Data Mining and Analysis: Crime rates in India

R Brahmankar, Master of Information Technology student, University of Auckland, New Zealand

Abstract - As we can see, Crime rates are increasing daily. In the modern world, where the growth of technology has led to many marvels, it is helping ill-minded people achieving their misdeeds as well. According to Indian Police, Crime Records, the crimes like burglary, theft, etc. have decreased a bit because with the help of technologies like CCTVs, GPS tracking, etc. case-solving has been simplified. On the other side, crimes like murder, rape, sex abuse, etc. have been increased. Even though we cannot predict who might be the victims of crime, but we can predict other factors like place, time of its occurrence, etc. I started my data mining process to contribute something towards sustainable development and hence ended up with the analysis of crime rate prediction to make the world a better place with my tiny contribution. With the vast increase in the availability of criminal data, we can predict the patterns of crimes such as the age group being targeted, place where the chances of being a victim of a crime are more, etc. But with the information of the criminal data and data mining on it, we also don't want criminals to get benefited from it. Determining the crime patterns will not only help people with safety but also help police with the information so that they can increase security for some places, police patrolling can be improved

———————◆———————

1. INTRODUCTION

Crime degrades living standards. It restricts travel and thus hinders access to future jobs and educational opportunities; it also discourages the accumulation of assets. It retards business and other economic activity because crime puts people at risk of being averse. Crime is often more 'costly' for poor people in developed countries, as it can lead to the loss of medical costs and productivity that developed people in developing countries are ill prepared to bear.

“In 2006, the maximum crime rate for crimes under Indian Penal Code was recorded in Puducherry (447.7), which is 2.7 times the national crime rate of 167.7. Kerala posted the highest crime rate among all of India's states at 312.5. The only metro cities of Kolkata (71.0) and Madurai (206.2) were the only Metropolis that recorded fewer crime rates than their West Bengal (79.0) and Tamil Nadu (227.6) respective states. Delhi, Mumbai and Bangalore accounted for 16.2 percent of the overall IPC crimes registered from 35 mega cities, 9.5 percent and 8.1 percent, respectively. Indore reported the highest rate of crime (769.1) among India's mega cities followed by Bhopal (719.5) and Jaipur (597.1)” (Accidents & Accidents, 2015)(Edwardes, 2008)

As the above statistics show the rise in the crime indifferent parts of India, I have encountered the attributes that can be the decision factors for such cases. And as discussed earlier, preventing crime can be the ultimate goal in refining the environment. Also, as discussed by the governing body in the united nations, the sustainable development is a result of overall development and hence for development in the all the fields of humanity it needs to be determined that what are the factors we still need to work on and what factors can be changed by using the mordent technology and bringing the world to a better place that what it is. The ultimate goal of my data mining part from my business objective mention

below would be to use the technology and many advanced inventions in the modern world to have a positive effect on humanity and working towards the betterment of the same, resulting into the advancement of the place where we live and contribute in any smallest way possible to make this world a better place.

The above issues incorporate to my business objectives:

- Detecting factors that affects increase or decrease in crime rate in India.
- Exploring attributes that can contribute in the change of crime rates.
- Using data mining to calculate level of importance of the attributes I chose that can affect crime rate.

As mentioned in the above points, my main objective is to detect the factors that affect in the increase or decrease of the crime rate in India. In achieving the same, I gathered data from various sources. I chose attributes that I thought might affect the crime rate and change analysis of the crime rate detection. In this process, I use data mining models for predicting the importance of the factors that I chose in my dataset and iteratively I can evaluate or discard if some factors do not prove useful according to my data mining and analysis.

2. LITERATURE REVIEW

3. METHODOLOGY

4. DATA HANDLING AND ACTIONS

4.1 DATA COLLECTION

Data needed for the criminal records is most of the time confidential as it involves the security threat of data leak, which can be misused by the ill-minded ones. Datasets for criminal records can be found online; hence I can use the existing data. To fulfil my data requirements, I had to collect the data from various sources. I used https://www.worldometers.info/world-population/india-population/ to get the population of the country, where I could find the population of the country from the time span of 1990 to 2020. Here, when explored further, I got the population for each month distributed across the page and had to sort the columns accordingly. For getting the data of average age of people in India and census of the same for the given period, I explored this website and similar to the population, I got https://www.statista.com/statistics/254469/median-age-of-the-population-in-india/ the data for average age distributed monthly. To collect the data regarding the crime rate and police rate, I gathered some data from the website https://www.macrotrends.net/countries/IND/india/crime-rate-statistics , where I got the data for the crime rate from the year 1995 to 2018.

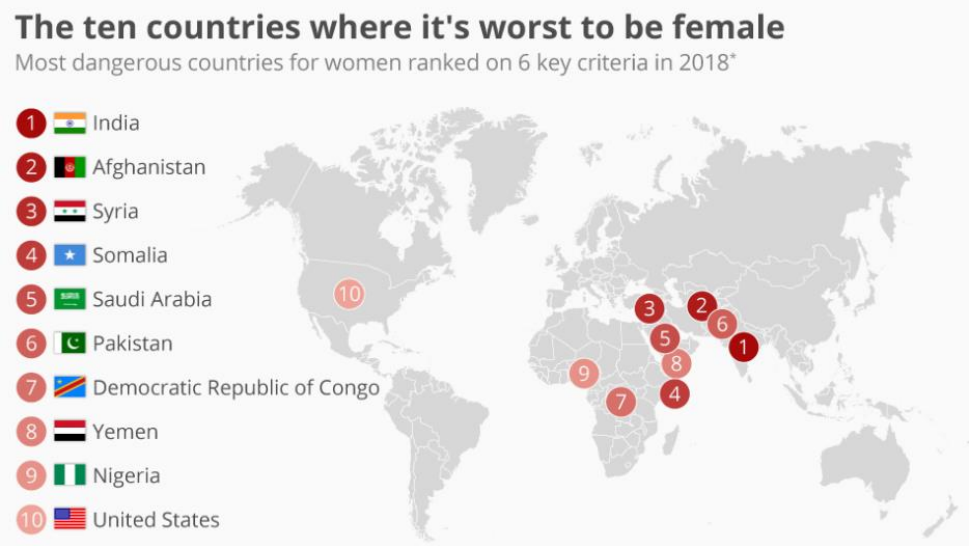
To collect further data, I approached the Cybersecurity office, which is easy to access and hence I asked them if they can avail me the information I need. Eventually, with the help of these people and Law Directors, I got the data fulfilling the data from the time period and distribution as my project demands. Law directors helped me getting the data on the crime rate. Another attribute I was searching for was the police rate that is the total number of police. I got this information with the use of police stations, and I succeeded in doing that after facing many questions and giving many statements of no data breaching to happen in future, with the data I am given. While merging the data from these many sources, fortunately, I have a primary key in terms of year and month, so that distribution is even and easy to merge. The only problem I faced was combining the police per 10000 with the other data. I had

to normalize the data and convert it into the police rate, and this factor then changed from total police to police rate.

Here is the representation of the data of crime in the various areas of India:

Fig. Map Showing Crime Prone Areas(*National Crime Statistics (page 196)*, 2013)

I have explored data regarding crimes and the factors affecting along with the statistics of the rate of these crimes happening in the country. Eventually, I realise that the crimes in my home country has made it miserable for every citizen, of course. But one thing that stuck to my mind and got me inspired to do this in the field and contribute in the least way possible is doing something to the betterment of humanity back in India. While collecting data, I was shocked by some facts and was struck by a few of them. Out of all such facts, one fact that shattered me was the place where I live is counted in the worst places to live for a woman. This is the situation of the world's biggest democracy. There are times in the world when there have to be steps taken towards the humanity and contribute by every person possible and then is the time when the world will be called a better place for every individual, no gender discrimination.



Source: Thomson Reuters Foundation

4.2 DATA UNDERSTANDING

4.2.1 Data description and exploration:

With the objective of data mining and making predictions about the crime rates in India, I started searching for the data everywhere, and getting such records can be hard and also easy at times; Since the data you need for prediction should be sufficient as well as meaningful. I came across various datasets online and finally found some that I could sense was good. Eventually, as I went forward, I started doubting my dataset selection technique and faced various problems while choosing the data I need and discarding, which I don't. Using IBM SPSS Modeler, I got to learn more about my data, and it helped me get my answers relating to my objectives and also cleared my goals.

