AUTO INSURANCE FRAUD DETECTION & INCOME PREDICTION

AUTHOR
Shekhar Jogawade
210268199

SUPERVISOR
Matthew Bickley

Aston University
Birmingham

Abstract

At the current time, Fraud insurance claims are a popular problem faced by numerous insurance companies. These frauds cause huge losses to these insurance companies. These frauds do not only bring adverse consequences to fraudsters but also to other parties regardless they may or may not be at fault. As a reaction to this act, the premium cost of the policyholder is increased by a big margin and they end up paying the extra premium for the time they drive the car. But this is not the only problem we are trying to solve. The insurance industry is a big market therefore companies don't want to spend resources on the wrong customers who might not end up not buying a single service from them therefore in the second phase of the project, We will also be independently predicting the income of customers so the insurance companies can target those customers with other types of insurance such as health, property or pet insurance based on their income. The point to remember is that we are using two separate datasets for this research because there is no dataset available online which includes the data for both insurance fraud detection and income prediction. We have chosen a different income dataset with minimum features because insurance companies can collect this data from customers through surveys. The Advancement in technology and machine learning made it possible to find the solution to the problem discussed above.

In the case of both Insurance Fraud detection and Income prediction, we used 7 Different algorithms from the sklearn library such as Adaptive Boost (AdaBoost), Extreme gradient Boosting (XGBoost), Support vector machine (SVM), Logistic Regression, Random Forest, Gaussian Naive Bayes(Gaussian NB), and Decision Tree. We have also used two different encoding techniques: One-hot encoding and the Weight of evidence method. We have obtained 14 different accuracy values for machine learning model for these 7 algorithms in two different encoding techniques. The model with higher accuracy will be an optimal solution to our problem which intends to independently detect insurance fraud and predict the income of customers by considering the different features from the dataset. The Results Show that, For Insurance fraud Detection, Naïve Bayes, Logistic regression, and SVM machine learning models managed to obtain a Maximum accuracy score. The Weight of Evidence encoding technique was more successful as than One-Hot Encoding. The decision Tree Model, in which data was encoded by One-Hot encoding, was the worst-performing model in insurance fraud detection. In the case of Income Prediction, the One-Hot encoding technique was more successful. Logistic regression, SVM, and AdaBoost machine learning models managed to obtain the best accuracy score while naïve bayes was the worst performing machine learning model in terms of accuracy score.

Acknowledgement

I want to start by offering my heartfelt gratitude towards the supervisor, Prof. Matthew Bickley for his support and encouragement through this master's dissertation, his compassionate mentoring, as well as for his tremendous expertise. I want to applaud him for pushing me to extend for this research as well as looking at it from multiple angles in addition to their constructive comments & motivation. Also, I am extremely thankful to Prof. Roberto Alamino for his timely and important announcements and expertise. This project would not have been possible without their help & Contribution.

1 List of Figures

2 List of tables

3 List of Abbreviations :

- USA : United States of America
- Woe : Weight of Evidence
- XGBoost : Extreme Gradient Boosting
- AdaBoost : Adaptive Boost
- SVM : Support Vector Machine

Table of Content

Contents

4 Introduction

4.1 Research Introduction

Insurance frauds cover the range of improper activities which an individual may commit to achieving a favorable outcome from the insurance company. This could range from staging the incident, misrepresenting the situation, including the relevant actors and the cause of the incident, and finally, the extent of damage caused. The insurance industry has grappled with the challenge of insurance claim fraud from the very start. On the one hand, there is the challenge of impact on customer satisfaction through delayed payouts or prolonged investigation during stress. Additionally, there are costs of investigation and pressure from insurance industry regulators. improper payouts cause a hit to profitability and encourage similar delinquent behavior from other policyholders. The first phase of this research helps insurance companies find out the fraud insurance claim from a large number of claims with the help of machine learning. The insurance industry is a billion dollars market right now in the united states of America. This research will give insurance companies an upper hand over fraudsters [5].
Subsequently, as we progress toward the second phase of the research where we will independently predict customers' income. We will be using predictive analytics to identify behaviors and trends so that we can decide on an unknown event. Income prediction is so important to analyze the best combination of age, education, marital status, etc. for a person's Income and some business purpose[13].
Our main focus of this research is to demonstrate how insurance companies can detect fraudulent insurance claims and also predict the income of customers if they have proper data of customers. Due to the Unavailability of the dataset which includes the necessary features to progress through both phases, we have decided to use two datasets.

4.2 Contextual Information

As we move forward with the execution, we would like to point out some surprising facts about insurance fraud to address the severity of the problem we are dealing with. In 2016, Insurance providers detected 125,000 fake protection claims esteemed at £1.3 billion. It is assessed that a comparable measure of fraud goes undetected every year. This is why insurers contribute at least £200 million each year to distinguish misrepresentations. 107,000 deceitful protection claims worth £1.2 billion were uncovered by insurance providers in 2019. For every 5th minute, new insurance fraud is detected in the USA. Before Moving Forward, let's understand how the insurance market works in the USA. Car Insurance assures drivers and passengers of financial safety from road accidents or any other car-related mishap. Insurance is a legal contract between a customer or a driver and an insurance company where the customer accepts the terms and conditions stated by the insurance company and after the accident insurance company pays a monetary sum if there is bodily injury or property damage if it's mentioned in the contract. The majority of the companies provide two basic coverage areas for liability coverage: bodily injury and property damage. As different states in the USA have different insurance laws therefore some states require a minimum amount of coverage for medical payments and/or personal injury protection (PIP). Therefore people from different states use different insurance fraud tactics to get the most money out of the pockets of insurance companies.

4.3 Motivation For Research

The primary motivation for this research is the result of me being a victim of insurance fraud. I was driving my car and commuting during city traffic and due to peak hours and bumper-to-bumper traffic, my car bumped into the next car in traffic. So to take advantage of the situation, the opposite party applied for a personal injury claim of 15000 pounds against my insurance company. I started researching how many insurance frauds occur in a single day and I came across fascinating numbers of insurance frauds. This was the first time I realized that insurance frauds are very common and we need to do something about it. The facts that I have found about insurance fraud present another motivation why the insurance industry needs help in identifying fraudsters, as they are causing insurance companies financial losses worth billions of dollars around the world. The main idea behind the project motivation is that this research will be not only helpful for car insurance fraud detection but also for the other type of insurance. In the Second phase of the project, we are predicting the income of customers. The main goal behind this concept is to provide insurance companies with maximum advantage to target the customers to offer them other services. As a business owner, I faced problems in marketing my business to the right group of people. For example, if I decide to launch a new car worth 75000 pounds then targeting the customers whose annual income is less than 25000 might not return the desired results. Similarly, In the case of the Insurance industry, if we can accurately predict the customer's income then insurance companies can make a good fortune if used properly.

4.4 Aims and Objectives

Aims :

- Primary Aim of this research project is the demonstration how we can detect fraudulent insurance claims and also prediction of the income of customers if the insurance companies successfully collect the necessary information from the customers.
- Given research progress through 2 independent phases such as insurance fraud detection and income prediction.
- Detecting Fraudulent insurance claims from a given set of data by using different machine learning algorithms from the sklearn library of python.
- Finding the important insights from the data that can be useful in decision-making.
- Predicting the income of customers using different machine learning algorithms using the sklearn library of python.
- Achieve maximum accuracy of Machine Learning models.

Objectives :

- Using Several Encoding techniques such as One-Hot-Encoding and Weight of Evidence to improve model performance.
- Developing and training different machine learning models in both of the independent phases of research.
- Assess the quality of the machine learning models at the end of insurance fraud detection and income prediction.

- Insurance fraud detection dataset belongs to the company in the USA. So a general understanding of the USA insurance market is necessary to make the sense of the data insights.

4.5 Benefits Of Research

Academic :

- Insurance fraud detection with income prediction is a new concept. There will be lots of new takeaways from the research we are doing and it will be helpful for future implementations as an academic reference.

- we will be making use of different methods which are totally apart from the literature survey we assessed. We are proving the importance of Data encoding and feature Engineering through our research which has not been used significantly before.

Technical :

- insurance fraud detection is better equipped to identify risk than any one person. Machine learning and predictive modeling systems can analyze massive amounts of data in fractions of a second, whereas a person would be combing through documents for days before they could identify the same patterns this technology can.

- Insurance fraud detection is more effective at looking for the peculiarity and red flags it to indicate potential fraud. These red flags help the Data Science team to build high-quality referrals for their fraud teams. These algorithms can also identify high-risk areas that should be included in a fraud risk assessment.

- Sometimes technology advances to catch up with fraud, criminals also find new ways to stay under the radar by adapting to new tactics like seeking small claim amounts but machine learning models can be configured to identify new abnormal data anomalies

Business

- We all agree on the fact that the faster the fraud gets detected, the faster insurance companies can respond to it and try to prevent any loss. Fraud detection using machine learning greatly increases the speed at which insurance companies are identifying fraudulent claims.

- The events which slip through the cracks and cost insurance companies major financial loss are low-incident events. They are hard to identify but with the help of machine learning these events can be easily flagged and referred to the fraud team for further analysis.

- Nowadays, income is the most important data a company can hold about its customers. Income can help predict which products the customer can buy and what services he can afford and for how much time.

Social

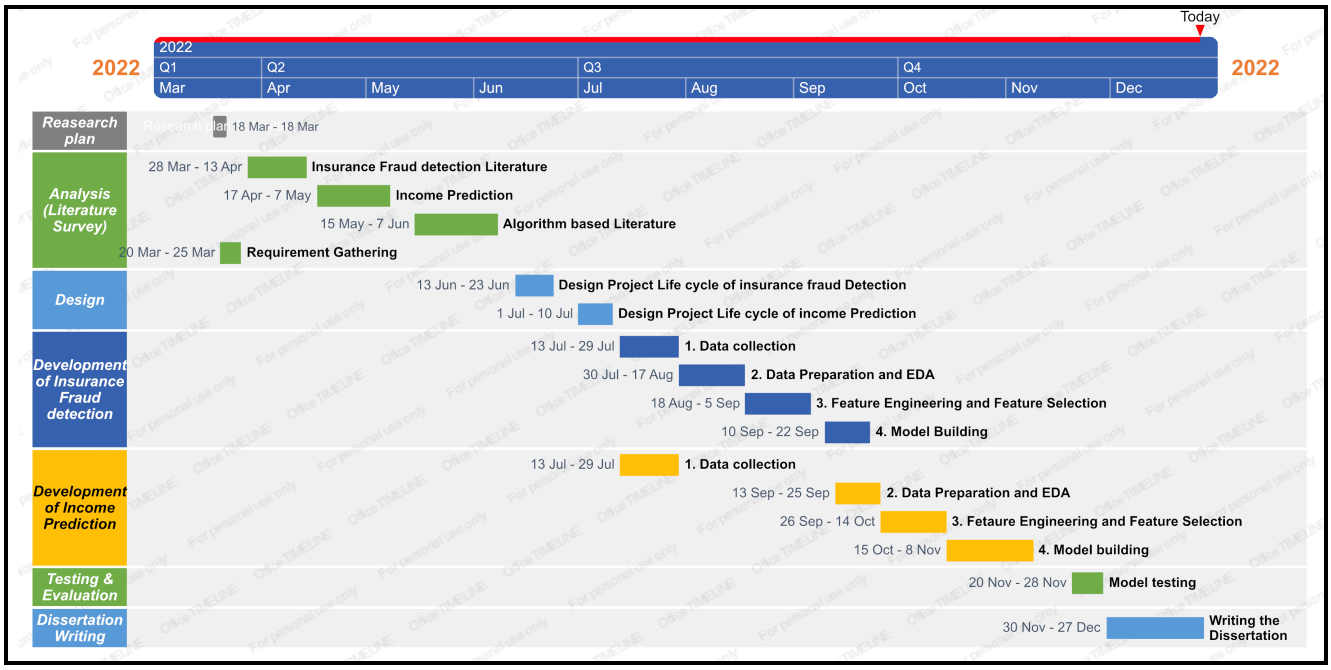
- We are using this research to make social change by reducing the number of frauds. We are trying to detect the most possible insurance fraud through our machine learning model. We are also trying to spread social awareness that machine learning made it possible to tackle fraudsters so the number of frauds can be reduced.
- With our model, we will be reducing the number of insurance fraud victims. As a victim of insurance fraud, I can relate to the situation where we have to invest our time in court proceedings and lawyer meetings. We are trying to make change happen socially by pointing out the fact that insurance laws need to be revised to tackle frauds more easily. [22].

4.6 Overview

We have Advanced through two stages of research. First phase is insurance fraud detection and the second phase is income prediction. We begin by introducing The problem we are solving with some contextual information. Similarly, we have also presented the Aims, benefits, and motivation for this research project. We have also given a brief explanation of the project timeline. In the case of the literature survey, We have assessed multiple studies for insurance fraud detection and income prediction and presented some similarities and patterns with our analysis. We have also explained algorithm-based literature where we have compared different algorithms used in different studies mentioned in the literature survey. In the end, We also explained the applications, requirements, and limitations of our research.

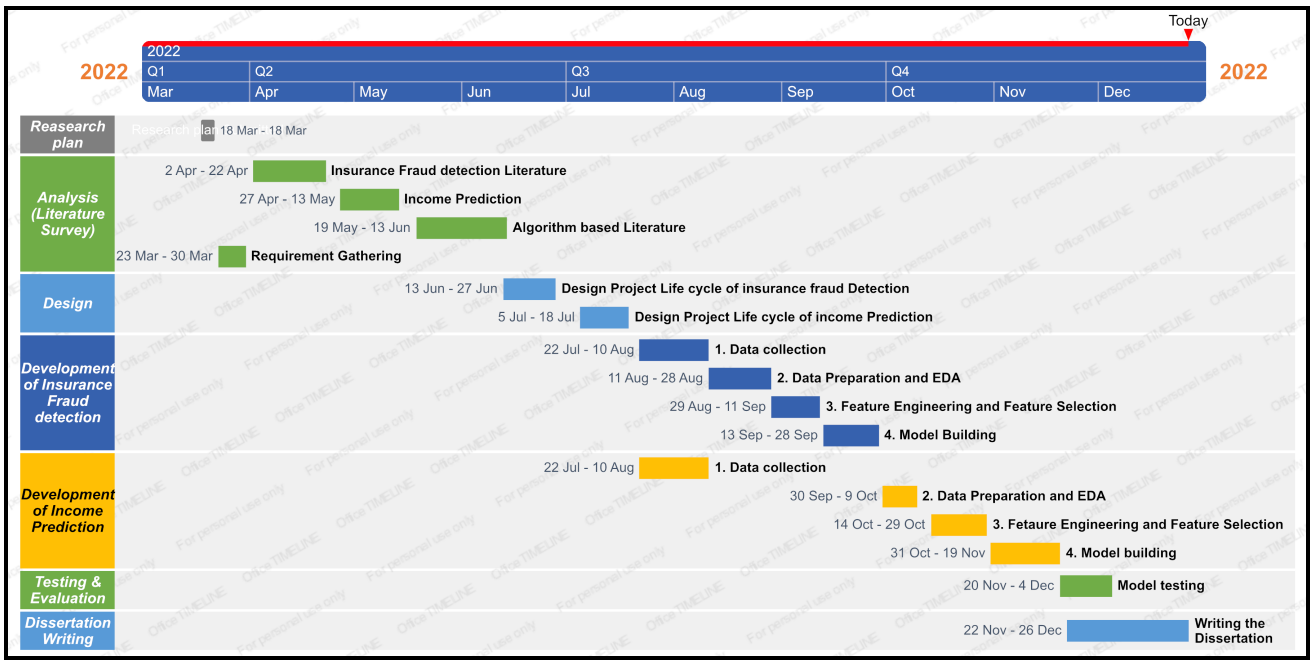
In the Methodology section, we have explained the methodology used in both phases of the project. In the Exploratory data analysis section, we have added some figures explaining the insights we got from our analysis. We have tried to make sense of the data with the help of facts and research. In the model Development phase, we compared the model performance of both phases. In the Evaluation section, we have given extensive information about how we evaluated the model. What was the main metric of score we used? how successfully we achieved the aims of our research and also talked about the limitations and ethical implications. We also discussed the practical approach of our research project. Concluding the research, we compared the performance of all the models with different encoding techniques. Again, we have tried to compare the results with the literature survey we assessed. I also suggested my opinion regarding the future work of this research project.

