A Machine Learning Approach to Predict Crime

By

K Ajay Kumar Reddy

A Krishna Vamshi

G Koushik

**Contents:**

**1.1-Introduction**

**1.2 Objectives of Research**

**1.3    Problem Statement**

**2. Review of literature**

**3.  Data Collection**

**4. Methodology**

**4.1 Exploratory Data Analysis**

**4.1.1 Figues and Tables**

**4.2 Data Modelling**

**5. Findings and Suggestion**

**6.Conclusion**1.1 Introduction

The project is developed using Python and Machine Learning. **Python** is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

**Machine** **learning** is the science of getting computers to act without being explicitly programmed. In the past decade, **machine learning** has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome.

1.2 Objectives of Research

Criminal activities are present in every region of the world affecting quality of life and socio-economical development. As such, it is a major concern of many governments who are using different advanced technology to tackle such issues. Crime Analysis, a sub

branch of criminology, studies the behavioural pattern of criminal activities and tries to identify the indicators of such events.

Machine learning agents work with data and employ different techniques to find patterns in data making it very useful for predictive analysis. Law enforcement agencies use different patrolling strategies based on the information they get to keep an area secure. A machine learning agent can learn and analyse the pattern of occurrence of a crime based on the reports of previous criminal activities and can find hotspots based on time, type or any

other.

1.3 Problem Statement

In this research, we have considered crime data from Chicago that contains attributes such as type of crime, hour, day of week, month, year and the district where the crime takes place. Objective behind the collected data is to predict the type of crime that takes place considering the following set of columns district, hour, day of the week, month and year.

2.Review of Literature

As combating criminal activity has always been a priority for governments around the world, many researches has been done to effectively find countermeasures and indicators of

crime prior to happening. Criminologists have been pursuing to identify hotspots that need major attention from law enforcement agencies.

Researches have been done to study the relation between criminal activities and socio-economic variables like unemployment, income level, race, level of

education.

Analysing the usage of mobile network infrastructure and demographic information of people living in different areas of Chicago, a group of researchers were able to predict if particular areas of Chicago would become a criminal hotspot. They have implied that anonymized data collected by mobile networks contain indicators for predicting crime levels.

3. Data Collection

Data Collection involves gathering, collecting and measuring information on targeted variables in an established system which then enables one to answer relevant question and evaluate outcomes.

Initially we import the required packages for data manipulation and exploration. The packages that are imported are :

1. Pandas: It is the basic package to handle data manipulation. It is also used to import our dataset into the python for data Analysis.

