

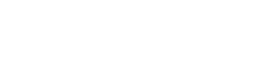
**CAPSTONE PROJECT**



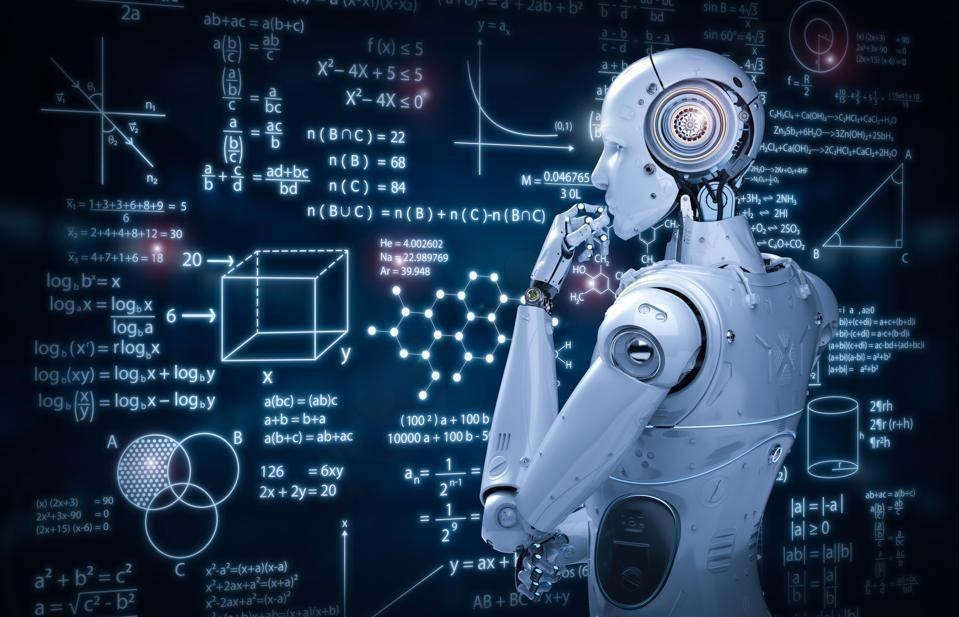
**-**



**S**



**YNOPSIS**



|  |  |
| --- | --- |
| **Batch details** | **PGPDSE-FT OCT\_22** |
| **Team members** | 1. **Bhavya Sharma** 2. **Siddhant** 3. **Gagan Bhatt** 4. **Pradum Kakran** 5. **Anmol Bhatia** 6. **Rohit Dahiya** 7. **Vivek Singh** |
| **Domain of Project** | **Insurance** |
| **Proposed project title** | **Insurance Claim Prediction using Machine**  **Learning techniques** |
| **Group Number** | **4** |
| **Mentor Name** | **MR. JAYVEER NANDA** |

**Date: 11/Feb/2023**

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Sl NO** | **Topic** | **Page**  **No** |
| **1** | **Overview** | **03** |
| **2** | **Business problem statement** | **04** |
| **3** | **Approach** | **05-08** |
| **4** | **Topic survey in depth** | **08-14** |
| **5** | **Methodology to be followed** | **14-17** |
| **6** | **Dataset** | **18-20** |
| **7** | **References** | **20** |
| **8** | **Notes for the Project team** | **21** |

# Overview

## Insurance Claim Prediction using Machine learning

**In** [**claim prediction,**](https://seon.io/resources/fraud-detection-and-prevention/) **machine learning is a collection of algorithms trained with the historical data to suggest factors affecting the insurance claim.**



**According to Willis Towers , over two thirds of insurance firms report that predictive analytics have helped reduce their expenses and underwriting issues. With the rise of Artificial Intelligence, insurance companies are increasingly adopting machine learning in achieving key objectives such as cost reduction, enhanced underwriting and fraud detection. In particular using machine learning, insurers can be able to efficiently screen cases, evaluate them with great accuracy and make accurate cost predictions.**

