International Journal of Advance Research in Science and Engineering
Volume No. 12, Issue No. 04, April 2023
www.ijarse.com
IJARSE
ISSN 2319 - 8354

Fraud Detection in Automobile Insurance Claims using Machine Learning Algorithms

[1]B. Sridharamurthy, [2]Anjali Priya Ugge,

[3]Tharun Pasupuleti, [4]Swetha Reddy Ganta, [5]Parvateswar Prasad Temburu

[1]Assistant Professor, Dept. of CSE, Kakatiya Institute of Technology and Science, Warangal-506015, Telangana

[2,3,4,5]Student, Dept. of CSE, Kakatiya Institute of Technology and Science, Warangal-506015, Telangana, India

ABSTRACT

There are thousands of firms in the insurance industry globally. And collect premiums of more than $1 trillion annually. The most common sort of insurance fraud is health, or vehicle, which is accomplished by filing a false accident claim. The goal of this research is to use machine learning techniques to identify any fraud. An ideal insurance agent would be able to investigate each claim and determine if it is real or not. However, this process is both time and money-consuming. It is just not practical to find and finance the competent manpower necessary to examine each of the thousands of claims that are submitted every day. Machine learning steps in to rescue the day in this situation. We have used K-Nearest Neighbor, Random Forest, KMeans, Decision Tree and Support Vector Machine algorithms. These algorithms were best suited for the dataset we had collected and gave the best accuracy of the results we were expecting.

Keywords *- Fraud Detection, Insurance Fraud, RandomForest Algorithm, Machine Learning, Performance analysis.*

1. INTRODUCTION

Any action taken to manipulate the insurance system is referred to as insurance fraud. When a claimant tries to receive a benefit, they really aren't entitled to or when an insurer willfully withholds a benefit that is owed, it happens. Insurance claims that are submitted fraudulently against an insurance provider are referred to as false insurance claims.

Since the commencement of insurance as a commercial business, there has been insurance fraud. A sizable fraction of all claims that insurers receive are fraudulent, and they cost insurers billions of dollars each year. All fields of insurance are subject to various kinds of insurance fraud. The seriousness of insurance offenses also varies, from modestly inflating claims to purposefully causing harm or accidents. Innocent individuals are harmed by fraudulent acts both directly via unintentional or purposeful harm or damage and indirectly through crimes that raise insurance costs. Governments and other organizations work to prevent insurance fraud since it is a serious problem.

Fraudulent insurance claimants include

1. Organized thieves who commit large-scale financial crimes using false business practices,
2. Experts and technicians who charge for services that aren't provided or exaggerate service rates, and
3. People in the general public who want to pay their deductible or see making a claim as a chance to make little money.

2. LITERATURE SURVEY

There were several attempts of previous works on detecting insurance fraud.

The first research paper was published in 2015 by Rekha Pal and Saurabh Pal entitled, Health Fraud Detection Using Data Mining Techniques. The classification algorithms used in this paper are, J48, ID3(Iterative Dichotomise 3), and Naive Bayes. The paper concluded that the decision tree ID3 was the most accurate of the three algorithms.

The next proposed work was by Richard A. Bauder and Taghi M. Khoshgoftaar in 2017. The title was Medicare Fraud Detection using Machine Learning Methods. The paper focuses on explaining supervised, unsupervised, and hybrid machine learning.

After exploring various Machine Learning methods, the paper concluded that supervised learning shows greater accuracy in detecting Medicare fraud.

K. Supraja and S.J. Saritha's research paper in 2017 on Detecting insurance frauds using Robust Fuzzy rule-based techniques is another attempt at detecting frauds. On the training dataset, the fuzzy rule-based method is used to estimate the level of fraud or legal activity based on the cases. Concluded that this method is utilized for huge, high-dimensional datasets with precision.

In 2018, T. Badriyah and Lailul Rahmaniah I. Syarif had come up with the research paper titled, Detecting Fraud in Auto Insurance using Nearest neighbor and Statistics Method. In order to identify fraud in the auto insurance data, this study uses the closest-neighbor approach and the interquartile method. The study's findings indicate that selecting the nearest neighbor yields the greatest outcomes.