4.7 Timeline Of Project



We have two figure which shows the project timelines. The first figure shows the planned timeline while the second figure shows the followed timeline. A project timeline is a breakdown of each project's commitments sorted by date. This helps to move along well by outlining what has to be done before starting a new activity. A Gantt chart is used in this research's scheduling to highlight the timeline, monitoring, but also end date. Whenever it concerns managing and monitoring project timelines, the Gantt chart is among the best tools, outperforming others such as PERT (Program Evaluation and Review Technique) as well as the CPM (Critical Path Method) (Geraldi, 2012).

This same Gantt chart representing this research timeline with key stages is shown in Fig.1 below. A literature review is done to analyze previous systems proposed by various researchers. In the design phase, first, we created a design specification to understand exactly what are the objectives of this research and decide on which machine learning algorithms to be implemented.



In the Development phase, there was a plan to use a single dataset but that dataset contained the values which could be simulated therefore we started searching for two new datasets. Therefore there was a delay of a few days in the actual Gantt chart than the planned Gantt chart. In the Testing Phase, We compared the machine learning models based on accuracy rate. We got a 14-accuracy rate for each phase.

5 Literature Survey

Researchers and practitioners have long been interested in insurance fraud detection and finance prediction. Numerous important research has been conducted regarding insurance fraud detection and income prediction. The following works are most closely linked to this proposed Research.

5.1 Literature Survey Related to Insurance fraud Detection

5.1.1 Insurance Claim Analysis Using Machine Learning Algorithms

Mrs. Rama devi burri and other Co-authors have presented several objectives of this research and decided on which machine learning algorithms to implement into the area of insurance fraud detection Automated and personalized product offerings, Improved Risk Assessment, and Enhanced Fraud detection. Mrs. Rama devi Burri also specified different resistances for adapting machine learning to their challenges which is the most important information that we've found in our research area. Another thing to notice in Mrs. Rama devi's research is that, with the accuracy of the algorithm, they have got more than 99 percent accuracy with 4 algorithms. Notably, It is quite possible to get 99 percent accuracy with the machine learning models but there was no mention of model overfitting in the research paper which will give us more insight into the accuracy of machine learning models. Therefore we decided not to follow the techniques mentioned in the research paper but the information related to use cases and limitations related to the research area of insurance fraud detection is quite impressive therefore we have only referred to this specific study for information and not the techniques[2].

5.1.2 Fraud Detection and Frequent Pattern Matching in Insurance claims using Data Mining Techniques

Mr. Pinak Patel and other co-authors have proposed a fraud detection framework for the healthcare industry. They have categorized the fraud based on the period of the claim, anomalies disease-based anomalies. Their framework is based on real-world medical data. Mr. Pinak Patel also carried out impressive literature research regarding their area of research. They have not only assessed the information related to data mining in health science but also about how efficient naïve byes technique is in identifying security bugs. Pinak Patel also successfully mentioned the dataset they are dealing with. They have provided all the information related to the CMs dataset and an important thing to remember is that Mr. Pinak Patel has used the One-Hot encoding technique for categorical variables. One of the most important things which are why we decided to add the research of Mr. Pinak Patel in the knowledge-based literature is because they did not mention any results they got through the machine learning models. There is quite unclear information about the results Mr. Pinak Patel got through his research [12].

5.1.3 Extreme Gradient Boosting Machine Learning Algorithm For Safe Auto Insurance Operations

Mr. Najmeddine Dhieb and other co-authors evaluated the performance of the XGBoost algorithm for detecting and classifying different types of car insurance-related fraudulent claims. They have proposed a mathematical approach to solve the problem. They have largely focused on Extreme gradient-boosting algorithms. They compared the results of the Extreme gradient boosting algorithm to other state-of-the-art solutions such as naïve byes, nearest neighbor, and decision tree algorithms. Mr. Najmeddine dhieb explained all the details about XGBoost very well, but they compared the results directly with the already executed results of other state-of-the-art solutions[4].

5.1.4 Analysis from Above Studies :

- Mrs. Rama Devi Burri's research does not mention anything about overfitting to clarify the accuracy of machine learning models because 4 models out of 6 got an accuracy of more than 99 percent. Undoubtedly, there is a possibility that machine learning models can achieve more than 99 percent accuracy but rama devi Burri also failed to mention briefly the nature of the dataset. Therefore we decided not to follow the techniques used by Rama Devi burri in her specific research.
- Mr. Pinak Patel's research is based on fraud detection in the healthcare industry. We have referred to his study to get an idea about how frauds are detected in other industries.
- Mr. Pinak Patel has mentioned the use of One-Hot-Encoding for converting numerical variables into categorical variables. It is a good idea to use this concept in our research.
- Given the fact, Mr. Pinak Patel has not mentioned proper details about the execution of his project such as the accuracy Score of all the machine learning models but the literature survey he conducted is worth looking at because of the methodology.
- XGboost Algorithm is getting maximum accuracy over the other machine learning algorithm. We will make sure to use XGboost Algorithm to check the hypothesis presented by Mr. Najmeddin Dhieb.
- Similarly, One more important question arises is that, Will XGBoost perform this well? if we were to use it with different encoding techniques.

5.2 Literature survey related to Income Prediction :

5.2.1 Prediction of Individual Level Income: A Machine Learning Approach

Mr. Michael Matkowski and other co-authors have compared the performance of several machine learning algorithms with the traditional OLS regression model. They have found that as compared to other machine learning models The traditional OLS model with a variable selection from literature resulted in the lowest R2 values and highest MAE (Mean Absolute Error) and MSE(Mean Squared Error) was the worse performing model. Mr. Michael Matkowski analyzed the results by R-Square and MSE (Mean Squared Error). Mr. Michael provided all the relevant details about all of the algorithms they have used in their respective research. The most important algorithms that we found might be

relevant and useful to our research studies are Gradient boost(XGboost), Random Forest, and Logistic Regression concerning the nature of the dataset[11].

5.2.2 Intelligent Income prediction model

Mr tarun and other co-authors compared 3 different mathematical machine learning algorithms to decide the best algorithm to predict income. They have used a naïve byes classifier to use probability-based approach, a Support Vector Machine algorithm to follow a vector algebra-based approach, and finally, a random forest classifier to follow a decision tree-based approach. One thing to keep in mind is that Mr. Tarun has considered a large dataset for the research while we are using a dataset with thousand entries. Mr. Tarun got maximum accuracy for the random forest classifier while Naïve byes performed worst out of all three[9].

5.2.3 Analysis of Income Prediction Literature :

- By the information provided by Mr.Michael, it is quite clear that traditional OLS models do not perform well as compared to other Machine learning models. Therefore it might not be a good idea to use the traditional OLS model for our research.
- Naïve byes algorithm performed worst in predicting the income.
- One thing is crucial a Support vector machine takes more training time than a random forest classifier.

5.3 Algorithm-Based Research :

While progressing with this research, One would ask "which machine learning model you will be using and why ?". To answer this question we would like to point out some considerate factors that would help us decide which machine learning model to use.

- Performance: The quality of the Model's result is an important factor to look into. One should prioritize the algorithm which might give us higher accuracy. However, a Single metric does not work in every single situation. For example, the Accuracy score is not an important metric when dealing with imbalanced data.
- Explainability: Explaining the results of the model is important in some situations. For example, Linear regression and random forest are really easy to explain but it's not the case with neural networks.
- Dataset Size: Dataset Size is one of the most important factors when choosing the machine learning algorithm. Dataset is the main reason why most of the research we assessed got different results from one another.
- Training time and Cost: Dataset we are dealing with in our research is a small dataset but the methodology in our research aims to solve big data problems, therefore, training time and cost play a crucial role.

In the case of insurance fraud Detection and Income prediction, We found that out of all the research-based studies, several machine learning algorithms were quite similar. Random Forest, Logistic Regression, XGBoost, Support Vector Machine, etc. Notably, in most of the cases, random forest and XGBoost algorithm provided the best results while Naïve byes was a total failure in Mr.Tarun Kumar's Research of Intelligent income prediction. On the other

hand, Mr. Michael also compared the performance of machine learning algorithms with the traditional OLS model but the performance of machine learning models are far more efficient as compared to the traditional OLS model. Therefore we decided not to follow the traditional OLS model for our research studies. Mr. Najmeddine Dhieb conducted thorough research on the XGBoost algorithm and found that it is the most efficient algorithm if we are detecting fraudulent insurance claims. To check this hypothesis we will make use of the XGBoost algorithm in insurance fraud detection and income prediction as well. Similarly, the choice of the algorithm also depends on the type of problem we are dealing with.

5.4 Application of insurance fraud detection and income prediction

- First part of Our research focuses on the detection of fraudulent insurance claims in the automobile industry but our research is also useful for detecting healthcare, and property insurance frauds as well.
- Credit card fraud is also a big problem the banking sector is facing nowadays. Our research aims to detect the anomalies in data, therefore, our research can be helpful in providing insight into tackling credit card frauds as well.
- In the second phase, We are predicting the income of customers. This part of the research can be useful to the banking industry to target customers based on their income.
- Our research uses Predictive analytics and it plays a crucial role in today's world. It can be used to predict the economy of the country based on certain features.
- We've assessed several research works to get an insight into how both industries (insurance and finance) work. We have carried out extensive research to detect fraud and predict income in a single research project. Our studies will provide a strong foundation for the collaboration of the insurance and finance industry in the future.
- The data we are dealing with is originated from an insurance company which is based in the USA. There is a chance that our research might return good results if they were to use our methodology in another country that has a similar set of rules as the USA.

5.5 Limitations

- As Artificial intelligence is making a stronger hold on market, fraudsters also upgrading the frauds they are making. Such frauds can slip through the gaps unnoticed.
- data unavailability is one of the biggest limitations of our research project. Insurance companies can collect data from customers which is essential for insurance fraud detection and income prediction but at an academic level, it is not possible to collect personal information which can be useful in both phases of the project.
- Every Country has a different set of insurance rules. The point to remember is that it is possible that our research is quite relevant to the insurance rules of the USA insurance market and it might be a big success in the USA but there is no guarantee that our insurance model will return the same results in any other country with a different set of insurance rules.
- It takes some amount of time to notice a new type of fraud. During that time, all the cases will go through the radar unnoticed.

- Authenticity of the data we are using is another concern. Simulated data can present imbalanced data and the randomness of the data might be the concern.
- Income data is one of the personal assets we are predicting. Although it might not be 100 percent accurate, it will definitely give companies a rough idea about the annual income of the customer. This brings a limitation as if this data is leaked, then it can cause a big loss for the customer and the company as well. Therefore data protection is one of the limitations and needs to be mentioned.

6 Problem Description :

The goal of this Research project is to build a machine-learning model that can detect auto insurance fraud. The main challenge behind fraudulent claim detection in machine learning is that fraud claims are more common as compared to legit insurance claims. Building a machine learning model that can detect fraudulent insurance claims is quite complex, given the variety of fraud patterns and a relatively small number of known frauds in a given sample. Insurance frauds cover a range of illegal actions which customers might commit to achieving monetary sums from insurance providers by staging the incident, falsely representing the situation, including the actors. Many of these cases go unnoticed by insurance companies. Most Companies in the USA use prediction analytics for price prediction, but there are more ways in which machine learning can help.
Our research will make use of machine learning algorithms and help detect insurance frauds and predict the income of customers independently. Through extensive research, we have found that many people have worked on insurance fraud detection, but no one has used more than five different machine learning models. Notably, we have found one more interesting thing is that no one has used different encoding techniques to convert the numerical variable into a categorical variable. From the literature review, We can see that only Mr. Pinak Patel has tried to use one -hot-encoding and managed to obtain good performance from the machine learning model. After extensive research, we have seen that other researchers have only worked on insurance fraud detection but we are providing more advantages to insurance companies by predicting the income of customers. Income prediction in the insurance industry is still uncommon.
There are several ways we could have proceeded with the research, such as a mathematical model or traditional OLS model, but we have decided to opt for python programming for developing a machine learning model. We also could have made use of R programming language but using python is efficient for us as there are already predefined libraries that might be useful for us. We have assessed several research papers mentioned above to get an idea about project implementation and project flow.

6.1 Requirements:

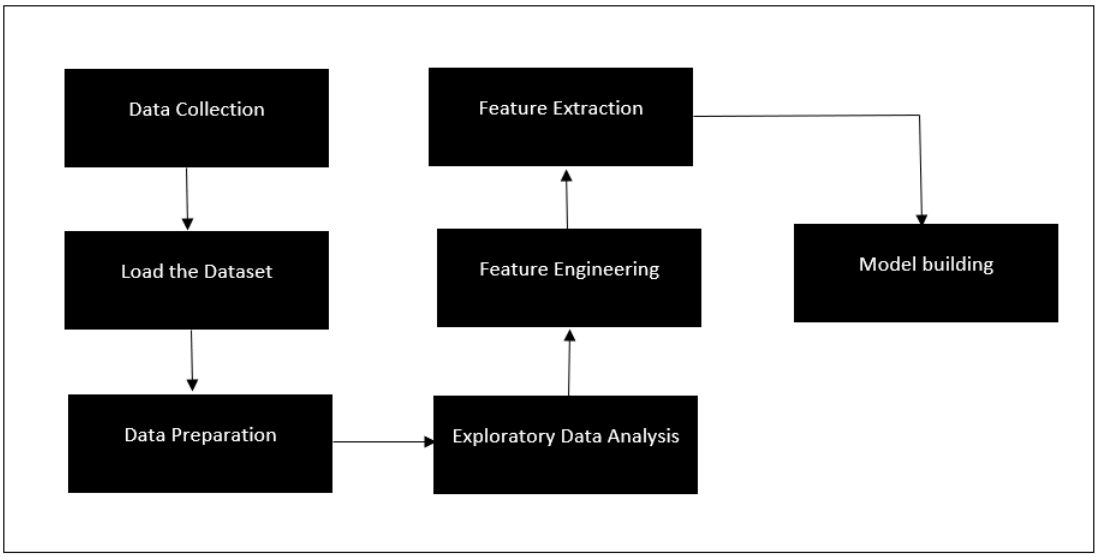
In the given research, we have decided to use the python programming language. We have used Jupyter notebook as a code editor but it can be executable in google collaboratory as well. We will load the data, and then it will split into training and testing datasets. We will train the model using the training dataset and the results will be obtained by the performance of the testing dataset. The following requirements are valid for both insurance fraud detection and income prediction.

The final requirements of the proposed system are as follows :

- Dataset must be read in a single CSV file and converted into a format that can be used by machine learning libraries.

- Removal of any feature from the dataset should not be a problem at any given time
- After Data processing, the environment must be able to use the available data to train the model and test its performance of it.
- Each machine learning model must be tested independently.
- Accuracy score of each machine learning model should be considered as a primary performance metric.

7 Theory Methodology



We have used the above methodology to execute insurance fraud detection and income prediction independently. We started with the data collection and loaded the data into the jupyter notebook. After, we performed data preparation operations by checking null values and unknown values. Similarly, we performed Feature engineering and Feature Extraction to improve the accuracy of the model and in the end, we executed a machine learning algorithm for model building.

7.1 Insurance Fraud detection

7.1.1 Data Collection

We have downloaded our insurance fraud detection dataset from the Kaggle website. Kaggle is an online community platform for data scientists and machine learning enthusiasts. We have added all the information about the missing values and unknown values in the data preparation part which will be helpful to know the nature of the dataset.

- months_as_customer: It denotes the number of months for which the customer is associated with the insurance company.
- age: continuous. It denotes the age of the person.
- policy_number: The policy number.