# BUSINESS PROBLEM STATEMENT

## Business problem understanding

**The problem is to recognize factors affecting insurance claim so that the customers of Insurance companies are not charged wrongfully**

**Main challenges involved in insurance claim prediction are:**

* **Enormous Data is processed every day and the model build must be fast enough to respond to the scam in time.**
* **Data availability as the data is mostly private.**
* **Misclassified Data can be another major issue, as not every claim is reported rightfully.**

## Business Objective

**Insurers are using machine learning to progress operational competence, from claims registration to claims settlement. Many carriers have already started to automate their claims processes, thereby enhancing the customer experience while reducing the claims settlement time. Machine learning and predictive models can also equip insurers with a better understanding of claims costs. These insights can help a carrier save millions of dollars in claim costs through proactive management, fast settlement, targeted investigations and better case management.**

## Approach

**Our Approach is to perform various methods on the Dataset which includes Understanding the Dataset, Data pre-processing, Handling Datatypes, Scaling & Transformation of Data wherever required, dealing with the null values, Handling of Outliers with help of EDA techniques & Visualizations, building classification Model, Evaluate & Deployment of the model.**





**Data preprocessing**

**It is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.**

**When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.**

**It includes the following:**

**Descriptive Analysis: Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data. Measures of variability help communicate the spread of distribution by describing the shape and spread of the data set.**

**Inferential Analysis: Validating the inferences which are found with the help of descriptive analysis (Graphs) with the help of respective statistical tests if needed. Treating Outliers: Outlier is an observation in the data that lies at an abnormal distance from other values. Presence of an outlier may skew the results. Hence it is necessary to remove them. The interquartile range is a measure of outlier treatment. It is the difference between the third quartile and the first quartile. The IQR gives the range of middle 50% of the data which is free from outliers**

**Treating missing values:**

**One way of handling missing values is the deletion of the rows or columns having null values. If any columns have more than half of the values as null then you can drop the entire column. In the same way, rows can also be dropped if having one or more columns values as null**

**It is important to handle the missing values appropriately.**

* **Many machine learning algorithms fail if the dataset contains missing values. However, algorithms like K-nearest and Naive Bayes support data with missing values.**
* **You may end up building a biased machine learning model which will lead to incorrect results if the missing values are not handled properly.**
* **Missing data can lead to a lack of precision in the statistical analysis.**

**Encoding Categorical Variables:**

**Encoding categorical data is a process of converting categorical data into integer format so that the data with converted categorical values can be provided to the different models. In the field of data science, before going for the modelling, data preparation is a mandatory task.**

**There are several different techniques that are used to encode categorical values**

* **Label Encoding or Ordinal Encoding**
* **One hot Encoding**
* **Dummy Encoding**
* **Effect Encoding**
* **Binary Encoding**
* **BaseN Encoding**
* **Hash Encoding**
* **Target Encoding**

**Dropping Unnecessary Columns: We are removing the columns which do not contribute to the model building or the columns which are of less, or of no importance.**

**Removal / Replacing of Special Characters (if any): Special characters such as ‘?’, ‘$’, ‘%’ should be replaced with Nan so that they are easier to treat and replace or to remove.**

**Data preparation**

 **Standardization of data: Standardization of data is a major important step that is required for machine learning algorithms to give good results. There are different scaling functions present in the preprocessing module of sci- kit learn. If data is not scaled and is passed to the algorithm the result might be wrong due to wrongly distributed data.**

 **Normalization of data: It is the process of scaling each sample to have a unit standard. These types of techniques are much more effective if you are computing the similarity between different pairs of samples or using a quadratic form like a dot product. This is the base of models used in text classifications. As discussing text classification, learn more about text mining and text mining techniques.**

 **Scaling: It helps to normalize the data within a particular range and as well as in speeding up the calculations in an algorithm.**