3. PROPOSED METHOD

Machine learning algorithms used in this project are:-

1. K-Nearest Neighbor
2. Support Vector Machine (SVM)
3. Decision Tree
4. Random Forest
5. K-Means

K-Nearest Neighbor Algorithm: K-Nearest Neighbor uses the supervised learning method. The K-NN technique places the new instance in the category that is most like the existing categories on the presumption that the new case and the old instances are comparable. After storing all the previous data, a new data point is categorized using the K-NN algorithm based on similarity. This indicates that new data may be reliable and rapidly categorized using the K-NN approach.

Support Vector Machine (SVM) Algorithm: Using supervised machine learning with support vector machines, problems with classification or regression may be resolved. Yet, categorization issues are when it is most frequently utilised. With the SVM approach, each data point is represented as a point in an n-dimensional space, where n is the number of features you have and each feature's value is associated with a certain position. A linear hyper-plane between these two classes may be easily accessible to the SVM classifier.

Decision Tree Algorithm: The decision tree algorithm is a popular machine learning method for tackling classification and regression issues. It recursively divides the input space into smaller and smaller subsets until they are homogenous with regard to the target variable. The characteristic that best separates the data is used to make a split at each node of the tree, resulting in a tree-like model that may be used to forecast the class or value of a new observation. Decision trees are common in a variety of industries due to their simplicity to comprehend, analyze, and depict. [3].

Random Forest Algorithm: The Random Forest algorithm is a component of the supervised learning approach that can be used to solve classification and regression problems in ML. It performs well, predicts outcomes with high accuracy even for large datasets, and can maintain accuracy even in the absence of a sizable amount of data. It also requires less training time due to its blending of a variety of trees to anticipate the dataset's class. When all the trees are taken into account, it accurately predicts the outcome.

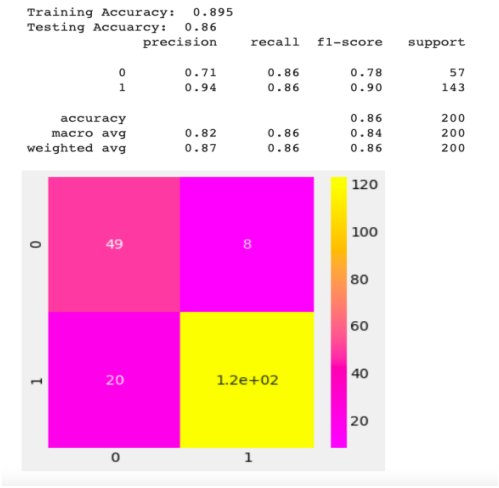

Fig.No.1 Confusion Matrix of Random Forest Algorithm

K-Means Algorithm: An unsupervised learning algorithm is K-Means clustering. Unlike supervised learning, this clustering does not use labelled data. Using K-Means, things are categorized according to their similarities and differences with other objects in the same cluster.

K is a numerical designation. The system needs to know how many clusters you need to build. For instance, K = 3 designates three clusters. It is possible to determine the optimal value of K given a set of data.

DATA COLLECTION

A researcher can assess their hypothesis using the data they have gathered. Regardless of the topic of study, data collecting is typically the first and most crucial phase in research. The method of data gathering varies for each field of study based on the information that is needed. Making sure that accurate and trustworthy data is

gathered for statistical analysis so that research decisions may be informed by data is the most important goal of data collecting [1].

DATASET AND ATTRIBUTES

The dataset that is being used in this project is taken from Kaggle and the following is the description of attributes that are used in this project.

