Mining Data to Extract Visualize Insights in Python

Manish Gupta[1], Abhishek Shukla[2], Abhishek Kumar[3], Mohd Asad Abbas[4], Alka Singh[5],
[1]*Assistant Professor, Department of Computer Science, Raj Kumar Goel Institute of Technology*
[2,3,4,5]*Student, Raj Kumar Goel Institute of Technology*

Abstract **- As we know, technology plays great role in our life with many conspectus. Machine learning is the most useful technology among all, which provides different type of algorithms, methods, and accurate classification of the dataset. we can broadly say this "a technology era", because in this generation we produce a lot of information every second and all those information is not useful. That is why in this project we are going to learn how to load and how to extract useful information from our dataset and learn how to clean our dataset. This project helps to analyze, summarize, provide concept description and the predictions are based on data collected from daily uses. In this project, we are using different techniques such as classification analysis, k-nearest neighbor, PCA and Random Forest Algorithm for prediction/inference and assumption. After that we can evaluate and compare the same in order to find those which provide the best performance.**

Index Terms **- Prospectus, Machine learning, algorithms, classification, summarize, concept description, 1prediction /inference, assumption, evaluate, performance.**

INTRODUCTION

In this project, our target to load and extract useful information from our dataset using Python, a free, open-source program that we can download. then learn how to clean our data set by removing unwanted whitespaces, columns containing empty values, rows containing empty column values and duplicated row entry.

Next, we will create various visualizations to identify patterns and outliers in our dataset and visualize correlations between different columns.

Lastly, we will learn how to visualize a highly dimensional dataset using principal component analysis (PCA). And visualization gives the patterns and insights which is unexpected. One should seek for insights which can be used to tell stories, and not just expecting the visualization to illustrate a story by itself and also, insights represent very important data like an error in data, nonvisible pattern, therefore, to ensure an effective approach in finding insights from data and visualizations, the followings steps are helpful and can be repeated Predicting the price of used cars. According to the National Transport Authority, the number of cars registered between the year 2003 to 2013 has witnessed a spectacular increase of 234% as the no. of registered car was 68,524 in the year 2003. This number has now reached 160,701 approximately. With difficult financial conditions, we can say that sales of second-hand imported cars and used cars will increase. It is reported in that the sales of new cars have shown a decrease of 8% in 2013. In many developed countries such as UK, US, Japan, Australia etc., it is common to charter a car rather than buying it. A lease is a binding contract between the parties to it such as a buyer and a seller (or a third party – usually a bank, insurance firm or other financial institutions) in which the buyer must pay fixed instalment for a pre-defined number of months/years to the financer. After the lease period is over, there is a possibility that the buyer may buy the car at its residual value. Thus, it is in the commercial interest to make the predictions. Keywords-car; price; machine learning; artificial intelligence and financers to be able to predict the salvage value (residual value) of cars with accuracy. If the residual value of the car is under-estimated by the financer at the beginning, the instalments will be big for the clients who will certainly then opt for another financer. If the residual value is over-estimated, the instalments of the car will be lower for the clients but then the financer may have much difficulty at selling these high-priced used cars at this over-estimated residual value. Thus, we can see that estimating the price of used cars is of highly commercial importance as well. Manufacturers from Germany made a loss of 1 billion Euros in US market due to miscalculation of the residual value of leased cars. Most of the individuals in Mauritius while buying a new car remains in apprehension about the resale value of their

cars after some years when they will sell it in the used cars merchandise. Predicting the resale value of a car is tough row to hoe. As it is widely known that the value of used cars depends on several factors. The most important ones are usually the age of the car, model no., the origin of the car (the manufacturer country), mileage (the number of kilometers it has run) and horsepower. Due to hike in fuel prices, fuel economy is also considered of prime importance. Unfortunately, most people do not know exactly how much fuel their car consumes per km driven. Other factors such as the what type of fuel it uses, the interior style, the braking system, acceleration, the volume of its cylinders (measured in cc), safety index, size, number of doors, quality and colour of paint, weight of the car, consumer reviews, prestigious awards won by the car manufacturer, physical state, whether it is a sports car, whether it has cruise control, whether it is automatic or manual transmission, whether it belongs to an individual or a company, air conditioner, sound system, power steering, cosmic wheels and GPS navigator all these factors may influence the price of its resale . Some special factors which in Mauritius buyer finds of equal importance are local of previous owners, whether the car had been involved in serious accidents and whether it is a lady-driven car and many other factor comes. The look and feel of the car certainly contribute a lot to its price. As we can see, the price of resale or used cars also depends on many factors. Unfortunately, information about all these factors are not always available and the buyer have to make his decision of purchase on few factors only. In this project, we are considering only a small subset of the factors mentioned above. In the next section, a review of related work is being provided. Section III describes the methodology while in section IV we are describing, evaluating and comparing different machine learning techniques to predict the price of used machines. Finally, end the paper here with a conclusion and some pointers towards future work.