**Dataset Splitting:**



**Scikit-learn alias sklearn is the most useful and robust library for machine learning in Python. The scikit-learn library provides us with the model\_selection module in which we have the splitter function train\_test\_split().**

 **The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.**

 **It is a fast and easy procedure to perform, the results of which allow you to compare the performance of machine learning algorithms for your predictive modeling problem. Although simple to use and interpret, there are times when the procedure should not be used, such as when you have a small dataset and situations where additional configuration is required, such as when it is used for classification and the dataset is not balanced.**

**Evaluate the Model**

* **Uses some metric or combination of metrics to "measure" objective performance of model** o **Test the model against previously unseen data**
* **This unseen data is meant to be somewhat representative of model performance in the real world, but still helps tune the model (as opposed to test data, which does not).**
* **Good train/eval split? 80/20, 70/30, or similar, depending on domain, data availability, dataset particulars, etc.**

**Testing the data:**

**Using further (test set) data which have, until this point, been withheld from the model (and for which class labels are known), are used to test the model; a better approximation of how the model will perform in the real world.**

**Conclusion**

**We can use the model to identify certain financial traits of future borrowers that could have the potential to recognize the customers.**

**TOPIC SURVEY IN DEPTH:**

**Insurance Claims**

**Price optimization procedure is a complex notion. Therefore it uses numerous combinations of various methods and algorithms. Despite the fact that it is still the disputable issue of applying this procedure for insurance, more and more insurance companies adopt this practice. This process supposes combining the data not related to the expected costs and risk characteristics and the data not related to the expected loss and expenses, and its further analysis. That is, it takes into consideration the changes in comparison to the previous year and policy. Thus, price optimization is closely related to the customers’ price sensitivity. In other words, historical costs, expenses, claims, risk, and profit are projected into the future. Special algorithms give the insurers the opportunity to adjust the quoted premiums dynamically. As a key positive feature, price optimization helps to increase the customers’ loyalty in long perspective. Along with this, comes the maximization of profit and income**

**Current solution to the problem:**

**Some of the currently used approaches to such problems are:**

* **Artificial Neural Network**
* **Fuzzy Logic**
* **Genetic Algorithm**
* **Logistic Regression**
* **Decision tree**
* **Support Vector Machines**
* **Bayesian Networks**
* **Hidden Markov Model**
* **K-Nearest Neighbour**

**Proposed solution to the problem**

**Our proposed solution is based on completely machine learning model.**

 **Machine learning is a core sub-area of Artificial Intelligence (AI). ML applications learn from experience (or to be accurate, data) like humans do without direct programming. When exposed to new data, these applications learn, grow, change, and develop by themselves. In other words, machine learning involves computers finding insightful information without being told where to look. Instead, they do this by leveraging algorithms that learn from data in an iterative process.**

 **At a high level, machine learning is the ability to adapt to new data independently and through iterations. Applications learn from previous computations and transactions and use “pattern recognition” to produce reliable and informed results.**

 **The Machine Learning process starts with inputting training data into the selected algorithm. Training data being known or unknown data to develop the final Machine Learning algorithm. The type of training data input does impact the algorithm, and that concept will be covered further momentarily.**

 **New input data is fed into the machine learning algorithm to test whether the algorithm works correctly. The prediction and results are then checked against each other**

 **If the prediction and results don’t match, the algorithm is re-trained multiple times until the data scientist gets the desired outcome. This enables the machine learning algorithm to continually learn on its own and produce the optimal answer, gradually increasing in accuracy over time.**

 **Supervised Learning**

**In supervised learning, we use known or labeled data for the training data. Since the data is known, the learning is, therefore, supervised, i.e., directed into successful execution. The input data goes through the Machine Learning algorithm and is used to train the model. Once the model is trained based on the known data, you can use unknown data into the model and get a new response.**

**Here is the list of top algorithms currently being used for supervised learning:**

* **Polynomial regression**
* **Random forest**
* **Linear regression**
* **Logistic regression**
* **Decision trees**
* **K-nearest neighbors**
* **Naive Bayes**