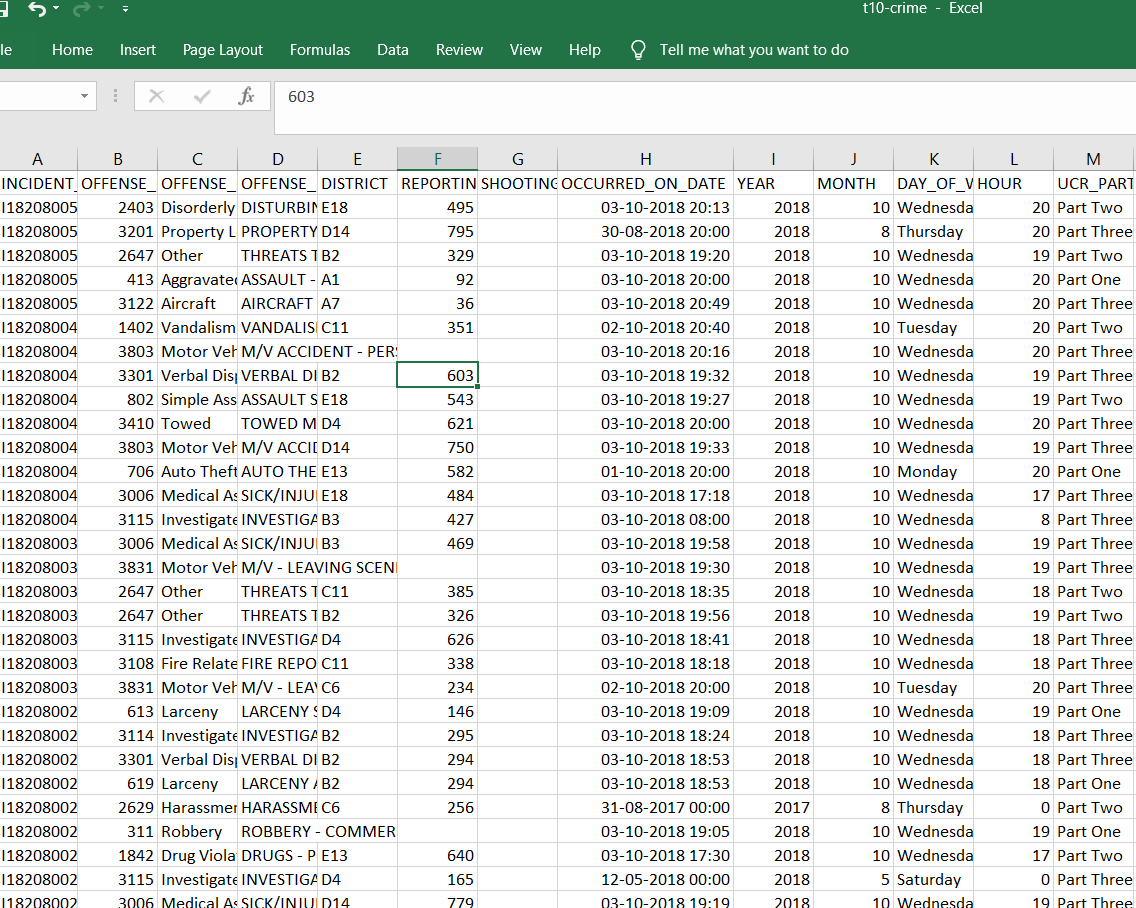
2.MatPlotLib: It is used for graphical representation of the data which would help in knowing how one dependent variable varies with independent variable.

3.Seaborn: It is advanced package of Matplotlib used to improve the representation between the predictors and target variables.

4.Numpy: It is used to create the arrays which only contains uniform data and contains various mathematical operation that can be applied on datasets.

As described above using pandas read\_csv() we import our dataset into python file. The dataset name that we used is ‘Crime\_rate \_Analysis’ which consist 372800 rows and 17columns.

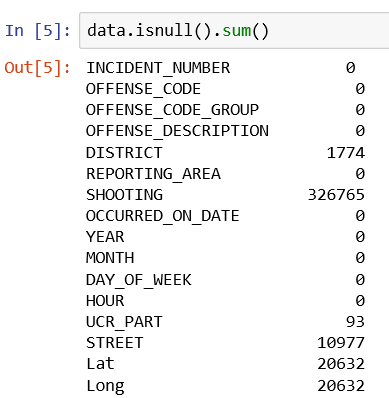
The following shows the columns and rows of the data set.



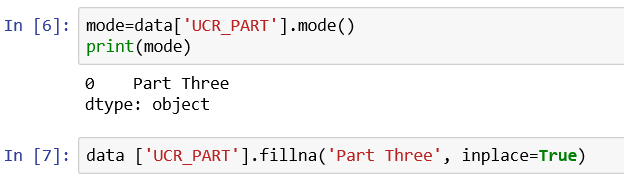
2.Methodology

The First in predicting the output using Linear Regression involves

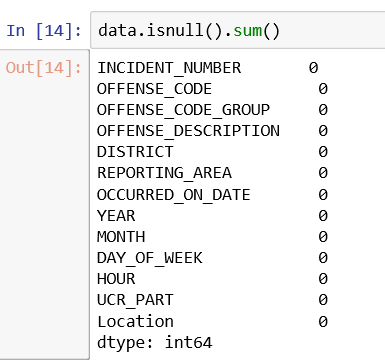
“Data Cleaning”. Data Cleaning involves removing null values, removing Outliers normalizing the data. The dataset contain null values which is to be refined.