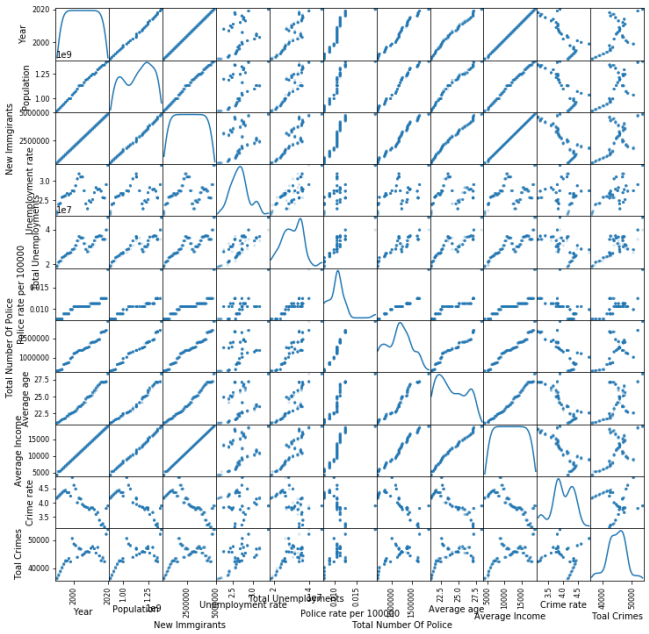


Fig. Data scatter plot after data reduction

Here, as my business objective focuses on predicting the crime rate, I have some fields focusing on the same, and eventually, I have some pieces of evidence to show the correlation between the several fields and crime rate, which can be demonstrated using the correlation graph as displayed above.

From the dataset collected, I have information regarding the census of Indian population and fields like average income, average age, Total unemployed people, New immigrants collected from the Indian government website https://censusindia.gov.in/Tables_Published/A-Series/pca_main.html . Detailed data description in the form of tables was provided by the government official online. It made the data collection a bit easier since much of the data was available in tabular format merging became simpler. Even though with these sources and police officials, I managed to collect the data for the years from 1990 to 2020, resulting in 380 rows. Resulting in the required amount of data, even though it seems less in number, the data mining objective of extracting the knowledge and my business objective of predicting the crime rate has successfully been achieved. Most of the value types are continuous since most of the fields represent the term depicting the "amount. Fields like population can have a huge number and results in a continuous type. Along with the total population, I have fields like total unemployment, total immigrants, total crimes, total police, average income, the average age in the continuous fields. I have several numeric fields in my data. Since it can be easily nominated that the month and year can be assigned as the Nominal data type. With months been classified as 12 different values and belongs to the set of unique 12 outcomes, it becomes evident to classify as the

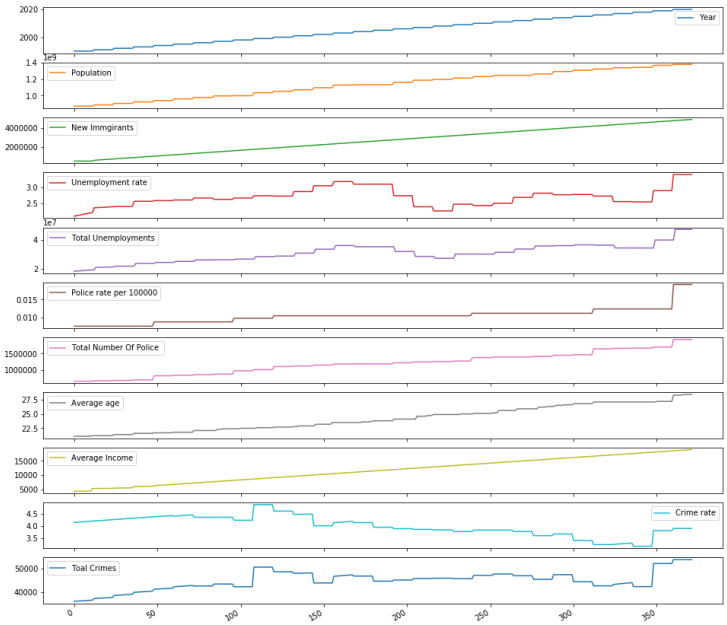
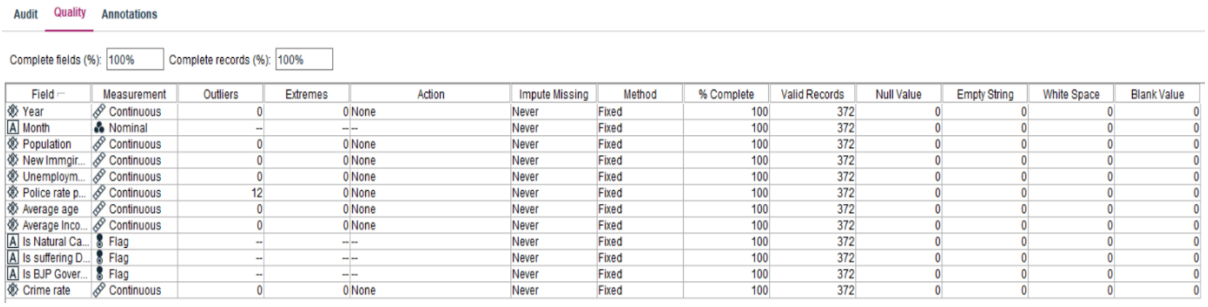


Fig. Data Visualization showing patterns

Nominal. With the fields like continuous and nominal in the line-up, I have attributes with the data type as Flag. I have the data of the country being affected by the natural calamity or not; this gets classified as a yes/no type and nominates itself into a flag data type. Along with this, I have a field to decide the government in India (from independence till now there has been either of the two parties) which constitutes two parties and hence I have cast these types as "IsBJPGovernment" to symbolize the yes/no data type. Fields constituting my dataset and their datatype with the role in modelling is explained using python as shown in the figure.

4.2.2 Data Quality:

As I gathered data, I came to know I need to merge two different datasets to get my desired prediction. While blending, I faced issues like there were NULL entries for the records, which were supposed to be continuous. To deal with it, I changed the NULL values with 0. Also, two different datasets had some fields common, but the value type of the fields was different, which made me change the values so that I get the merging right. I also had to eradicate the coding consistencies, which consisted of the different value types referring to the same output, such as False and 0, indicating a negative response. I have data of the attributes derived from the already existing attributes. As the derived attributes contribute to the misleading data analysis and eventually leading to data conflicts, I decided to filter out the derived attributes with the purpose of eradicating the error values in my data.



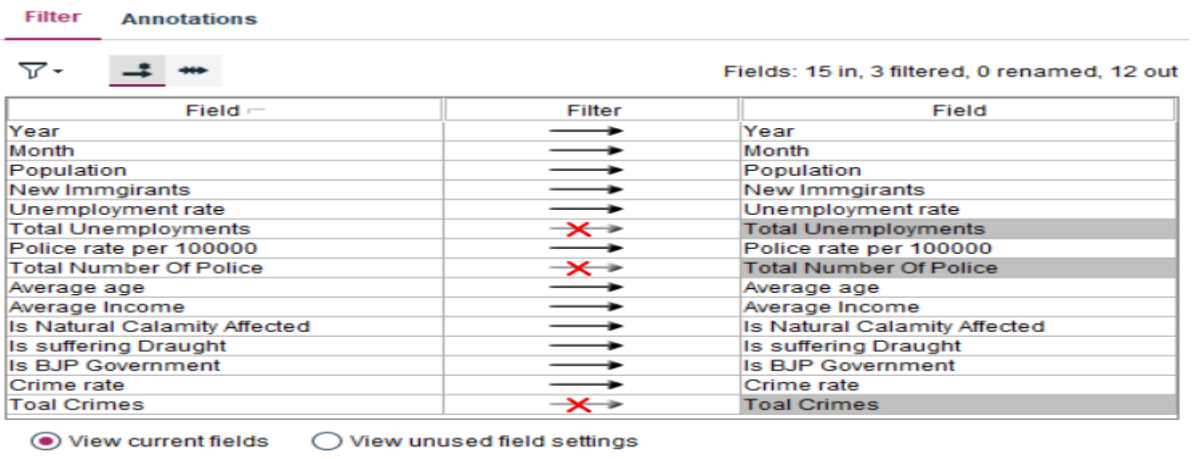
Audit | Quality | Annotations

Complete fields (%): 100% Complete records (%): 100%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
Year	Continuous	0	0	None	Never	Fixed	100	372	0	0	0	0
Month	Nominal	--	--	--	Never	Fixed	100	372	0	0	0	0
Population	Continuous	0	0	None	Never	Fixed	100	372	0	0	0	0
New Immgir...	Continuous	0	0	None	Never	Fixed	100	372	0	0	0	0
Unemploym...	Continuous	0	0	None	Never	Fixed	100	372	0	0	0	0
Police rate p...	Continuous	12	0	None	Never	Fixed	100	372	0	0	0	0
Average age	Continuous	0	0	None	Never	Fixed	100	372	0	0	0	0
Average Inco...	Continuous	0	0	None	Never	Fixed	100	372	0	0	0	0
Is Natural Ca...	Flag	--	--	--	Never	Fixed	100	372	0	0	0	0
Is suffering D...	Flag	--	--	--	Never	Fixed	100	372	0	0	0	0
Is BJP Gover...	Flag	--	--	--	Never	Fixed	100	372	0	0	0	0
Crime rate	Continuous	0	0	None	Never	Fixed	100	372	0	0	0	0

Fig. Data Quality

From the above quality table, it can be seen that all my attributes have the values with utmost perfection. It gave me the extra benefit of not removing the fields relating to the values which do not lie in the correlation with the other values. Even though the entire data is related to the different values with the proper correlation and distribution of this data is linear, there is one column 'police rate per 100000' where I needed to make the decision for the extreme values. This is the decision yet to be taken, after selecting my data and seeing how much data is appropriate for my objective, I shall make the decision about the outliers in my data. As you can see from above, there is no other deficiency in my data like missing values or null values.



