**Insurance claim Underwriting**

**Table of Contents**

[1.0 Abstract 3](#_Toc49116989)

[2.0 Introduction 5](#_Toc49116990)

[2.1 Background 5](#_Toc49116991)

[2.2 Research Rationale 7](#_Toc49116992)

[2.3 Research aims 8](#_Toc49116993)

[2.4 Research Objectives 9](#_Toc49116994)

[2.5 Research Questions 9](#_Toc49116995)

[2.6 Research Hypothesis 10](#_Toc49116996)

[2.7 Conclusion 10](#_Toc49116997)

[3.0 Literature Review 11](#_Toc49116998)

[3.1 Introduction 11](#_Toc49116999)

[3.2 Conceptual Framework 11](#_Toc49117000)

[3.3 Empirical Study 14](#_Toc49117001)

[3.4 Models and theories 19](#_Toc49117002)

[3.5 Literature Gap 22](#_Toc49117003)

[4.0 Methodology 23](#_Toc49117004)

[5.0 Finding Analysis 24](#_Toc49117005)

[5.1 Insurance Claim Data 24](#_Toc49117006)

[5.2 Exploratory Data Analysis 25](#_Toc49117007)

[5.3 Data preparation & Data Pre-processing 31](#_Toc49117008)

[5.4 Machine Learning Algorithms 33](#_Toc49117009)

[5.5 Evaluation of Metrics 34](#_Toc49117010)

[6.0 Conclusion 35](#_Toc49117011)

[6.1 Results and Discussion 35](#_Toc49117012)

[6.2 Future Scope 35](#_Toc49117013)

[7.0 Reference list 36](#_Toc49117014)

# 1.0 Abstract

This research paper is for the purpose to understand the risks that are involved in delivering insurance to the customers and hence solving those risks with the help of the underwriting process. There are some major aims of this research like understanding the risks those are involved in delivering insurance to the consumers by the insurance companies, developing a statement for profit or loss by understanding the scenario based on prediction through available parameters and thereby providing insurance to the irrespective individuals and implementing the use of different techniques that are present in machine learning which makes use of artificial intelligence and hence implementing algorithms for better predictions through calculations.

Question based approach is developed to get the desired objectives and moreover to get the ideas about the whole range of questions that are being developed to perform this research. Some of the questions are like, what are the types of risks that are involved in providing insurance to individuals by some respective insurance company? What are the methods or techniques that are involved in machine learning that will help in the underwriting process? What models and theories are being followed for getting ideas about the risk factors that are involved in delivering insurance? What are the independent and dependent variables present in the insurance claim underwriting process? What limitations are being present in the usage of machine learning approaches in performing underwriting?

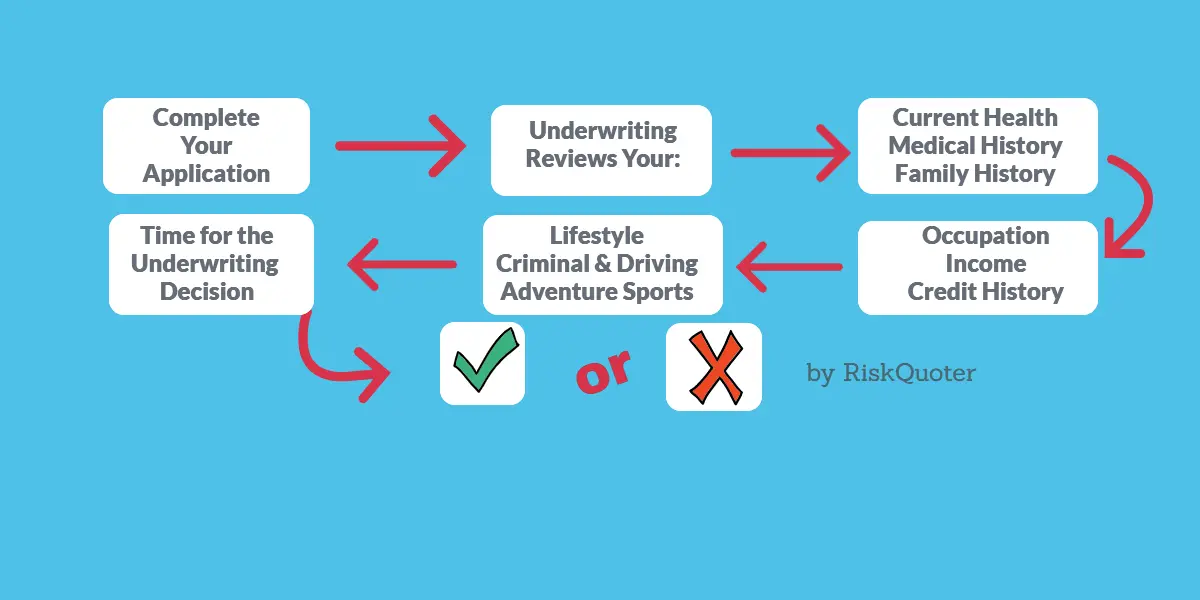
The machine learning model used is XgBoost using the tool Python. For the comparative analysis, we have used traditional classification model support vector machine, decision trees and naïve Bayes. For evaluation of the models, the performance metrics used are accuracy score, f1 score and classification report.

# 2.0 Introduction

Underwriting process has been proved to be of utmost importance which is being utilised on a large scale in the insurance fields for getting the valuable data from the applicant and also proves to be helpful in the management of different types of risks that are involved in the delivery of the insurance to the customers. Basically, with the help of this underwriting process, the loss of profitability of different financial organisations are being judged by providing the facilities of insurance to the applicants and thereby measuring the range of risks involved in delivering insurance to the consumers. Implementation of different techniques that are present in machine learning and hence make use of artificial intelligence for suitable mitigation of those risks by the detection of available faults in the system.

## 2.1 Background

Financial and insurance companies mainly involve business with customers and thereby they need to obtain the personal details of the customers for making accounts in their companies for further proceedings in business. Problems mainly arise when these details are unavailable to these companies in a correct pattern which leads to the development of risks for these companies (Besman and Wells, 2017). Prediction is the major factor that needs to be developed by these companies based on the present scenario for understanding the probability of loss or profit that might affect them in future. Hence underwriting is the only process which helps in making accurate predictions based on the parameters available at present and hence implementing suitable strategies with the help of machine learning technologies for understanding the probability of risks in future. Underwriting process has been also common and used in the sanction of loans for understanding the capability of the borrower in “repayment of loans” on time. Whenever the risk is generated in an insurance company, then proper linking of that developed risk should be done with the correct premium. This tends to be of utmost importance as a failure in proper linking might result in the enhancement of the level of toxicity as the maintaining balance gets disturbed and hence leading to severe problems of the insurance companies (Howe, 2020). The insurance is being provided based on age, personal habits, family income and other facts that proved to be beneficial for the insurance companies in generating a term-insurance policy for a particular customer.

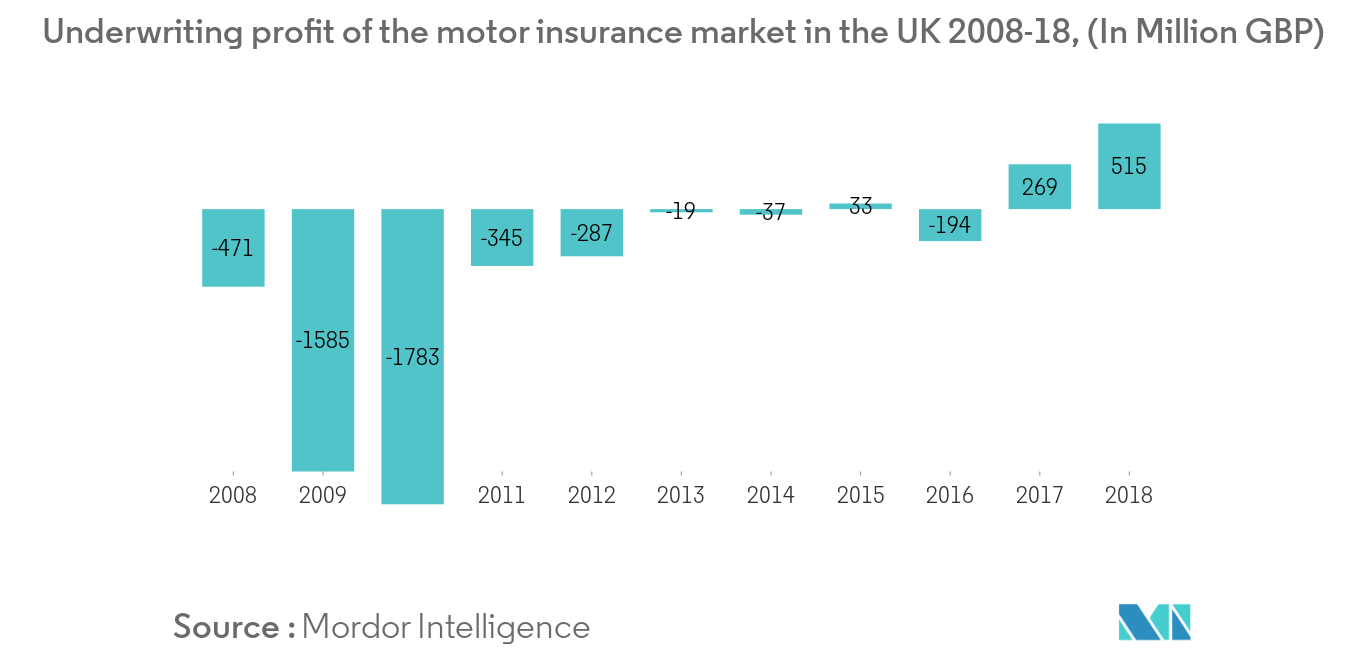


**Figure 1: Life-insurance underwriting**

(Source:https://www.riskquoter.com/underwriting/)

Insurance states that the burden that is being taken by the insurance companies for some respective individuals for delivering security to those individuals as their emergency conditions. This process is basically done by analysing the risks that will be developed in future and hence taking the responsibility of providing protection to them by the insurance companies. The problems mainly occur when these risks are not being predicted by these companies and thereby leading to loss on a great extent (Nayak and Yung, 2019). Legal measures are involved in providing services to the customers on their emergency conditions. Failure in providing services at those times which generally happens due to miscalculations in corrections prediction might lead to severe losses through legal measures. So, an underwriter helps in delivering justice through accurate predictions by machine learning methods for understanding the gravity of that situation and thereby providing insurance to that particular individual and hence providing risk management through underwriting.