LITERATURE SURVEY

Data is categorized into dimensions in terms of the 3Vs, which are referred to as volume, velocity and variety. Also, speed at which the data comes has become so fast that traditional data analytical tools cannot handle them properly and may breakdown when used. Also, the increase in volume has made the extraction, storage, and pre- processing of data more difficult and challenging as both analytical algorithms and system must be scalable in other to handle it. Data has been coined to represent this outburst of massive data that cannot fit into traditional database management tools or data processing applications. These data are available in three different formats such as structured, semi-structured and unstructured format and the sizes are in scales of terabytes and petabytes. Lastly, the ever-changing variety of data and its numerous sources of integration makes the storage and analysis of data difficult. If we look in our today's world, over 90% of the data in the world was generated in approximately two years. This shows that data has really come to stay and therefore new research and studies must be carried out to fully understand the massive data. This means there must be a significant shift from past theories, technologies, techniques and approaches in data mining and analysis in order to fully harness these data. Thus, we need proper data visualization and mining to cope with these large amounts of data generated every day. The growth of data has been exponential, and from the perspective of information and communication technology, it contain key to better and robust products for businesses and firms.

LITERATURE REVIEW

Many statistical tools like SPSS, STATA, etc., are popular for data analysis but the choice of which to use usually varies among data analysts. The SPSS (Statistical Package for the Social Sciences) is widely used tool in the field, though it was originally built for the social sciences. SPSS software is now used by researchers, survey companies, health researchers and others. Data produced is quite large, hence the information resident in it must be mined, captured, aggregated, and visualized by data scientists. In order to fully carry out these tasks effectively, data analysts are expected to have a specific kind of knowledge and to leverage powerful data analytics tools. There exists lots of tools for big data mining and these are broadly categorized into three groups, statistical tools, programming languages and visualization tools.

Programming language Python is famous for data analysis and data mining. The dynamic and interactive nature of these languages combined with the abundance of scientific libraries make them a

preferred choice for analytical tasks While R is still the most popular language for traditional data analytical tasks, the usage of Python language is also high.

An important aspect of data analysis is visualization, to this end, numerous tools and software has been built to aid effective data visualization. Most of the programming languages like Python and R has their own plotting packages some of which are R's ggplot, Python's matplotlib, seaborn, bokeh, plotly, etc.

There has also been massive development of GUI visualization tools, and some popular ones are Tableau, Power Bi, QlikView, etc.

Data visualization also uses a technique Principal component analysis (PCA). It is generally a process of computing principal component and using them for various purposes.

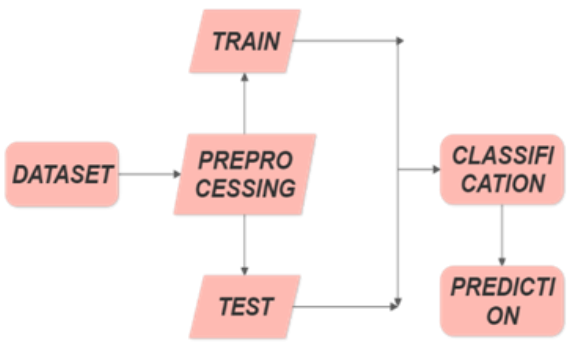
Data analysis is the process of inspecting, transforming, and modelling data with the purpose of discovering useful information, getting actionable insights, and informing conclusions. From these techniques, we can visualize our data using python.

Methodology

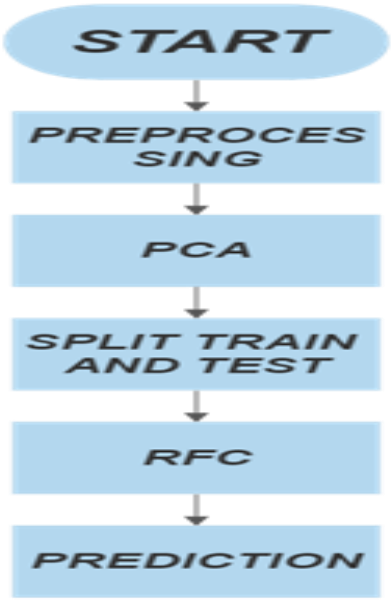
In this project we are done different works like data mining, extraction, data visualization, insights etc.so that after analyzing a complexity and flexibility of different machine learning algorithms and also surveying a number of research papers , the following methodologies have been identified.