The following method can be used to refine the data i.e fill the null values.



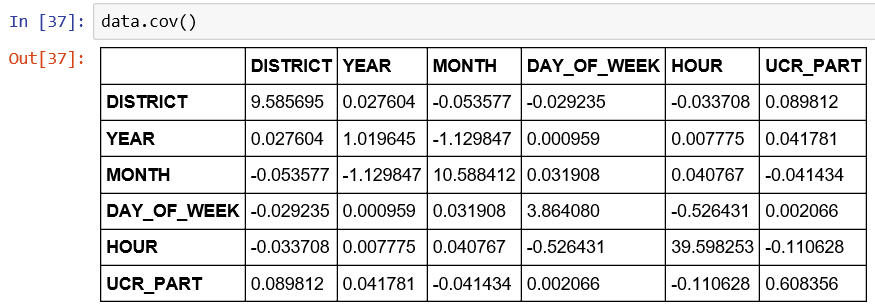
After redefining the Null values with appropriate values.



Exploratory data analysis

It is most important and useful step in predicting the model and creating the model. Exploratory data analysis involves finding the relation between various columns and target so as to decided which columns need to considered as predictors. First step in Exploratory Data Analysis is to find Co Variance. 1. Co Variance Co variance helps us to find whether there exists a relationship between the columns. It only gives magnitude of dependency not the direction

Command used to find Co Variance



2. Co-Relation: It gives how the two columns are dependent on each other that is magnitude and direction of dependency. Following inferences can be obtained from Co Relation Values

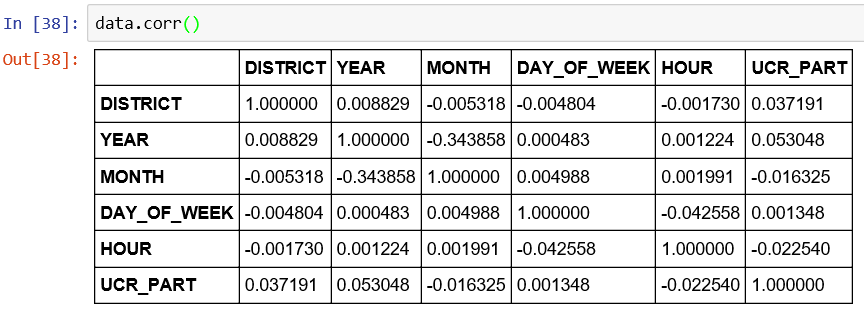
1. 0.0 -0.5 Weak Positive Dependency

2. 0.5-1 Strong Positive Dependency

3. -0.5 – 0.0 Weak Negative Dependency

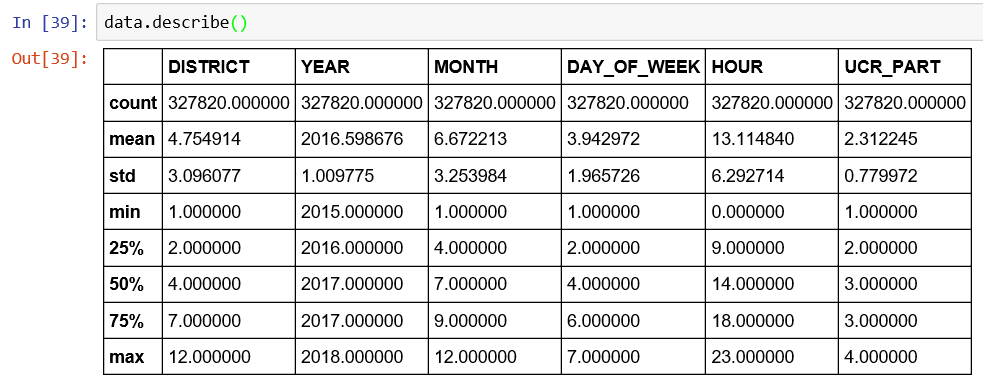
4. -1 - -0.5 Strong Negative Dependency

5. 0 no Relation Command used to calculate Co Relation



3. Descriptive Statistics: these statistics refer to displaying basic information like minimum, maximum, mean, standard deviation and quartile information.

