**Prediction of Insurance Claim with Machine Learning**

****

**What is Machine Learning?**

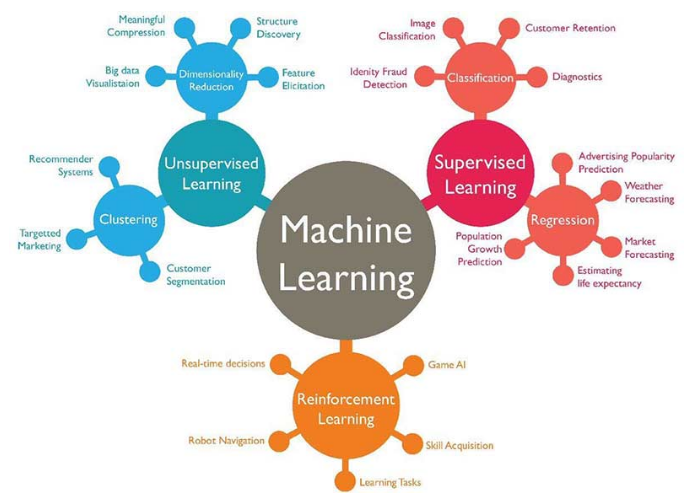
Arthur Samuel, an expert of artificial Intelligence and computer gaming, once stated about Machine Learning as “Field of study that gives computers the capability to learn without being explicitly programmed”.

In another very simple language Machine Learning is automating and improving the learning process of computers based on their experiences without being actually programmed i.e. without any human assistance. The process starts with feeding good quality data and then training the machines by building machine learning models using the data and different algorithms. The choice of algorithms depends on what type of data we have and kind of task we are trying to automate.

**Type of Machine Learning?**

It is classified into 2 types, which is as followed:

1. Supervised Learning:
2. Unsupervised Learning
3. Reinforcement Learning

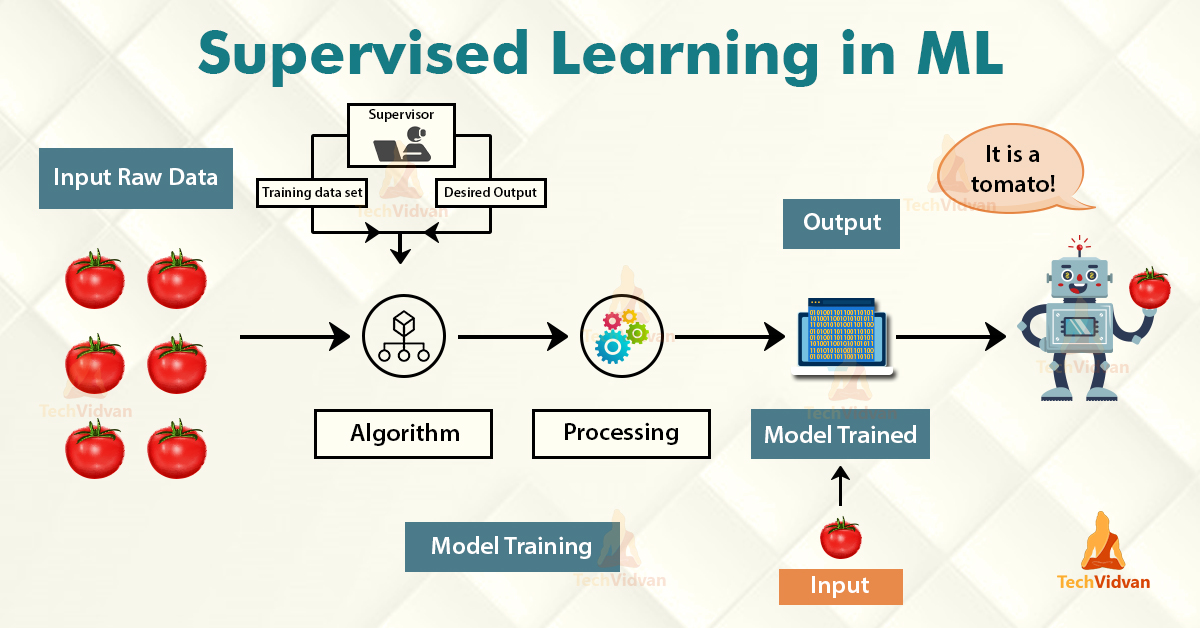


**Overview of Supervised Learning Algorithm**

In supervised Learning, Machine is trained by using Labelled data. Our dataset having both target and independent variable are called labelled. We can understand this more clearly by an example, as in class we have teacher to supervised likewise in this we have one training data to learn and understand the dataset and then as per learning of training data we try to predict the output of test data. This algorithm is very much useful to solve various real world problems.