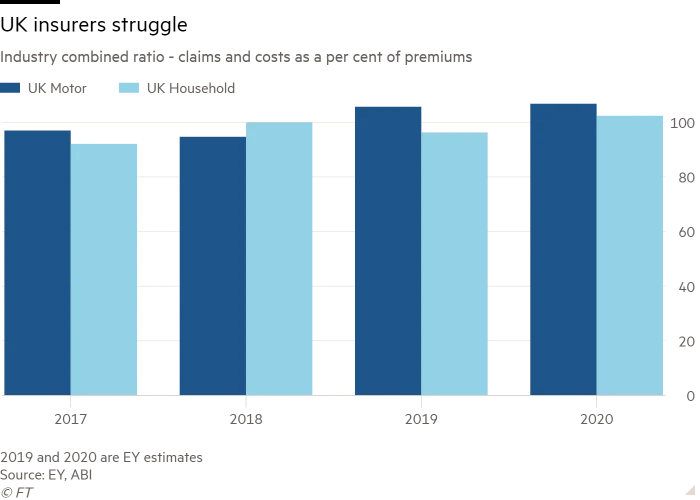
## 2.2 Research Rationale



**Figure 2: Insurance market for motor in the United Kingdom**

(Source:https://www.mordorintelligence.com/industry-reports/united-kingdom-motor-insurance-market)

The report shows that there has been a significant rise in the growth in 2018 for motor insurance which tends to be around 515 million (GBP), which has been around 258 million in 2017. This increment has mainly occurred in the United Kingdom due to the implementation of underwriting by the insurance companies and hence making suitable predictions through machine learning. Liechtenstein has recorded an increase of 0.21%, which tends to the rise by 7.46% in providing insurance to the consumers and hence the total increment has led to the figure of 14.3% in 2018. This has been possible due to the increased use in the rising trends in the regional markets and segments in the UK, and hence making use of different application types and market dynamics while providing underwriting by the insurance companies (Hashim, 2017). As per Figure 2, it can be said that in previous years of 2008 to 2016, the insurance companies in the UK have suffered a loss but with the increasing use in the machine learning technologies which provided algorithms for making suitable predictions, there has been considerable growth in these sectors.



**Figure 3: Estimates provided for insurance claims in the UK**

(Source: https://www.ft.com/content/eceff0ba-129c-11ea-a7e6-62bf4f9e548a)

As per the estimates that are being provided by EY, there has been increasing demands in getting insurance by the people of the United Kingdom which is majority due to the costs that are on the rise and hence bothering the people to get insurance for getting security and safety in future. Hence, the use of underwriting proves to be beneficial in providing insurance based on the risks like increasing prices of goods in future. From figure 3, it can be established that from 2017 to 2020, there has been an equitable amount of demands of insurance that is being required by the people of the United Kingdom and the significant amount of insurance that is being provided by UK motor during this time period (bus.wisc.edu, 2016).

## 2.3 Research aims

Research is being performed for understanding the risks that are involved in delivering insurance to the customers and hence solving those risks with the help of the underwriting process (Mwangi, 2017). The major aims are as follows:

* To understand the risks those are involved in delivering insurance to the consumers by the insurance companies.
* To develop a statement for profit or loss by understanding the scenario based on prediction through available parameters and thereby providing insurance to tg irrespective individuals.
* To implement the use of different techniques that are present in machine learning which makes use of artificial intelligence and hence implementing algorithms for better predictions through calculations.

## 2.4 Research Objectives

Based on the major aims that will be accomplished with the help of theoretical research in the latter section, the objectives of providing an underwriting process to the insurance companies for delivering insurance are being accomplished with the help of following points (Nurse and Creese, 2020).

* To provide a proper risk management plan through better predictions and hence making proper calculations with the help of algorithms in machine learning.
* To understand the concept of underwriting with the help of models and theories that will provide an idea about this process.
* To identify the dependent and independent variables that are present in this research topic and thereby deriving a suitable relationship for the proper mitigation of all the risk factors available in delivering insurance.
* To make use of Python as a tool for coding the details and hence generating a program for making further calculations.

## 2.5 Research Questions

Questions are developed for getting the desired objectives and also for getting ideas about the range of queries that are being developed for performing this research. The questions are as follows:

1. What are the types of risks that are involved in providing insurance to individuals by some respective insurance company?
2. What are the methods or techniques that are involved in machine learning that will help in the underwriting process?
3. What models and theories are being followed for getting ideas about the risk factors that are involved in delivering insurance?
4. What are independent and dependent variables present in the insurance claim underwriting process (Wang, 2017)?
5. What limitations are being present in the usage of machine learning approaches in performing underwriting?

## 2.6 Research Hypothesis

Research is being done for understanding the importance of underwriting while providing insurance to the applicants by taking into account their personal details for risk management (Ross*et al*. 2019). However, by performing the hypothesis, the advantages and disadvantages of performing this research can be well understood. The hypothesis is made below.

H1: Underwriting helps in providing insurance to the consumers based on the personal and family details that are being taken from them previously and hence making suitable utilization through prediction.

H0: Underwriting, however, creates obstacles in making extra income, as through prediction, many consumers are not provided insurance by insurance companies which ultimately lead to their loss.

H2: Predictions are made with the help of machine learning which includes algorithms for making calculations through available parameters and hence providing good probability value.

H0: Predictions which are developed often leads to wrong results if parameters are not taken correctly (www.researchgate.net, 2017).

## 2.7 Conclusion

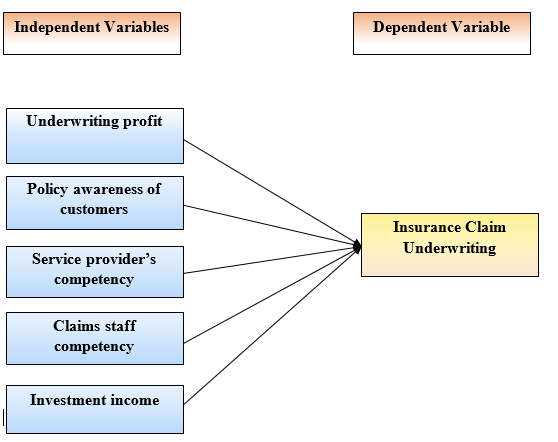
To conclude that underwriting involves the use of machine learning technologies for making suitable predictions through algorithms in artificial intelligence. To conclude that insurance is being provided to the consumers by analysing the risks that include basic parameters. The major objectives and aims are being listed for developing the research context and questions are developed for getting the clarification of doubts in the latter section. Hypothesis helps in getting the positivity and negativity of this research and rationale helps in getting an idea about the different cases of insurance policies that are registered till date and highlighting the concept of underwriting.

# 3.0 Literature Review

## 3.1 Introduction

Research is developed by taking the opinions and useful suggestions of different authors which helps in getting the usefulness and advantages of the underwriting process that helps in getting the details for analysing the situations for providing insurance to different applicants. As insurance itself defines the safety and security that is being provided in future, hence the possibility of different risks are being measured through different models and exclusive theories that involve machine learning approaches for making predictions and hence understanding the range of profit before providing insurance to applicants. However, many limitations are present in this research which will be discussed later in the literature gap.

## 3.2 Conceptual Framework

****

**Figure 4: Conceptual framework**

(Source: Self-created)

A conceptual framework is being developed based on the range of independent and dependent variables that are available and hence developing a suitable relationship between these two types of variables for understanding and hence performing suitable research. This structure will help understand the necessity of the underwriting process in insurance companies of the United Kingdom.

**Dependent Variable**

**Insurance “claim” underwriting:** For making claims for insurance by the applicant, it is necessary to check the details which are being represented with the help of independent variables which are being directly linked with this attribute of the dependent variable (Soye and Adeyemo, 2018). Based on these variables, the underwriting process is carried out for getting an idea for providing efficient risk management while providing security to different applicants. All these variables are being used for performing the research by an underwriter and hence developing predictions for understanding the profitability range of respective insurance companies. Also based on the independent variables, the machine learning approaches are being well utilized for understanding the necessity of this dependant variable and hence making suitable use out of it.