**Unsupervised Learning:**

**In unsupervised learning, the training data is unknown and unlabeled - meaning that no one has looked at the data before. Without the aspect of known data, the input cannot be guided to the algorithm, which is where the unsupervised term originates from. This data is fed to the Machine Learning algorithm and is used to train the model. The trained model tries to search for a pattern and give the desired response. In this case, it is often like the algorithm is trying to break code like the Enigma machine but without the human mind directly involved but rather a machine.**

**With the help of these methods, we can predict future risks.**

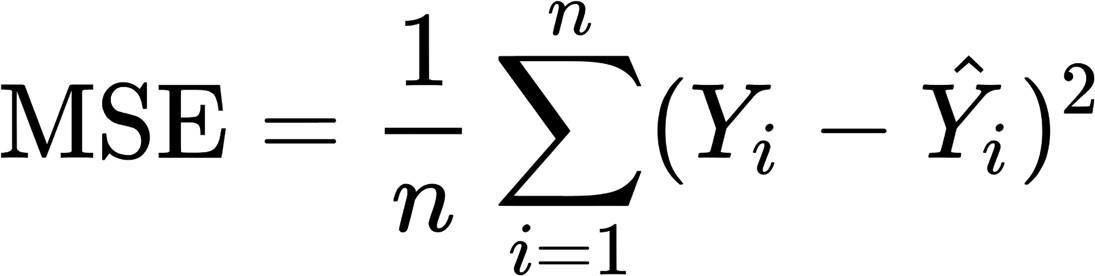
* **The top 7 algorithms currently being used for unsupervised learning are:**
* **Partial least squares**
* **Fuzzy means**
* **Singular value decomposition**
* **K-means clustering**
* **Hierarchical clustering**
* **Principal component analysis**

**Reinforcement Learning:**

**Like traditional types of data analysis, here, the algorithm discovers data through a process of trial and error and then decides what action results in higher rewards. Three major components make up reinforcement learning: the agent, the environment, and the actions. The agent is the learner or decision-maker, the environment includes everything that the agent interacts with, and the actions are what the agent does.**

**Gradient Boosting:**

**Gradient boosting algorithm is one of the most powerful algorithms in the field of machine learning. As we know that the errors in machine learning algorithms are broadly classified into two categories i.e., Bias Error and Variance Error. As gradient boosting is one of the boosting algorithms it is used to minimize bias error of the model. Gradient boosting algorithm can be used for predicting not only continuous target variable (as a Regressor) but also categorical target variable (as a Classifier). When it is used as a regressor, the cost function is Mean Square Error (MSE) and when it is used as a classifier then the cost function is Log loss. We want our predictions, such that our loss function (MSE) is minimum.**



**CRITICAL ASSESSMENT OF TOPIC SURVEY:**

**METHODOLOGY TO BE FOLLOWED:**

**The customers are always willing to get personalized services which would match their needs and lifestyle perfectly well. The insurance industry is not an exception in this case. The insurers face the challenge of assuring digital communication with their customers to meet these demands. Highly personalized and relevant insurance experiences are assured with the help of the artificial intelligence and advanced analytics extracting the insights from a vast amount of the demographic data, preferences, interaction, behavior, attitude, lifestyle details, interests, hobbies, etc. The consumers tend to look for personalized offers, policies, loyalty programs, recommendations, and options. The platforms collect all the possible data to define the major customers` requirements. After that, the hypothesis on what will work or won`t work is made. Here comes the turn to develop the suggestion or to choose the proper one to fit the specific customer, which can be achieved with the help of the selection and matching mechanisms. The personalization of offers, policies, pricing, recommendations, and messages along with a constant loop of communication largely contribute to the rates of the insurance company**