Command to display Descriptive statistics

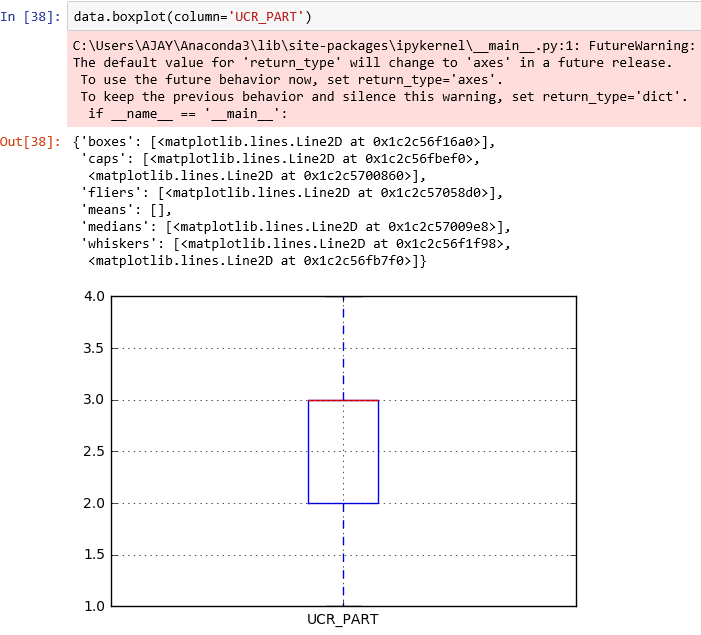


4. Exploration Data Analysis using Data Visualization Tools

4.1 Figures and Tables

Boxplots: The range of the column is represented in form of box with two ends. To read Boxplot, first there is a line at the centre, this indicates median value of all the data points. Median is the value at the centre when you sort the data from the smallest to the largest. To deal with high records or observations, boxplot gives the clear information. In this project, when we compare two features we will get lower whisker which points to lower scale and upper whisker which corresponds to maximum sales value. In below boxplot lower whisker and upper whisker is increasing slightly when compare to other quartiles and lower hinge, median and upper hinge is increases.

Boxplots of various predictors

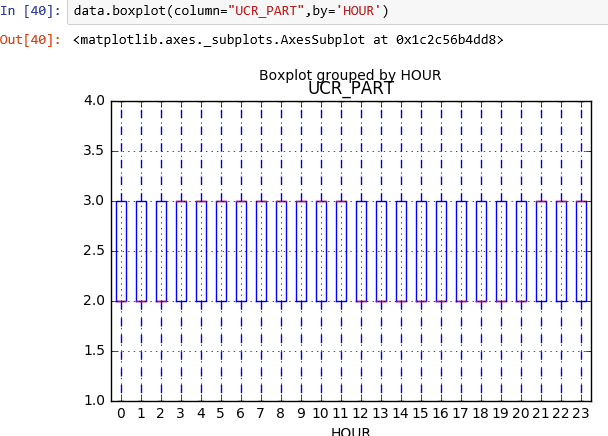


Inferences:

1.Most of the crimes are in the category of UCR part Three

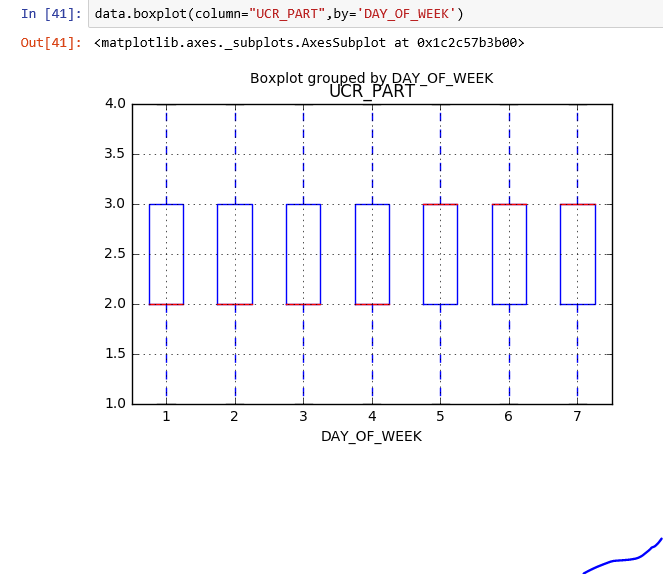
and hence the median.