**Independent variables**

**Underwriting profit:** According to Howe, 2020, it has been stated that the insurance sectors implement the process of underwriting in order to develop the profit or loss statement before providing insurance to different applicants. Hence, this attribute proves to be the major factor which governs the dependent variable to a great extent. The claims and expenses that are involved in providing insurance against the premiums which are being earned by the respective insurance companies help in making calculations for understanding the level of profit developed in underwriting (Howe, 2020). The premiums that are being charged by the insurance sectors helps in making a profit by covering different types of risk which are targeted for consumers. This process is being accomplished with the help of the underwriting process.

**Policy awareness of customers:** This attribute helps in making judgemental decisions based on the type of proposal that arrives at a particular instance and thereby analyzing it instantly for acceptance or denial with the help of the dependent variable that is available and hence making profit through suitable decisions. The scope, terms and conditions and price are being developed are the acceptance of that particular proposal through the above-mentioned process of the dependent variable and hence making suitable satisfaction to both the customers and insurance companies in making such transactions. The process of underwriting and claim management both are interrelated as the risks are being underwritten by these insurance companies for making claims in payments during the materialization of different risks and also the documents for the particular policy are being followed for making proper deals with the customers.

**Service provider’s competency:** This attribute proves to be essential in delivering service to different customers as per the claims that are being made by them. This attribute is being linked to the process of underwriting by increasing the reliability and trust level of the applicants in getting suitable services from the insurance sectors at the time of emergency conditions that might develop in future due to some mishap and accidents for the applicants (Biddle and Xu, 2018). The risk coverage rate is being denoted with the help of services that are being provided by underwriters by making exact predictions which lead to the generation of high “benchmarking” values for providing integrity to the consumers. Service providers which majorly involve the employees of different insurance sectors who play an important role in making appropriate selection of suitable term policies which will be suitable for different customers based on the risks that might be faced by different customers in future.

**Claims staff competency:** This attribute helps in increasing the efficiency and performance rate of different insurance sectors with the help of skilled underwriters who are experts in performing the process of underwriting in providing insurance claims to different applicants. With the help of skilled underwriters, the insurance sectors have been able to provide suitable analysis of different risks and hence make perfect calculations in getting the range of income and profit that will be available to both the applicants and insurance sectors based on the range of premiums that are being provided (Trainor*et al*. 2017). This attribute has helped in increasing the satisfaction level of different applicants of insurance through proper treatment and thereby increasing the profit of these companies in the UK.

**Investment income:** This attribute helps in denoting the level of income for both the applicants and insurance companies while delivering lifetime policies for providing insurance to different applicants and hence increasing income from those policies. This attribute is being linked to insurance "claim" underwriting in which the underwriters analyse the range of income they can earn from the premiums that are being delivered to the customers and hence making a huge profit by analysing the present conditions that are available from the personal details that are being recorded for making suitable predictions for future conditions.

The above mentioned two types of variables help in getting the idea about the structure and the range of formalities that are involved in insurance sectors in providing suitable insurance term-policy to different customers (Devereauxand Swain, 2017). It also highlights the importance of insurance "claim" underwriting processes for performing suitable management of different risks and covering those risks with the help of term-policies. The relationships between the two types of parameters are nicely explained, and the different parameters are being considered for increasing the performance rate of underwriting in different insurance sectors. Based on the parameters, the appropriate model is being established for making calculations with the help of machine learning which will be discussed in the later section (d1wqtxts1xzle7.cloudfront.net, 2017).

## 3.3 Empirical Study

According to Ekin, 2020, the trade-off between the simplified claim handling process and the risk management of insurance underwriting and machine learning is a persistent concern of the Insurance Industry. According to the report of the Federal Bureau of Investigation (FDI), the estimations record over an average of fund mismanagement of $40 billion lost every year that lead to the coverage of the non-health insurance products. Thus 95% of insurance companies are anti-fraud. The costs that are incurred in this whole process are passed onto the policyholders of the claim that count the money in the increasing premiums. Automation of the underwriting of the insurance claim can happen by grouping the customers, and thus the machine learning can also increase the cutie satisfaction and depends on the determination of fraud investigation cases that can also lead to the maximum amount of the resource allocation measures and the different adaptations that are brought to change. The author states here that a lot of insurance companies deliberately look over the results that are not in the healthcare benefits and thus are paid on a suspicious basis that can also lead to health insurance fraud. A recent survey proved that the estimations in the number of false claims received are 15% of the total claims and thus the insurance companies lose around 30 billion of USD on an annual basis due to the health care frauds. Some of the possible frauds are as mentioned below:

·**Billing Services are not looked into properly.**

Billing Services are ensured for the insurance purposes that can result in the forging of the signature that is involved in getting the bills and thus if these services are not looked into properly then it would be a hugely problematic situation for the organization.

·**“Upcoding of Services”**

Billing Services are much more costly than the procedures in the discussion of the sessions which are ensured in making a 45-minute session into a 60-minute session.

·**“Duplicate Claims”**

The insurance claim of the false bills submitted also ensure a different level of duplicity and prejudices that result in the development of the charge that is levied on the insurance company.

·**Unnecessary Services**

Filing claims no way also helps in developing the conditions of the person who is the claimant (Spedicato*et al.* 2018). The author also stated that anomaly detection calculates the probability that is tested with the claims and the fraudulent activities are also ensured in the deterioration of the insurance cases which is also the fundamental classification to generate the disco boundary between the classes of the estimate and the fraudulent claims. The technique of ensuring the underwriting insurance with the help of insurance claims also helps in the detection of the major sense of the duplicate claim of the fraudulent class and the detection measures.

According to (Raines et al.) a method or an apparatus was proposed for underwriting and rating a particular insurance policy. Moreover by the help of the apparatus proposed one can generate information for underwriting the subject policy and further rating the policy in the scanner. This system and method, which is proposed, starts with recognizing the vehicle, further the determination of minimum single vehicle history data variable or maybe the group of variables. Further by the data variable of at least a single vehicle or more the basis is decided for the process of policy underwriting and rating the policy. This paper proposes a system and also a method that can amalgamate the generation of a score based on the value of a minimum of single data history variable or can be a group of data variables.

This paper claims a proper method of underwriting a vehicle insurance policy. The method is consists of:

Moreover, an overall process for underwriting insurance is been proposed by (Hele et al.) The proposed model process comprises of automated underwriting using norms Related U.S. Application Data from a wide variety of insurance carriers. These norms can be general, carrier Specific, and product specifications. This method also includes as it is an auto responsive feature process for giving a questionnaire to a user that is interested in purchasing life insurance. This model or invention hands us a method, software and system completely based on the user profile for underwriting life insurance. To determine the ratings of the insurance by the customer or user information and to underwrite the rules of a plurality of carriers, this model can be used easily. This invention also has some additional features like a multi-pass underwriting process which calculates the price of life insurance as the dependent of user profile and underwriting norms of carriers. This invention showcases methodology that amalgamates:

* Receiving the first set of user information, showing the first range of price information of life insurance as per the first set of information suitability of the user to the user.
* Receiving the second set of user information, showcasing the second range of life insurance pricing as per the available information of the user up to the second set of information to the user.
* Receiving quotes for the life insurance plan based on the plurality of insurance carriers and sharing the quotes to the user.

Customer experience also has a call for improvement and thus better compliance and the legal documents help in the determination of the customer's understanding and experiences. Underwriting and insurance claims are the two pillars of risk management for life insurers. These elements help in the determination of the profitable portfolio that helps in the materialization of the investment and the underwriting of the investments (Ekin, 2020). Automated and digitized underwriting or claims are the handling keys to success. The capitalization of the aspects helps in the detection of different experiences of the global expatriation that ensures a competitive edge and thus also ensures the delivery of the securities and the profitability measures.

The author claims that the underwriting of the insurance claims helps in determining the risk of the insured of the particular claims that also ensures the difference of the classes of the asset and the premium pricing of the insurance policies. Machine learning measures hell in the determination of the factual data helps in determining the magnitude of the payout according to the magnitude of the life policy. The author also states that the day to day work of an underwriter ranges from the research of the data that is also ensured in the determination of the pricing risk which is ensured in the diligence of the underwriting that helps in the development of the different approval of the machine learning measures (Ticconi, 2019). The different tasks of the insurance underwriting and machine learning outcomes are discussed below:

·**Analyzing the claim of the individuals**

When an underwriter adds a new policy in the book of business, they take a risk. The likelihood and the claim of this policy help in the determination of the underwriting measures that help in developing an uncertainty and the claims in the nearby future (Jamal and Shah, 2019). During this process, the underwriters help in assessing the risk that is based on the determination of the major predictions and the different perspectives of the experiences that are evolved.

·**Predicting the claim of the incident**

There is a claim in the determination of the severity and the expenditures that are stored in the determination of the different claims. In the case of medical liability, the insurance is listed in the determination of the different claims that are personally involved in the incidental accidental coverage (Liu *et al.* 2017). Frivolous claims incur the different legal expenses that are ensured in claiming a legitimate right over the flag of the massive development of the insurance underwriting measures in terms of the machine learning outcomes.

·**Generating a Quote**

The quote that is generated in the determination of the fine-grain analysis helps in ensuring the different opportunities of the underwriting process, which helps in determining re-speciality in practice behind determining the claims.