- policy_bind_date: Start date of the policy.
- policy_state: The state where the policy is registered.
- policy_csl : combined single limits. How much of the bodily injury will be covered from the total damage.
- policy _deductible: The amount paid out of pocket by the policyholder before an insurance provider will pay any expenses.
- policy_annual_premium: The yearly premium for the policy.
- umbrella_limit: An umbrella insurance policy is extra liability insurance coverage that goes beyond the limits of the insured's homeowners, auto, or watercraft insurance. It provides an additional layer of security to those who are at risk of being sued for damages to other people's property or injuries caused to others in an accident.
- insured_zip: The zip code where the policy is registered.
- insured_sex: It denotes the person's gender.
- insured_education_level: The highest educational qualification of the policyholder.
- insured_occupation: The occupation of the policyholder.
- insured_hobbies: The hobbies of the policyholder.
- insured_relationship: Dependents on the policy-holder.
- capital_gain: It denotes the monitory gains by the person.
- capital_loss: It denotes the monitory loss by the person.
- incident_date: The date when the incident happened.
- incident_type: The type of the incident. collision _type: The type of collision that took place.
- incident_severity: The severity of the incident.
- authorities_contacted: Which authority was contacted.
- incident_state: The state in which the incident took place.
- incident_city: The city in which the incident took place.
- incident_location: The street in which the incident took place.
- incident_hour_of_the_day: The time of the day when the incident took place.
- property_damage: If any property damage was done.
- bodily_injuries: Number of bodily injuries.
- Witnesses: Number of witnesses present.
- police_report_available: Is the police report available?
- total_claim_amount: Total amount claimed by the customer.

- injury_claim: Amount claimed for injury
- property_claim: Amount claimed for property damage.
- vehicle_claim: Amount claimed for vehicle damage.
- auto_make: The manufacturer of the vehicle
- auto_model: The model of the vehicle.
- auto_year: The year of manufacture of the vehicle.

Our Target Column is Fraud_Reported

- **Fraud_reported**: Fraud detected Y or N.

7.1.2 Load the Data

From the data collection process, we got the dataset in CSV format. Therefore we have used the pandas library from python to load the dataset into jupyter notebook. We have also used df.describe() to check normal parameters such as count, mean, maximum value, minimum value, standard deviation, 25%, 50%, and 75% values from the data.

7.1.3 Data Preparation

Data preparation is an important step and some researchers spend about 80% of the time. If data preparation is not done properly then the machine learning model may not provide good accuracy, and we need to start all over again from the data preparation stage.
Regarding our research project, we have decided to check the datatype null values of each column to get an idea about which type of data we are dealing with. Furthermore, we have also checked for several unique values in our dataset so that we can remove those features from the dataset that are not useful in knowledge extraction. As a result, we decided to drop 8 features from the dataset because they were not contributing to the knowledge extraction. The name of those features was 'policy_number', 'policy_csl', 'policy_bind ', 'incident_location', 'insured_hobbies', 'incident_date', 'incident_city', and 'auto_model'. We have performed a drop operation on the dataset to remove all these columns.
In Contrast to our research project, We checked for any null values in the dataset, but there were no null values or missing values in any of the columns. We also checked for different values in a different column and we came across crucial information that there are many '?' present in the 'Collision_type', 'property_damage', and 'police_report_available' columns. It can't be considered as one of these three. The value is not missing but it's unknown. Therefore we decided it's best to replace '?' with 'not known' in the given three columns. So here is the output before and after replacing '?' with 'not known'.

7.1.4 Exploratory data Analysis

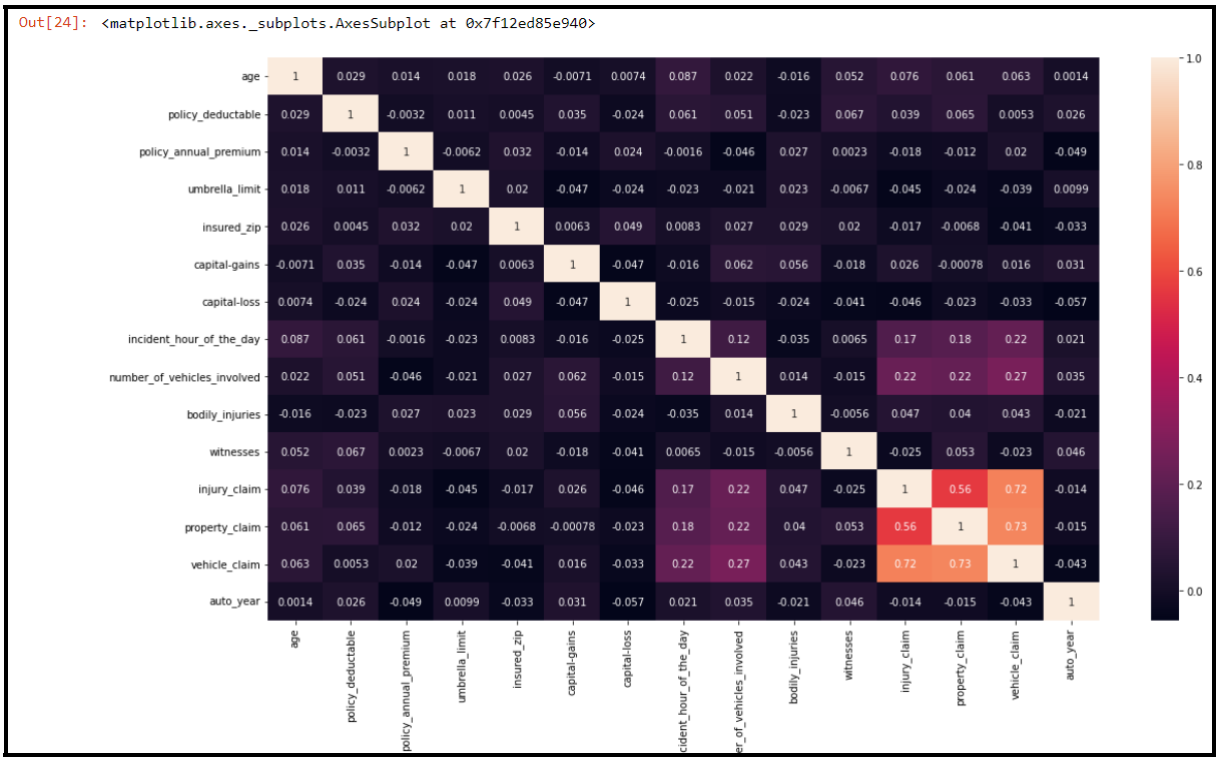
Exploratory data analysis is a stage where we will try to extract meaningful information from the data. Exploratory data analysis is a crucial stage where we analyze the data to find its characteristics of it. It is impossible to look at a column of data and discover its important characteristics. It also provides the context needed to develop an appropriate model for the problem at hand and to correctly interpret its results[15].

We have used pandas Profiling to get an idea about each column of the dataset.The Primary goal of pandas profiling is to support consistent and fast solution experience by providing one-line data exploratory analysis. Pandas profiling automatically detects the datatype of columns. It also provides a summary of problems in the data in the form of warnings. It also provides informative visualization in the form of histograms. We call it univariate analysis. Similarly, It also provides interactive information on correlation, missing data, duplicate rows, etc. it is also called bivariate Analysis. We have used pandas profiling to derive crucial insights from the data.

"Correlation is an analysis of the co-variation between two or more variables"—(A.M Tuttle)

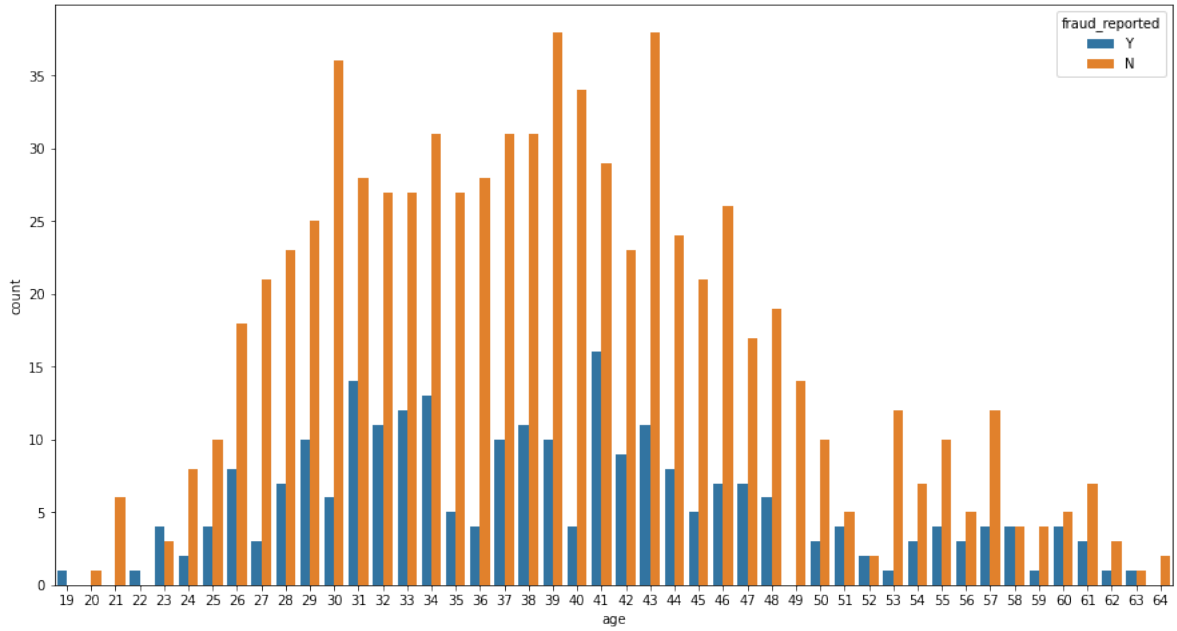
Correlation is used to find the relationship between two variables. Finding a relationship between two variables is really important because it helps us predict the value of one variable which is correlated to another. There are basically two types of correlation. When two variables move in the same direction, such as when there is an increase in one variable results in an increase in another variable. Also, if a decrease in one variable results in a decrease in another variable, then both variables are positively correlated. In Contrast, from our dataset, the months_as_age column and age column are positively correlated. When the values of two variables move in opposite directions,then those two variables are negatively correlated. There are no negative correlated values in our dataset. We have used Heatmap correlation to find out if two variables are correlated to each other or not. As seen in the figure below, when the correlation is +1.0 that means variables are positively correlated and when the correlation is -1.00 it means two variables are negatively correlated.

From the analysis, we have got following Correlation :
Months_as_customer is highly correlated to age
Total_claim_amount is highly correlated with injury_claim
Total_claim_amount is highly correlated with Property_claim
Total_claim_amount is highly correlated with Vehicle_claim



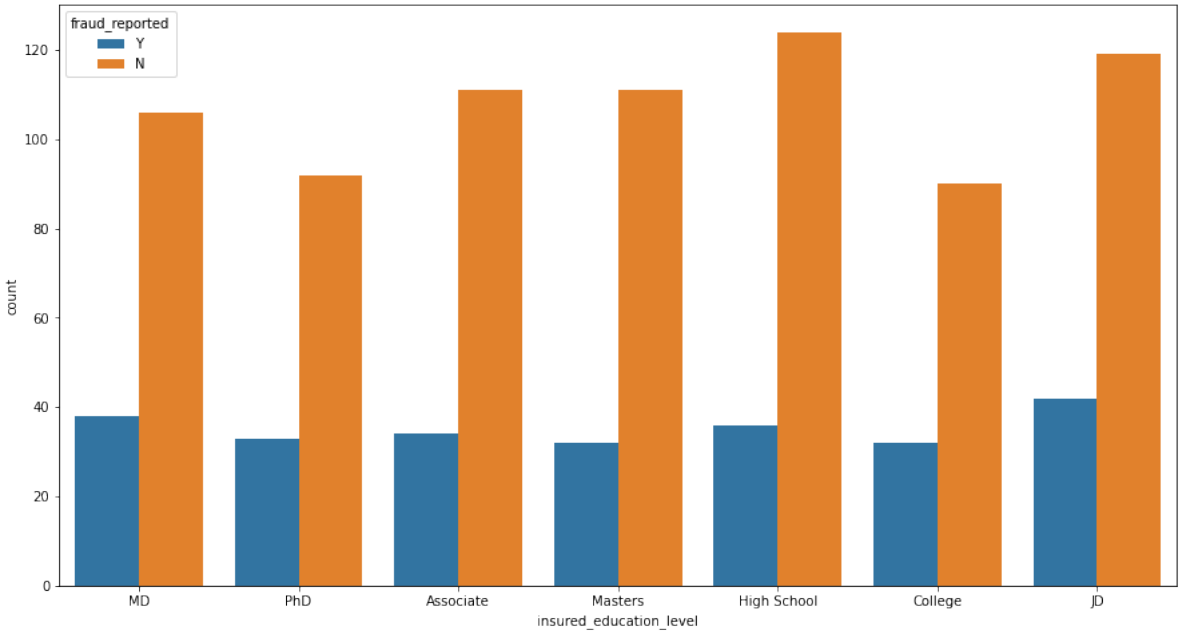
Hypothesis Based on EDA :

1. Is the age of a person have any prominent relationship with insurance fraud ?



We have plotted a graph with age and fraud distribution to obtain any crucial insight from the data. we have a data available from age group 19 till age group 64. From the graph, it is quite clear that age group 26-48 have reported maximum frauds. Customer aged 40 have reported a maximum number of frauds while customers aged 20,21, and 64 have not reported any frauds.

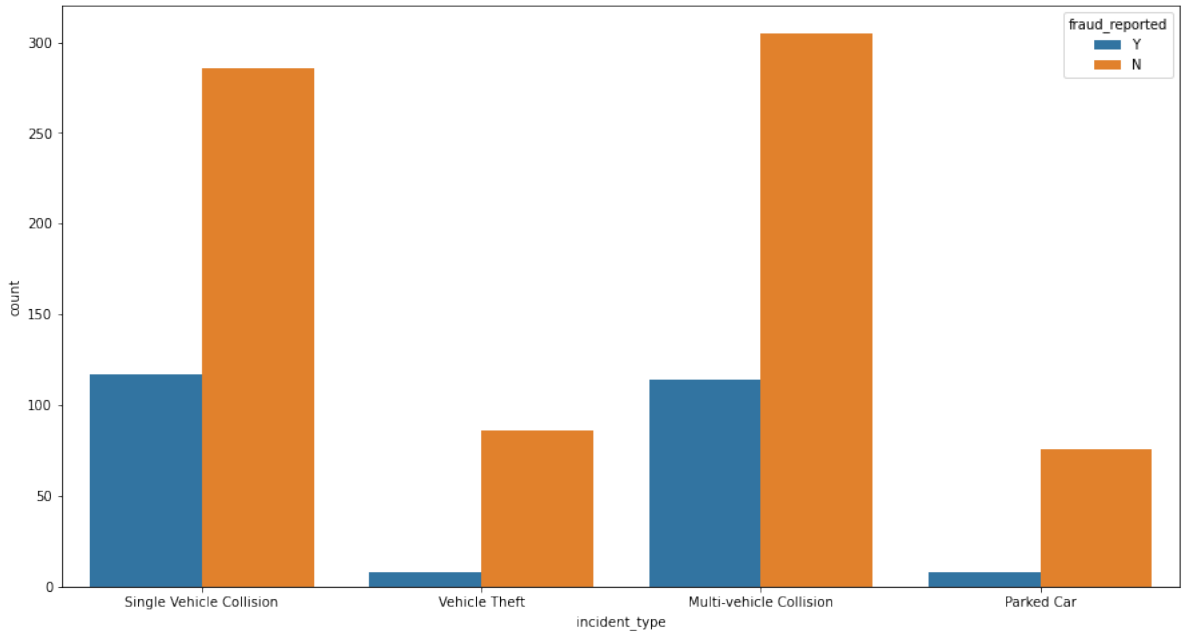
2.Did highly educated policyholder are involved in fraud on a high scale?