Filter | Annotations

Fields: 15 in, 3 filtered, 0 renamed, 12 out

Field	Filter	Field
Year	→	Year
Month	→	Month
Population	→	Population
New Immgirants	→	New Immgirants
Unemployment rate	→	Unemployment rate
Total Unemployments	✕	Total Unemployments
Police rate per 100000	→	Police rate per 100000
Total Number Of Police	✕	Total Number Of Police
Average age	→	Average age
Average Income	→	Average Income
Is Natural Calamity Affected	→	Is Natural Calamity Affected
Is suffering Draught	→	Is suffering Draught
Is BJP Government	→	Is BJP Government
Crime rate	→	Crime rate
Toal Crimes	✕	Toal Crimes

View current fields View unused field settings

Fig. Data filtering table

As it can be seen from the above screenshot that I have filtered out “Total unemployment, Total Number of Police, Total Crimes”, since the data in these attributes was dependant on the data of

the other columns. Often, we see that there are NULL values in the forms of many other non-uniform data as 0, -, and many other forms. Fortunately, I collected the data from various source, and I got the data clean and ended up with the data not containing such values.

4.3 DATA PREPARATION

With the data for every given field in every row, I did not need to discard the missing data. Since I had no missing data. I have all the rows filled. Along with the wholly filled data, I also have some data with the errors (outliers). With the process of discarding the data would lead my data to shrink more. Hence, I decided to coerce in order to salvage the data and reduce my data losses to increase the data for my training model. With the coercing of the outliers, it made my outliers with the values close to the remaining values in the range of all the data. In this way, I achieved my data reduction. Along with this, I actually was worried if I am manipulating the original data records obtained from the official sites. With the coercing of some data, I found out that some attributes as defined earlier were derived from the other attributes from the same dataset, which could result in the discrepancy in creating the training model with the accuracy needed to verify the exact prediction values. Hence, the best way was not to include these values in the decisive factor, and I eventually decided not to use these fields in my data analysis and model training.

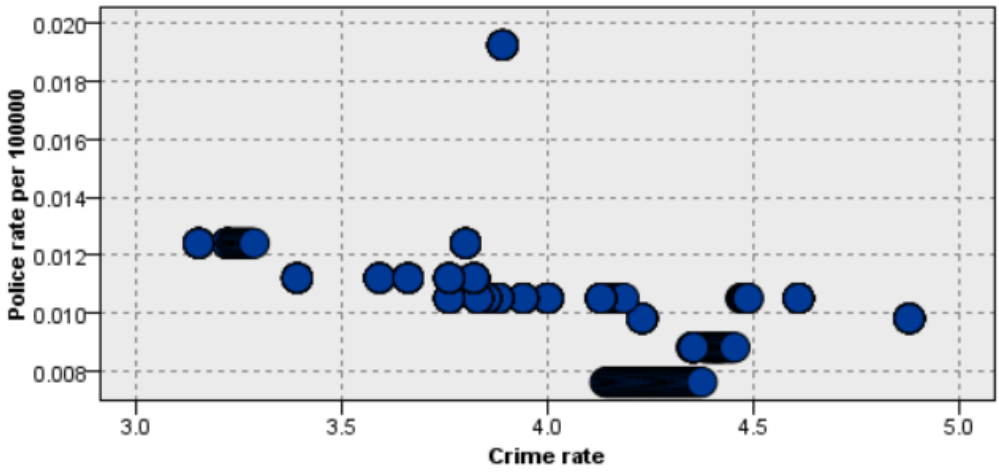


Fig. Data before Coercing

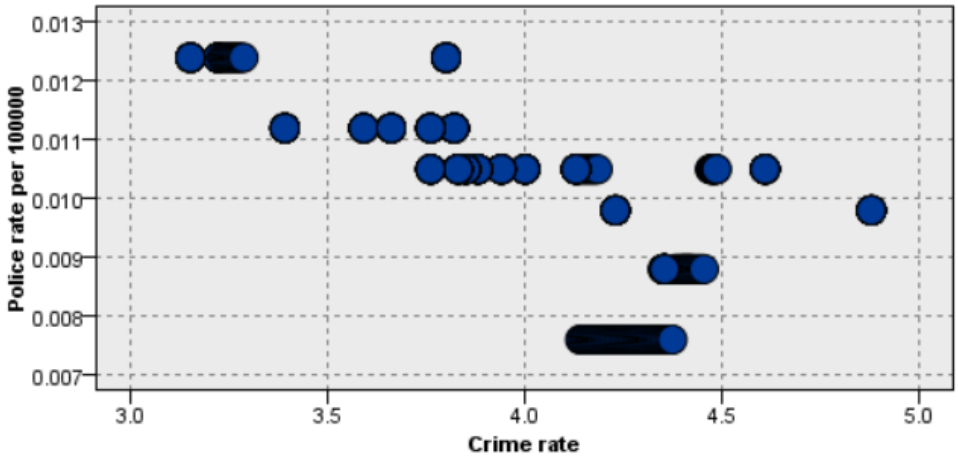


Fig. Data after Coercing

As I do not have any null values or missing values, I did not need to create a missing value node. But as it can be seen from the figure above, I had to manage the outliers in my data. I did the coercing of the data and the results are shown above. With this, cleaned my data by eradicating the values that do not lie in the distribution of the other values in the dataset. To do this, I went to data audit, I open the data quality tab and it can be seen in the fig 2.4.2. Going forward I had to take the decision about what to do, whether to discard the values or do the coercing. I chose coercing the values, in order to fulfil my second business objective as to make the most of the data I have and since I have less amount of data. Try not to decrease the data. But make the most of the data I have in my dataset. Hence, I chose to coerce the data. The following model shows the node generated for coercing the outliers.

Field	Measurement	Outliers
Year	Continuous	0
Month	Nominal	--
Population	Continuous	0
New Immgir...	Continuous	0
Unemploym...	Continuous	0
Police rate p...	Continuous	12
Average age	Continuous	0
Average Inco...	Continuous	0
Is Natural Ca...	Flag	--
Is suffering D...	Flag	--
Is BJP Gover...	Flag	--
Crime rate	Continuous	0

Fig. Outliers in my data

As it can be seen from the fig 2.4.2, I do not have any missing vale, null values, extremes, White spaces, blank spaces, empty string, I did not have to deal with that portion. This made it easier for me to show my data consistency, since there was no other noise in my data. The only noise in my dataset was the outliers related to 'Police Rate'. Apart from this, my data was clear form all the data distortion. This can be the outcome of my data collection and good sources leading me to the clear dataset.

4.4 DATA TRANSFORMATION

My principal business objective of prediction of the crime rate asks me to show the crime rate in various forms of the data and represent the relation of the various fields with the crime rate. In doing so, I have to answers of questions like how is this attribute related to the prediction attribute? What makes the particular attribute eligible to be the dependent one for prediction attribute? And this list goes on. With the answer to these questions, I bring the representation into the picture for the data projection. With my data, the inaccuracy lies in the merging of the data from various sources; here it can be seen that I have 'police rate' column where it is described that what is the rate of police in an area of 10000. But with the error of the data collection, it can be seen that the highlighted part shows the values in one row, where the values differ by a huge margin. With the sense of data understanding, I got to learn that the data was not the data related to the police rate, but it was actual police in that area for that particular period of time.