The digitalization of the business is in high gear. Some of the sectors have embraced the change that has resulted in the implementation of the advances and thus has also ensured the different category of the risk-averse culture measures which have been specifically assumed with the continuation of the investment eye of the technology of the different concepts of the machine learning measures outcome (Rosén, 2020). The job of an insurance underwriter helps in the analysis of the potential risks which are used in the determination of the risks and thus the assurance of machine learning outcomes depending on the integration of the sources of information which helps in the determination of the integration process and the provision of the relevant information which also is developed according to the perceived notion according to the combat of the fraudulent filings in the determination of the insurance claims of machine learning. The insurance sector is constantly afflicted with fraudulent claims that make the solution of the different algorithms to be correlated with the different levels of human intelligence. One of the major deficiencies in the human-based system helps in the determination of the inherent techniques that help in the pruning of the errors that are an inefficiency in the operations. In the insurance claim underwriting the artificial intelligence geckos in the addition of the enhanced insurer, which is developed in the development of accurate information according to the artificial intelligence measures of the digital revolutions. The most important concept of machine learning is to ensure the different method of the facilitation of the datasets of the insurance claims that helps in the predictive accuracy of the different information that results out of the different level of the technological advancements (Naujoks*et al.* 2017). The assess of the major level risk is to be insured in the different preferable outcomes and; thus it also ensures the supervised learning outcome of the improved measures of the risk assessment measures. Fraud is growing in every sector that helps in the development of the different level estimations according to the insurance companies being relied upon over the falsified actions that happen if the concept of the chain learning measures is not taken into control Thus the implications of the machine earning outcomes are important to get an idea about the applications that are used in the development of the field of the insurance businesses the underwriting processes that the policies and the risks are managed according to the different level of the tasks that are damaged in the development of a different measure of the learning algorithms in the development of the standard product offering of the reinforcement of the learning process (Patil*et al.* 2017). The entire process of data in the process is universal that helps in the detection of the share of the investment mechanisms across the public entities to develop the capability of the understanding that is in fact used in the development of the client service outcomes. This is much important to understand the importance of the acceptance of the claim registration prices what is ensured in the different development of the predictive models on the determination of the claim costs. Thus it is much important to get an idea about the different implication of the profit that is issued in the different articulation of the solutions that are developed by the machine learning measure of the claims processing measures depending on the evidence that is ensured in the automation of the employment of the monitoring outcomes which help in the endurance of the useful set of information in the determination of the different protocols of the advanced machine learning outcomes (Sabharwal, 2018). The most important utilization of the measures helps in the determination of the outcomes of the technology in the determination of the modifications according to the dissimilar combinations of the changes that are ensured in the different set of goals according to the machine learning outcomes. Machine Learning helps in the determination of different level outcomes and thus, the risk of the insurance claims as per the machine learning outcomes depending on the falsified actions. The fraud detection in the determination of the insurance underwriting measures helps in the determination of the estimations, which are generally put to use to improve the conceptualizations of the different aspects of the machine learning outcomes.

## 3.4 Models and theories

Risks that are mainly involved and are being faced by the insurance sectors are financial risks, investment risks, legal risk and technological risks. Financial risk is being experienced by insurance companies after providing the term-policies to different applicants without making suitable predictions based on the available parameters. Investment risks are being experienced mainly by consumers who are directly involved in investing money based on the premium rate and term policy that is being provided by different insurance sectors. Legal risk is developed whenever there is a breach in the term for a particular policy that is being provided to the applicants at the time of issuance of policy in the initial stages. Technological risks are mainly involved in the use of technologies for making proper calculations of risks for measuring the profitability of insurance sectors.

Based on the availability of the above risks, efforts have been made for developing suitable models and hence following theories for proper risk management.



**Figure 5: Theories**

(Source: Self-created)

The theories that are being used for conducting this research are the theory of potential risk management which proves to be beneficial for the mitigation of the above-mentioned types of risks and helps in proper analysing of those risks for understanding the profit of the insurance companies in future before providing insurance to the applicants (Denning and Perschy, 2020). According to Bonissone and Russell, 2018, it has been stated that this type of theory has been successful in the determination of different types of risks that might be faced by insurance companies in future and providing suitable risk management strategies for overcoming those risks immediately. This theory basically involves four basic steps which are as follows:

**React:** Based on the range of data that is being available to the underwriters, they react to understanding the possibility of providing insurance to the applicants and hence making efficient risk management while providing term insurance (Bonissone and Russell, 2018).

**Reduce:** After analyzing the different risk factors, if the underwriters are able to figure out the type of risks that will affect their profitability in future then they will reject the customer in providing the policy and thereby safeguarding the insurance sectors from falling in the trap in future.

**Alter:** Depending on the necessity and therefore analyzing the emergency conditions of the applicants, the underwriters might alter the type of term policies for providing insurance that might vary from the previous insurance plans up to a certain extent. Therefore this factor helps in better management through mutual satisfaction between insurance companies and customers.

**Transfer:** Transferring is done for transmitting the insurance policy at certain times and thereby managing certain risk periods tactfully without suffering a total loss. This factor helps in increasing the reliability of different applicants.

Another theory that is involved is smart thinking theory which is done through machine learning. This theory also involves three steps which are as follows:

**Sense:** Techniques of AI are being used for sensing the data that are being available in the present scenario which are being provided by underwriters for getting suitable results.

**Think:** The thinking process is done by performing the calculations with the help of different algorithms available in machine learning and thereby providing exact results through calculations on existing parameters (Perschy and Ford, 2019).

**Act:** After suitable calculations, data is obtained which help in making predictions for the underwriters for getting the idea about the possibility of risk that will be faced in future and thereby calculating the profitability rate based on this data.



**Figure 6: Model**

(Source: Self-created)

The model that is being used for conducting this research is data mining which involved the use of AI for making calculations. Data mining is defined as the process which helps in making predictions automatically with the help of different types of data that is being available and hence generating results for getting ideas about unknown data. The two types of methods involved in this model are unsupervised and supervised learning.

**Supervised learning:** This type of learning makes use of AI for the generation of patterns based on the legitimacy that is being developed while providing insurance. The patterns are generated based on the models that are already known.

**Unsupervised learning:** This type of learning makes use of artificial intelligence for developing different patterns especially on the data that are not available at present and hence using this data for detecting the type of risks that might arise in future.

Different methods are available in the above two-mentioned process, which are as follows:

**Anomaly detection:** This method helps in performing calculations based on the probability range of a particular claim that is obtained by reviewing the “insurance claim” that was done previously.

**SVM:** This method helps in the detection of different types of risks that are being generated in future and hence making use of artificial intelligence for performing different calculations based on the available data for developing decisions through suitable judgements.

**K-means Algorithms:** This method helps in the performance of algorithms by clustering methods for making use of data to understand the similarity level amongst the different types of clusters generated (Hanson and O'halloran, 2020). Based on these similarities, the underwriters are able to develop decisions for providing term insurance to different applicants. These algorithms make the job of the underwriters easier for making the proper calculation of data (patentimages.storage.googleapis.com, 2019).

## 3.5 Literature Gap

Research is being done for understanding the probability of different types of risks that are being generated while providing insurance policies to the different applicants and thereby understanding the profitability range of the insurance sectors (Nikodem, 2019). Based on the objectives of this research topic, different models and theories are developed for making ultimate use of them for the accomplishment of the goals and hence providing a suitable plan for risk management. Moreover, the opinions and suggestions of different authors are taken for understanding their way of approach in understanding the essentiality of underwriting in providing claims for insurance to different applicants and hence meeting their satisfaction level with respect to the profit that is being earned from the premiums deposited by the consumers. However, in spite of so many advantages of underwriting which is being done with the help of machine learning, there have been few limitations that are available in the adoption of this method which leads to the generation of the gap in providing an academic review for making detailed study. The limitations are basically generated with the use of supervised learning which lacks in developing a proper classification of new types of risks that will be generated in future while providing insurance to different applicants. The problem is that insurance is being basically provided for making the future secure from different types of risks, whereas this model only helps in making predictions based on known types of risks. Hence, due to this factor, new risks remain unpredicted, which leads to the generation of serious problems in future. Again, limitations also might arise with the use of unsupervised learning in machine learning in which "duplicate claims" cannot be detected, which might lead to fraudulent activities in future. Therefore, the use of this method might not be successful in getting a proper risk management plan and hence increasing the insecurity level of the insurance sectors in future. Furthermore, clustering methods that are being used for making calculations with the help of K-means algorithms prove to be deficient in making proper management of new parameters that will probably be created in future and hence unable to give proper predictions with the help of calculations that are developed through this type of algorithms. Moreover, drawbacks might also arise due to human errors which occur mainly while recording the measurable parameters for making predictions through calculations. The input of wrong data might lead to the generation of wrong data through calculations and hence making the wrong prediction of different types of risks (Akkor and Ozyukse, 2020). This will lead to an increase in the generation of problems of insurance sectors in future and thereby lead to the application of legal measures by the applicants when their demands are not being fulfilled in future (patentimages.storage.googleapis.com, 2019).

# 4.0 Methodology

The design flow of the application is shown in Figure 9 below:

Figure 9: Proposed Methodology

To start with, the model requires the Life Insurance claims data and the same dataset has been taken from https://www.kaggle.com/easonlai/sample-insurance-claim-prediction-dataset. As the dataset is not big data, we have a small number of variables which includes age, gender, bmi, children, smoker, region, charges, insurance claim etc. Now the next step is to do the exploratory data analysis. In this stage, we will segregate the data set as per different variable like the segregation of the data on the basis of different important attributes like gender counts, smoker statistics, BMI data statistics, age statistics (by defining the age groups) is done. Next step is to prepare the data and data pre-processing for the training and testing purpose. Data Pre-processing is also considered the most crucial phase in data analytics, where the unrefined information is cleaned and converted into an understandable format. Data cleaning, data enrichment and data normalization is the part of data pre-processing techniques. In other words, data preparation is the step by step process of readying data for the training, testing, and implementation of an algorithm. This process can be comprised of NA Treatment, Missing values Imputation, Treatment of outliers etc. (if feasible with our data). After that, the machine learning part comes and for that Python, the tool is used to develop an algorithm or a model. As in this case is XgBoost and this model is trained with the dataset and further this model is tested on the dataset remaining using the evaluation metrics like F1 score, accuracy, recall and precision as used in this case.