We have used a dataset from the USA insurance company which contains the education level of customers ranging from high school, college, master, Ph.D., Associate, MD, and JD.

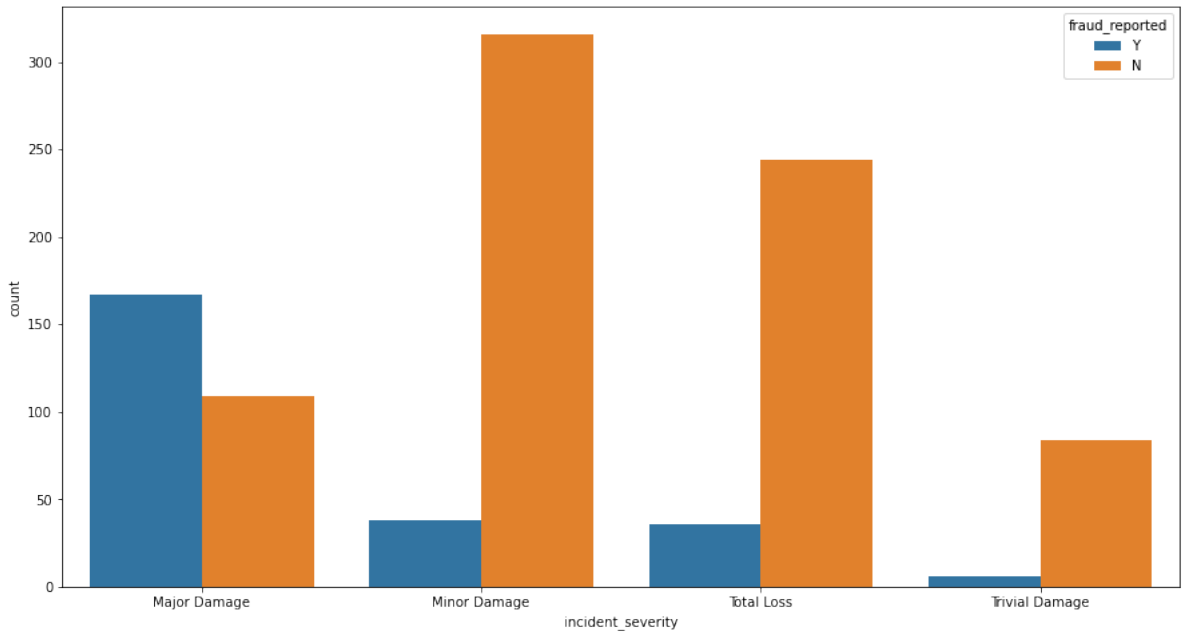
out of all these, JD (Juris Doctor) has reported the most fraud. Notably, the least fraud is reported by college and master's degree holders.
From the figure, we can see that most claims which are reported are associated with customers who have completed juris doctor degrees. It is an undeniable fact that Juris doctors are good law practitioners. It might be a possibility that the juris doctors are involved in fraudulent activities.

3. which incident type has more probability of fraud ?



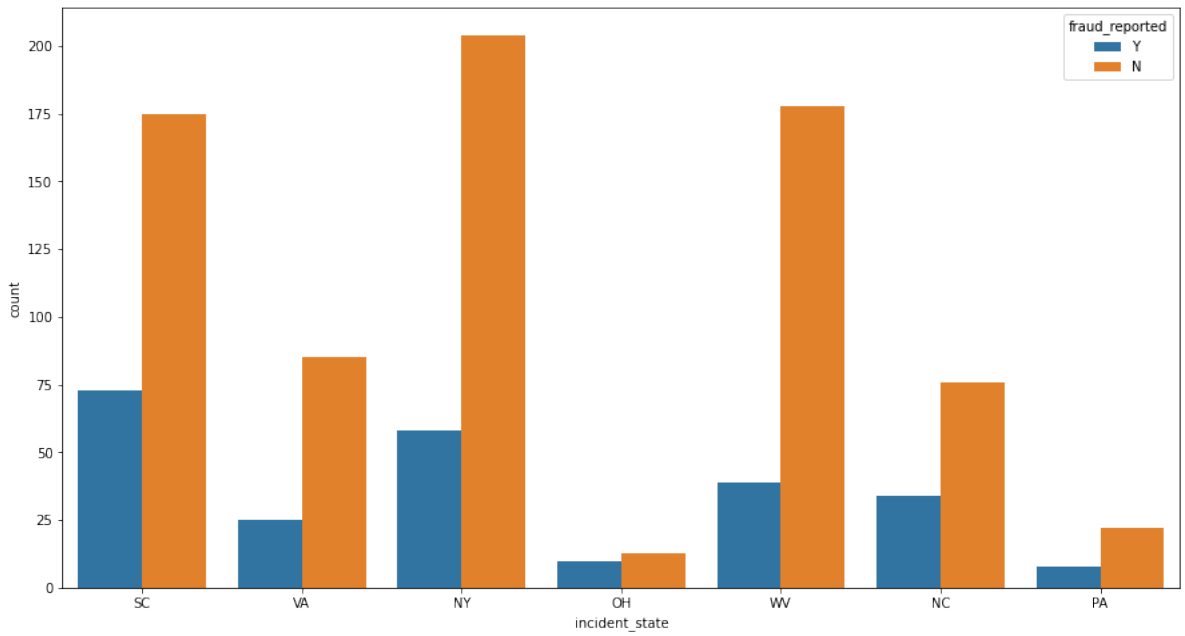
Our dataset includes four types of incident types, namely single-vehicle collision, Vehicle Theft, Multi-vehicle Collision, and Parked Car., As we can see from the figure, Most frauds that are reported are associated with single-vehicle collisions and Multi-vehicle collisions.
Single-vehicle collisions could include cases like banging on a tree or any other object. Cases like these have a high chance of being fraudulent because these types of accidents could be planned. On the other hand, Multi-Vehicle collisions have a low chance of being fraudulent because cases like this could include witnesses and lots of other things which are really hard to manage for fraudsters, but there are some cases of multi-vehicle collisions being fraudulent as well.

4. which incident_severity have more impact on fraud?



Incident Severity refers to a type of loss or damage. How severe the accident can depends on the loss and the damage of the car. We don't have a data here on personal injury. As we can say, There are lots of cases where a car has suffered major damage. The main aim of fraudsters is to make maximum profit out of the pockets of insurance companies through frauds. Therefore it is quite clear- in order to do so, fraudsters may plan the incident in such a way that it will make them more profitable. In case of Major damage, fraudsters may get a monetary sum or will be given a new/used car by the insurance company. Therefore we can see the most number of major damage cases in the dataset. We have already talked about the authenticity of the data. This insight may point out the fact that this data is genuine.

5. which state incidents leads to fraud?

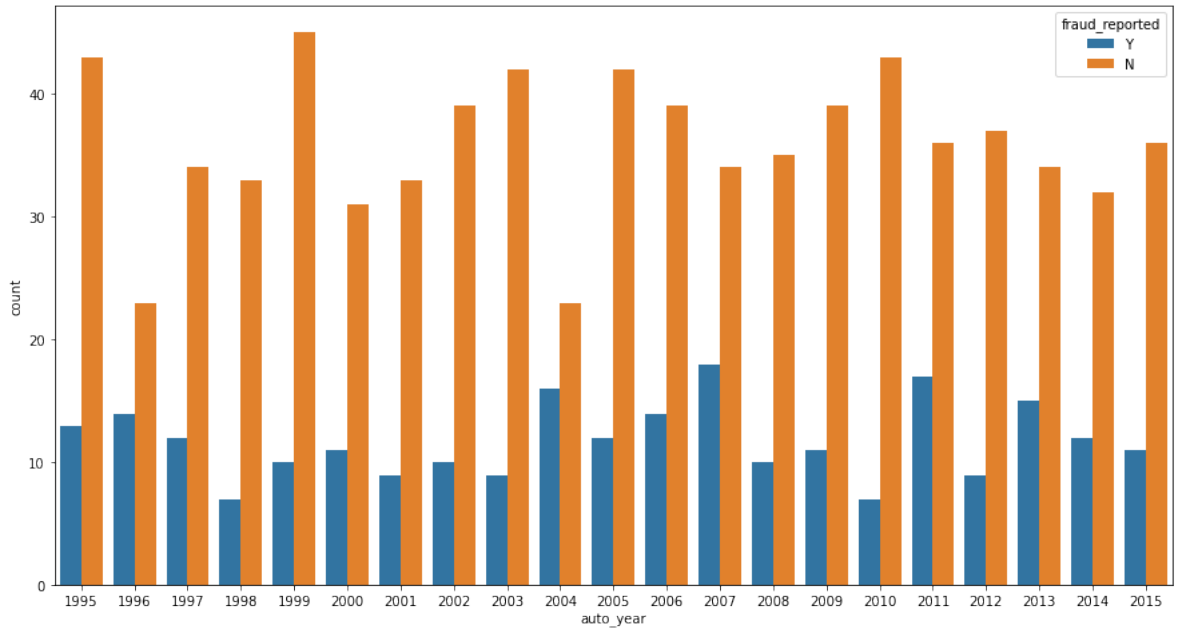


Our dataset contains following states from USA in our Dataset :
SC = South Carolina
VA = Virginia

NY = New York
OH = Ohio
WV = West Virginia
NC = North Carolina
PA = Pennsylvania

South Carolina is a state where most of the fraudulent cases are reported. New York is second on the list of most fraudulent cases. Ohio has reported the least number of fraud cases. But if we take a closer look at the ratio of Fraudulent vs, Non-Fraudulent cases, we can see that South Carolina and North Carolina top the chart, while New York has registered fewer fraudulent claims.
South Carolina has the most fraudulent claims reported. The main reason behind that is the law in South Carolina might unintentionally favor the fraudsters, and it's a bit soft as compared to other states of the USA. We have analyzed one research study from Jeffrey J. Wiseman and Joseph W. Rohe, which specifically pointed out the jurisdiction loopholes for fraudsters. Mr. Wiseman pointed out one law under the name of "burden of proof"." In contrast, fraud must be proven by the elevated burden of clear, cogent, and convincing evidence. See Brown v. Stewart, 348 S.C. 33, 42, 557 S.E.2d 676, 680 (Ct. App. 2001). Similarly, there are several significant evidentiary issues that may arise in a fraud case. The issue of whether the insured was criminally prosecuted is often raised. There are so many serious issues that are affecting insurance companies to take down fraudsters. Therefore it explains the reason why the dataset has so many fraudulent claims from South Carolina [21].

6. does older auto/vehicle leads to fraud ?



From the above observation, It's quite difficult to say that customers with old cars are often involved in fraudulent activities. Distribution is uniform, and there is no fixed pattern that will support this theory. Our dataset Represents the data from the years 1995-2005 (a 20-year period). The maximum number of fraud cases are registered by customers whose car model belongs to the year 2007, followed by the year 2011. Relatively, Customers with car models from 1995-2003 have reported fewer fraudulent cases.

7.1.5 Feature Engineering

Feature engineering is a machine learning technique that holds data to create new variables which are not in the training set. It can also generate new features for both supervised and unsupervised learning. The goal of feature Engineering is to simplify and speed up data transformations and also enhance model accuracy. Feature Engineering is important when working with machine learning models. Regardless of data and architecture, the terrible feature will affect the accuracy of the model[6].

Firstly, let's handle the row values. We will try to convert all the object type values in categorical type by using 0 and 1.

- We have one column named as fraud_reported. fraud_reported is our target column in string format such as "Yes", and "No" we will convert it to numeric indications "Yes" = 1 and "No"= 0.
- insured_sex also has a gender column in a format such as "male" and "female" we will convert that to numeric, indicating "male"=1 and "female"= 0.
- education_level and incident_severity is replaced with Ordinal Encoding such as MD =4, phD=5, Associate =3, masters =4, High-School = 2, College =3 , JD = 1.
- We will also be replacing values in incident_severity if trivial_damage =1, minor_damage =2, Major_Damage =3, total_loss = 4.

What is the problem with categorical data?

Some algorithms can learn directly from categorical data. A decision tree is one of those algorithms that can be trained with categorical data directly with no data transformation needed but there are some machine learning algorithms that cannot be trained with categorical data. They need numeric data in input and output. In general, this is mostly a constraint of the efficient implementation of machine learning algorithms rather than hard limitations on the algorithms themselves.

We have Used Two Techniques For Data Encoding :
1. One-hot Encoding
2. Weight of Evidence

1. One-Hot Encoding

The input to this transformer should be an array-like of integers or strings, denoting the values taken on by categorical (discrete) features. The features are encoded using a one-hot (aka 'one-of-K' or 'dummy') encoding scheme. This creates a binary column for each category and returns a sparse matrix or dense array (depending on the sparse_output parameter). By default, the encoder derives the categories based on the unique values in each feature[7]. Alternatively, you can also specify the categories manually. For example, there are three colors blue, green, yellow, and red. And if the customer's car is of red color, then one-hot-encoder will display this as follows

Green	red	yellow
0	1	0

Applying One hot Encoding to our dataset, this is the output we have got.

```
In [34]: # Applying one hot encoding using get_dummies method
         one_hot_df = pd.get_dummies(processed_df, drop_first=True)
         one_hot_df.head(10)
```

Out[34]:

d_relationship_not-in-family	insured_relationship_other-relative	insured_relationship_own-child	insured_relationship_unmarried	insured_relationship_wife	incident_type_Parked Car	inciden Veh
0	0	0	0	0	0	
0	1	0	0	0	0	
0	0	1	0	0	0	
0	0	0	1	0	0	
0	0	0	1	0	0	
0	0	0	1	0	0	
0	0	0	0	0	0	
0	0	0	1	0	0	
0	0	1	0	0	0	
0	0	0	0	1	0	

2. Weight of Evidence :

The weight of evidence tells the predictive power of an independent variable in relation to the dependent variable. Weight of Evidence (WOE) helps to transform a continuous independent variable into a set of groups or bins based on the similarity of dependent variable distribution, i.e., number of events and non-events. We have applied the Weight of Evidence, and this is the output we got[3].


```
In [34]: # Applying one hot encoding using get_dummies method
         one_hot_df = pd.get_dummies(processed_df, drop_first=True)
         one_hot_df.head(10)
```

Out[34]:

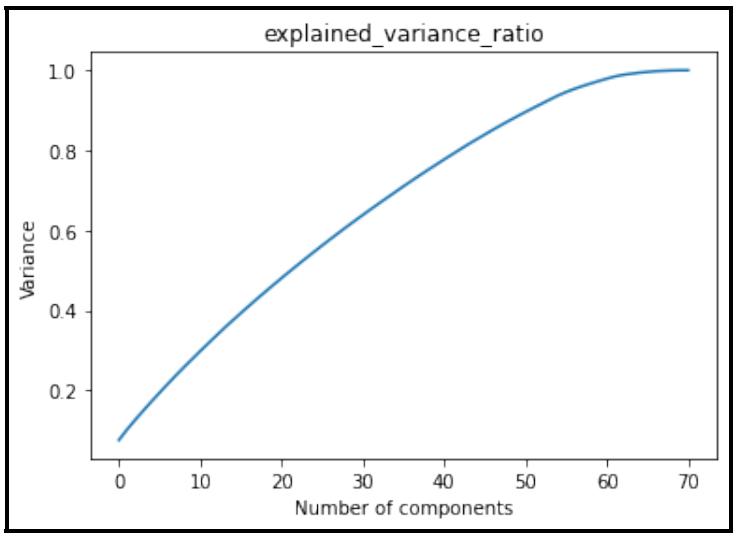
d_relationship_not-in-family	insured_relationship_other-relative	insured_relationship_own-child	insured_relationship_unmarried	insured_relationship_wife	incident_type_Parked Car	inciden Veh
0	0	0	0	0	0	
0	1	0	0	0	0	
0	0	1	0	0	0	
0	0	0	1	0	0	
0	0	0	1	0	0	
0	0	0	1	0	0	
0	0	0	0	0	0	
0	0	0	1	0	0	
0	0	1	0	0	0	
0	0	0	0	1	0	

7.1.6 feature Extraction :

Principle component analysis :

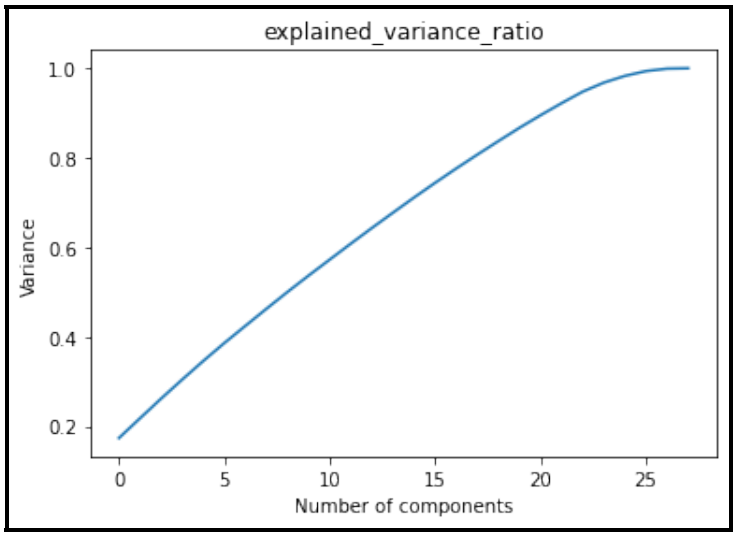
Principal component analysis, or PCA, is a statistical procedure that allows you to summarize the information contained in large data tables by means of a smaller set of "summary indices" that can be more easily visualized and analyzed. We have used 2 Encoding techniques therefore, we've got PCA for One hot encoding and the weight of evidence method.