- System Architecture
- Flow Diagram
- Data Selection and Loading
- SPLITTING DATASET INTO TRAIN AND TEST DATA
- PCA (Principal Component Analysis)
- Random Forest Algorithm

System Architecture



Flow Diagram



Data Selection and Loading

The data selection is the process of selecting the Dataset. which contains both useful and non-useful data. After the selection of data, we load the data by using pandas and extract the useful data therefore dataset contains the information about test, train valid data.

```
import pandas
filename = 'data.csv'
names = ["Year', 'Engine HP', 'Engine Cylinders', 'Number of Doors', 'highway MPG', 'city mpg', 'Popularity', 'MSRP"]
data = pandas.read_csv(filename, names=names)
print(data.shape)
```

SPLITTING DATASET INTO TRAIN AND TEST DATA

Data splitting is act of partitioning present data into two portions, usually for cross-validator purposes. One Part of the data is used to develop a predictive model and the other part is used for evaluating the model's performance. Separating dataset into training and testing sets is an important part of evaluating data mining models. after separation of a dataset into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing. Train Dataset is Used to fit the machine learning model and test dataset is Used to evaluate the fit machine learning model.

```
>>> import numpy as np
>>> from sklearn.model_selection import train_test_split
>>> X, y = np.arange(10).reshape((5, 2)), range(5)
>>> X
array([[0, 1],
       [2, 3],
       [4, 5],
       [6, 7],
       [8, 9]])
>>> list(y)
[0, 1, 2, 3, 4]

   >>> X_train, X_test, y_train, y_test = train_test_split(
   ...     X, y, test_size=0.33, random_state=42)
   ...
   >>> X_train
   array([[4, 5],
          [0, 1],
          [6, 7]])
   >>> y_train
   [2, 0, 3]
   >>> X_test
   array([[2, 3],
          [8, 9]])
   >>> y_test
   [1, 4]

>>> train_test_split(y, shuffle=False)
[[0, 1, 2], [3, 4]]
```

PCA (Principal component analysis)

Principal component analysis is a statistical procedure that allows you to summarize the information content in large dataset tables. The smaller set of "summary indices" that can be more easily visualized and analyzed.

The underlying data can be measurements describing properties of production samples of dataset, reactions, process time points of a continuous process, batches from a batch process.

Organizing information in principle component analysis this way, will allow you to reduce dimensionality and reduce complexity without losing big amount of information, and discarding the components with low information and considering the remaining components as your new variables of dataset.

The principal component analysis are less interpretable and do not have any real meaning since they are constructed as linear combinations of the initial variables.in other word principal components represent the directions of the data that explain a maximal amount of variance, that to say, the lines that capture most information of the data.

```
import matplotlib as mpl
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from sklearn import base
from sklearn import preprocessing
from sklearn import utils

from . import plot
from . import svd


class PCA(base.BaseEstimator, base.TransformerMixin):

  def __init__(self, rescale_with_mean=True, rescale_with_std=True, n_components=2, n_iter=3,
         copy=True, check_input=True, random_state=None, engine='auto', as_array=False):
    self.n_components = n_components
    self.n_iter = n_iter
    self.rescale_with_mean = rescale_with_mean
    self.rescale_with_std = rescale_with_std
    self.copy = copy
    self.check_input = check_input
    self.random_state = random_state
    self.engine = engine
    self.as_array = as_array

  def fit(self, X, y=None):

    # Check input
    if self.check_input:
      utils.check_array(X)

    # Convert pandas DataFrame to numpy array
    if isinstance(X, pd.DataFrame):
      X = X.to_numpy(dtype=np.float64)

    # Copy data
    if self.copy:
      X = np.array(X, copy=True)

    # scikit-learn SLEP010
    self.n_features_in_ = X.shape[1]

    # Scale data
    if self.rescale_with_mean or self.rescale_with_std:
      self.scaler_ = preprocessing.StandardScaler(
        copy=False,
        with_mean=self.rescale_with_mean,
        with_std=self.rescale_with_std
      ).fit(X)
      X = self.scaler_.transform(X)
```