**Type of supervised Learning**

1. Regression: In this we have problems which are continuous output data or numerical output. For example: age, weight, temperature, salary etc.
2. Classification: In this we have problems which are categorical output data or character output. For example: weather, class, colour etc.

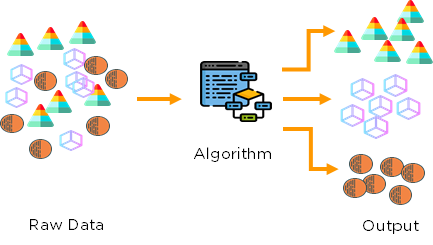


**Overview of Unsupervised Learning Algorithm**

Unsupervised learning does not have labels to work off of, resulting in the creation of hidden structures. Relationships between data points are perceived by the algorithm in an abstract manner, with no input required from human beings. The creation of these hidden structures is what makes unsupervised learning algorithms versatile. Instead of a defined and set problem statement, unsupervised learning algorithms can adapt to the data by dynamically changing hidden structures. This offers more post-deployment development than supervised learning algorithms.

**Type of unsupervised Learning**

1. **Clustering**: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behaviour.
2. **Association**: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.



**Overview of reinforcement learning Algorithm**

It features an algorithm that improves upon itself and learns from new situations using a trial-and-error method. Favourable outputs are encouraged or ‘reinforced’, and non-favourable outputs are discouraged or ‘punished’. We can understand reinforcement learning more clearly by an example as finding the shortest route between two points on a map, the solution is not an absolute value. Instead, it takes on a score of effectiveness, expressed in a percentage value. The higher this percentage value is, the more reward is given to the algorithm. Thus, the program is trained to give the best possible solution for the best possible reward.

**Introduction of our Problem**

Insurance fraud is a huge problem in the industry. It's difficult to identify fraud claims. Machine Learning is in a unique position to help the Auto Insurance industry with this problem.

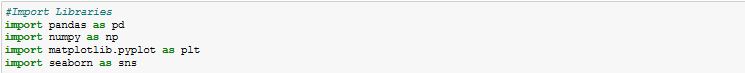
In this project, we are provided a dataset which has the details of the insurance policy along with the customer details. It also has the details of the accident on the basis of which the claims have been made.

In this example, we will be working with some auto insurance data to demonstrate how we can create a predictive model that predicts if an insurance claim is fraudulent or not.

Dataset:

<https://github.com/dsrscientist/Data-Science-ML-Capstone-Projects/blob/master/Automobile_insurance_fraud.csv>

First, we must import all the important libraries which we are going to use for pre-processing of data



After importing the libraries, we will load the dataset on which we have to work on (Load data from above link given as dataset)

2.JPG

3.JPG

From here the job of data scientist starts. Now we start working on dataset. We check the features present in our data and then we will look at their datatypes.

4.JPG

5.JPG

6.JPG

Then we can check different values total counts in column. And also check the null values or missing value.

7.JPG

8.JPG

8.JPG

9.JPG

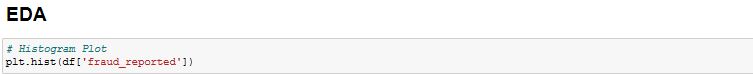
11.JPG

12.JPG

13.JPG

14.JPG

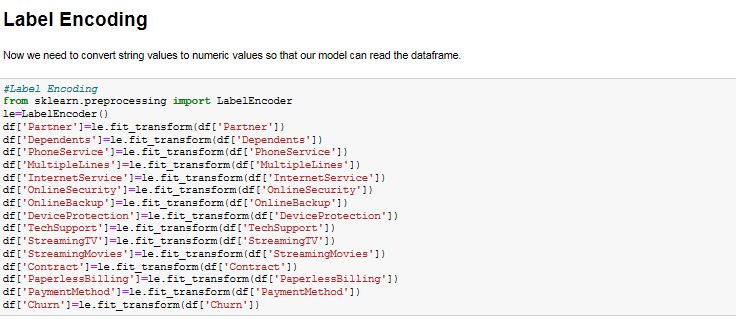
After checking this much we can now plot histogram, bar graph and many more for exploratory data analysis. Plotting of Histogram graph is known as univariate analysis and plotting of Bar Graph, Scatter Graph etc are called bivariate analysis.