PCA for One hot Encoding :



We can see that around 99% of the variance is explained by 66 components. So instead of giving all columns as input in our algorithm, let's use these principle components.

PCA for Weight of Evidence :



We can see that around 99% of variance is being explained by 25 components. So instead of giving all columns as input in our algorithm, let's use these principle components.

7.2 Income Prediction

Some of the stages of Insurance fraud prediction methodology are similar to Income prediction because in terms of our research project, income prediction is also a classification problem. Therefore we will be using same machine learning models which are used in insurance fraud detection.

7.2.1 Data Collection :

We have Usedthe Kaggle website to download this dataset. This dataset contains 10 columns and 9000 rows.

Information about the features of dataset :

- Work: Explains the type of work such as private, self-employed etc.
- Marital Status : Explains if customer's marital status.
- Hours_Week: Number of hours customers work per week
- Country: Country of Citizenship of Customer.
- Above 50K: If its salary above 50k or not.
- Education: Education level of Customer.
- Gender: Gender of Customer.
- Occupation: Occupation of Customer.
- Age: Explains the Age of Customer.

7.2.2 Loading the data :

We have a dataset in CSV format. Therefore we will be making use of pandas library to load the data into the jupyter notebook Environment.

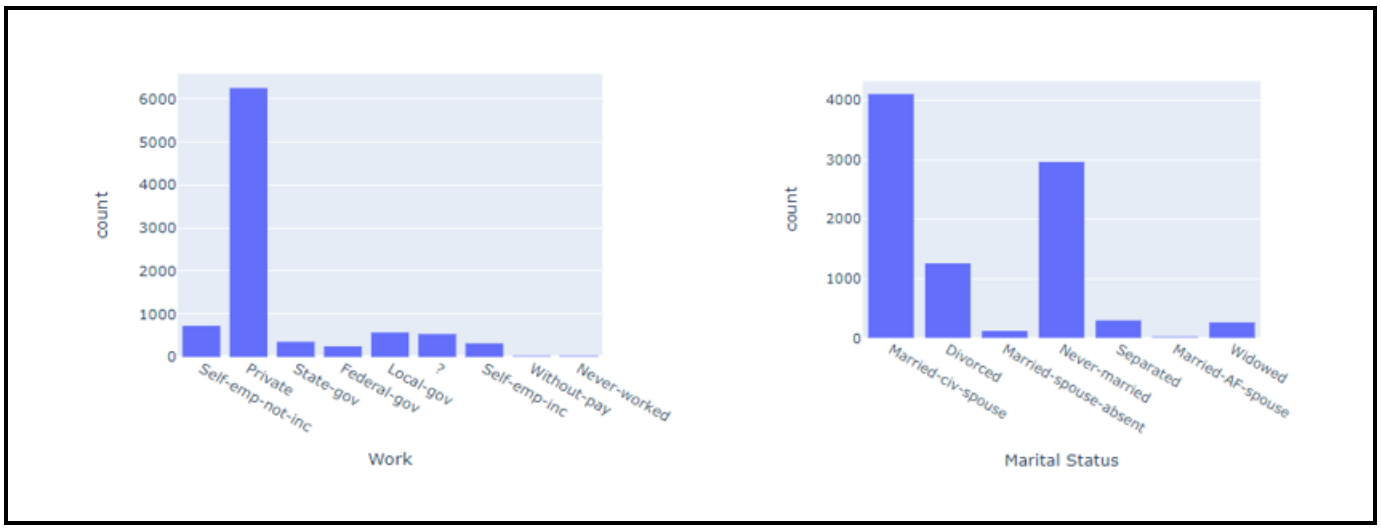
7.2.3 Data Preparation :

We have analyzed the data, and we came to know that all the feature datatypes are good to proceed with. Notably, the dataset does not contain any Missing values as such. But there are '?' values present in the country of citizenship and work columns. We decided to deal with these values in the exploratory data analysis phase. We also decided to remove Customer_id because it is not helpful for knowledge extraction.

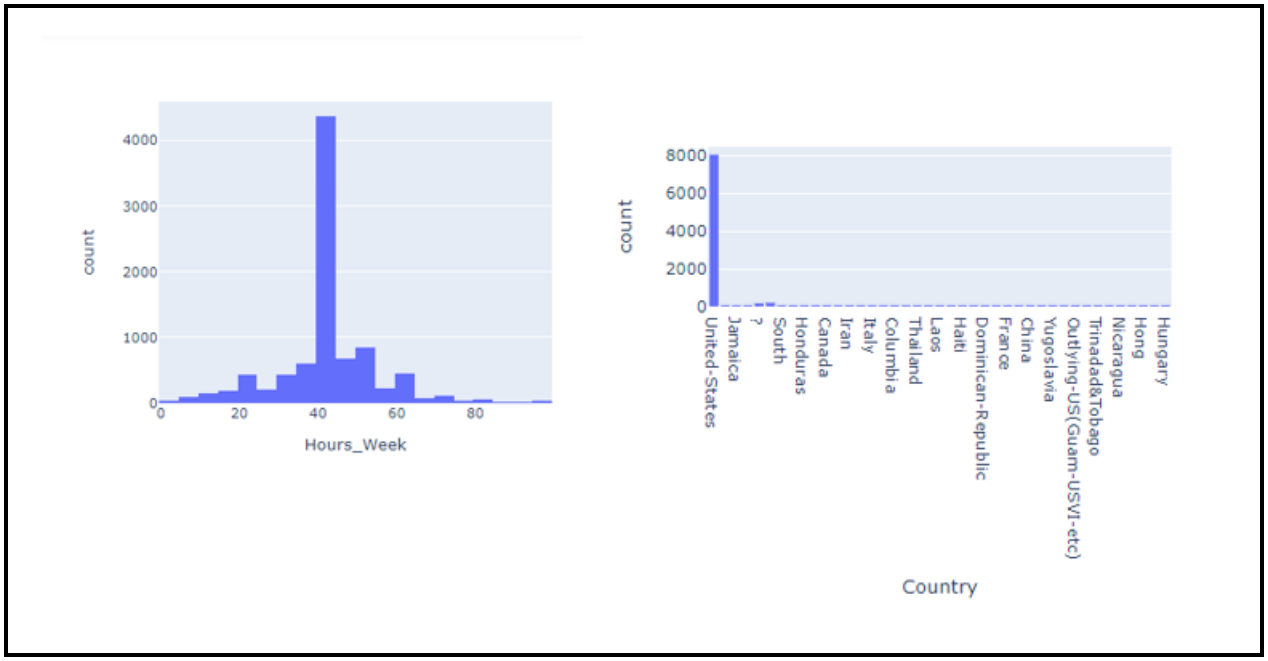
7.2.4 Exploratory data Analysis :

Before proceeding to exploratory data analysis, We decide to change the name of the Above_below 50K column. We renamed the column name to Above 50K. We also decided to change its values to 1 and 0. If the income is above 50k, then it will be shown as 1 and if income is less than or equal to 50k, then it will be 0. This stage will help us in exploratory analysis and feature engineering.

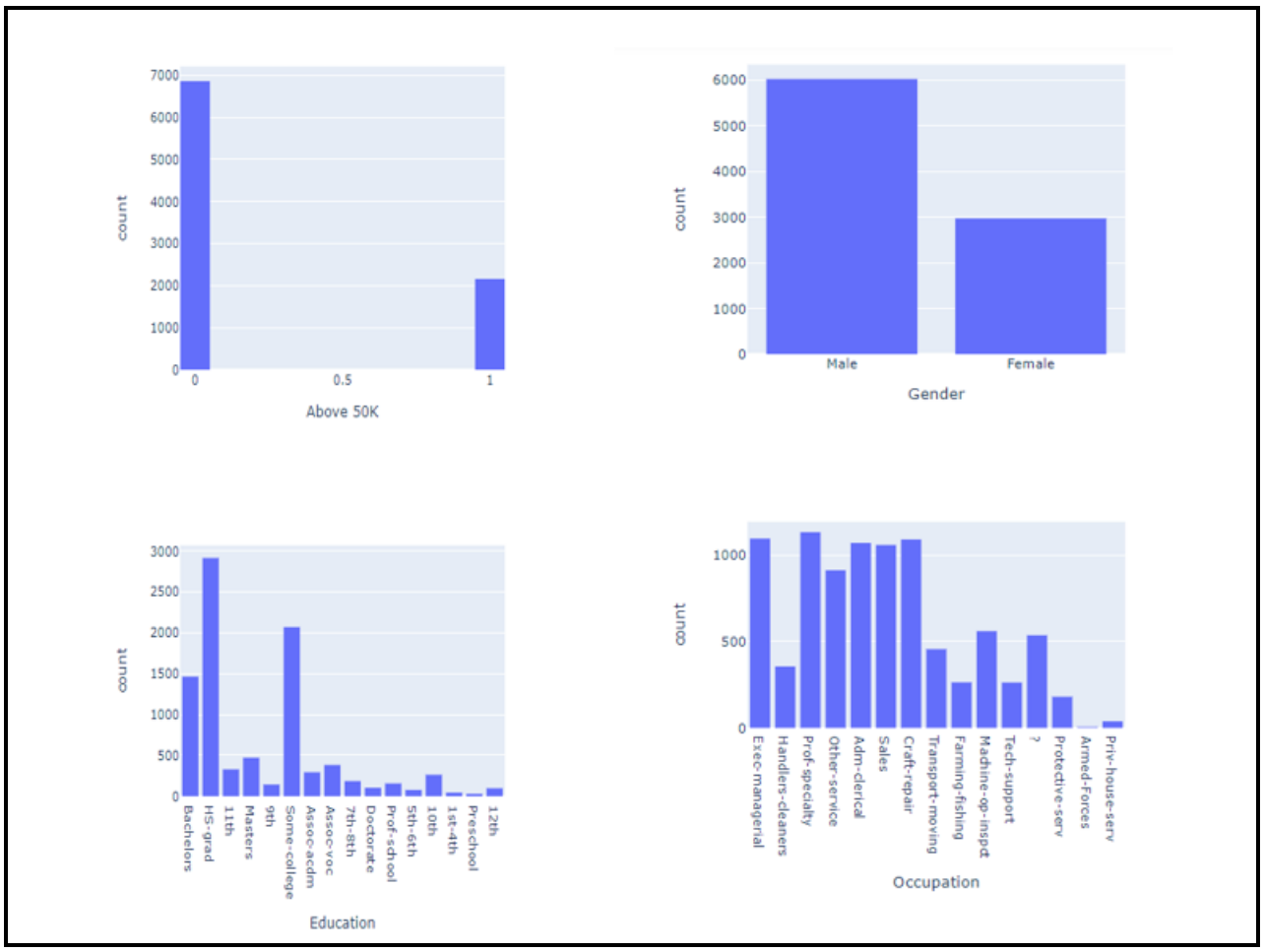
Data Distribution Using Histogram :



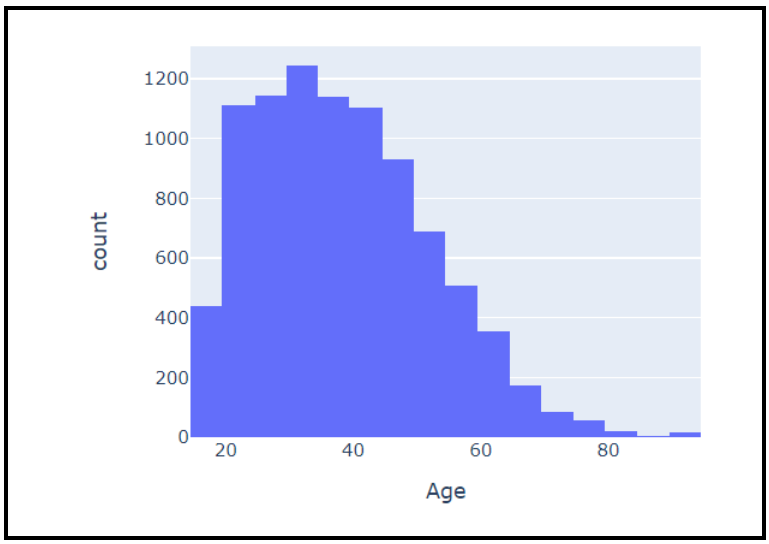
- From the histogram we can clearly see that more than 6000 employees are working in the private sector followed by self-employed customers ranging approximately 800.
- From another histogram of marital status, we can clearly see that we have more than customers who are married. We have approx. 3000 customers who are not married. Similarly, we have around 1200 customers who are divorced.



We have more than 4000 customers who are working more than 40 hours per week. It appears that, we have a few customers who are working 100 hours per week as well. This value is an outlier but we decided to keep it in the data because it's a valid value. There are some cases where customers work 100 hours a week. Furthermore, In the Case of Country of citizenship, We have 8000 customers who are citizens of the United States of America. This data does make sense because this data belongs to a company from the USA.