For comparative analysis, Support Vector Machine, Decision Trees and Naïve Bayes. These models are used only for comparison of the XGBoost score. Support Vector machine tries to find a hyperplane in an n-dimensional space where n is the number of the features where the data points are separated into their distinct classes. Decision Tree is a tree representation of the data points that leads to the final leaf nodes which are classes in the data needs to be classified. Naïve Bayes model works under the assumption that there is a conditional independence present in between the class variables for every pair of features considering the class. Accuracy score is the score of the data points that the model has identified correctly. F1-Score is the based on precision and recall rate of the model. Precision is the measure of how exact the model was to predict the class. A low precision of the model means that the model is predicting more true classes which are in real data negative. A Recall is a measure of the completeness of the model. A low recall rate means that the model is predicting more false negatives which are in real data true.

Mathematically it can be represented as shown in figure 10 below.

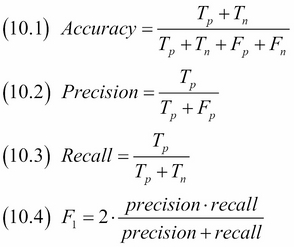


Figure10: Accuracy, precision, recall and F1-Score formula

(Source: https://subscription.packtpub.com/book/big\_data\_and\_business\_intelligence/9781785282287/10/ch10lvl1sec133/computing-precision-recall-and-f1-score)

Further, the results are analysed and reviewed for the conclusion.

# 5.0 Finding Analysis

## 5.1 Insurance Claim Data

In this paper, we have used 1338 number of rows with the data of the customers who already had bought or intended to get the life insurance. This data includes age, gender, bmi, children, smoker, region, charges, insurance claim etc., as shown in the Table 1. To do the data exploratory analysis, we have used the Python tool. Firstly, we have segregated the data on the basis of different important attributes like gender counts, smoker statistics, BMI data statistics, age statistics (by defining the age groups).

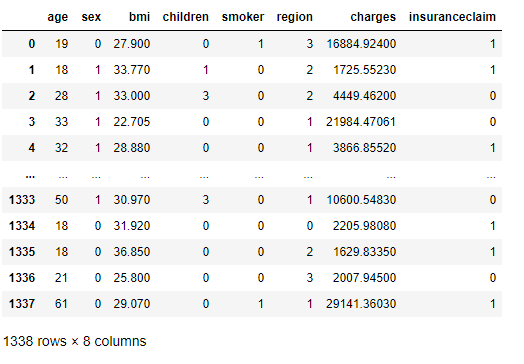


Table:1 Sample Data

## 5.2 Exploratory Data Analysis

Now this data is summarized by using related statistics data like mean, minimum, maximum, also to get a fair sight of data, it also classified on the basis of quartiles like 25%, 50%, 75% etc. and maximum values we have got in the dataset is also displayed to get the idea of the data. One can see the maximum age is 64, bmi is almost 53 and the charges are 63770 while the mean of the data says age, bmi and charges as 39, 31 and 13270. The probability of getting the insurance claim is almost 60% in the mean data while more than 50% of the data is showing 100% chances of getting the insurance claim. Using this data the underwriter can review the charges of the insurer to get a check on the losses in the future. This summary can be seen in the Table 2.

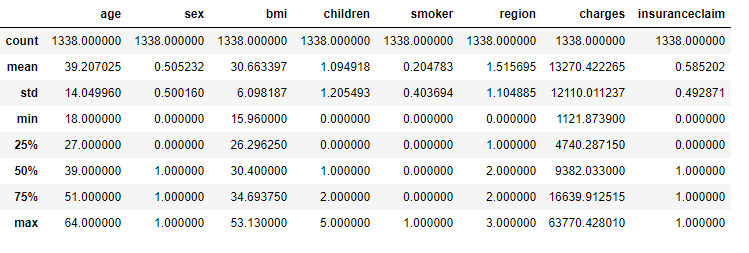


Table 2: Sample Data Statistics

In most of the statistical case analysis, the fundamental task is to characterize the dataset according to its location and its variation. One can go some step ahead by characterizing further like analysing skewness and kurtosis. Now what are skewness and kurtosis of a particular dataset and what the permissible limits of it are?

**Skewness** is a measure of symmetry, or we can say more accurately, the deficit of symmetry. In other words, if we plot the dataset on a graph the distribution of the data are not even or regular, the degree of un-symmetry of the data is called skewness. A distribution, or data set, is symmetric if it looks the same to the left and right of the centre point.It is the degree of distortion from the symmetrical bell curve or the normal distribution. It measures the lack of symmetry in data distribution.It differentiates extreme values in one versus the other tail. A symmetrical distribution will have a skewness of 0. Basically two types of Skewness: Positive and Negative can be seen.

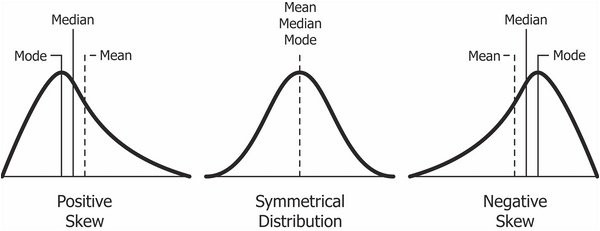


Figure11: Skewness

**Positive Skewness** means when we see the data plot graphically the tail on the right side of the distribution is longer or fatter. Also the mean and median will be greater than the mode.

**Negative Skewness** means when we see the data plot graphically the tail of the left side of the distribution is longer or fatter than the tail on the right side. Moreover the mean and median will be less than the mode.

Now the question arises, when is the skewness too much?

The thumb rule says:

* If the skewness is between -0.5 and 0.5, the data are fairly symmetrical.
* If the skewness is between -1 and -0.5(negatively skewed) or between 0.5 and 1(positively skewed), the data are moderately skewed.
* If the skewness is less than -1(negatively skewed) or greater than 1(positively skewed), the data are highly skewed.

Mathematically,

g1=∑Ni=1(Yi−Y¯)3/Ns3

Where ‘Y¯’ is the mean, ‘s’ is the standard deviation, and ‘N’ is the number of data points. Note that in computing the skewness, the‘s’ is computed with ‘N’ in the denominator rather than ‘N – 1’.

The value zero belongs to the normal distribution for skewness while any data that is symmetric or close to symmetric should have a value close to zero of skewness.Left skewed relates to the negative value of the skewness which means data that are skewed left will have negative value and talking the other way, data which indicated on the plot that are skewed to right means that the data is having positive value of the skewness. As one can see in the relevant images the left skewed data plot is that which is showing as the left tail is long with respect to right tail and the other way for the other result i.e. Right skewed means that if we compare the length of the right tail to the left tail in the data plot, we will see the right tail is long in comparison which eventually means that the data is skewed right. If the data are multi-modal, then this may affect the sign of the skewness.Some measurements have a lower bound and are skewed right.

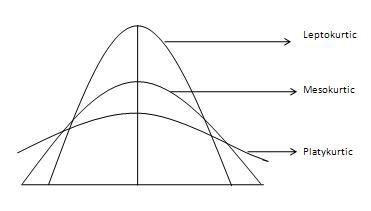
**Kurtosis** is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers. A uniform distribution would be the extreme case.

**Kurtosis** term goes around the tails of the distribution — not the peaked-ness or flat-ness of the distribution. The showcasing of extreme values with respect to the other tail is also the Kurtosis of the distribution of the data. It is used to describe the extreme values in one versus the other tail. It is actually the measure of outliers present in the distribution.

**High kurtosis** in a data set is an indicator that data has big tails or outliers. If there is a high kurtosis, then, we need to find out why we have so many outliers. It can be the symptoms of many problems that is wrong in the data like wrong data entry, or other things.

**Low kurtosis** in a data set is an indicator that data has small tails or lack of outliers. If we get low kurtosis, more often doesn’t seems to be true, like too good to be true, then also we need to find out the reason and trim the dataset of unwanted results which is causing low kurtosis.

**Mesokurtic**: This type of distribution has kurtosis statistic more or less similar to that of the normal distribution. It indicates that the extreme values of the distribution are very similar to that of a normal distribution characteristic. This definition suggests that the standard normal distribution has a *kurtosis of three.***Leptokurtic (***Kurtosis > 3***):**Distribution is longer, tails are fatter. Peak is higher and sharper with respect to Mesokurtic. In return this means that data are heavy-tailed or there are intrusions of outliers in the dataset.  
Outliers stretch the horizontal axis of the histogram graph, which makes the bulk of the data appear in a narrow (“skinny”) vertical range, thereby giving the “skinniness” of a leptokurtic distribution.  
**Platykurtic: (***Kurtosis < 3***):**Distribution is shorter; tails are thinner than the normal distribution or with respect to Mesokurtic. The peak is lower and broader than Mesokurtic, which means that data are light-tailed or lack of outliers.  
The reason for this is because the extreme values are less than that of the normal distribution.