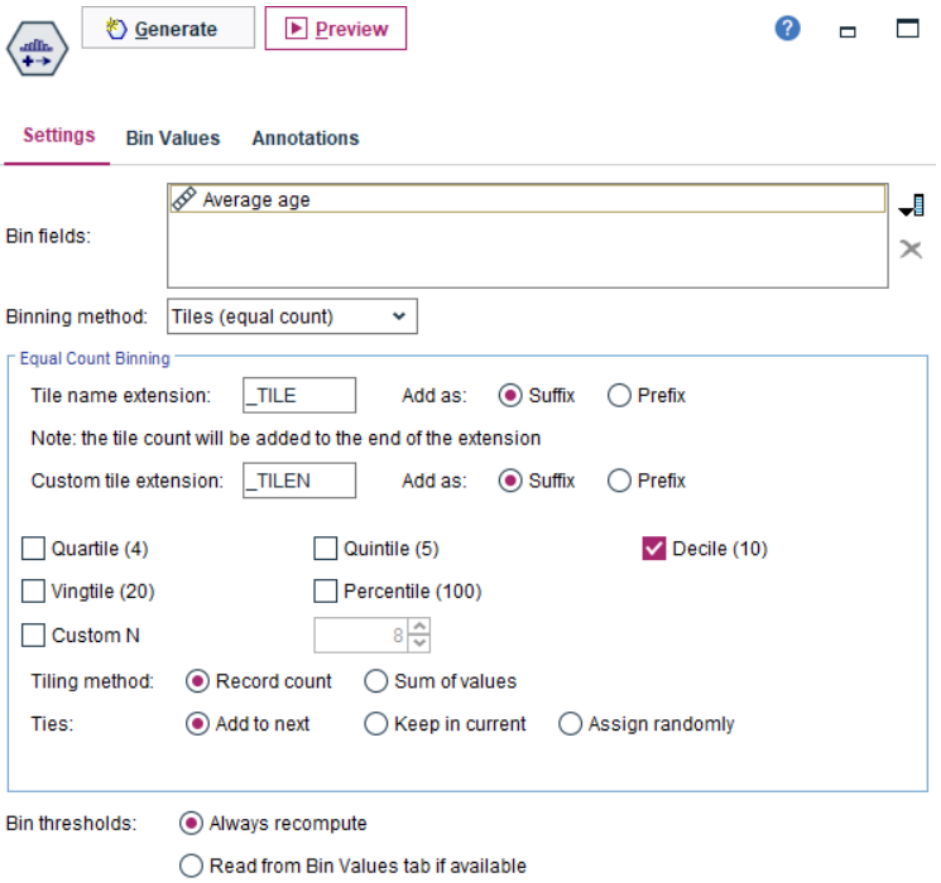
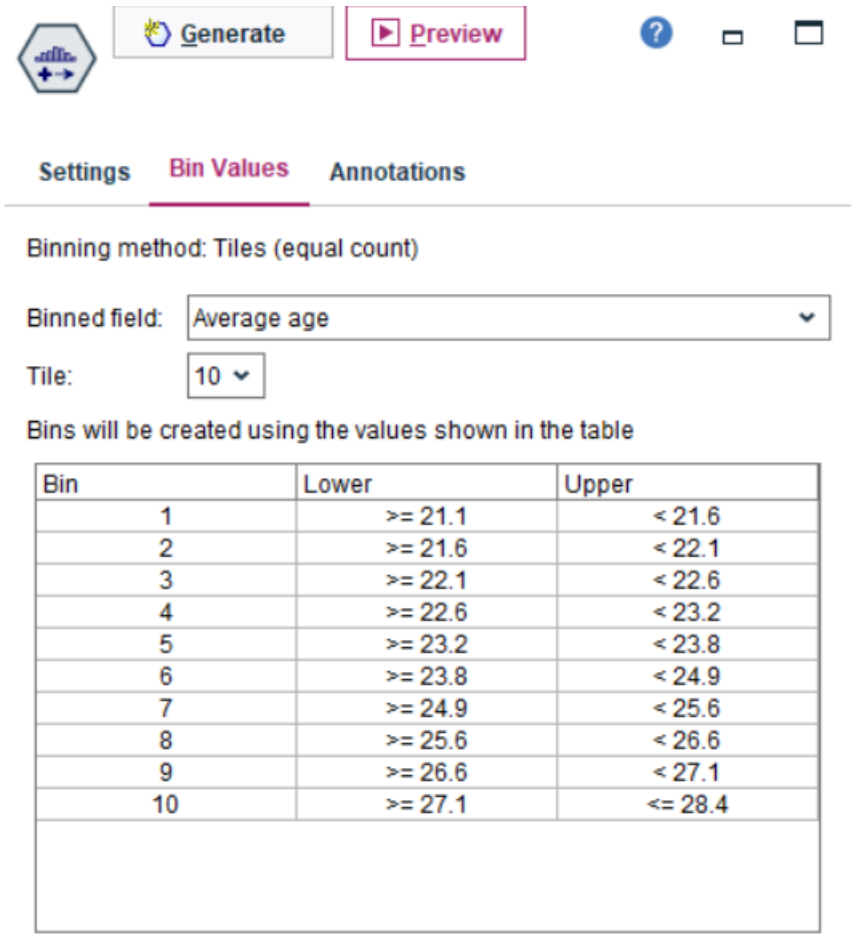


Fig. Binning Example

For projecting the data, I already had normalized data and had to work for the visual binning problem with my data. As explained in IBM SPSS modeler, the Bin Values tab in the Binning node allows you to view the thresholds for generated bins. Using the Generate menu, you can also generate a Derive node that can be used to apply these thresholds from one dataset to another. You can use the Generate menu to create a Derive node based on the current thresholds. This is useful for applying established bin thresholds from one set of data to another. Furthermore, once these split points are known, a Derive operation is more efficient (meaning faster) than a Binning operation when working with large datasets. Going further, it shows the tab for selecting the field that is to be selected for binning. And as show in the above figure I have used ‘Average age’ for this purpose and testing my knowledge on binning the attribute for the projecting the data. Since I find ‘Average age’ as more important attribute and while iterating through my process of data mining, I chose to select the methods to do more certainty on my data and get the accurate results with the better outcome for my target variable to achieve the prediction as I mentioned in business objective and hence complete the prediction as achieving the goal of predicting the ‘crime rate’ of the places in India.



Bin	Lower	Upper
1	>= 21.1	< 21.6
2	>= 21.6	< 22.1
3	>= 22.1	< 22.6
4	>= 22.6	< 23.2
5	>= 23.2	< 23.8
6	>= 23.8	< 24.9
7	>= 24.9	< 25.6
8	>= 25.6	< 26.6
9	>= 26.6	< 27.1
10	>= 27.1	<= 28.4

Fig. Binning 'Average Age'

Figure 4.2.2 shows how I have binned 'Average age' into 10 bins, with each having a lower and upper. So, the first bin is any value greater than 21.6 and less than 21, and so on. This created a new attribute 'AverageAge_transformed_bin'. This is just so the original attribute is not directly modified. Now that the binning has been completed, I would remove the Average Age_transformed attribute (Age attribute (continuous measurement)) so that we are only examining one the bins. I would do this by making that attribute's input as NONE.

5. DATA MINING AND ANALYSIS

With the information regarding the study of the crime reports all over the country, it came into my notice that I had to give more attention to the cities. Different areas give rise to different sets of the crime rate. There are some areas in India like Delhi, where we can see that the crime rate goes to the pick of them all. The reason here, as found by the various sources, is not a factor of money, immigration as I have used in my datamining attributes. Instead, there are cases of murder, rape, kidnapping showing the brutality of the people in the capital city of the country. This gives rise to the thought that is there a police inefficiency causing this to happen. Then there is another side to it, where it shows that the capital city not only has the protection by the Delhi police but also consists of some of the military persona with the armed moto. From the statistics it is seen that the places affected the most by this kind of brutality mostly includes places where people go out for having fun, making chill and doing the extravaganza activities apart from the daily routines. The reason behind this can be the habits of drinking and many other illegal sources to make people unaware of the situation and lose their conscience so that the motives of the culprits are achieved and that too not with much of the efforts. This shows the lack of awareness in public regarding these behaviours and extra efforts need to be taken to get the safety as

a priority and not to keep the luxury above the security of any person around them. In some case, there are places like airports, bus stations, train stations being the areas of crime. These public places are supposed to be surrounded by some extra care regarding any behaviour. No matter what, there should be some ground rules based for the sake of this safety and at least the safety at public places.

As I discussed all the business objectives above, it is clear that my goal here is to predict the crime rate on the basis of the factors that I have decided to use for the prediction. We know that there are various methods for data mining. Some of them are listed below:

- Tracking patterns
- Classification.
- Association.
- Clustering.
- Regression.
- Prediction.

We shall discuss some of them which can be taken into consideration to achieve my goal. Before selecting any particular method for the mining process, it is always suggested that all the possible solutions should be taken into account. Considering the same aim, I have decided to evaluate some of the methods discussed above for fulfilling my mining process.

