2. **Initial Data Inspection**:

* List all uploaded files and read the Excel file.
* Display the content of the DataFrame and inspect its structure.

A screen shot of a computer code

Description automatically generated

Figure 32(Initial Data Inspection)

### 5.2.3 Data Preprocessing

1. **Train-Test Split**:
   * Split the data into training and testing sets using train\_test\_split.
2. **Encoding and Scaling**:

* Install and use category\_encoders to apply TargetEncoder for encoding categorical features.
* Define a pipeline with a TargetEncoder and XGBClassifier.

### 5.2.4 Hyperparameter Optimization

1. **Bayesian Hyperparameter Optimization**:
   * Install scikit-optimize and define a search space for hyperparameters.
   * Use BayesSearchCV for Bayesian optimization.

### 5.2.5 Model Evaluation

1. **Model Performance**:
   * Retrieve the best estimator and evaluate its performance on the test set.
   * Make predictions and calculate accuracy.

A screenshot of a computer code

Description automatically generated

Figure 33(Model Performance)

1. **Feature Importance**:

* Plot the feature importance of the best XGBoost model.

A graph with blue bars

Description automatically generated with medium confidence

Figure 34 (PCA Feature)

A graph with numbers and letters

Description automatically generated

Figure 35 (Feature importance)

Figure 36 (Best cross valdition score )

1. **Accuracy Calculation**:

* Calculate and print the accuracy of the model.

A number on a white background

Description automatically generated

Figure 37 (XGBoost Accuracy)

# Chapter 6

**6.1) Conclusion:**

### 6.1.1 Overview of the Project

The project aimed to enhance the efficiency and accuracy of car insurance risk assessment and claim prediction. The primary goal was to develop a robust model that can categorize insurance customers based on their risk profiles and predict the likelihood of them filing a claim within a specific period. To achieve this, various machine learning algorithms and data analytics techniques were employed. The key components and features of the project include:

### Features:

**6.1.2.1 Risk Assessment Module:**

* Utilized machine learning algorithms to evaluate the risk associated with each car insurance customer.
* Factors considered included occupation, number of home children, motor points, demographical location, and age.
* Provided a risk score for each customer, aiding in categorizing them into low, medium, and high-risk categories.
  + - 1. **Risk Score Classification**:
* Defined thresholds to classify customers based on their risk scores.
* Ensured that the classification criteria were transparent and easily adjustable based on new insights or changing conditions.
  + - 1. **Claim Prediction Model**:
* Employed regression and classification algorithms to predict whether a customer is likely to file a claim within the next 6 months.
* Analyzed historical claims data to train the model, ensuring accurate predictions.

### 6.1.3 Limitations:

**6.1.3.1 Data Limitations**:

* **Difficulty in finding data**: It was hard to find real data in Egypt suitable for our model.
* **Limited Data Availability**: Which could impact the accuracy and robustness of the risk assessment and predictive models.
* **Data Quality Issues**: Incomplete data and the necessity for extensive data cleaning and processing might lead to potential biases or inaccuracies in the models.

**6.1.3.2 Model Limitations**:

* **Predictive Model Accuracy**: We aim to create predictive models for risk assessment and claim forecasting, but there is a lack of in-depth discussion regarding potential limitations in model accuracy and the uncertainty of the predictions.

**6.1.3.3** **Computational Resources**:

* **High Computational Demand**: Training and deploying machine learning models, especially complex ones, can require significant computational power and memory, which may not be feasible for all organizations.
  1. **Future Work Recommendations:**

### 6.2.1 Enhanced Data Collection:

Collaborate with telematics companies to collect real-time driving data from customers' vehicles. This data could include speed, braking patterns, acceleration, and time of day when driving. Such detailed information can help build a more accurate risk profile for each customer.

**ex:** If a driver frequently drives late at night at high speeds, the risk assessment model can flag this behavior as high risk, leading to a more accurate classification.

* + - 1. **Incorporate Advanced Features:**

Integrate weather data into the risk assessment model. For instance, if a customer lives in a region prone to heavy snowfall, this can be factored into the risk calculation.

