Fraud Detection and Analysis for Insurance Claim using Machine Learning

ABSTRACT

Insurance Company working as commercial enterprise from last few years have been experiencing fraud cases for all type of claims. Amount claimed by fraudulent is significantly huge that may causes serious problems, hence along with government, different organization also working to detect and reduce such activities. Such frauds occurred in all areas of insurance claim with high severity such as insurance claimed towards auto sector is fraud that widely claimed and prominent type, which can be done by fake accident claim. So, we aim to develop a project that work on insurance claim data set to detect fraud and fake claims amount. The project Implement machine learning algorithms to build model to label and classify claim. Also, to study comparative study of all machine learning algorithms used for classification using confusion matrix in term soft accuracy, precision, recall etc. For fraudulent transaction validation, machine learning model is built using PySpark Python Library.

**EXISTING SYSTEM**

Machine learning is usually abbreviated as metric capacity unit. The study of machine learning includes computers with the implicit capability to be trained whereas not being expressly programmed. This capacity unit focuses on the expansion of pc programs that has enough capability to alter, that square measure once unprotected to the new information. Metric capacity unit algorithms square measure generally classified into 3 main divisions that square measure supervised learning, unattended learning and reinforcement learning. Data processing a neighborhood of machine learning has advanced considerably within the current years.

Data mining focuses at analysing the whole data obtained. Furthermore data processing makes an attempt to seek out the realistic patterns in it. On the contrary, within the different of getting the knowledge for world understanding is within the processing applications like machine learning, it uses the knowledge to locate patterns in information and improvise the program actions thereby. Mainly within the supervised machine learning is that the objective of deducing which means from label on the information used for the coaching.

The coaching information consists of a group of coaching samples. Just in case of supervised learning, every instance are often a base which incorporates Associate in Nursing input object that’s considered the vector and also the output features a worth that acts as an indicator to run the model. A supervised learning rule initially accomplishes a groundwork task from the sample information then tries to construct a short lived perform, therefore it will plot new input vectors.

The supervised learning algorithms square measure conspicuously employed in large choice of application areas. Associate in Nursing best setting altogether the chance assist the rule to accurately mark the class labels for close instances and therefore a similar aspires supervised learning rule to chop back from the knowledge to the enclosed objects in terribly good manner.

**Disadvantages**

* The system is not implemented Convex-NMF based Supervised Spammer Detection with Social Interaction (CNMFSD).
* The system is not implemented any ml classifier for test and train the datasets.

Proposed System

The influence of the feature engineering, feature choice parameter modification area unit explored with an aim of achieving superior prophetic performance with superior accuracy. The assorted machine learning techniques area unit utilized in the development of accuracy of detection in unbalanced samples. As a system, the info are divided into 3 completely different segments. These area unit loosely coaching, testing and validation.

The algorithmic program is trained on partial set of knowledge and parameters. These area unit later changed on a validation set. This may be studied for evaluation and performance on the particular testing dataset. The high acting models area unit formerly tested with numerous random splits of knowledge. This helps to confirm the consistency in results the approach discussed above comprises of three layers.

• Data Pre-processing step: In this step, the data is ready in order that are often employed in code with efficiency. Extraction of the dependent and freelance variables from the given dataset. Then the dataset is split as coaching and checking victimisation train test split module from sklearn library. Feature scaling is completed therefore on get correct results of predictions

• Fitting Logistic Regression to the Training set: LogisticRegression category of the sklearn library is employed. Classifier object is made and accustomed work the model to the supply regression Predicting the test result: The model is well trained on the training set, the result is predicted by using test set data.

• Test accuracy of the result: Confusion matrix is employed to judge the check accuracy. In this model of fraud detection, the prediction is completed therefore on check if deceitful dealings is claimed as deceitful and the other way around.

• Visualizing the test set result: Adjust the model fitting parameters, and repeat tests. Adjust the model fitting parameters, and repeat tests. Adjust the options or machine learning algorithmic program and repeat tests.

**Advantages**

• Different models are tested on the dataset once it is obtained and cleaned.

• On the basis of the initial model performance, different features of the model are engineered and tested again.

• Once all the options area unit designed, the model is made and run victimisation completely different completely different values and victimisation different iteration procedures.

• A predictive model is created that predicts if an insurance claim is fraudulent or not.

• Binary Classification task takes place which gives answer between YES or NO. This report deals with classification algorithm to detect fraudulent transaction.

**SYSTEM REQUIREMENTS**

➢ **H/W System Configuration:-**

➢ Processor - Pentium –IV

➢ RAM - 4 GB (min)

➢ Hard Disk - 20 GB

➢ Key Board - Standard Windows Keyboard

➢ Mouse - Two or Three Button Mouse

➢ Monitor - SVGA

**SOFTWARE REQUIREMENTS:**

* **Operating system :** Windows 7 Ultimate.
* **Coding Language :** Python.
* **Front-End :** Python.
* **Back-End :** Django-ORM
* **Designing :** Html, css, javascript.
* **Data Base :** MySQL (WAMP Server).