- Tracking patterns: The ability to identify patterns in your data sets is one of the most common strategies in data mining. Normally this is a recognition of any aberration in your data that occurs at regular intervals or an ebb and flow of a particular time variable. You might see, for example, that your sales of a specific product tend to increase just before the holidays, or you might find that hot weather brings more people to your website. Here, in my case I have I want to track whether with the change in the other attributes and other factors how much my data is affected, with the basis of the same I can find the pattern in my data with the changes in these factors. But this does not give me the liberty of predicting my predictor variable, which is ‘crime rate’, hence this method is not suitable for my business objectives.
- Classification: Classification is a more complex technique of data mining which forces you to collect different attributes into discernible categories, which you can then use to draw more conclusions or to serve some purpose. With the data I have it is possible to classify the crimes int various sorts like classification on the basis of the type of crime, area of the crime, and year of the crime. But even with this my objective is not achieved.
- Clustering: Clustering is somewhat similar to sorting, but it includes grouping together pieces of data based on their similarities. For example, you could opt for different clusters items like spoon and fork can be your items and you can train your model in two ways supervised or unsupervised learning. It determines whether your model has created its own distinguishing patterns or you have explicitly mentioned these characteristics for the model. Here, I don’t want to cluster ant thing hence this is also not a suitable method for me.
- Regression: Regression, mainly used as a method of planning and modelling, is used to classify a certain variable's probability, in the existence of other variables. You may use it for example to project a certain price, based on different factors like availability, consumer demand, and competition. More specifically, regression’s primary focus is to help you uncover the exact relationship between two (or more) variables in a given data set. This is the exact thing I want to achieve where I have different attributes like ‘population’, ‘police rate’ etc. and depending on these variables derive a relationship between these variables with my predictor variable. This relationship between my variables will decide the accuracy with which I can predict the ‘crime rate’ in future trends.
- Prediction: Prediction is one of the most useful techniques in data mining, as it is used to forecast the types of data you will see in the future. For certain cases, it's enough to know and consider past patterns to make a fairly reliable forecast of what will happen in the future. For example, you could analyse credit history of customers and past transactions to determine whether they will be a credit risk in the future. I have enough data to predict the ‘crime rate’. So, this can be one of the models I

would like to focus to predict ted at from the historical data I have for the previous years and guess the future increase or decrease in the crime rate.

In this case where I have single predictor variable, I can explain the prediction using the regression. To make it easier to understand and interpret, statistical technique which uses several explanatory variables to predict the outcome of a variable response. Multiple linear regression (MLR) is intended to model the causal relationship between the explanatory (independent) variables and the variable response (dependent). Only when one has two continuous variables, an independent variable, and a dependent variable, can linear regression be used. The independent variable is the parameter used to determine the equation or consequence depending upon it. A multiple regression model refers to multiple describing variables.

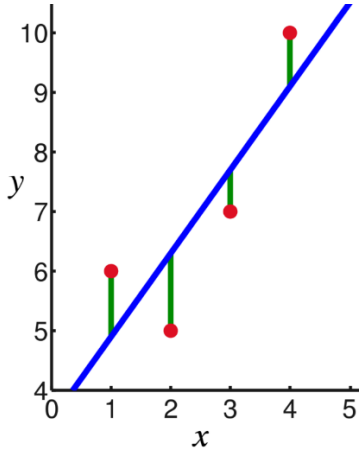


Fig. Regression Model type

In linear regression, the observations (red) are assumed to be the result of random deviations (green) from an underlying relationship (blue) between a dependent variable (y) and an independent variable (x)[10]. With multiple attributes that are dependant variables, I can have various regression for numerous variables. I have independent variables like 'Population', 'new immigrant', 'average age', 'average income', 'unemployment rate'. It can have several plots of every variable against the dependent variable which is 'crime rate'.

Following are the implementation of various data mining algorithms using IBM SPSS modeller to find out the patterns and results using these algorithms. In every algorithm the dataset after transformations and appropriate data preparations are supplied. This gives the modelling techniques varied data to get the outcome for the algorithm trained model.

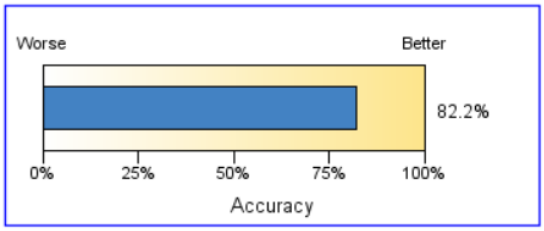
5.1 The linear model (with Forward Stepwise Model selection method)

Linear models predict a continuous target based on direct relationships between the target and one or more predictors. Linear models are relatively simple and give an easily interpreted mathematical formula for scoring. The properties of these models are well understood. They can typically be built very quickly compared to other model types (such as neural networks or decision trees) on the same dataset. The field requirements in this modelling is that the target must be continuous (scale). There are no measurement level restrictions on predictors (inputs). In my case, the objective is 'Crime Rate' and the information are all the other fields mentioned in the dataset. Using this, I am going to get test the accuracy of my dataset and calculate the error in my dataset. Along with the skill which is measures in "Adjusted R Square", I have the priorities set for the time limit. The time taken by the model to compline and execute can be a matter of importance, and I take it with the same level and priority.

Model Summary

Target	Crime rate
Automatic Data Preparation	On
Model Selection Method	Forward Stepwise
Information Criterion	-1,248.505

The information criterion is used to compare to models. Models with smaller information criterion values fit better.



In the fig above, it can be seen that I have set Crime rate as my target variable and others as the input to this target predictor. The linear model follows the **'Forward Stepwise'** method where it goes on selecting the attributes based on the importance in the prediction of the target and with the forward approach goes on choosing the ones which have the most influence in ascending order. This starts with no effects in the model and adds and removes effects one step at a time until no more can be added or removed according to the stepwise criteria. In this way, it goes on selecting few and eliminating a few of the attributes in the process of prediction importance. Here, one add-on this modelling provides is that there is data manipulation and using the means of various methods. As we can see in the table above, '_Transformed' field is appended to the attributes in the dataset where there has been some modification done by the model itself.

Automatic Data Preparation

Target: Crime rate

Field	Role	Actions Taken
(Average age_transformed)	Predictor	Trim outliers
(Average Income_transformed)	Predictor	Trim outliers
(New Immgirants_transformed)	Predictor	Trim outliers
(Police rate per 100000_transformed)	Predictor	Trim outliers
(Population_transformed)	Predictor	Trim outliers
(Unemployment rate_transformed)	Predictor	Trim outliers
(Year_transformed)	Predictor	Trim outliers

If the original field name is X, then the transformed field is displayed as (X_transformed). The original field is excluded from the analysis and the transformed field is included instead.
One or more records were excluded because of a predictor or target that is missing, a frequency weight that is missing or less than one after rounding, or a regression weight that is missing, negative, or zero.

Along with the fields, there are fields like action taken, which represents the actions taken to navigate the data from the source to modifiers. The example of action taken is shown, and the table is 'Trim Outliers'. This action is made for all the fields in the input as mentioned earlier attributes.

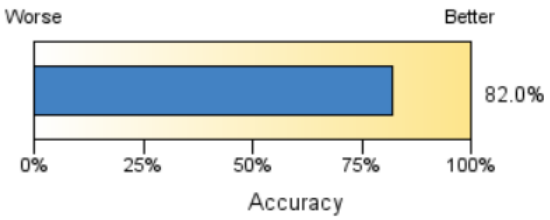
5.2 The linear model (with Include all predictors Model selection method)

Linear modelling with the inclusion of all the input fields decides the selection of the model prediction on the basis of the entire input values. Unlike the forward stepwise, where the input and the predictors are determined one by one, and all the input values are not considered in deciding the predictions of the model. With different models, different rate of accuracy is observed for the modelling. Here as we can see in the screenshot attached below, the accuracy is different as compared to the one above. The only change in the two models is that the modelling method changes for the inclusion of input variables. We can see that the change in the two modelling steps also determines the changes in the 'Information Criteria'. These variables just determine the smallest possible value or the information that fits better. Here the fitting of the values is not similar to the concepts of underfitting or overfitting, it relates to the values where the prediction based on these inputs allows the model to get the values from the original vales or there is some change required in the input values to fit the values better with the model so as to create a better chance of the model prediction and the value usage in the future.

Model Summary

Target	Crime rate
Automatic Data Preparation	On
Model Selection Method	None (All Predictors Entered)
Information Criterion	-1,238.691

The information criterion is used to compare to models.
Models with smaller information criterion values fit better.



Even though the model accuracy of the two models shows that the accuracy of the model is better and that the model is sufficient to answer the changes in prediction target variable, I decided to explore the decision tree concept and model to get more ideas of the other model.