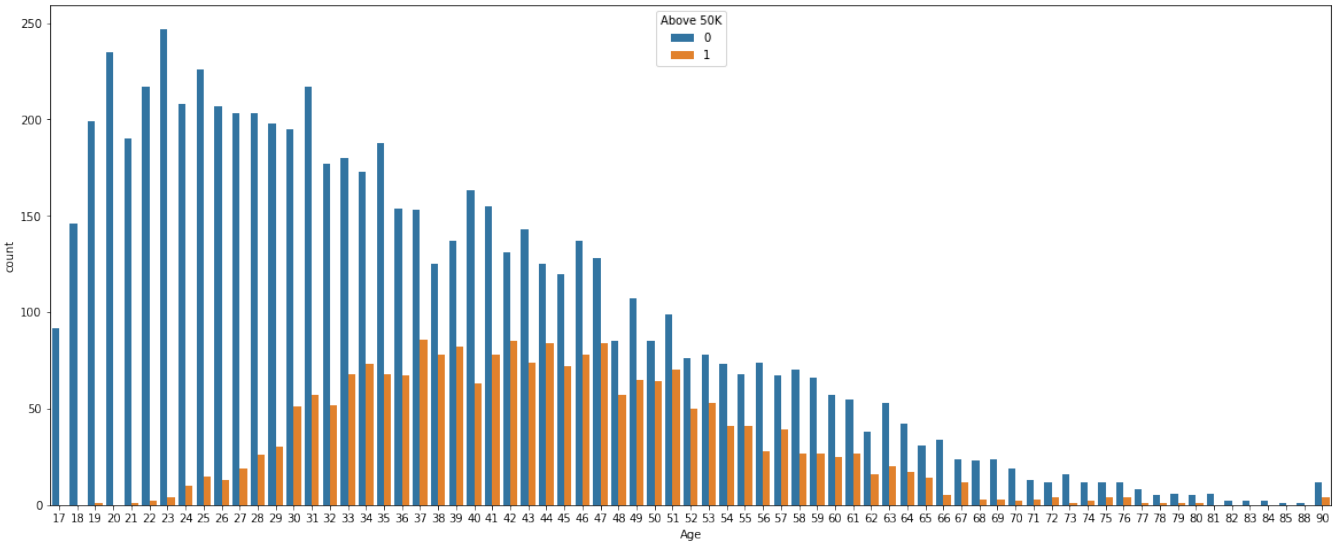
In our dataset, We have more than 6500 customers whose annual income is less than 50 thousand dollars, while customers whose yearly income is greater than 50 thousand dollars is slightly above 2000. In the case of Gender, We have more than 3000 male customers as compared to female customers. In terms of Education, our dataset contains 2900 customers who have completed high school, followed by 2100 customers who have completed some college degree, and finally, 1400 customers who have completed a bachelor's degree. There are other types of educational qualifications as well, which are shown in the figure above. At last, the Occupation of customers describes the distribution such as, we have more than 5000 customers working in Executive Managerial positions, craft repair, sales, and other services. Comparatively, there are lots of fewer customers from Armed-forces.



We have customers aged from 18 to 100 in our dataset. We have a maximum number of customers from the age group 20-60 years. We have around 1200 customers whose age is about 35 years in our dataset.

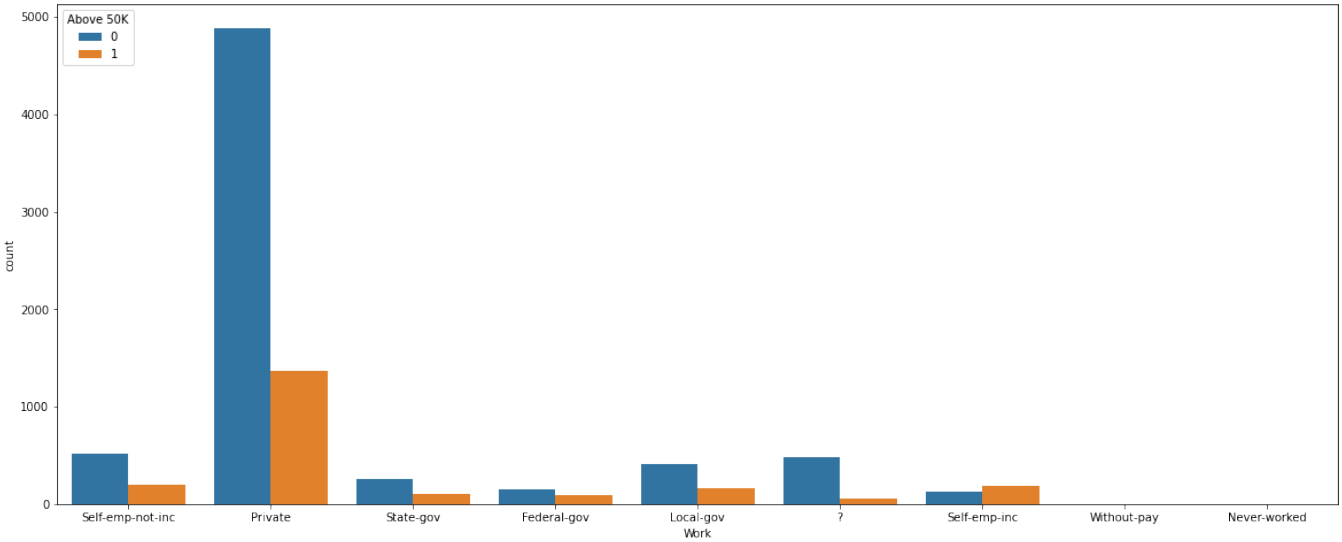
Hypothesis Based on EDA :

1.Is the age of a person have any prominent relationship with Income ?



We have lots of young customers who does not earn more than 50 thousand dollars a year. From the above bar plot, we can clearly see that there is a relationship between age and income. We have lots of customers from age group 25 to 67 who are earning more than 50 thousand dollars a year.

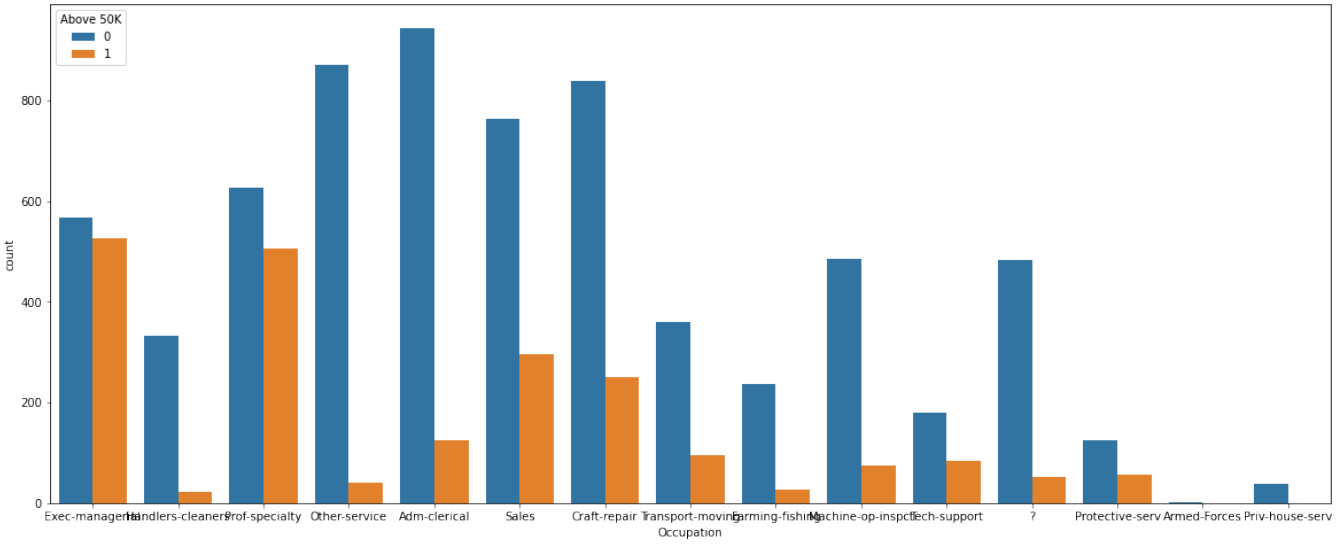
2.Does Work Class affects the income range ?



We have another type of customer who are working in the private sector. Therefore, it is quite relevant that there are lots of private employees who will be earning more than 50

thousand dollars a year. Only In the case of incorporated self-employees, Customers who are earning more than 50 thousand dollars are comparatively higher than the customers who do not earn more than 50 thousand dollars. We also have '?' in data and have the lowest number of customers who earn more than 50k; therefore, we decided not to consider it.

3. Does Occupation affects the income range ?



From the above Distribution, it is quite clear that Occupation does affect the income range. If customers are working as an executive manager or professional specialty or similar kind of role, then there are higher chances of the customer earning more than 50 thousand dollars a year. We also have '?' in the data where there are lots of customers who are not earning the specified income mark therefore, we are not considering this specific analysis.

7.2.5 Feature Engineering :

We have already presented all the necessary details about Feature Extraction, one hot encoding, and the Weight of evidence above. Firstly, we converted all the '?' values into 'not known'.

One- Hot Encoding :

Result After One-Hot-Encoding

In [105]:
```
# Drop the Education columns

processed_df.drop('Education', axis=1, inplace=True)

# Work, Country, and Occupation column have '?' values. We will introduce the new category for the same.
df['Work'].replace('?','Not Known', inplace=True)
df['Country'].replace('?','Not Known', inplace=True)
df['Occupation'].replace('?','Not Known', inplace=True)


# Applying One Hot Encoding to the Dataframe
one_hot_df = pd.get_dummies(processed_df, drop_first=True)

one_hot_df.head(5)
```

Out[105]:

Without-pay	Marital Status_Married-AF-spouse	Marital Status_Married-civ-spouse	Marital Status_Married-spouse-absent	Marital Status_Never-married	Marital Status_Separated	Marital Status_Widowed	Country_Cambodia	Country_Canada	Country_China
0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0

Weight of Evidence Encoding :

In [129]:
```
# Applying WOE to the Dataframe
from category_encoders.woe import WOEEncoder
WOE_encoder = WOEEncoder()

categorical_cols = ['Work', 'Marital Status', 'Occupation',  'Gender', 'Country']
for column in categorical_cols:
    processed_df[column] = WOE_encoder.fit_transform(processed_df[column], processed_df['Above 50K'])

wef_df = processed_df.copy()
```

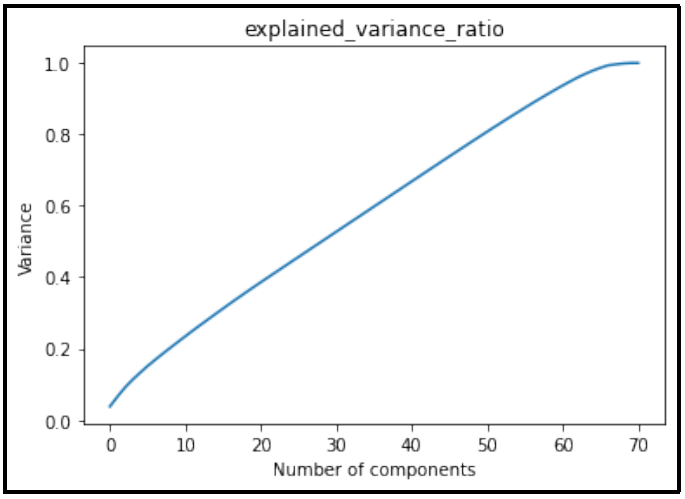
In [130]:
```
wef_df.head()
```

Out[130]:

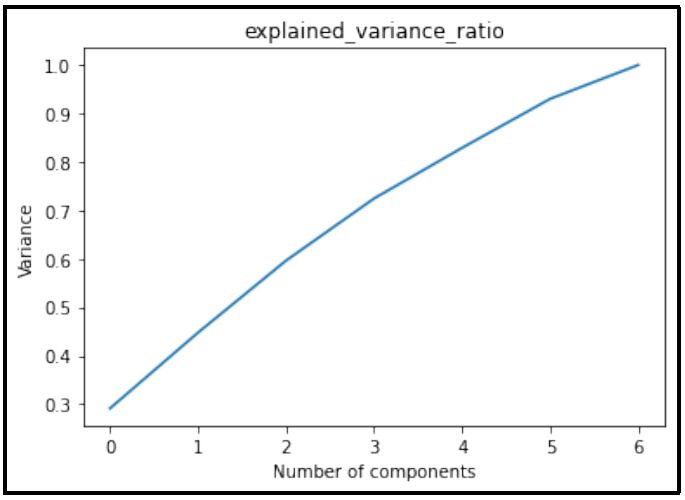
	Work	Marital Status	Hours_Week	Country	Above 50K	Gender	Occupation	Age
0	0.209936	0.948163	13	0.024828	0	0.313434	1.079632	50
1	-0.120254	-1.037379	40	0.024828	0	0.313434	-1.519335	38
2	-0.120254	0.948163	40	0.024828	0	0.313434	-1.519335	53
3	-0.120254	0.948163	40	-0.635448	0	-0.888444	0.941902	28
4	-0.120254	0.948163	40	0.024828	0	-0.888444	1.079632	37

7.2.6 Feature Extraction :

For One-hot Encoding :



For Weight Of Evidence :



We perform PCA for both one-hot_encoding and Weight of evidence. In the case of One hot Encoding PCA, We can see that around 99% of the variance is explained by 66 components. So instead of giving all columns as input in our algorithm, let's use these 66 principal components. On the other hand, In the case of Weight of Evidence, We can see that around 95% of variance is being explained by five components. So instead of giving all columns as input in our algorithm, let's use these principle components.

7.3 Model Building for Insurance Fraud detection and Income Prediction :

K-Fold Cross-validation for Training and Testing :

We have used Cross- validation here. Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation. It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split[8].

In our case k=10, which means we are splitting the dataset into 10 folds and running the train and tests. So during each run, only one fold will be considered for testing and the other 9 will be for training and moving the iteration.

We have decided to use the following machine learning algorithms For our research because we are dealing with a classification problem. As our Research project progress through 2 stages (i.e., insurance fraud detection and income prediction), both are Classification problem.

- Random Forest
- Decision tree

- Logistic Regression
- Extreme gradient Boosting (XGboost)
- Adaptive Boost (Adaboost)
- Support Vector Machine (SVM)
- Gaussian naive bayes (Gaussian NB)

1. Random Forest Classifier:

The Random forest technique is considered as a meta-estimator. Meta estimators usually fit various decision tree classifiers on various samples of datasets. Random Forest also uses averaging to control overfitting. Sometimes the whole dataset is used to build each tree, but to prevent that, sub-sample size can be controlled with max_samples parameter[18].In terms of the random Forest classifier, we have got maximum accuracy for income prediction in weight of the Evidence Method.

Name	One-Hot- Encoding	Weight of Evidence (WoE)
Insurance fraud detection	*0.75*	*0.77*
Income Prediction	*0.795*	*0.798*

2. Decision Tree:

Decision tree is a type of supervised learning method used to solve classification and regression problems. The primary goal of a Decision tree is to develop a model that predicts the value of a target variable by learning simple decision rules from various features of data. A tree can often be seen as a piecewise constant approximation [17]. Income Prediction in one hot encoding was the best-performing model as compared to insurance fraud detection.

Name	One-Hot- Encoding	Weight of Evidence (WoE)
Insurance fraud detection	*0.65*	*0.71*
Income Prediction	*0.769*	*0.763*

3. Extreme gradient Boosting (XGBoost) :

XGboost is an optimized and distributed gradient boosting library. XGBoost is highly flexible, efficient, and portable. It uses Gradient boosting framework to implement a machine-learning algorithm. XGboost solves many data science problems by providing parallel tree boosting. It is fast and accurate as well. The same code can solve problems beyond billions of examples because it runs on big distributed environments like Hadoop, SGE, etc .[19]. XGBoost algorithm Managed to achieve 80% accuracy in income prediction (one-hot Encoding) while in the case of insurance fraud detection, its only 72%. Notably, income prediction and insurance fraud detection both in weight of evidence encoding managed to return approximately the same accuracy 78% and 79% respectively.

Name	One-Hot- Encoding	Weight of Evidence (WoE)
Insurance fraud detection	*0.72*	*0.78*
Income Prediction	*0.80*	*0.79*

4. Adaptive Boost (AdaBoost) :

An AdaBoost classifier is a meta Classifier. AdaBoost starts by fitting a classifier on the original dataset, and then it fits extra copies on the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted in such a way that it gives extra focus to difficult cases [16]. In the case of Insurance Fraud detection, the Weight of Evidence managed to return good accuracy while in the case of income prediction, One-hot encoding is the best performer.

Name	One-Hot- Encoding	Weight of Evidence (WoE)
Insurance fraud detection	*0.69*	*0.78*
Income Prediction	*0.81*	*0.79*

5. Logistic Regression:

Logistic regression is a classification algorithm. It is used to predict a binary outcome based on a set of independent variables. Logistic regression is the correct type of analysis to use when working with binary data. we know we are dealing with binary data when the output or dependent variable is dichotomous or categorical in nature; in other words, if it fits into one of two categories (such as "yes" or "no", "pass" or "fail", and so on)[20]. In the case of Insurance Fraud detection, the Weight of evidence has performed much better than one-hot encoding in terms of accuracy score. Similarly, One-hot encoding successfully managed to return higher accuracy than weight of Evidence.