Mathematically,

kurtosis=∑Ni=1(Yi−Y¯)4/Ns4−3

Where ‘Y¯’ is the mean, s is the standard deviation, and ‘N’ is the number of data points. Note that in computing the kurtosis, the standard deviation is computed using ‘N’ in the denominator rather than ‘N – 1’.

This definition is used so that the standard normal distribution has a kurtosis of zero. In addition, with the second definition positive kurtosis indicates a "heavy-tailed" distribution and negative kurtosis indicates a "light tailed" distribution.

Skewness and Kurtosis of the data for its normality is also checked and data is pre-processed by removing the void or null data to make the data ready to use in the model for testing.

Data has also been plotted on the graph to get a better representation for understanding the data as shown.

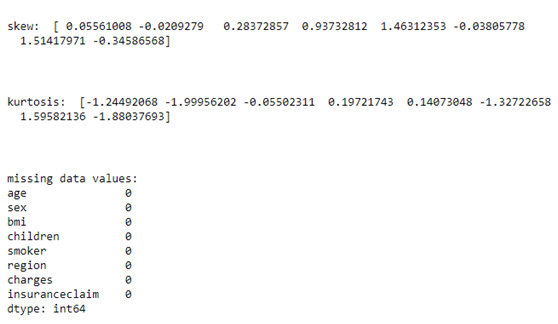
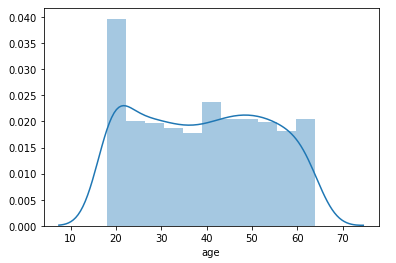
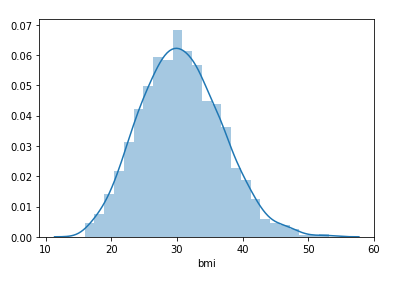


Fig.12 Skewness, Kurtosis and null value check



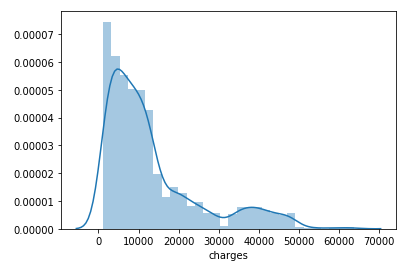


Fig.13 Data plot

This plotting can roughly predict the best fit line which in turn can tell us the average, minimum, maximum of the particular attribute in the dataset e.g., age, bmi, charges etc.,

Pair-plotting of all the attributes is also done, to predict any relation in these attributes which can be helpful in the process of underwriting.

One can see a relation between age & charges, age & insurance claim, smoker & charges as these are also evident in the industry also. In short one can say if a particular person is having age more than 50, is a smoker, having chronic illness etc., other logical factors than the particular has all the chances to bear a high premium charges for the insurance policy as the person is more expected to claim and use the policy benefits or in other words will be more cost to the insurance company.

Further the data is prepared to train our models for making prediction for the purpose of benefiting the underwriter.

## 5.3Data preparation &Data Pre-processing

After exploratory data analysis the step that is performed is data Pre-processing and it is also considered most crucial phase in data analytics, where the unrefined information is cleaned and converted into the understandable format. Data cleaning, data enrichment and data normalization is the part of data pre-processing techniques. In other words data preparation is the step by step process of readying data for the training, testing, and implementation of an algorithm. This is a multi- step process which comprises of:

* data collection,
* cleaning & pre-processing,
* feature engineering, and
* labeling.

Some of the major steps in data preparation are discussed below:

**NA Treatment**

The one important step involves in the process is NA Treatment, which normally can be seen in all the data which is available always have NA. In this process, we have to carefully find out these values and fill or replace them in an appropriate manner. There are so many techniques or methods to do the same. There are many packages also available in the industry in R and Python. On the basis of some default logic, these package deals with this issue automatically. Although, resolving this issue manually is always being a better option as doing this manually can give a better understanding of the data even more and also can replace this issue with one's own understanding of the business understanding and requirement.

**Missing values Imputation**

This is also a very important step to treat Null, in the process. Below can be the steps for the Null’s treatment. Note these steps are totally depending on the data in hand type but certainly these steps will help a bit though in every aspect.

i. If we have found the variable as being continuous, continuous variable, or we can say numerical variable. The use of Mode/mean/Median can be seen for the treatment of missing value.

 ii.     If we have found the variable as being categorical, imputation of the nulls with “Unavailable” as those are null or unavailable values can amalgamate in contributing significantly in the creation of the model and one cannot lose many important characteristics.

**Treatment of outliers**

The treatment of the outliers in the dataset must be the further step. In this, we confirm the attendance of the outliers, and imputation is done using mode/mean/median methods. This imputation process can be followed by several other methods i.e., imputation or assigning of the same outliers from those data that are either not an outlier or placed at borderline. One may show these types of variables by utilizing Boxplot of characteristics. One more display methods are Bin, which can also be used for the showcase of the variables as a part of the outlier. The box plot for our data is shown in figure below.

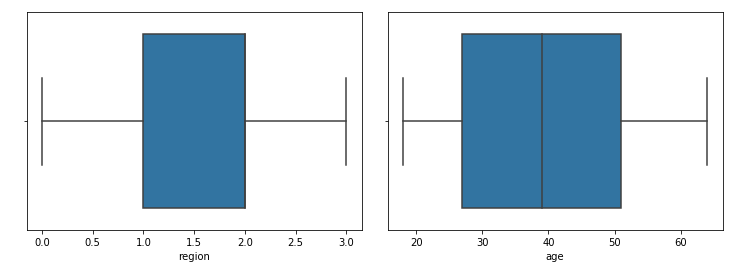


Fig.9 Box plot

In our experiment we have also performed the feature extraction method, where we have extracted the age, bmi, charges, smoker and insurance claim as the features. Our dataset was already encoded and we did not have to use any encoding technique to convert it.

## 5.4 Machine Learning Algorithms

The tool used for developing machine learning algorithms is Python. The experimentation process involved 70% of the data for the purpose of training and remaining for testing the model. The models are implemented starting from initializing the model and then the model is fitted with the training data. Once the models are trained, the prediction is performed on the test feature set to predict the target set.

The model’s performance is then evaluated by using the predicted values of the test data and the actual known values of the test data. The accuracy score, f1 score is generated and saved into a separate data frame for comparative analysis. The classification report having the precision and recall is also generated for the model. These steps continue for all the models and the performance is tracked.

## 5.5 Evaluation of Metrics

The XgBoost algorithm has performed with 99% accuracy and 99% F1 score. The classification report for the model is shown in figure 10 below. From the figure, we can see that model identified class 0(no-claim) with 99% precision and 98% recall, also the model identified class 1(claim) with 98% precision and 98% recall.

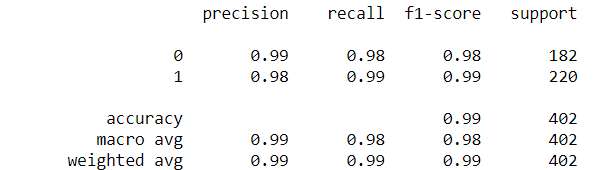


Fig.10 Classification report of XGboost model

The Decision Trees algorithm has performed with 69% accuracy and 71% F1 score. The classification report for the model is shown in figure 11 below. From the figure, we can see that model identified class 0(no-claim) with 81% precision and 43% recall, also the model identified class 1(claim) with 66% precision and 92% recall. The model very low measure for all the evaluation metrics.

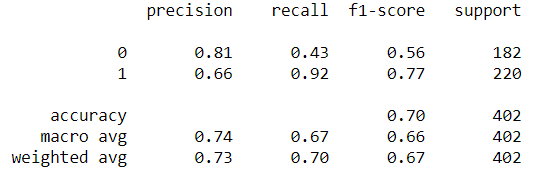


Fig.11 Classification report of Decision Trees model

The Support Vector Machine algorithm has performed with 89% accuracy and 89% F1 score. The classification report for the model is shown in figure 12 below. From the figure, we can see that model identified class 0(no-claim) with 90% precision and 86% recall, also the model identified class 1(claim) with 89% precision and 92% recall.

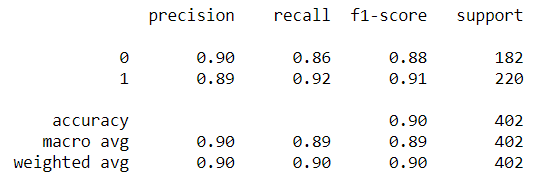


Fig.12 Classification report of Support Vector Machine model

The Naïve Bayes algorithm has performed with 89% accuracy and 89% F1 score. The classification report for the model is shown in figure 13 below. From the figure, we can see that model identified class 0(no-claim) with 90% precision and 86% recall, also the model identified class 1(claim) with 89% precision and 92% recall.