5.3 CR Tree

Decision trees work by recursively partitioning the data based on input field values. The data partitions are called branches. The initial branch (sometimes called the root) encompasses all

data records. The source is split into subsets, or child branches, based on the value of a particular input field. Each child branch can be further divided into sub-branches, which can, in turn, be divided again, and so on. At the lowest level of the tree are branches that have no more splits. Such offices are known as terminal branches (or leaves). As we can see that, in the following approach the decision tree is used where the 'crime rate' is used as the root node and the 'year' is used as the decision making attribute based on the selected fields in the input, model decided the inputs it needs for the decision making attribute. The working of the decision tree is explained with the here types of nodes as follows:

- Predictor node (root node) with many incoming edges but with no outgoing edge
- Intermediate nodes, which are the following decision attribute which has one incoming node based on decision of the previous higher version node and can have two outgoing edges as yes/no
- Last node, which has the last decisive power and where the tree stops with the decision of yes/no. Some intermediate nodes go further with a 'yes' while some are carried further with a 'no'. With the end of this decision comes the last node (leaf node).

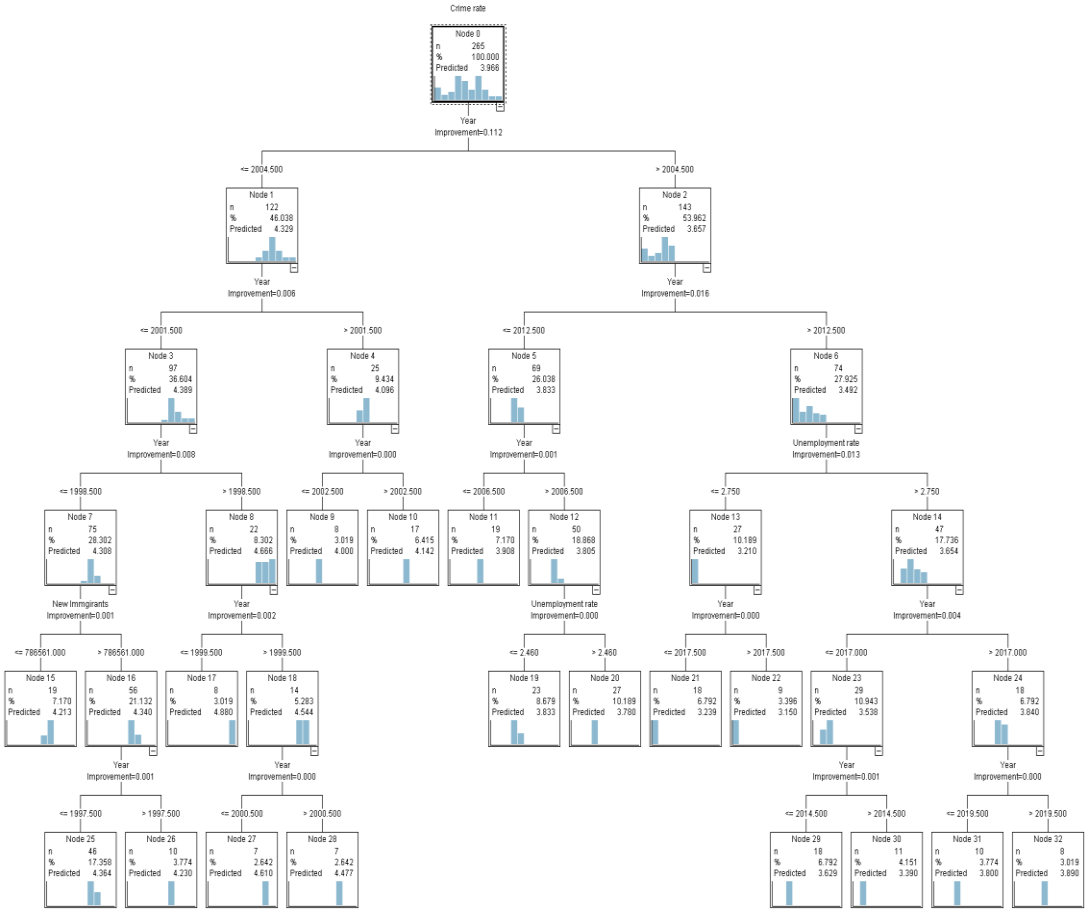


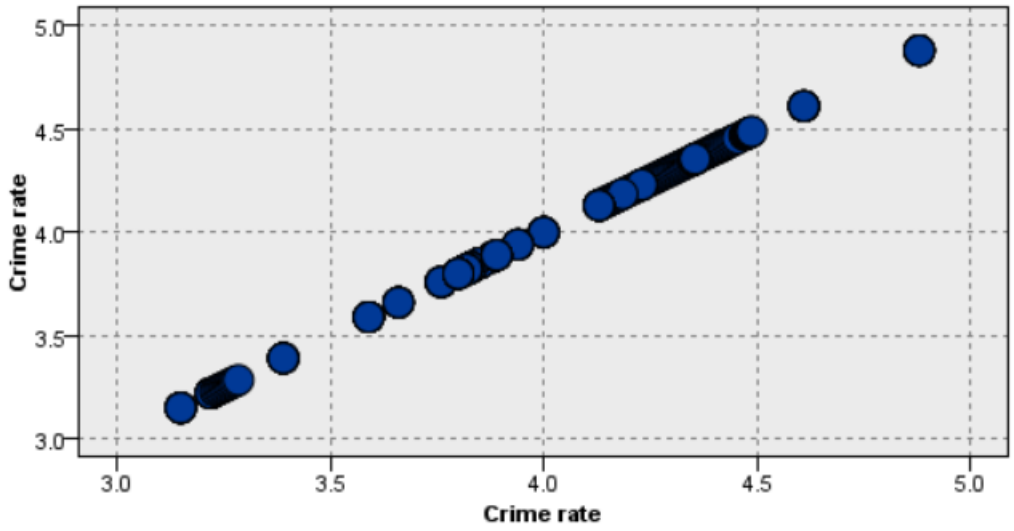
Fig Decision tree for Crime rate using CRT mode

With the use of decision tree concept as the reference to my prediction, I carried on my prediction work using the same approach of splitting the dataset into two training and testing dataset. Analyzing the scenario with the result obtained after testing the model with the testing dataset. I got two occurrences of the conclusion where I can see that the data mining method, I

used is sufficient to give me the answers. I have explained to them in detail in my Interpretation. I have listed some of the statistics that I got while using the IBM SPSS modeller for the mining problem. I got to summarize the filed I used in the dataset. It is taken as a prominent notice that the time your model takes to execute is of more importance. Even as I heard from my instructors, try to make the model as fast as possible. In the meantime, I wondered why is it that the time for model compiling is taken into consideration when your model is giving you the desired result. This was a question I have raised to my tutor. But then, not knowing the scenario where for each and every step of the mining, you have to build your model, compile and run it several times during the entire process. I figured out this while building my model, and I used some variables which were derived from other attributes of the same dataset. I eventually read that it is a good practice to use the independent variable, and hence I turned to filter out the variable that caused these discrepancies. Eventually, my decision turned out to be a good one, it not only increased the speed of my building the model but also increased the significance of my attributes leading to the better training model and hence I can nor use the testing dataset to check the efficiency of my model and my training dataset along with it

5.4 Linear Regression

In stats, the study of regression is used to identify patterns in results. Regression analysis can provide you with a graph equation, so you can predict your results. Linear regression helps in the independent analysis of the variable in its distribution and linearity within these values of these variables are distributes across the dataset.

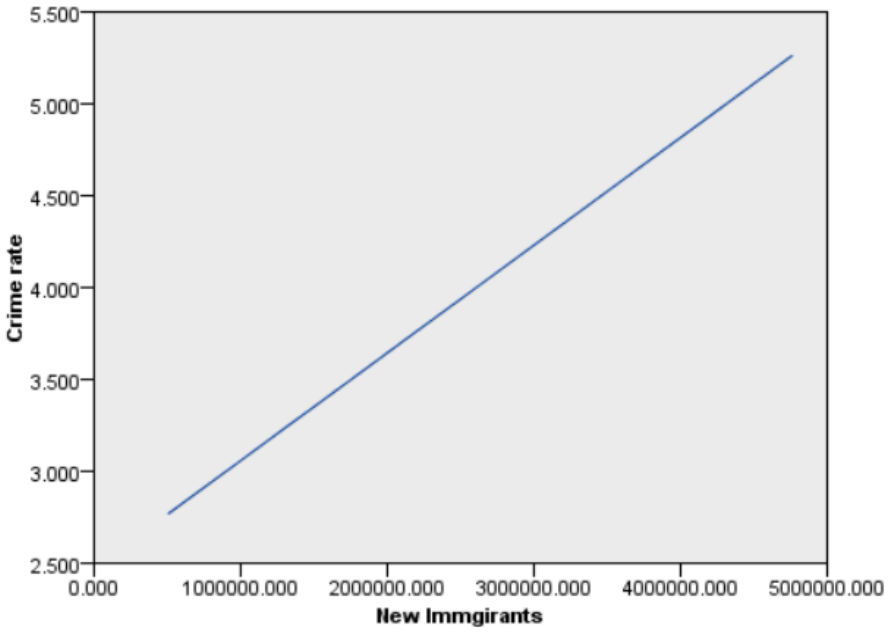


Here as we can see there is the distribution graph shown for the ‘crime rate’. This infers that only the distribution of a single variable is the outcome. But I want my distribution to be in relation to the multiple variables, and it can be achieved by using multilinear regression.