Name	One-Hot- Encoding	Weight of Evidence (WoE)
Insurance fraud detection	*0.73*	*0.80*
Income Prediction	*0.81*	*0.79*

6. Support Vector Machine (SVM) :

SVM stands for Support vector machine. It is a type of supervised learning. We use SVM when we are dealing with classification or regression problems. Our research problem is a classification problem. therefore, we have decided to use it as well. Similarly, it is also useful in outlier detection. SVM is highly efficient in high-dimensional spaces. SVM does not provide probability estimates directly. These are calculated using expensive five-fold cross-validation[10]. In the case of Income prediction, both the encoding managed to return exactly the same accuracy percentage. On the other hand, the Weight of evidence was more proficient than one-hot encoding in terms of accuracy.

Name	One-Hot- Encoding	Weight of Evidence (WoE)
Insurance fraud detection	*0.75*	*0.80*
Income Prediction	*0.818*	*0.815*

7. Gaussian Naive Bayes (Gaussian NB) :

Gaussian Naive Bayes is a variant of Naive Bayes that follows Gaussian normal distribution and supports continuous data. Naive Bayes is a group of supervised machine learning classification algorithms based on the Bayes theorem. It is a simple classification technique but has high functionality. They find use when the dimensionality of the inputs is high. Complex classification problems can also be implemented by using Naive Bayes Classifier[14].
in the case of income prediction, Both the encoding techniques managed to return the same accuracy score. However, In the Insurance fraud detection phase, the Weight of evidence managed to achieve good accuracy but one-hot encoding did not perform well.

Name	One-Hot- Encoding	Weight of Evidence (WoE)
Insurance fraud detection	*0.33*	*0.79*
Income Prediction	*0.795*	*0.798*

8 Evaluation :

We have used 7 Machine learning algorithms and two data encoding techniques. therefore, we have got 14 results each in both the insurance fraud detection and income prediction phases. We have used Accuracy score metrics to evaluate the final product. Whenever we deal with the classification problem in data science, we use an accuracy score to evaluate the performance of machine learning algorithms in fraction format.

We have used 2 Encoding techniques. If we recall the literature survey used for this research, Only Mr.Pinak Patel used the encoding Technique for their Research and got a satisfactory result. Notably, The dataset that we have used is the a small dataset. In terms of reliability, Although we are working on a smaller dataset this methodology can be successful with a big dataset as well because we have used principle component analysis in our research which will help summarise the information contained in large data tables by means of a smaller set of "summary indices" that can be more easily visualized and analyzed. Regarding the usability of the end product, we just need to pass the data to machine learning to detect fraudulent claims or another way is the deployment or production.

Regarding the aims related to the research, we have successfully managed to achieve the aims specified above. We have successfully executed the research through 2 independent phases as a result shown above. We have also managed to achieve 80% accuracy in both phases of the research project. We also managed to find insights from the data and logical explanations regarding it in the hypothesis of the exploratory data analysis section.

Even though we had executed the income prediction model independently due to the unavailability of the dataset, which will include insurance fraud detection data columns and income prediction data columns, We Managed to get 80% accuracy in both phases of our research project. **"Good accuracy in machine learning is subjective. But in our opinion, anything greater than 70% is a great model performance. In fact, an accuracy measure of anything between 70%-90% is not only ideal, but it's also realistic"[1].** This is also consistent with industry standards."

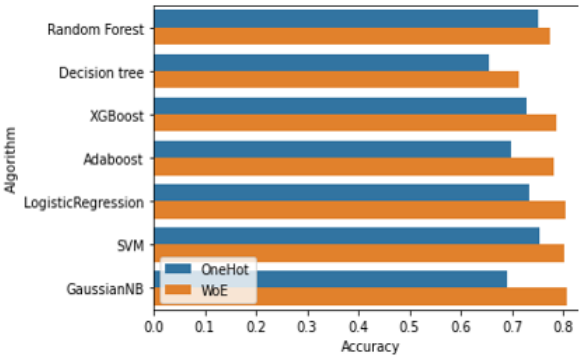
Talking about the limitations, We have divided the limitation in two types, i.e., Technical and Business. In terms of technical Limitations, the Machine-learning model does take time to identify a new type of fraud that was not introduced to it. We provided some data to the machine-learning model, and then the ML model made a prediction based on that. Technical limitations also include the performance of the model. There were no Missing values in our dataset; therefore dealing with missing values should also be important as it directly affects the model accuracy. One of the most important business limitations is that there are various different rules in different states of the USA. As we have seen earlier, There are lots of fraudulent cases due to the loopholes in the law enforcement system of south Carolina.

In the case of Ethical Implication, one thing to remember is that the Purpose of this research is to help Insurance companies detect fraudulent cases and predict the income of customers to target them in the future with promotional offers. We have used the dataset from open-source repositories, but Insurance companies should keep ethical aspects in mind while collecting the data from the user with their consent. Even though insurance companies are predicting the income of customers. The value might not be 100% accurate but insurance companies will definitely get the idea about the income range of customers. Income is a personal attribute, and not every customer would like to share but if this data gets leaked then customers might go through unnecessary trouble. therefore, Ethical implications should be kept in mind by insurance companies to avoid any future loss.

Practically speaking, In Our research project, we did not use a single dataset that will include the insurance fraud data columns and income prediction data columns because such type of data is not available online. But the methods used in our research are quite relevant and accurate. We managed to get realistic accuracy for the machine learning model. Due to the methodology; we have used, the practical value of our end product is really Auspicious.

9 Conclusion :

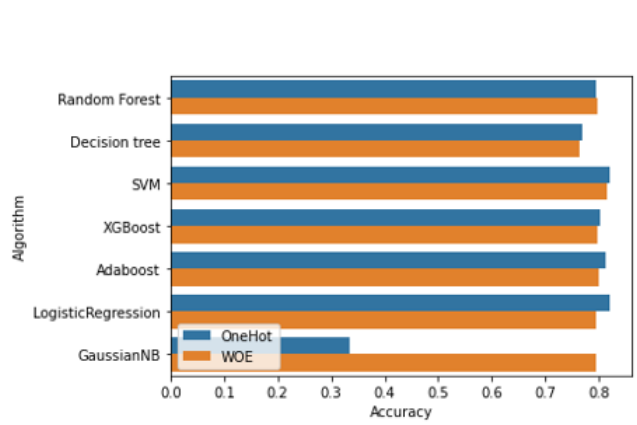
In this research, We have considered various studies related to insurance fraud detection and income prediction. This research attempted to achieve maximum accuracy for the machine learning models that have been used in the research. We managed to obtain 80% accuracy for insurance fraud detection and income prediction with the combination of data encoding techniques and principal component analysis so that these machine learning models can efficiently work with large datasets as well. We have performed exploratory data analysis to get important insights from the data. We have tried to present those insights with logic to make the sense of the data.



Out[78]:

	Algorithm	Encoding	Accuracy
13	GaussianNB	WoE	0.806
11	LogisticRegression	WoE	0.804
12	SVM	WoE	0.802
9	XGBoost	WoE	0.787
10	Adaboost	WoE	0.781
7	Random Forest	WoE	0.773
5	SVM	OneHot	0.753
0	Random Forest	OneHot	0.751
4	LogisticRegression	OneHot	0.732
2	XGBoost	OneHot	0.727
8	Decision tree	WoE	0.714
3	Adaboost	OneHot	0.699
6	GaussianNB	OneHot	0.691
1	Decision tree	OneHot	0.654

The Above figures show the performance of machine learning models of insurance fraud detection. We have got 14 results because we used 2 data encoding techniques on Seven machine learning algorithms. We have arranged the accuracies in ascending order. We got maximum accuracy on the naïve bayes algorithm, which is 80% when we encoded the data in the Weight of evidence method. Indeed, it is clearly visible that encoding techniques do make a difference in the performance of machine learning models because Another naïve bayes algorithm in which data is encoded with one hot encoding has got an accuracy of 69%. We have got the second highest accuracy for logistic regression where data was encoded with the weight of evidence method. Mrs. Rama Devi Burri mentioned in her studies that she got less accuracy for naïve bayes, which was 91% ,but in our case, naïve bayes is the best algorithm in terms of accuracy. According to the research of Mr. Najmeddin Dhieb, They have got maximum accuracy for XGboost Algorithm, but in our case, XGBoost with the weight of evidence method is the 4th best algorithm. Notably, the results are different from each of the literature studies we assessed. There are so many factors that can cause this such as the nature of the dataset, Methods used, etc. On the other hand, The worst-performing model was Decision Tree, with only 65% accuracy.



Out[147]:

	Algorithm	Encoding	Accuracy
5	LogisticRegression	OneHot	0.819556
2	SVM	OneHot	0.818778
9	SVM	WOE	0.815222
4	Adaboost	OneHot	0.811222
3	XGBoost	OneHot	0.803000
11	Adaboost	WOE	0.799111
7	Random Forest	WOE	0.798000
10	XGBoost	WOE	0.797111
0	Random Forest	OneHot	0.795444
13	GaussianNB	WOE	0.794778
12	LogisticRegression	WOE	0.793667
1	Decision tree	OneHot	0.769111
8	Decision tree	WOE	0.763333
6	GaussianNB	OneHot	0.334778

The above figures represent the performance of machine learning models used for income prediction. if we recall the studies Prediction of Individual Level Income: A Machine Learning Approach BY Michael Matkowski, In their research, they have got good accuracy with

logistic regression, XGBoost, and random Forest algorithm. Our results are also identical to Mr.Michael Matkowski's results. We have got maximum accuracy for logistic Regression in the one-hot encoding technique which is 81.9%. Similarly, XGboost and Random Forrest also managed to return approx. 80% model accuracy. Notably, The second highest accuracy we got is for the SVM model in One-Hot Encoding whose accuracy is 81.8%. The worst performing model was naïve bayes which is 33% only.

In the upcoming time, I would like to suggest the use of the Encoding technique as it does make a significant difference in similar kind of research project where categorical variables needs to be encoded into numerical ones. The use of Principal component analysis is also a good idea as it sets the platform for big datasets. Exploratory data analysis is one of the important stages to get important insights from the data. Our suggestion is to spend more time on the exploratory data analysis phase.

Reference For dataset:
1. Insurance Fraud detection :
https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/4954
2. income Prediction : https://www.kaggle.com/datasets/knowledgearoundyou/insurance-data

References

[1] Kirsten barkved. How To Know if Your Machine Learning Model Has Good Performance . https://www.obviously.ai/post/machine-learning-model-performance: :text=But%20in%20our%20opinion%2C%20anything,not%20only%20ideal%2C%20it's% 2022. Accessed december 20th, 2022.

[2] Rama Devi Burri. Insurance Claim Analysis Using Machine Learning Algorithms . https://www.ijitee.org/wp-content/uploads/papers/v8i6s4/F11180486S419.pdf, 2019. Accessed december 6th, 2022.

[3] deepanshu bhalla. WEIGHT OF EVIDENCE (WOE) AND INFORMATION VALUE (IV) EXPLAINED . https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html, 2015. Accessed december 19th, 2022.

[4] Mr. Najmeddine Dhieb. Extreme Gradient Boosting Machine Learning Algorithm For Safe Auto Insurance Operations . https://www.researchgate.net/publication/337508754_Extreme_Gradient_Boosting_Machine_Learning 2019. Accessed december 10th, 2022.

[5] R Guha. Comparative Analysis of Machine Learning Techniques for Detecting Insurance Claims Fraud . https://www.wipro.com/analytics/comparative-analysis-of-machine-learning-techniques-for-detectin/, 2020. Accessed november 21st, 2022.

[6] harshil patel. What is Feature Engineering . https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10, 2021. Accessed december 19th, 2022.

[7] jason brownlee. Why One-Hot Encode Data in Machine Learning? . https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/, 2017. Accessed december 19th, 2022.

[8] jason brownlee. A Gentle Introduction to k-fold Cross-Validation . https://machinelearningmastery.com/k-fold-cross-validation/, 2018.

[9] Tarun kumar. Intelligent Income prediction model . https://www.jetir.org/papers/JETIR2008084.pdf, 2020. Accessed december 13th, 2022.

[10] Scikit learn. Support Vector Machine . https://scikit-learn.org/stable/modules/svm.html, 2022. Accessed december 20th, 2022.

[11] Michael Matkowski. Prediction of Individual Level Income: A Machine Learning Approach . https://digitalcommons.bryant.edu/cgi/viewcontent.cgi?article=1038context=honors_economics, 2021. Accessed december 10th, 2022.

[12] Pinak patel. A Survey Paper on Fraud Detection and Frequent Pattern Matching in Insurance claims using Data Mining Techniques . https://www.irjet.net/archives/V6/i1/IRJET-V6I1104.pdf, 2019. Accessed december 8th, 2022.

[13] Dipika pawar. Person's Income Prediction and Analysis using Python . https://medium.com/clique-org/persons-income-prediction-and-analysis-using-python-97f939d19694, 2021. Accessed november 29th, 2022.

[14] prateek majumdar. Gaussian Naive Bayes . https://iq.opengenus.org/gaussian-naive-bayes/, 2018. Accessed december 22nd, 2022.

[15] Mel Restori. What is Exploratory Data Analysis . https://chartio.com/learn/data-analytics/what-is-exploratory-data-analysis/, 2020. Accessed december 15th, 2022.

[16] scikit learn. ADABoost Classifier . https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html, -. Accessed december 19th, 2022.

[17] scikit learn. Decision Tree . https://scikit-learn.org/stable/modules/tree.html, -. Accessed december 19th, 2022.

[18] scikit learn. RandomForestClassifier . https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html, -. Accessed december 19th, 2022.

[19] scikit learn. XGBOOSt . https://xgboost.readthedocs.io/en/stable/, -. Accessed december 19th, 2022.

[20] Anamika Thanda. What is Logistic Regression? A Beginner's Guide . https://careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/, 2022. Accessed december 19th, 2022.

[21] Jeffrey J. Wiseman. Insurance Fraud in South Carolina . https://www.ycrlaw.com/wp-content/uploads/fraud1-wiseman-article-pdf.pdf, -. Accessed december 17th, 2022.

[22] Andy yohn. 8 Benefits of Insurance Fraud Analytics . https://www.duckcreek.com/blog/insurance-fraud-analytics/, 2021. Accessed decemmber 4th, 2022.