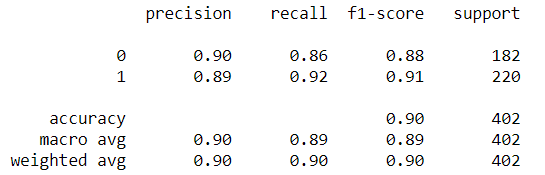


Fig.13 Classification report of Naïve Bayes model

# 6.0 Conclusion

## 6.1 Results and Discussion

To evaluate the results, we have calculated Accuracy, Recall, Precision and F1 score for all the models implemented XgBoost, Support Vector Machine, Decision trees and naïve Bayes algorithms. The outcomes obtained from the evaluation in table 3 shows the different type of aspects of the experiment which highlight that with the accuracy score of the test is 99% and F1 score of the same is coming around 99% XGBoost outperformed all the models. All the evaluation metric used showed very good result and consistency in the results and also even if we talk in regards of the sample quantum and it will definitely help the underwriter to decide which claim to process and which claim is not feasible to process.

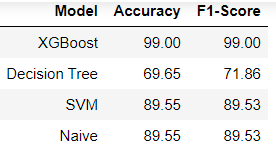


Table 3: Model Result

Figure 14 below shows the plot of accuracy score of the models. Its evident from the graph that XGBoost has the highest accuracy score, followed by SVM and Naïve Bayes and the lowest score belonged to Decision Trees.

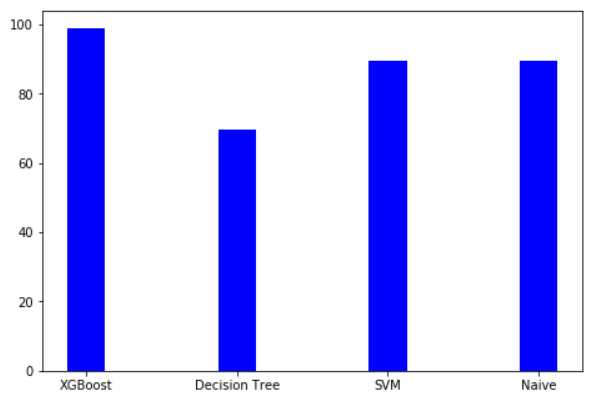


Fig.14 Accuracy plot of the models

Figure 15 below shows the plot of F1-score of the models. Its evident from the graph that XGBoost has the highest F1-score, followed by SVM and Naïve Bayes and the lowest score belonged to Decision Trees.

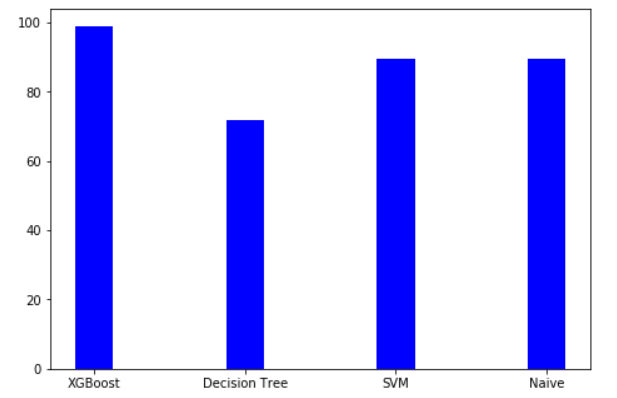


Fig.15 F1-Score plot of the models

To conclude that different authors have suggested the use of underwriting process proves to be beneficial for providing insurance to the different applicants by the insurance sectors. To conclude that with the help of a conceptual network, a good relationship has been established between the independent and dependent variables for getting ideas about the different factors generated while providing insurance. This research has helped in getting a proper plan for risk management with the help of different theories which implies machine learning approaches for making suitable predictions. To conclude that there are numerous drawbacks which might affect the use of machine learning in performing underwriting by insurance sectors. The size of the data for training the model needs to be huge to put the model to use to predict underwriting in real-world, also the resources(system capability) and time required for the model pay a major role in models efficiency. If any of these factors go offtrack the performance of the model goes haywire.

## 6.2 Future Scope

In my view, there is the future scope of working this same issue by using Artificial neural network in deep learning techniques. As this model (Artificial Neural Network) can lead to better results using big data with a large number of variables while XgBoost is a better model for tabular data having the small number of variables using different evaluation metrics for accuracy check. These results can be made more realistic by using big data for the training purpose of the model and further for experimentation using different evaluation metrics.

**7.0 Reference list**

**Journals**

Akkor, D.G. and Ozyukse, S., 2020. The effects of new technologies on the insurance sector: a proposition for underwriting qualifications for the future. *Eurasian Journal of Business and Management*, *8*(1), pp.36-50.

Bonissone, P.P., Messmer, R.P., Durham, W.M., Yang, D., Pavese, M. and Russell, D.M., Pacific Life Insurance Co, 2018. *System for rule-based insurance underwriting suitable for use by an automated system*. U.S. Patent Application 15/836,689.

Denning, R.O., Moin, M.M. and Perschy, D.L., Hartford Fire Insurance Co, 2020. *System and method for proactive underwriting using social data*.U.S. Patent 10,672,077.

Devereaux, R., Franck, D.S., Moritz, K., Read, S., Wheeler, D., Vaith, E., McCaw, E., Norman, C., Healy, J.J., Prasad, B. and Swain, K.L., 2017. *Utilizing credit and informatic data for insurance underwriting purposes*.U.S. Patent 9,818,158.

Hanson, R.M., Ledder, J.A. and O'halloran, C.R., Allstate Insurance Co, 2020. *Communication of insurance claim data*.U.S. Patent 10,685,401.

Hashim, S.H., Oracle International Corp, 2017. *Techniques For Underwriting Insurance Policies Using Web-Centric Insurance Management System*. U.S. Patent Application 15/369,269.

Hibler, G.T., Thomas, J.B. and Ford, P., Traffk LLC, 2019. *Insurance risk evaluation systems and methods*. U.S. Patent Application 16/376,712.

Howe, J., 2020. Predicting the Unexpected: Applying Advanced Underwriting to Accurately Predict Early Duration Claims in Life Insurance.

Nikodem, E., 2019. Understanding underwriting: A look inside commercial real estate insurance—breaking down the driving differentiators behind property insurance pricing. *Corporate Real Estate Journal*, *8*(3), pp.228-234.

Ross, G. and Walker, T., Massachusetts Mutual Life Insurance Co, 2019. *Methods and systems for improving the underwriting process*.U.S. Patent 10,489,861.

Soye, Y.A. and Adeyemo, D.L., 2018. Underwriting Capacity and Income of Insurance Companies: (A Case of Nigeria). International.

Trainor, C. and Vavrasek, D., Sensormatic Electronics Corp, 2017. *Sensor Based System And Method For Augmenting Underwriting Of Insurance Policies*. U.S. Patent Application 14/870,463.

Ekin, T., 2020.Discussion of “Machine learning applications in non-life insurance”. *Applied Stochastic Models in Business and Industry*.

Jamal, S. and Shah, K., 2019.Predictive Analysis for Claims in Insurance Industry using Machine Learning.

Naujoks, H., Mueller, F. and Kotalakidis, N., 2017. Digitalization in insurance: the multibillion dollar opportunity. *Boston, MA: Bain & Company. Preuzeto*, *30*.

Patil, K., Kulkarni, M., Sriraman, A. and Karande, S., 2017, December. Deep learning based car damage classification. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 50-54).IEEE.

Rosén, H., 2020. Automation of Medical Underwriting by Appliance of Machine Learning.

Sabharwal, C.L., 2018. The rise of machine learning and robo-advisors in banking. *IDRBT Journal of Banking Technology*, p.28.

Spedicato, G.A., Dutang, C. and Petrini, L., 2018. Machine learning methods to perform pricing optimization.A comparison with standard GLMs.

Ticconi, D., 2019. Individual claims reserving in creditor protection insurance using machine learning. *Available at SSRN 3458826*.

Nurse, J.R., Axon, L., Erola, A., Agrafiotis, I., Goldsmith, M. and Creese, S., 2020. The Data that Drives Cyber Insurance: A Study into the Underwriting and Claims Processes.

Mwangi, M., 2017. Effects of Underwriting and Claims Management on Performance of Property and Casualty Insurance Companies in East Africa.

Nayak, N., Olsen, J., Sree, N., Abrokwah, S.O., Martin, J. and Yung, J., Swiss Reinsurance Co Ltd, 2019. *Life insurance system with fully automated underwriting process for real-time underwriting and risk adjustment, and corresponding method thereof*. U.S. Patent Application 16/165,370.

Howe, J., 2020. Predicting the Unexpected: Applying Advanced Underwriting to Accurately Predict Early Duration Claims in Life Insurance.

Besman, E.D., Hay, J.T., Onnen, D.J., Reale, P.S. and Wells, I.J.E., Hartford Fire Insurance Co, 2017. *Calibrated underwriting system*.U.S. Patent 9,786,012.

Wang, Y., 2017. Study on the Underwriting Circle of Property-Liability Insurance Market under Competition. *Advances in Social Sciences Research Journal*, *4*(6).

Liu, S., Xu, G., Zhu, X. and Zhou, Z., 2017, October. Towards simplified insurance application via sparse questionnaire optimization. In *2017 International Conference on Behavioral, Economic, Socio-Cultural Computing (BESC)* (pp. 1-2).IEEE.

Biddle, R., Liu, S., Tilocca, P. and Xu, G., 2018, May. Automated Underwriting in Life Insurance: Predictions and Optimisation.In *Australasian Database Conference* (pp. 135-146).Springer, Cham.