**Modern technologies have brought the promotion of products and services to a qualitatively new level. Different customers tend to have specific expectations for the insurance business. Insurance marketing applies various techniques to increase the number of customers and to assure targeted marketing strategies. In this regard, customer segmentation proves to be a key method. The algorithms perform customers’ segmentation according to their financial sophistication, age, location, etc. Thus, all the customers are classified into groups by spotting coincidences in their attitude, preferences, behavior, or personal information. This grouping allows developing attitude and solutions especially relevant for the particular customers. As a result, target cross-selling policies may be developed and personal services may be tailored for each particular segment.**

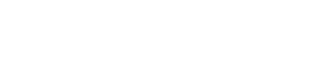
**Insurance companies lose an estimated US$30 billion per year to falsified claims. Machine learning helps them identify potential fraudulent claims faster and more accurately, and flag them for investigation. Machine learning algorithms are superior to traditional predictive models for this application because they can tap into unstructured and semistructured data such as claims notes and documents as well as structured data, to identify potential fraud. Chola MS, one of India’s fastest-growing insurance companies, has adopted mobile technology for its claims survey process. The company’s vehicle surveyor application uses the voice, camera and data connectivity capabilities of the Samsung Galaxy Tablet to capture and store auto survey data in one database. In the past, loss adjusters had to manually match survey notes with e-mail and photos saved in other databases before making a decision on a claim. This initiative helped to speed up the claims settlement process, increased surveyor productivity and improved fraud prevention**

**Insurers use machine learning to predict premiums and victims for their policies. Detecting risks near the beginning in the process enables insurers to make improved exploit of underwriters’ time and gives them a vast competitive improvement. Progressive Insurance is evidently leveraging machine learning algorithms for predictive analytics based on data collected from client drivers. The car insurer claims that its telematics mobile app,**

**Snapshot, has collected 14 billion miles of driving data. To encourage the use of Snapshot, Progressive offers ―most drivers‖ an auto insurance discount averaging US$130 after six months o**