Table 1. Description of Attributes

ATTRIBUTES	DESCRIPTION
months_as_customer	Months as a customer (Number)
age	The age of the customer (Number)
policy_deductible	The amount a policyholder must pay before the insurance provider begins to provide benefits (Number)
policy_annual_premium	The total amount paid in a year's time to keep the policy in force (Number)
insured_gender	The gender of the customer (String)
insured_education	The education qualification of the customer (String)
insured_occupation	The occupation of the customer (String)
incident_type	The type of incident (String)
incident_severity	The severity of the incident (String)
collision_type	The type of collision (String)
authorities_contacted	The authorities that are informed right after the incident occured (String)
number_of_vehicles_involved	The number of vehicles that are involved in the incident (Number)
property_damage	Is there any damage incurred towards the property (YES/NO)
witnesses	The witness in the incident (Number)
police_report_available	Is the police report available (YES/NO)
total_claim_amount	The amount that is claimed on the incident (Number)
injury_claim	The amount claimed for injuries (Number)
property_claim	The amount claimed for property damages (Number)
vehicle_claim	The amount claimed for the damages of the vehicle (Number)

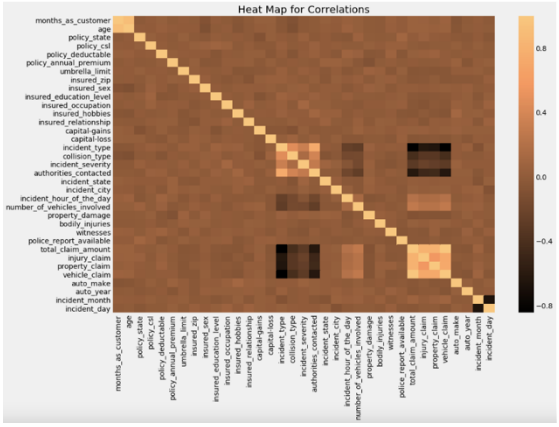

Fig.No.2 Heat Map for Correlations

DATA PREPROCESSING IN MACHINE LEARNING: Data preparation is the process of modifying raw data to be utilized with a machine learning model. It is both the first and most crucial phase in creating a machine-learning model. Finding clean, properly structured data is not always the case while developing a machine learning project. The majority of machine learning systems include tools and APIs for filling in gaps in data or balancing it out. The mean, median, and k-nearest neighbors (k-NN) of the data in the given field, along with the standard deviation, are frequently used when imputing missing values [2].

By producing extra observations or samples using techniques such as repetition, bootstrapping, or Synthetic Minority Oversampling Technique, and then bias or imbalance in the dataset may be corrected by adding them to the underrepresented classes and after that. How much memory and computation are needed for training rounds depends on the size of the dataset. Additionally, each data-related procedure must first be cleaned and prepared. Normalization reduces the quantity of the data by decreasing its order and magnitude.

EDA (Exploratory Data Analysis): Data scientists utilize EDA, which uses data visualization techniques, to examine, analyze datasets and summarize their key properties. Determining how they should change data sources to get the results they need makes it easier for data scientists to detect patterns, spot anomalies, test hypotheses, or validate assumptions.

EDA helps with understanding the variables in the data collection and their relationships and is usually used to investigate what the data might disclose beyond formal modeling or testing [2]. EDA methods, which were first created by John Tukey in the 1970s an American mathematician, are being used frequently even today.

MODEL BUILDING: Building machine learning models that can generalize effectively to future data necessitates careful examination of the data at hand and of assumptions made about the different training procedures that are now available [7]. The final assessment of the quality of a machine learning model necessitates the selection and proper interpretation of the assessment criteria. Algorithms used in machine learning can automate the development of analytical models.

MODEL EVALUATION: Model evaluation tries to estimate the accuracy of a model on future data. Model evaluation can be done using cross-validation. It's not advised to use the model's construction data to assess it. This is so that our model will always predict the right label at any point in the training set because it will just remember the entire training set. Overfitting is the term for this[8].

Metrics for model evaluation are needed to measure model performance. The assessment criteria to use depend on the machine learning job at hand. Certain measures, like precision recall, are helpful for a variety of activities. Most machine-learning applications use supervised learning tasks. We concentrate on metrics for supervised learning models in this paper. The primary observation dataset is split using the cross-validation approach into a training set for the model's development and an independent set for the analysis's assessment.

MODEL SELECTION: The process of choosing a single machine learning model out of a group of potential candidates for a training dataset is known as model selection. Fitting models are very simple but choosing one among the available models is the real issue. Given the limitations of each individual model, the incompleteness of the data sample, and the statistical noise in the data, all models have some prediction inaccuracy[6].