Raines, R., Nesson, E., Lackey, D., Fritts, T. and Bush, I., Carfax Inc, 2012. System and method for insurance underwriting and rating.U.S. Patent 8,255,243.

Hele, J. and Serflek, C., WORLDINSURE Ltd, 2002. Underwriting insurance. U.S. Patent Application 09/993,186.

**Online Articles**

bus.wisc.edu, 2016, A Model of Underwriting and Post-Loss Test without Commitment in Competitive Insurance Market, Available at: https://bus.wisc.edu/-/media/bus/knowledge-expertise/academic-departments/asrmi/events/2016-2017/risk-theory-society/huang\_kamiya\_schmit.pdf, [Accessed on: 23.07.2020]

d1wqtxts1xzle7.cloudfront.net, 2017, Analysis of Factors Affecting Outstanding Claim Provision of Non-Life Insurance Firms in Ghana, Available at: https://d1wqtxts1xzle7.cloudfront.net/54191013/B0804032634.pdf?1503275784=&response-content-disposition=inline%3B+filename%3DAnalysis\_of\_Factors\_Affecting\_Outstandin.pdf&Expires=1596290315&Signature=TMJAMMn~LSpJ3dGGZjytlYeFXQ9zto-M1QSOIIK6XfejLzHZ2zpACczW2A43JMteipA33AxkJ5a-uQKXyeJFk4w-sQsjaDysY6xhftesofAM7~NnRJwRhfZZ94wqZgV4Hs9tqdgkD30TFBxYGPdO0XJs825jxKS3BIyqk11JWmAkQGbBt50EtwtjiL-McDda2VkKNvIwr3hH341Ggrd4VUiuzkIos2xg0dI~s7a0VGhsk-nrP9Ikpk17rIGIIywH0L9-eILyj51sRqlDZj1hutc2bzNxZ2pz1OO3g4OFf5yP~OijRIS~cMPRDc6pWoms3-mGh6VrlzZlNpJIPP2SFQ\_\_&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA, [Accessed on: 23.07.2020]

patentimages.storage.googleapis.com, 2019, DETERMINING AND INITIATING INSURANCE CLAIM EVENTS, Available at: https://patentimages.storage.googleapis.com/8f/54/3d/b19482f4079f00/US10181159.pdf, [Accessed on: 23.07.2020]

patentimages.storage.googleapis.com, 2019, INSURANCE DATA MANAGEMENT SYSTEM, Available at: https://patentimages.storage.googleapis.com/fc/21/6c/b59b4dba1b0a62/US10453148.pdf, [Accessed on: 23.07.2020]

www.researchgate.net, 2017, The Insurer Factors and the Determination of Retention to Improve the Performance of Insurance Companies in Indonesian Insurance Industry, Available at: https://www.researchgate.net/profile/Bernardus\_Nugroho/publication/317745104\_The\_insurer\_factors\_and\_the\_determination\_of\_retention\_to\_improve\_the\_performance\_of\_insurance\_companies\_in\_Indonesian\_insurance\_industry/links/5a8e2b0caca272c56bc40fcc/The-insurer-factors-and-the-determination-of-retention-to-improve-the-performance-of-insurance-companies-in-Indonesian-insurance-industry.pdf, [Accessed on: 23.07.2020]

**Appendix**

**Code**

#!/usr/bin/env python

# coding: utf-8

# ## Importing Libraries

# In[1]:

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from scipy.stats import kurtosis, skew, stats

from xgboost import XGBClassifier

from sklearn.svm import SVC

from sklearn.tree import DecisionTreeClassifier

from sklearn.preprocessing import StandardScaler, MinMaxScaler

from sklearn.pipeline import Pipeline

from sklearn.naive\_bayes import GaussianNB

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import accuracy\_score, classification\_report,f1\_score

# ## Data Import

# In[2]:

data = pd.read\_csv("datasets\_26475\_38092\_insurance2.csv")

data

# ## Exploratory Data Analysis

# ### Gender counts

# In[3]:

female = int (data.sex.value\_counts()[0])

print (data.sex.value\_counts())

print ("Prob Male: %.3f,\tProb Female: %.3f" % (1-(female/1338), (female/1338)))

# ### Smoker statistics

# In[4]:

non\_smoker = data.smoker.value\_counts()[0]

print (data.smoker.value\_counts())

print ("Prob Non-smoker: %.3f,\tProb smoker: %.3f" % ((non\_smoker/1338), 1-(non\_smoker/1338)))

# ### BMI data statistics

# In[5]:

underweight = data.bmi.between(0,18.499999, inclusive=True).value\_counts()[True]

normal = data.bmi.between(18.5, 24.999, inclusive=True).value\_counts()[True]

overweight = data.bmi.between(25, 29.99999, inclusive=True).value\_counts()[True]

obese = data.bmi.between(30, 10000, inclusive=True).value\_counts()[True]

# print (underweight, normal, overweight, obese)

print ("underweight: \t%d,\t%.3f" % (underweight, underweight/1338))

print ("normal: \t%d,\t%.3f" % (normal, normal/1338))

print ("overweight: \t%d,\t%.3f" % (overweight, overweight/1338))

print ("obese: \t\t%d,\t%.3f" % (obese, obese/1338))

# ### Age stats

# In[6]:

underage = data.age.between(0,17.999, inclusive=True).value\_counts().get(True, 0)

adult = data.age.between(18, 39.999, inclusive=True).value\_counts().get(True, 0)

overadult = data.age.between(40, 59.999, inclusive=True).value\_counts().get(True, 0)

old = data.age.between(60, 200, inclusive=True).value\_counts().get(True, 0)

print ("0-18: \t%d,\t%.3f" % (underage, underage/1338))

print ("18-40: \t%d,\t%.3f" % (adult, adult/1338))

print ("40-60: \t%d,\t%.3f" % (overadult, overadult/1338))

print ("60-: \t%d,\t%.3f" % (old, old/1338))

# ### Summary Statistics

# In[7]:

data.describe()

# In[8]:

print("skew: {}".format(skew(data)))

# In[9]:

print("kurtosis: {}".format(kurtosis(data)))

# In[10]:

print("missing data values: \n{} ".format(data.isnull().sum()))

# In[11]:

sns.distplot(data['bmi'])

# In[12]:

sns.distplot(data['age']) # Distribution of age

# In[13]:

sns.distplot(data['charges']) # Distribution of charges

# In[14]:

sns.pairplot(data)

# ## Data Preparation

# In[15]:

X = data.drop(columns=['insuranceclaim'])

Y = data['insuranceclaim']

# In[16]:

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, Y, test\_size=0.3)

# In[17]:

# for comapritive analysis

score = pd.DataFrame()

# ## XGBoost

# In[18]:

model = XGBClassifier()

# In[19]:

model.fit(X\_train, y\_train)

# In[20]:

# make predictions for test data

y\_pred = model.predict(X\_test)

# In[21]:

# evaluate predictions

acc= np.round(accuracy\_score(y\_test, y\_pred)\*100,2)

acc

# In[22]:

f1= np.round(f1\_score(y\_test, y\_pred)\*100,2)

# In[23]:

print(classification\_report(y\_test, y\_pred))

# In[24]:

score= score.append([["XGBoost", acc, f1]])

# ## Decision Trees

# In[25]:

dt = DecisionTreeClassifier(max\_depth=1, random\_state=28)

dt.fit(X\_train, y\_train)

# In[26]:

y\_pred= dt.predict(X\_test)

# In[27]:

acc= np.round(accuracy\_score(y\_pred,y\_test)\*100,2)

acc

# In[28]:

f1= np.round(f1\_score( y\_pred, y\_test,average="weighted")\*100,2)

f1

# In[29]:

print(classification\_report(y\_test, y\_pred))

# In[30]:

score=score.append([['Decision Tree', acc, f1]])

score

# ## Support Vector Machines (SVMs)

# In[31]:

# SVM model intialization and generating pipeline for scaling variable for svm model

svm = Pipeline([('scaler', StandardScaler()),('svc', SVC(kernel='linear'))])

# In[32]:

svm.fit(X\_train, y\_train)

# In[33]:

y\_pred = svm.predict(X\_test)

# In[34]:

acc= np.round(accuracy\_score(y\_pred,y\_test)\*100,2)

acc

# In[35]:

f1= np.round(f1\_score(y\_test, y\_pred, average="weighted")\*100,2)

f1

# In[36]:

score=score.append([['SVM', acc, f1]])

score

# ## Naive Bayes

# In[37]:

gnb = GaussianNB()

# In[38]:

# fit

gnb.fit(X\_train, y\_train)

# In[39]:

# predict

y\_pred\_2 = gnb.predict(X\_test)

# In[40]:

acc= np.round(accuracy\_score(y\_pred,y\_test)\*100,2)

acc

# In[41]:

f1= np.round(f1\_score(y\_test, y\_pred, average="weighted")\*100,2)

f1

# In[42]:

score=score.append([['Naive', acc, f1]])

score

# ## Comparision

# In[43]:

score.columns=['Model',"Accuracy",'F1-Score']

# In[44]:

score

# In[49]:

fig = plt.figure()

ax = fig.add\_axes([0,0,1,1])

ax.bar(score['Model'],score['Accuracy'], color = 'b', width = 0.25)

plt.show()

# In[46]:

fig = plt.figure()

ax = fig.add\_axes([0,0,1,1])

ax.bar(score['Model'],score['F1-Score'], color = 'b', width = 0.25)

plt.show()