**Flowchart**





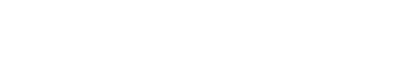
Data Acquisi



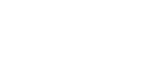
tio



n



Understanding



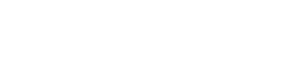
Data



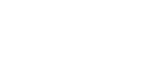
Pre



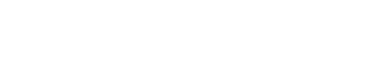
-



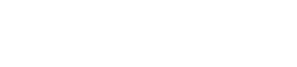
processing



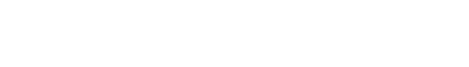
Data



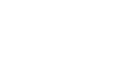
Applying EDA



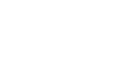
techniques



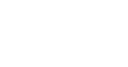
Splitting the Data



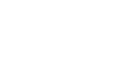
Train



Data



Test



Data

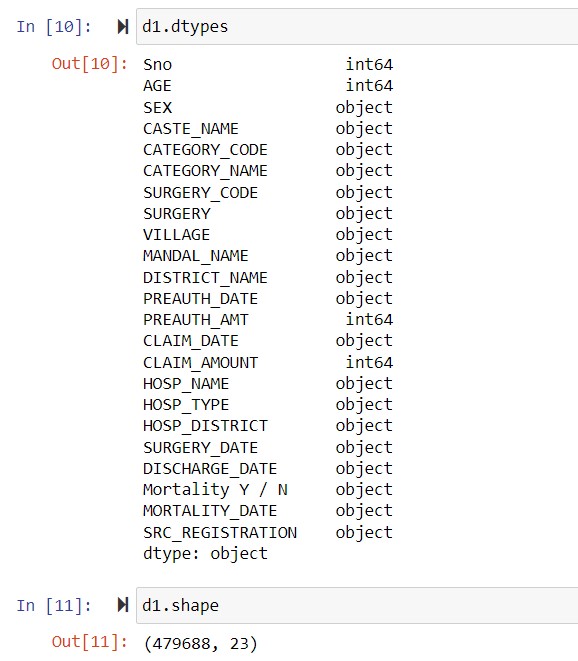


### Benefits of using the model

* **Faster and more efficient detection: The system gets to quickly identify patterns and behaviour’s that might have taken human agent’s months to establish.**
* **Reduced manual review time: Similarly, the amount of time spent on manually reviewing information can be drastically reduced when we let machines analyse all the data points.**
* **Better predictions with large datasets: The more data we feed a machine learning engine, the more trained it becomes. That is to say, while large datasets can sometimes make it challenging for humans to find patterns, it’s actually the opposite with an AI-driven system.**
* **Cost-effective solution: A machine learning system is a great ally to scale up the company without increasing risk management costs drastically at the same time.**

## Dataset used

**Reference** [**- https://www.kaggle.com/datasets/kartik2112/fraud-detection**](https://www.kaggle.com/datasets/kartik2112/fraud-detection) **Datatypes of the labels**



### Dataset attribute

|  |  |  |
| --- | --- | --- |
| **S.No.** | **Variable Name** | **Description** |
| **1** | **Serial Number** | **Serial Number given to the rows** |
| **2** | **Age** | **Age of the patient** |
| **3** | **SEX** | **Gender of the patient** |
| **4** | **CASTE\_NAME** | **Caste of the patient. (BC= Backward Caste, OC= Other Caste** |
| **5** | **CATEGORY\_CODE** | **Administrative Data (Code assigned to the Disease Category)** |
| **6** | **CATEGORY\_NAME** | **Administrative Data (Category of Disease)** |
| **7** | **SURGERY\_CODE** | **Administrative Data (Code assigned to the Surgery Name)** |
| **8** | **SURGERY** | **Name of the surgery operated.** |
| **9** | **VILLAGE** | **Village of the patient** |
| **10** | **MANDAL\_NAME** | **Mandal (sub-division of the district) of the patient** |
| **11** | **DISTRICT\_NAME** | **District of the patient** |
| **12** | **PREAUTH\_DATE** | **Date of pre-authorization of the treatment** |

|  |  |  |
| --- | --- | --- |
| **13** | **PREAUTH\_AMT** | **Amount pre-authorized** |
| **14** | **CLAIM\_DATE** | **Date of Insurance claim** |
| **15** | **CLAIM\_AMOUNT** | **Amount claimed (post-treatment)** |
| **16** | **HOSP\_NAME** | **Hospital Name** |
| **17** | **HOSP\_TYPE** | **Type of hospital (Government or Private)** |
| **18** | **HOSP\_DISTRICT** | **District where the hospital is located** |
| **19** | **SURGERY\_DATE** | **Date of Surgery** |
| **20** | **DISCHARGE\_DATE** | **Date of Discharge of Patient from hospital post surgery** |
| **21** | **Mortality Y / N** | **If the patient died in the process** |
| **22** | **MORTALITY\_DATE** | **Date of Death if dead** |
| **23** | **SRC\_REGISTRATION** | **Administrative Data (Source of Registeration)** |

**REFERENCES** 

[**https://www.kaggle.com/datasets/phiitm/andhra-pradesh-health-data**](https://www.kaggle.com/datasets/phiitm/andhra-pradesh-health-data) [**https://www.ijitee.org/wp-content/uploads/papers/v8i6s4/F11180486S419.pdf**](https://www.ijitee.org/wp-content/uploads/papers/v8i6s4/F11180486S419.pdf)

**Notes For Project Team**

|  |  |
| --- | --- |
| Original owner of data | Andhra Pradesh Government |
| Data set information | The Data contains information about personal details and personal claims of Individuals |
| Any past relevant articles using the dataset | NA |
| Reference | Claim Detection Using Machine Learning |
| Link to web page | [www.kaggle.com](http://www.kaggle.com/) |