5.5 Multi Linear Regression

Ordinary linear regression usually isn’t enough to take into account all of the real-life factors that have an effect on an outcome. For prediction, regression is the ultimate thing that I can perform. I used multiple models for the decision, but I came with the regression after a lot of research. I used multilinear regression with the predicting variable as the same in all the previous models. With the purpose of prediction, I chose the input of all the variable that I had remained with after the filtration applied to give me all the attributes except the three that were filtered during the phase of data cleaning. This model gave me the performance desired and showed the output with the best way possible. As I had discussed with my tutors and the fellow classmates, u had come to the conclusion that the models with the accuracy as well as the better

time performance is what makes it to the better modelling termed in the data mining. Since it takes many times that one needs to build the model to compile the model and run it with the sake of training and testing the dataset. In this procedure, it cannot be ignored that the time taken by your model is effective or not. The below-mentioned figure clearly indicates that the model built using multilinear regression technique was built in just 2 seconds which is efficient and can be tolerated up to a certain extent. The usage of the model is what is going to define the modelling and the algorithm used for the data mining selection.



Here, we can have multiple variables to show the regression for a particular instance of the variable. For example, I have used the 'new immigrants' as one of the variables and the other variable is 'crime rate'. This shows the multilinear regression for the variable of one variable with each other.

6. RESULTS AND DISCUSSION

With the aim of mining the data and interpreting the result, I have conducted the model mining till now and have come up with the information regarding mining algorithm, and I am going to use. With this, I am going to analyse the work done in the process of mining and validated the same using the following equations. As I have chosen regression for my data mining and the algorithm that I am going to use is multilinear regression, I have come up with to some knowledge of my data from the mining am going to use the same the validate the model. A linear regression model with two predictor variables can be expressed with the following equation:

$$Y = B_0 + B_1*X_1 + B_2*X_2 + e.$$

The variables in the model are:

- Y, the response variable;

- X1, the first predictor variable;
- X2, the second predictor variable; and
- e, the residual error, which is an unmeasured variable.

The parameters in the model are:

- B0, the Y-intercept;
- B1, the first regression coefficient; and
- B2, the second regression coefficient.

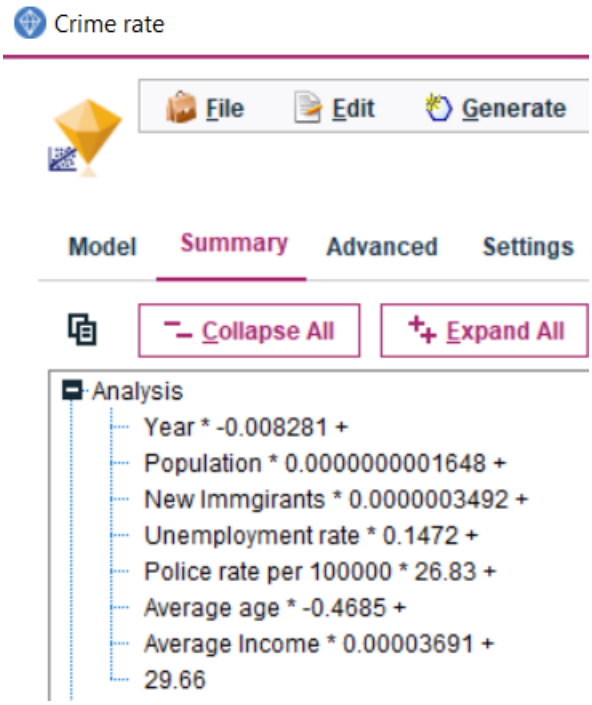


Fig. Data Analysis summary

From the above figure and the analysis done, I can get the values for the expression and use the same method to get the results while interpreting the data. Now I shall assign the variables to the values I have got and formulate the equation with the values I have got after the analyses of my dataset using the regression model. Here,

e = 29.66

B0 = coefficient for Year

B1 = coefficient for Population

B2 = coefficient for New Immigrants

B3 = coefficient for Unemployment rate

B4 = coefficient for Average age

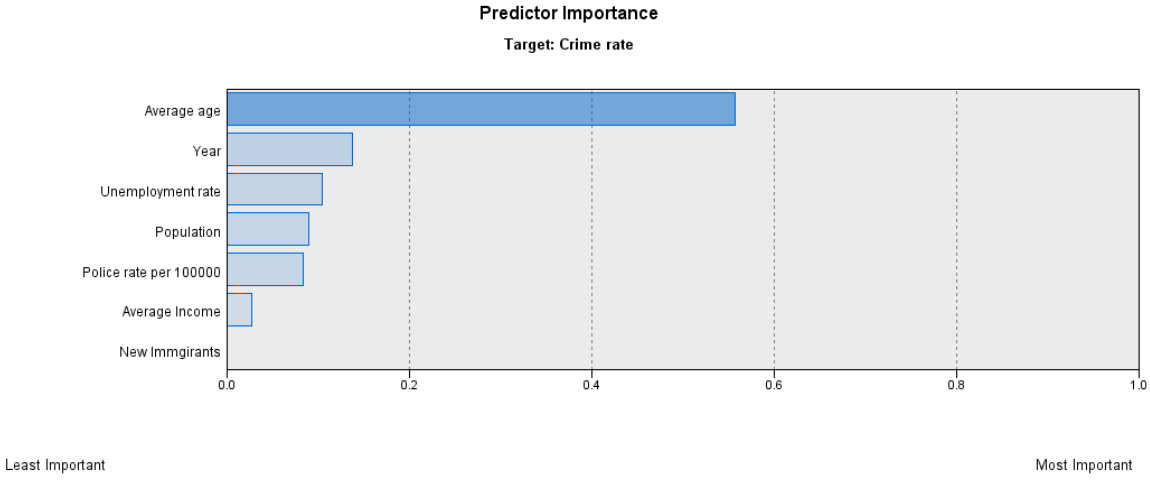
B5 = coefficient for Average Income.

Hence when I formulate the equation of regression with the values of coefficients and the names of the values representing the coefficients, it looks like this:

$$Y = 29.66 – 0.008281(Year) -0.0000000001648(Population) + 0.000003492(New\ Immigrants) + 0.1472(Unemployment\ rate) - 0.4685(Average\ age) – 0.00003691\ (Average\ Income)$$

From the above variable, I can calculate the Y – predicted Crime Rate. This gives the way to estimate the importance of a particular factor on the change of my predictor variable that is the crime rate.

My most important focus, which was tacit during the entire project was the effect of the 'unemployment rate' and 'new immigrants' on the 'crime rate'. As it is evident and can be logically understood that as the unemployment increases there can be a significant increase in the crimes like burglary, murders, kidnapping, in the sort of seeking money with one way or the other means. This is evident that yes, it is the deciding factor for the crime rate, but it is not the most effective factor in the increase in the crime rate. Another factor that I was focusing on was the 'New Immigrants'. Since the people who come from the land outside the country are new to the culture and new to the surroundings, there might be two cases possible that these people become a victim of the criminals who find these people new in the area and try to exploit their situation by giving false guidance and leading into the mis happenings with them. The other side of that can be that these people with the culture of their own land, can have a destructive mindset and can lead to disasters in the new country.