As a result, the idea of a perfect or optimal model is useless. We must instead look for a model that is "good enough." The best model selection technique requires "enough" data, which may be virtually infinite depending on how complex the problem is. The data would be split into training, testing, and validation sets in an ideal world. The performance of the selected model would then be reported on the test set after the candidate models had been fitted on the training set, assessed, and finalized.

MODEL DEPLOYMENT: Deploying a machine learning model into an existing production environment is known as deployment, and it allows you to use data to make useful business choices. This is the last step of the machine learning life cycle.

This poses a significant challenge because there is frequently a language barrier between the programming language used to create a machine-learning model and the languages that your production system can understand. Re-coding the model can also cause the project to take weeks or months longer to complete. Machine learning models must be smoothly deployed into production for businesses to use them to start making useful judgments. This will maximize their utility. We are implementing Flask, a microweb framework for developing web-based applications.

FLASK Framework: Python is used to create the Flask microweb application framework. Armin Ronacher was the person who created it. For creating web applications, Flask is utilised. The Werkzeug WSGI toolkit and Jinja2 template engine serve as the foundation for Flask.

A well-liked Python templating engine is Jinja2. To create dynamic web pages, a web templating system combines a template with a specific data source.

4. IMPLEMENTATION

As part of this project, we propose the following objectives as our primary solutions to the problem of auto fraud insurance fraud:

We firstly create a classification methodology using a machine learning technique to determine whether a customer is placing a fraudulent insurance claim by using historical data. Then, we designed a web application using flask framework which gives accurate results. It is also accessible and easy to use.

Later, we deployed our application to the backend platform. Here is the interface we created. First, the admin (in real life - the company employee) can login to the portal.

International Journal of Advance Research in Science and Engineering
Volume No. 12, Issue No. 04, April 2023
www.ijarse.com
IJARSE
ISSN 2319 - 8354

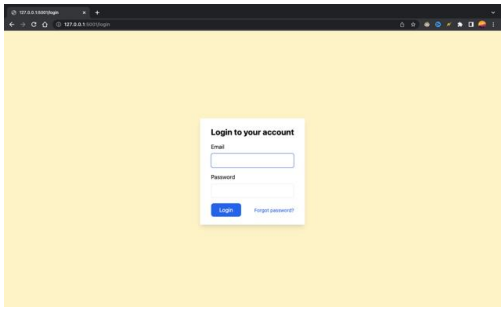

Fig.No.3 Login page

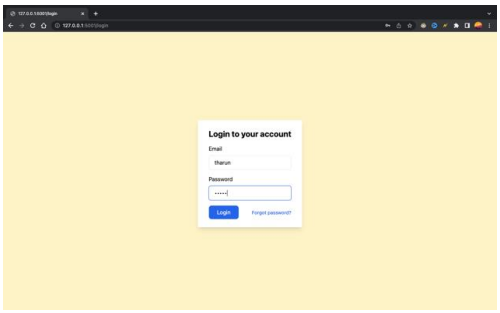

Fig.No.4 Admin logging into the website

Then the admin can upload the csv file which contains the data of all the customers with multiple parameters.

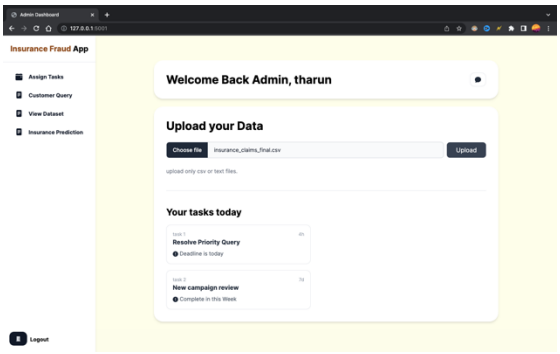

Fig.No.5 Dashboard

Then the admin has to train the data.

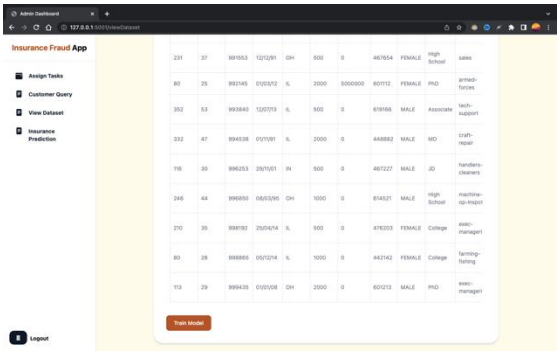

Fig.No.6 Training the ML model.