Random Forest Algorithm

A Random Forest is an ensemble learning technique capable to performing both regression and classification tasks with the help of multiple decision trees and a technique called Bootstrap and Aggregation, called bagging. The basic idea behind this algorithm is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

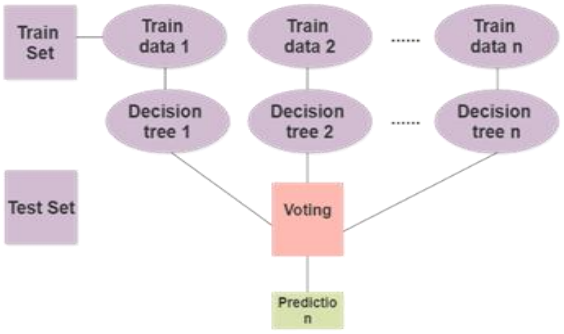
Random Forest contains the many decision trees as base learning models. We randomly execute row sampling and feature sampling from the dataset forming sample datasets for every model called Bootstrap.

It output's the class which is mode corresponding to the classes or average prediction or mean of the different trees. This methodology is better as compared to decision trees. Random forest algorithm is used as Black Box models especially in business. because they retrieve reasonable inference traverse a

huge data range while requiring small configuration in a scikit learn.

The main advantage of random forest is it can be utilized for classification as well as regression issues that form a major part of the machine learning systems. The hyper parameters in random forest are either utilized to enhance the predictive power and enhance the speed of the model.

Working of RFT



As we know random forest is a combination of multiple trees to predict the class of the dataset, it is possible that some decision trees may be gives the correct output, or not. But together, all the trees give the correct output. So that, there are two assumptions for a better Random Forest classifier

1. There should be some actual values present in the variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
2. The predictions should be each tree must have very low correlations.

CONCLUSION

In this project we have done many tasks like data mining, data extraction, data visualization etc. Basically, this project contains many information like how to load your data and how to extract your data, analysis, summarize, provide concept description, predict the success rate of cars with help of machine learning algorithms. After evaluating model on test dataset, each of these algorithms obtained a precision rate between 70% and 80%.

REFERENCES

[1] Pudaruth, Sameerchand. "Predicting the price of used cars using machine learning techniques." Int. J. Inf. Computer.

[2] Monburinon, Nitis, Prajak Chertchom, Thongchai Kaewkiriya, Suwat Rungpheung, Sabir Buya, and Pitchayakit Boonpou. "Prediction of prices for used car by using regression model." In 2018 5th International Conference on Business and Industrial Research (ICBIR). IEEE, 2018.

[3] Gegic, Enis, Becir Isakovic, Dino Keco, Zerina Masetic, and Jasmin Kevric. "Car price prediction using machine learning techniques." TEM Journal 8, 2019.

[4] Noor, Kanwal, and Sadaqat Jan. "Vehicle price prediction system using machine learning techniques." International Journal of Computer Application 167, no. 9 (2017): 27-31

[5] NATIONAL TRANSPORT AUTHORITY 2014. Available from: http://nta.gov.mu/English /Statistics /Pages/Archive.aspx [Accessed 15 January 2014].

[6] MOTORS MEGA. 2014.http://motors.mega. mu/news/2013/12/17/auto-market-8-decrease-sales-newcars/ [Accessed 17 January 2014].

[7] LISTIANI, M., 2009. Support Vector Regression Analysis for Price Prediction in a Car Leasing Application. Thesis (MSc). Hamburg University of Technology.

[8] RICHARDSON, M., 2009. Determinants of Used Car Resale Value. Thesis (BSc). The Colorado College.

[9] WU, J. D., HSU, C. C. AND CHEN, H. C., 2009. An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. Expert Systems with Applications. Vol. 36, Issue 4, pp. 7809-7817.

[10] DU, J., XIE, L. AND SCHROEDER S., 2009. Practice Prize Paper - PIN Optimal Distribution of Auction Vehicles System: Applying Price Forecasting, Elasticity Estimation and Genetic Algorithms to Used-Vehicle Distribution.

[11] Data Mining: A Brief Overview and Recent IPSI Research Radivojevic, Zaharije, Cvetanovic, Milos and Milutinovic, Veljko (2006)

[12] Development of Data Mining Systems to Analyze Cars using TkNN Clustering Algorithm International Journal of Advanced Research IN Computer Engineering & Technology (IJARCET) Volume 3, Issues 7, (July 2014).