16..JPG

17.JPG

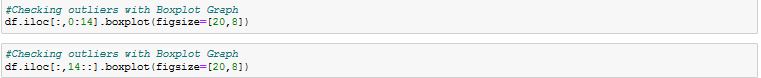
If our dataset has string values and sometimes it is important to convert these string data into numeric values to pass the data from Machine Learning models. In this project, we have many columns having string values which need to convert to numeric values. This process of conversion is known as Label Encoding.

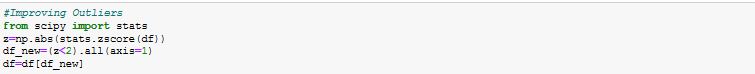


Now we should check correlation between each column, and it is also called Multivariant Analysis. Correlation of any column with itself is always 1.



Now, we can check the outliers present in our dataset by plotting boxplot graphs of each column.





After this much of analysis, we will split our dataset into independent variables X and target variable Y.

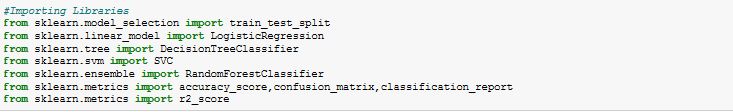
23.JPG

Now we check the skewness of independent variables X. If skewness of each column is between -0.5 to 0.5 then we can take our data forward, otherwise we need to improve that by using any one method among power transformation, log transformation and square root transformation. In our case, we have used Power Transformation.

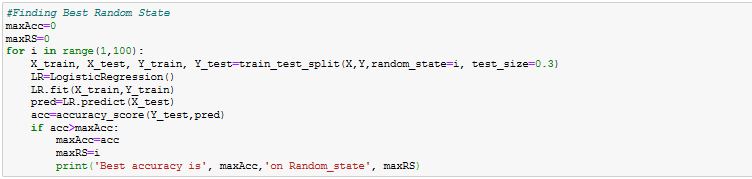
26.JPG

27.JPG

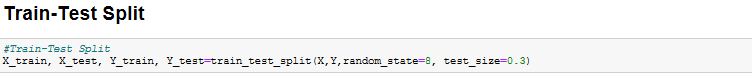
Now we will import all the modules which we are going to use in models.



For getting maximum accuracy we will need to find the best random state. In our case best random state we get is 8.

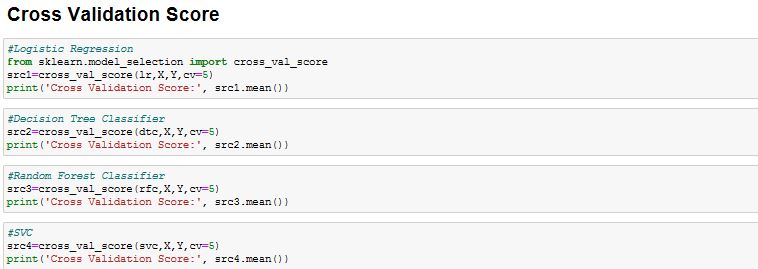


We will use train test split method to predict target values by calling different ML algorithms here we are calling only 4 algorithms, that is Logistic Regression, DecisionTree classifier, RandomForestClassifier and SVC.





After predicting the target values we have to check the overfitting, underfitting of actual data and the predicted data. For this we did the cross validation of all algorithm. And differentiate the accuracy of actual value and the accuracy found during checking cross validation. Algorithm have the minimum value will be expected to find algorithm for predict the test data.



We will save the best ML algorithm for future use.

**34.JPG**

**Conclusion**

From the exploratory Data analysis, we could generate insight from the data. How each of the features relates to the target. Also, it can be seen from the evaluation of 4 models that RandomForestClassifier performed better than other models.

**Summary**

In this blog, I have presented you a modern Data science problem with the basics concepts of Machine Learning and I hope this blog was helpful and would have motivated you enough to get interested in the topic.