**ex:** A customer who drives frequently during snowy conditions can be classified as higher risk due to increased chances of accidents, thus adjusting their risk score accordingly.

### 6.2.3 **Deployment** and Integration:

Develop a web-based dashboard that insurance agents can use to input customer data and instantly receive risk assessment and claim prediction results.

**ex:** An insurance agent inputs a customer's details into the dashboard and receives a real-time risk score and claim prediction, enabling them to make informed decisions about policy pricing and coverage options.

### 6.2.4Continuous Monitoring and Updating:

Implement a feedback loop where the model is continuously updated with new claims data to improve its accuracy over time.

**ex:** If the model initially classifies a customer as low risk but they file multiple claims within a short period, this new data can be used to retrain the model, enhancing its ability to predict future claims accurately.

**Sources and references**

* Insurance Explained

<https://www.cibeg.com/en/learning-center/general-topics/insurance-explained>

<https://tameenonline.net/car-insurance/>

* Evolving Catastrophe Losses to Pressure 2021 Property/Casualty Underwriting Profitability, Triple-I/Milliman Predict

<https://www.iii.org/press-release/evolving-catastrophe-losses-to-pressure-2021-property-casualty-underwriting-profitability-triple-i-milliman-predict-081721>

* The impact of rising car insurance costs

<https://www.motorfinanceonline.com/features/the-impact-of-rising-car-insurance-costs/>

* Insurers face challenges in battle against high premiums

<https://www.propertycasualty360.com/2023/05/18/insurers-face-challenges-in-battle-against-high-premiums/>

* Common Auto Insurance Industry Challenges that Personal Carriers Face

<https://sambasafety.com/blog/personal-auto-insurance-industry-challenges/>

* Egypt's insurance sector prepares for a new comprehensive law

<https://oxfordbusinessgroup.com/reports/egypt/2022-report/economy/protection-drive-the-insurance-sector-is-readying-for-a-new-comprehensive-law>

* All you want to know about types of car insurance in Egypt

<https://www.tokiomarine.com.eg/types-of-car-insurance-in-egypt/>

* How insurers can respond to higher interest rates

<https://www.pwc.com/us/en/industries/financial-services/library/higher-interest-rates.html>

* Insurance: Definition, How It Works, and Main Types of Policies

<https://www.investopedia.com/terms/i/insurance.asp>

* 2024 global insurance outlook

<https://www2.deloitte.com/uk/en/insights/industry/financial-services/financial-services-industry-outlooks/insurance-industry-outlook.html>

* Does family background impact driving attitudes and risky behaviors? An investigation on Chinese young drivers

<https://www.sciencedirect.com/science/article/abs/pii/S0001457516302214>

* <https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62#:~:text=What%20is%20Confusion%20Matrix%20and,of%20predicted%20and%20actual%20values>.

Different types of Encoding

* <https://ai-ml-analytics.com/encoding/>

What Is Logistic Regression? Equation, Assumptions, Types, and Best Practices

* <https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/#:~:text=Logistic%20regression%20is%20a%20supervised%20machine%20learning%20algorithm%20that%20accomplishes,1%2C%20or%20true%2Ffalse>.

Insurers face challenges in battle against high premiums

* <https://www.propertycasualty360.com/2023/05/18/insurers-face-challenges-in-battle-against-high-premiums/?slreturn=20240707-43821>

Common Auto Insurance Industry Challenges that Personal Carriers Face

* <https://sambasafety.com/blog/personal-auto-insurance-industry-challenges/>

All you want to know about types of car insurance in Egypt

* <https://www.tokiomarine.com.eg/types-of-car-insurance-in-egypt/>

Why XGBoost Still Beats Deep Learning At The Tabular Data Game

* <https://forecastegy.com/posts/gradient-boosting-vs-deep-learning-tabular-data/>

2024 global insurance outlook

* <https://www2.deloitte.com/us/en/insights/industry/financial-services/financial-services-industry-outlooks/insurance-industry-outlook.html>

Mastering XGBoost Parameters Tuning

* [XGBoost Parameters Tuning | Complete Guide With Python Codes (analyticsvidhya.com)](https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/)