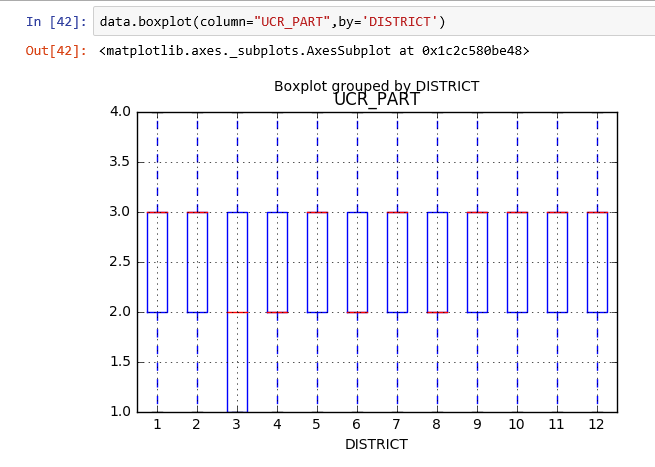
Box plot on more than a column

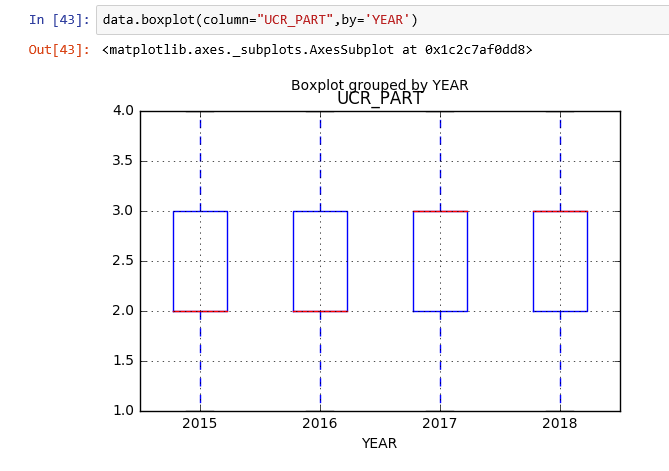


Inference:

1.Most of the median is either part two or part three



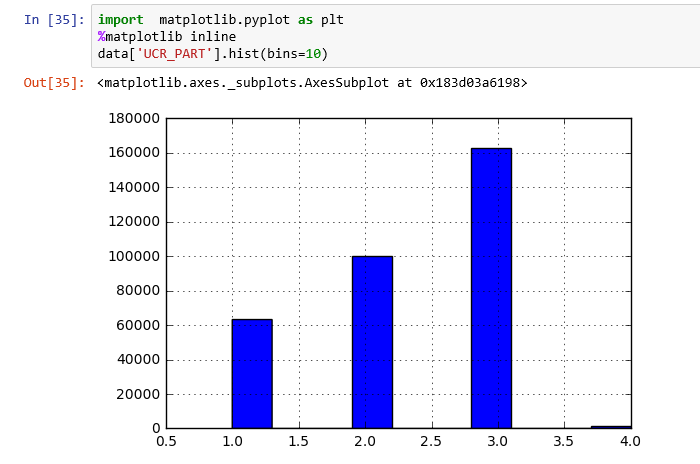




Inference:

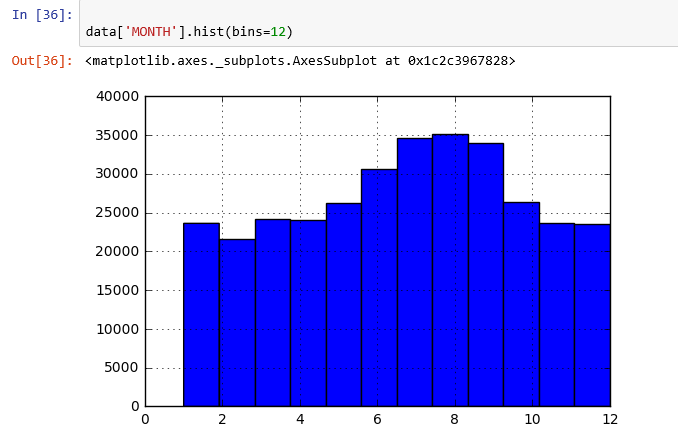
1. Median is either part two or part three i.e The crimes that mostly occur belong to either category two or three.
2. Histograms

A histogram is an accurate representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable (CORAL) and was first introduced by Karl Pearson. It differs from a bar graph, in the sense that a bar graph relates two variables, but a histogram relates only one. To construct a histogram, the first step is to "bin" (or "bucket") the range of values—that is, divide the entire range of values into a series of intervals—and then count how many values fall into each interval. The bins are usually specified as consecutive, non-overlapping intervals of a variable. The bins (intervals) must be adjacent, and are often (but are not required to be) of equal size.



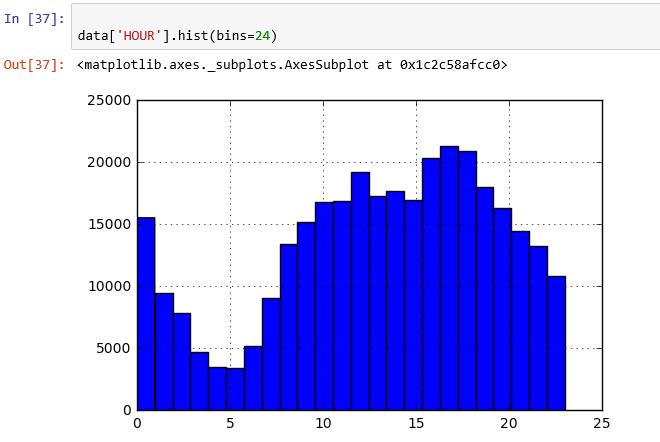
Inference:

1.As the UCR part increases count also increases but decreases for UCR part four



Inference:

1.There is even distribution of count with the months.

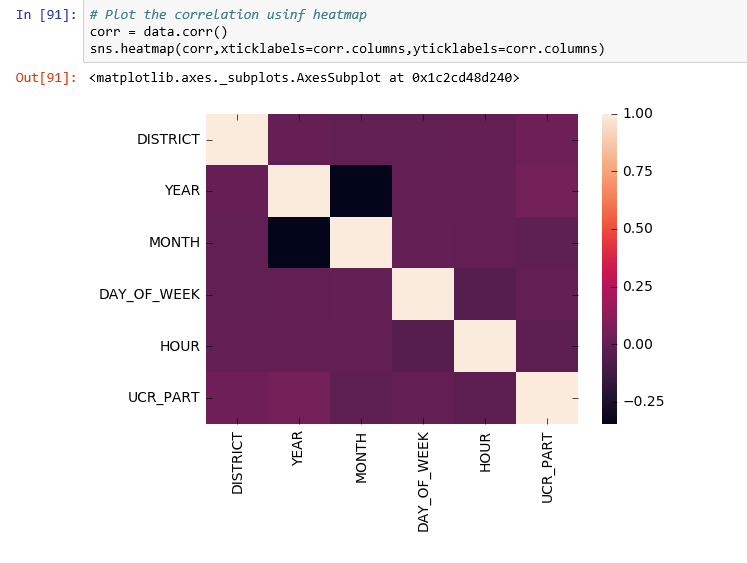


Inference:

1.There is uneven distribution i.e continuous raise and fall of count with the months.