10 Appendix :

Code Snapshots :

Data Preparation

- We prepare the data column wise, as every column would require different preprocessing

In [10]:
```
def display_basic_information(df, column_name):
    """
        Function is responsible for printing out basic information
        about the dataset such as Null values, Data types, etc.
    """
    data_type = df[column_name].dtype
    null_values = df[column_name].isna().sum()
    print("####################")
    print("Basic Information:")
    print("####################\n")
    print("column name:", column_name)
    print("data type:", data_type)
    print("total number of null values:", null_values)
    print()

column_headers = list(df.columns)
for i in column_headers:
    display_basic_information(df,i)
```

```
####################

column name: policy_bind_date
data type: object
total number of null values: 0

####################
Basic Information:
####################

column name: policy_state
data type: object
total number of null values: 0

####################
Basic Information:
```

Checking the quality of data

In [14]:
```
# Checking the null values
df.isna().sum()
```

Out[14]:
```
months_as_customer            0
age                           0
policy_state                  0
policy_deductable             0
policy_annual_premium         0
umbrella_limit                0
insured_zip                   0
insured_sex                   0
insured_education_level       0
insured_occupation            0
insured_relationship          0
capital-gains                 0
capital-loss                  0
incident_type                 0
collision_type                0
incident_severity             0
authorities_contacted         0
incident_state                0
incident_hour_of_the_day      0
number_of_vehicles_involved   0
property_damage               0
bodily_injuries               0
witnesses                     0
police_report_available       0
total_claim_amount            0
injury_claim                  0
property_claim                0
vehicle_claim                 0
auto_make                     0
auto_year                     0
fraud_reported                0
dtype: int64
```

Data Distribution using Histogram

In [22]:
```
def count_plot(df):
    """
        Function is resposible for printing out the count plot on sys.out
    """
    for col in df.columns:
        fig = px.histogram(df, col, nbins=20, width=500, height=400)
        fig.show()
count_plot(eda_df)
```

###Correlation

In [23]:
```
# Correlation heatmap for all numeric values
plt.figure(figsize = (18,10))
sns.heatmap(eda_df.corr(), annot=True)
```

Out[23]: <AxesSubplot:>

Following observations were made

1. `months_as_customer` is highly correlated with `age`
2. `total_claim_amount` is highly correlated with `injury_claim` and `vehicle_claim` and `property_claim`

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Markdown

XGBoost

In [50]:
```
from xgboost import XGBClassifier
result= ["XGBoost", "OneHot"]
# on WoE data
model = XGBClassifier()
result.extend(evaluate_model(model,X, y))
data.append(result)
```

```
########
Results
########

Average accuracy score:  0.7270000000000001
```

Adaboost

In [51]:
```
from sklearn.ensemble import AdaBoostClassifier
model = AdaBoostClassifier(random_state=0)
result= ["Adaboost", "OneHot"]
result.extend(evaluate_model(model, X, y))
data.append(result)
```

```
########
Results
########

Average accuracy score:  0.6990000000000001
```

Logistic Regression

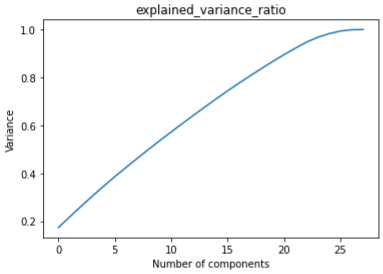
In [52]:
```
from sklearn.linear_model import LogisticRegression
model = LogisticRegression()
result= ["LogisticRegression", "OneHot"]
result.extend(evaluate_model(model, X, y))
data.append(result)
```

```
########
Results
########

Average accuracy score:  0.732
```

24	0.983067
25	0.993794
26	0.999207
27	1.000000

In [66]:
```
plt.figure()
plt.plot(np.cumsum(pca.explained_variance_ratio_))
plt.xlabel('Number of components')
plt.ylabel('Variance')
plt.title('explained_variance_ratio')
plt.show()
```



We can see that around 99% of variance is being explained by 25 components. So instead of giving all columns as input in our algorithm let's use these principle components instead

In [67]:
```
pca = PCA(n_components=25)
pca_woe_df = pca.fit_transform(x)

column_name = [f'PC-{i}' for i in range(1, 26)]

principal_woe_df = pd.DataFrame(pca_woe_df, columns=column_name)
principal_woe_df.head()
```

Out[67]:

	PC-1	PC-2	PC-3	PC-4	PC-5	PC-6	PC-7	PC-8	PC-9	PC-10	PC-11	PC-12	PC-13	PC-14	PC-1
0	-0.997570	0.780926	-0.533583	0.184909	-0.679628	-1.819256	-1.283710	-0.288213	0.726652	-1.200167	0.265428	-0.874040	-1.084496	-0.592400	0.0645
1	4.026343	0.842083	1.913565	0.716060	-0.067698	1.320849	0.535631	1.558847	0.755646	0.518444	1.281708	-0.793231	-0.203205	-0.601991	0.6357

Decision Tree

In [72]:
```
from sklearn.tree import DecisionTreeClassifier
model = DecisionTreeClassifier()
result= ["Decision tree", "WoE"]
result.extend(evaluate_model(model, X, y))
data.append(result)
```

```
########
Results
########

Average accuracy score:  0.7140000000000001
```

XGBoost

In [73]:
```
from xgboost import XGBClassifier
result= ["XGBoost", "WoE"]
# on WoE data
model = XGBClassifier()
result.extend(evaluate_model(model,X, y))
data.append(result)
```

```
########
Results
########

Average accuracy score:  0.787
```

Adaboost

In [74]:
```
from sklearn.ensemble import AdaBoostClassifier
model = AdaBoostClassifier(random_state=0)
result= ["Adaboost", "WoE"]
result.extend(evaluate_model(model, X, y))
data.append(result)
```

```
########
Results
########

Average accuracy score:  0.781
```

Saving Results for Insurance Fraud Detection

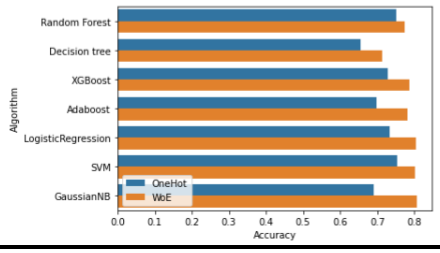
In [78]:
```
results_df = pd.DataFrame(data, columns=columns)
results_df = results_df[['Algorithm', 'Encoding', 'Accuracy']]
results_df.sort_values(by=['Accuracy', 'Encoding'], ascending=False)
```

Out[78]:

	Algorithm	Encoding	Accuracy
13	GaussianNB	WoE	0.806
11	LogisticRegression	WoE	0.804
12	SVM	WoE	0.802
9	XGBoost	WoE	0.787
10	Adaboost	WoE	0.781
7	Random Forest	WoE	0.773
5	SVM	OneHot	0.753
0	Random Forest	OneHot	0.751
4	LogisticRegression	OneHot	0.732
2	XGBoost	OneHot	0.727
8	Decision tree	WoE	0.714
3	Adaboost	OneHot	0.699
6	GaussianNB	OneHot	0.691
1	Decision tree	OneHot	0.654

In [79]:
```
sns.barplot(data=results_df, x='Accuracy', y='Algorithm', hue='Encoding')
plt.legend(loc='lower left')
```

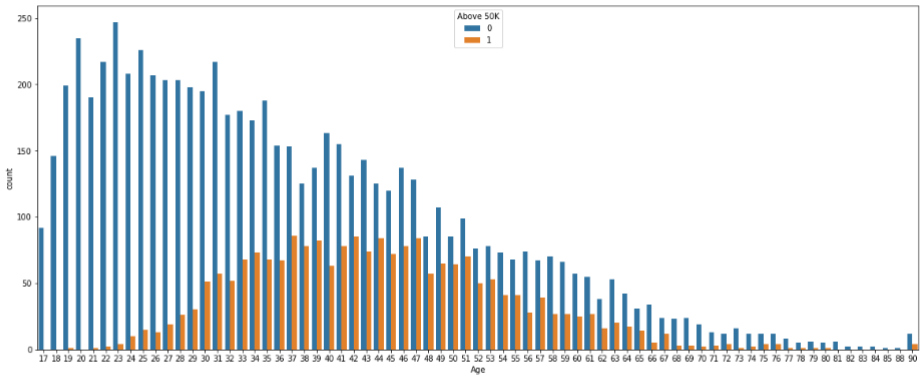
Out[79]: <matplotlib.legend.Legend at 0x183052474c0>



In [97]:
```
#Is the age of person have any prominent relationship with  insuarnce fruad  ?
from matplotlib import pyplot as plt
import seaborn as sns

plt.figure(figsize=(20,8))
sns.countplot(data=eda_df, x='Age', hue='Above 50K')
```

Out[97]: <AxesSubplot:xlabel='Age', ylabel='count'>

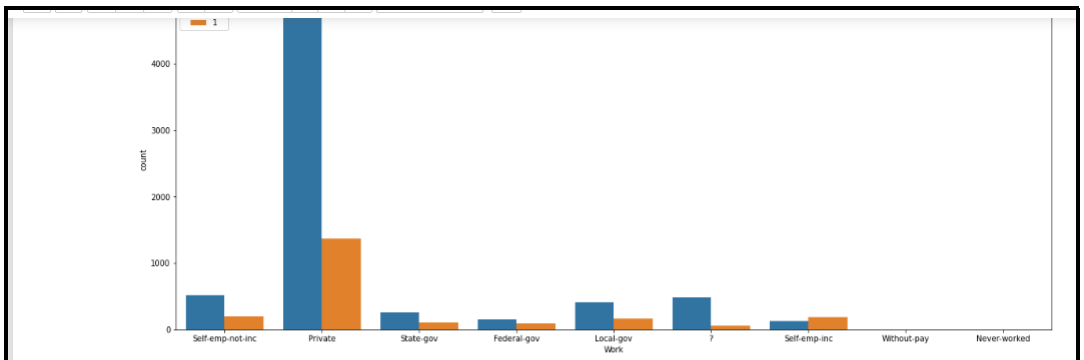


In [98]:
```
# Does WorkClass affects the income range

from matplotlib import pyplot as plt
import seaborn as sns

plt.figure(figsize=(20,8))
sns.countplot(data=eda_df, x='Work', hue='Above 50K')
```

Out[98]: <AxesSubplot:xlabel='Work', ylabel='count'>

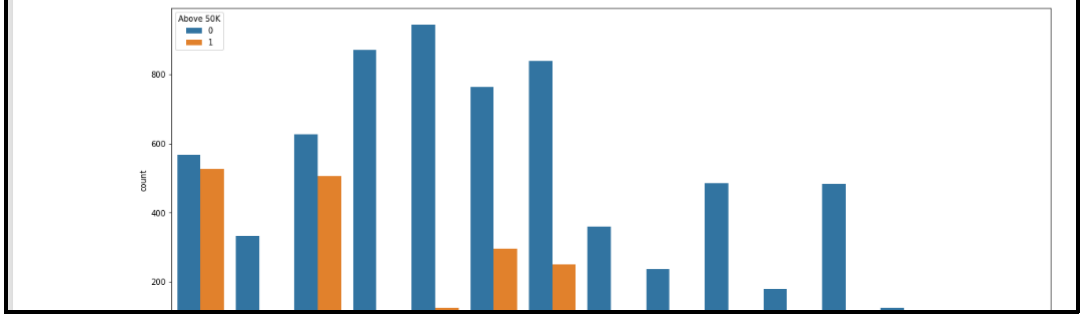


In [99]:
```
# Does Occupation affects the income range

from matplotlib import pyplot as plt
import seaborn as sns

plt.figure(figsize=(20,8))
sns.countplot(data=eda_df, x='Occupation', hue='Above 50K')
```

Out[99]: <AxesSubplot:xlabel='Occupation', ylabel='count'>



SVM

In [118]:
```
from sklearn.svm import SVC

model = SVC()
result= ["SVM", "OneHot"]

result.extend(evaluate_model(model, X, y))
data.append(result)
```

```
########
Results
########

Average accuracy score:  0.8187777777777777
```

XGBoost

In [119]:
```
from xgboost import XGBClassifier
result= ["XGBoost", "OneHot"]

model = XGBClassifier()
result.extend(evaluate_model(model,X, y))
data.append(result)
```

```
########
Results
########

Average accuracy score:  0.8029999999999999
```

Adaboost

In [120]:
```
from sklearn.ensemble import AdaBoostClassifier
model = AdaBoostClassifier(random_state=0)
result= ["Adaboost", "OneHot"]
result.extend(evaluate_model(model, X, y))
data.append(result)
```

```
########
Results
########

Average accuracy score:  0.8112222222222222
```

Pandas Profiling

In [18]:
```
from pandas_profiling import ProfileReport

# Generating reports for the dataset
profile = ProfileReport(df, title= 'Insurance Fraud Claimns Report', html={'style':{'full_width':True}})
profile.to_notebook_iframe()

# Saving the output in HTML format to view it
profile.to_file(output_file = '\Insurance_Prediction_Report.html')
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]
Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]
Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]
```

Overview

Overview | Alerts 18 | Reproduction

Dataset statistics

Number of variables	31
Number of observations	1000
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	242.3 KiB

Variable types

Numeric	13
Categorical	17
Boolean	1

In [18]:
```
from pandas_profiling import ProfileReport

# Generating reports for the dataset
profile = ProfileReport(df, title= 'Insurance Fraud Claimns Report', html={'style':{'full_width':True}})
profile.to_notebook_iframe()

# Saving the output in HTML format to view it
profile.to_file(output_file = '\Insurance_Prediction_Report.html')
```

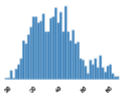
```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]
Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]
Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]
```

Insurance Fraud Claimns Report — Overview | Variables | Interactions | Correlations | Missing values | Sample

More details

age
Real number (ℝ)

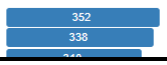
Distinct	46	Minimum	19
Distinct (%)	4.6%	Maximum	64
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	38.948	Memory size	7.9 KiB



More details

policy_state
Categorical

Distinct	3
Distinct (%)	0.3%



Data Encoding(WoE)

In [68]:
```
import sklearn.metrics as sm
from sklearn.model_selection import KFold

# Function to evaluate model
def evaluate_model(model, x, y):
    accuracy_scores = []
    precision_scores = []
    recall_scores = []
    f1_scores =[]

    kf = KFold(n_splits=10, random_state=0, shuffle=True)
    for train_index, test_index in kf.split(x):
        #setting up the data
        x_train, x_test = x.values[train_index], x.values[test_index]
        y_train, y_test = y.values[train_index], y.values[test_index]

        #Training model
        model.fit(x_train,y_train)

        #Evaluating model
        y_pred = model.predict(x_test)

        accuracy_scores.append(sm.accuracy_score(y_test,y_pred))
        precision_scores.append(sm.precision_score(y_test,y_pred))
        recall_scores.append(sm.recall_score(y_test,y_pred))
        f1_scores.append(sm.f1_score(y_test, y_pred))


    #displaying average results
    print("########")
    print("Results")
    print("########\n")
    print("Average accuracy score: ", (sum(accuracy_scores)/len(accuracy_scores)))
    results = [(sum(accuracy_scores)/len(accuracy_scores))]
    return results
```

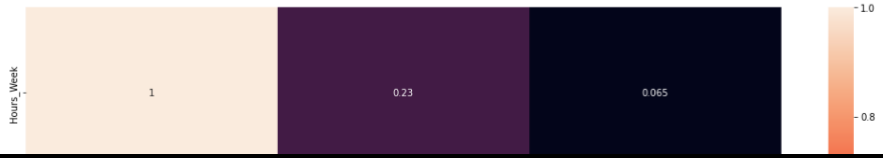
Data Distribution using Histogram

In [95]:
```
def count_plot(df):
  for col in df.columns:
    fig = px.histogram(df, col, nbins=20, width=500, height=400)
    fig.show()
count_plot(eda_df)
```

###Correlation

In [96]:
```
# Correlation heatmap for all numeric values
plt.figure(figsize = (18,10))
sns.heatmap(eda_df.corr(), annot=True)
```

Out[96]: <AxesSubplot:>



Saving Results for Income Prediction

In [147]:
```
results_df = pd.DataFrame(data,columns=columns)
results_df = results_df[['Algorithm', 'Encoding', 'Accuracy']]
results_df.sort_values(by=['Accuracy', 'Encoding'], ascending=False)
```

Out[147]:

	Algorithm	Encoding	Accuracy
5	LogisticRegression	OneHot	0.819556
2	SVM	OneHot	0.818778
9	SVM	WOE	0.815222
4	Adaboost	OneHot	0.811222
3	XGBoost	OneHot	0.803000
11	Adaboost	WOE	0.799111
7	Random Forest	WOE	0.798000
10	XGBoost	WOE	0.797111
0	Random Forest	OneHot	0.795444
13	GaussianNB	WOE	0.794778
12	LogisticRegression	WOE	0.793667
1	Decision tree	OneHot	0.769111
8	Decision tree	WOE	0.763333
6	GaussianNB	OneHot	0.334778

In [148]:
```
sns.barplot(data=results_df, x='Accuracy', y='Algorithm', hue='Encoding')
plt.legend(loc='lower left')
```

Out[148]:
```
<matplotlib.legend.Legend at 0x1836e7509a0>
```