Now, after the data is trained, the admin can now cross-check the customer's queries and find out whether the customer is a genuine one or a fraud.

International Journal of Advance Research in Science and Engineering
Volume No. 12, Issue No. 04, April 2023
www.ijarse.com
IJARSE
ISSN 2319 - 8354

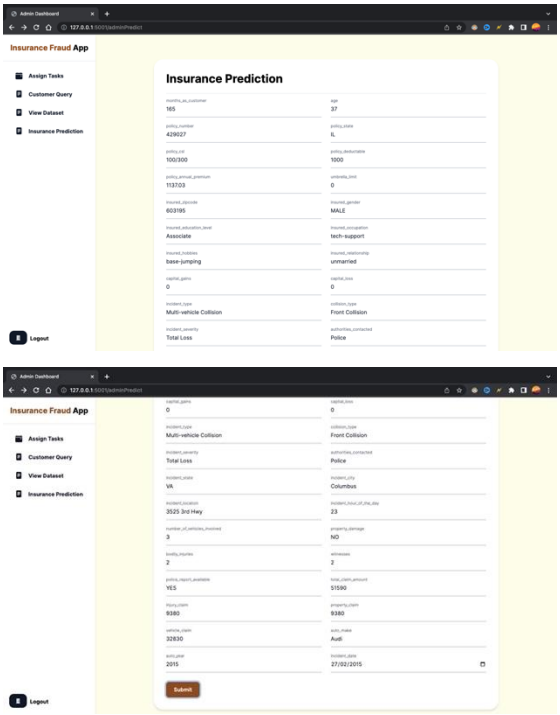

Fig.No.7 Insurance fraud prediction.

5. CONCLUSION

Here, we compared the performance scores of the four methods used to create machine learning (ML) models and make predictions on our dataset experimentally. By using an assessment method, we discovered that Random Forest's accuracy (89.375%) is higher than KMeans (84%), SVM's (75%), DT's (77%) and KNN's (74%) Despite the fact that the confusion matrix revealed an unbalanced distribution of data, we performed several model evaluation parameters.

Based on the results of the assessment phase, we selected Random Forest as the model for this study rather than SVM and other machine learning models. By using the flask framework, a web application has been developed with the selected machine learning model. This web application is hosted on the cloud platform.

6. FUTURE SCOPE

As there is always room for innovations, improvements are a never-ending process in technology, therefore some of the improvements of this Fraud detection. Advanced machine learning algorithms can be used to increase accuracy. The dataset that was uploaded to the system's front end ought to be able to be processed by the system. The system should automatically create a thorough report and notify the in-charge immediately if it discovers fraud. Then, until the issue is rectified, it should halt any further transactions with the consumer. The system should be able to show all previous client transactions together with accompanying notes and graphs.

REFERENCES

[1] Ghorbani, A. and Farzai, S. (2018). Fraud Detection in Automobile Insurance using a Data Mining Based Approach.

[2] Khaled Gubran Al-Hashedi (2019). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019.

[3] Admel Husejinovic (2020). Credit card fraud detection using naive Bayesian and C4.5 decision tree classifiers.

[4] Imane Sadgali, Nawal Sael & Faouzia Benabbou (2019). Fraud detection in credit card transactions using neural networks.

[5] A. Maria Nancya), G. Senthil Kumar, S. Veena, N. A. S Vinoth, and Moinak Bandyopadhyay (2020). Fraud detection in credit card transactions using a hybrid model.

[6] Dejan Varmedja, Mirjana Karanovic (2019). Credit Card Fraud Detection - Machine learning methods

[7] Lakshika Sammani Chandradeva (2020). Monetary Transaction Fraud Detection System Based on Machine Learning Strategies

[8] I.Sadgalian, Saelaf & Benabboua (2019). Performance of machine learning techniques in the detection of financial frauds.