HeatMap:

A heat map (or heatmap) is a graphical representation of data where the individual values contained in a matrix are represented as colors. "Heat map" is a newer term but shading matrices have existed for over a century.

Inference: 1. All independent attributes show a lesser relation with the dependent variable .

5.1.6

Multiple Linear Regression Model

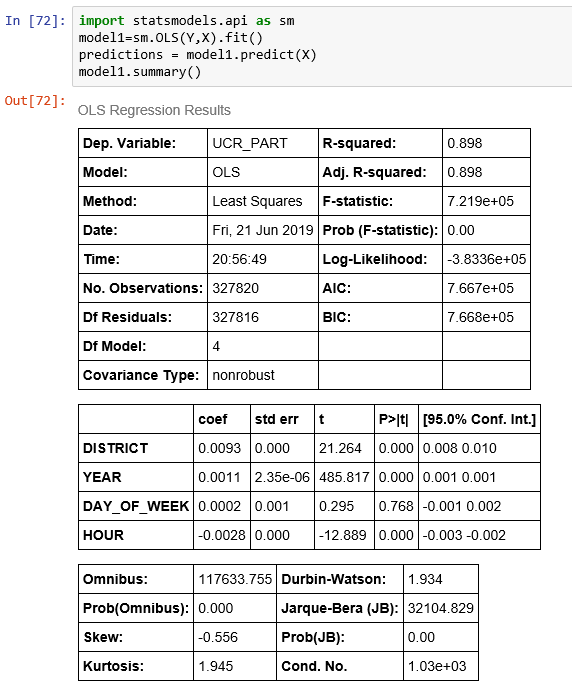
1. Predictors: District, Year, Day of Week, Hour

Research Target: UCR part

Method Stats Model

Inference:

The input variables show affect on the output variable.

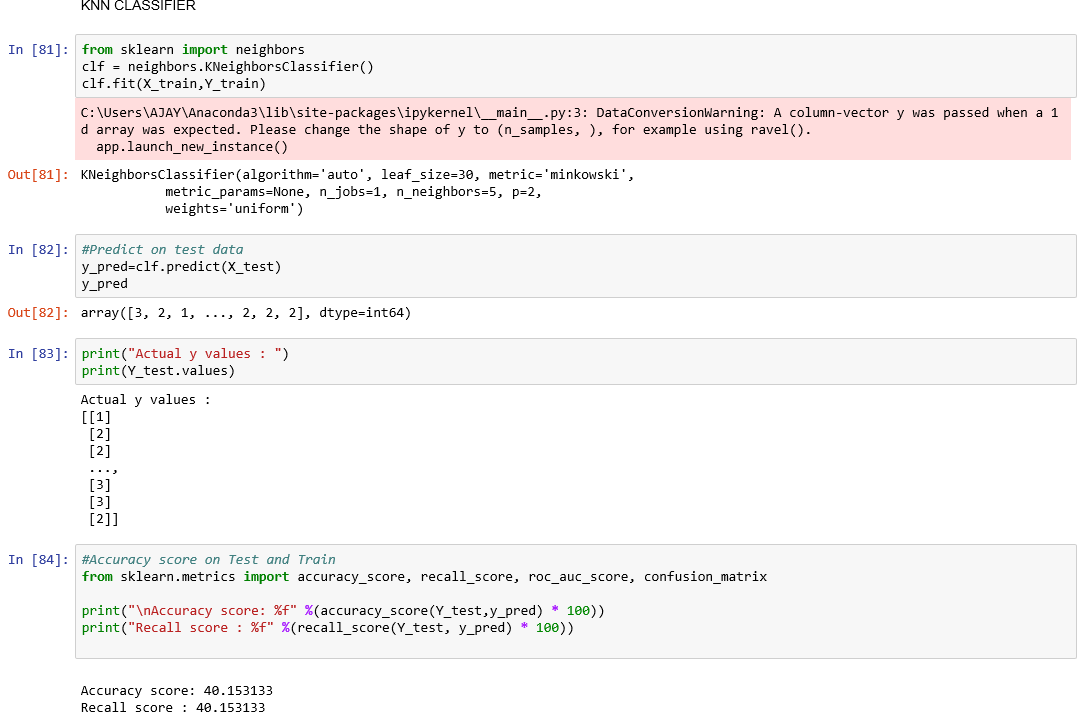


DATA MODELLING

1.KNN Classifier

In pattern recognition, the k-nearest neighbors algorithm is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression: In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours.

Training and testing the Model



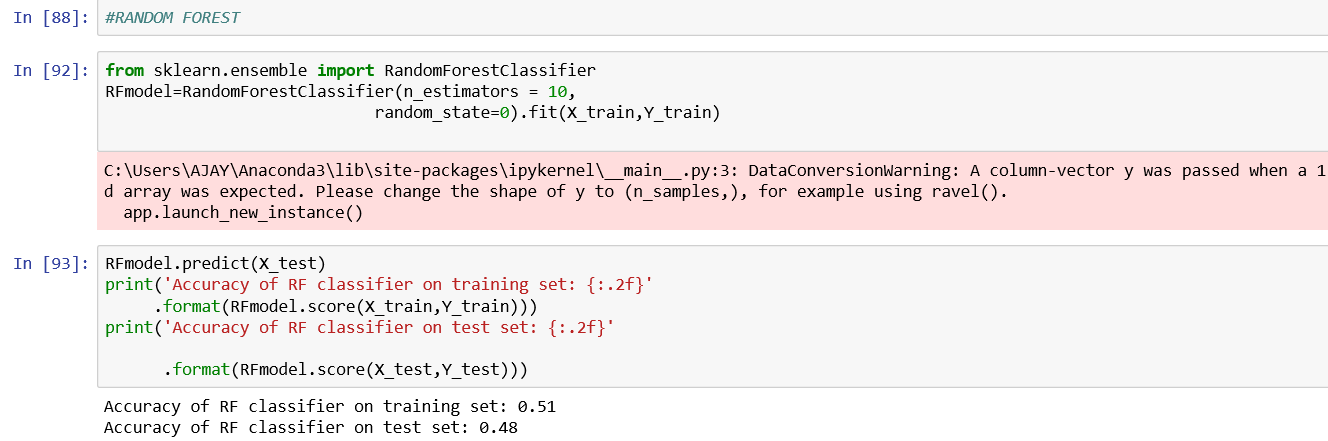
Inference :

The accuracy of KNN model is 40%.

Random Forest Classifier

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

Training and Testing the model



Inference:

The accuracy of the Random Forest Model is nearly 51%.

So, We conclude that the accuracy among KNN, Logistic Regression, Random Forest Classifier the accuracy is greater for Random Forest with an accuracy of 51 percent

Findings and Suggestions:

Actually we are given a dataset which tells us about “the factors which predict the crime in Chicago ”. The factors are ‘hour, day of week, month, year, UCR part etc. Actually total dataset is both numerical and categorical we convert required categorical to numerical and easily we can predict the input and output variables by seeing data so it is “SUPERVISED LEARNING”.

First we need to ‘import OS’ in order to get the current working directory to read the data from the given dataset. After reading data from the given file we checked whether there are null values or not using ‘LAMBDA’ function. There are no empty values in the given dataset so ‘Data cleaning’ is not required. Later we checked what are the features and labels in the given dataset and removed unwanted variable ‘Location’ and ‘incident number’ .

Conclusion

With the help of machine learning technology, it has become easy to find out relation and patterns among various data. The work in this project mainly revolves around predicting the type of crime which may happen if we know the time, day and district of where it has occurred. Using the concept of machine learning we have built a model using training data set that have undergone data cleaning and data transformation. The model predicts the type of crime with accuracy of 0.51.

Data visualization helps in analysis of data set. The graphs include bar, histograms and boxplot each having its own characteristics. We generated many graphs and found interesting statistics that helped in understanding Chicago crimes datasets that can help in capturing the factors that can help in keeping society safe.