In the ANOVA table below, we can see that the significance value is 0.000 (i.e., p = .001), that's < 0.05. And thus, there is a statistically significant difference in the mean length of time between the different courses taken to complete the spreadsheet problem. It's nice to say, but I do not know which of the specific groups differed. The ANOVA table is built by some terms that are derived from the data like:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.871[a]	.759	.755	.202866

a. Predictors: (Constant), Average Income, Unemployment rate, Average age, Population, Year, New Immgirants

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	47.214	6	7.869	191.206	.000[b]
	Residual	15.021	365	.041		
	Total	62.236	371			

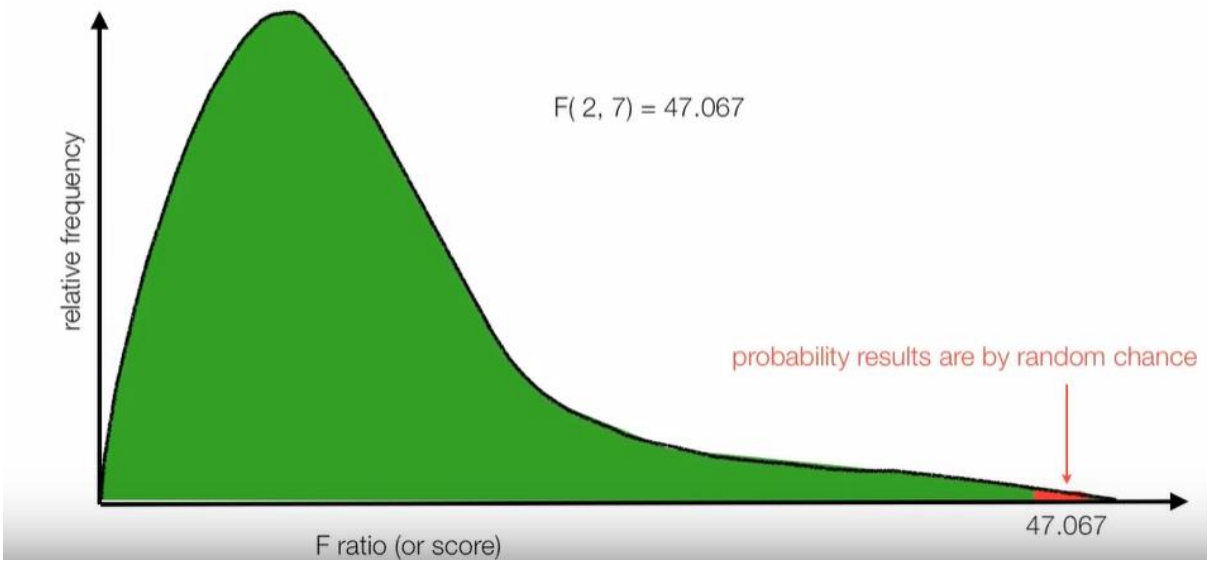
b. Predictors: (Constant), Average Income, Unemployment rate, Average age, Population, Year, New Immgirants

Total Sum of Squares is the sum of the values from the first row till the last of the difference between the mean crime rate and the actual crime present in my dataset. To get the value of the regression, calculate Y using the equation given below for the linear regression with multiple affecting variables. With this, calculate Y for every row and then this gives the estimated value for the crime rate for each row. Now, to get the regression value, simply find out the difference between the estimated crime rate and the actual crime rate. The sum of all the value of the differences gives the value for Regression sum of squares. Now, to understand the residual is the simplest thing. It is as the name suggests, just the residual between the total and the regression that is the difference between the regression value and the total value of the sum of squares. Degrees of freedom that can be abbreviated as df is the difference between the regressor and the predictor. Here the independent variables are six and the predictor is 1. Hence the value 6 in the first row.

Luckily, I can find this out in the Multiple Comparisons table. With the ANOVA table we know that the expression that comes out of this is F (x, y) where x – the degree of freedom for regression and y – degree of freedom for residual.

In my case x= 6 and y = 365.

The graph for F (6,365) = 191.206 can be seen below as



Source: Regression ANOVA explanation

I have used this image to describe the above expression here the values that are shown for the probability results by random chance, in my case is 191.206, and it accounts to the 0 significance or p-value. Using

the equation, I have mentioned in the datamining explanation; I can use it to define the above constants and the coefficients.

Y = 29.66 – 0.008281(Year) -0.0000000001648(Population) + 0.000003492(New Immigrants) + 0.1472(Unemployment rate) - 0.4685(Average age) – 0.00003691 (Average Income)

I derived the values for each row using the same equation and forum out the predicted crime rate for each row. This gave me the estimated crime rate that my model can provide with the training that I have given till now. Now I see how much my model results vary with the actual data I have in my rows ad this provides the deviation for each row with the real data in my dataset. Now the difference for the practical and estimated row is calculated for all the rows and mean deviation can be derived using the same. In, the next column that is Std. Error, It shows that how much confident am about the prediction of the extreme values and forecast of the lower values and accordingly the failure in the benefits can introduce the change in my confidence values and can add into the absolute costs and lower the low costs to show the error in estimation.

As I went forward this ode towards the variables increase and also, I started developing the interest in knowing the information about each attribute contributing to the prediction rate success. For example, I was quite interested in knowing the effect of 'unemployment rate' on the crime rate as it is evident that this factor speaks for itself and there is no big rocket science to see that this factor has led to the crimes. But the task I wanted to achieve was how much does this factor cause, and the results blow me that other factors lead to the sins which are more significant than this factor and I did not even put that more profound thought in it. But with this, I got to explore not only about various factors, but also, eventually, I learnt that going with some precise prediction never helps in this task. My learning through this would be to be open to all the possibilities and pay attention to every factor that you get in your dataset and never underestimate the effectiveness of some attribute in your mining process.

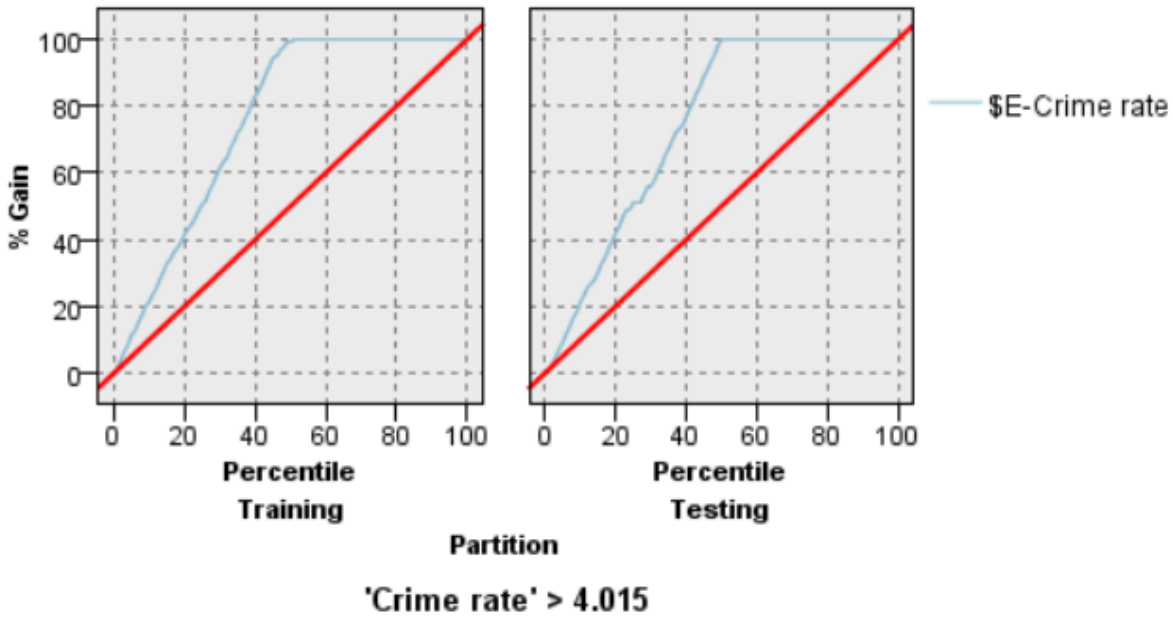


Fig Evaluating training model against testing model

From the figure above, I can say that my model is trained particularly good for the dataset. The model predicts the values for training dataset in an excellent manner where the difference is not much for the estimates values and the actual values for the crime rate in case of the training dataset. But as expected it would deflect a little when it comes to the testing dataset. Still the output does not vary a lot in accordance with the actual data. The testing dataset showed a good correlation for the actual values had the estimated values when related to the dataset.

7. CONCLUSION

As discussed in the objective, this research and data mining was conducted to get an idea of the factors contributing to the increase in the crime rates in India. Even though it limits to the specific region, the prediction aims to reduce the crime rate and help the disciplinaries in understanding the factors that affect the crimes and growth in crime across the country. From the factors taken into consideration and used for evaluating the predictions, 'Average Age' becomes a critical factor in the crime rate increase. As it can be seen from the dataset and the report, the average age for the country lies between the age of 22 to 29 which accounts to the younger generation. In this age, where the graph of development can be expected to be grown, somehow an unwanted graph and prediction is affected by this factor and a proper education and guidelines can be given to the youth of the country to be aware of their actions and nurture them into the development of the nation.

8. REFERENCES

Accidents, N., & Accidents, U. (2015). SNAPSHOTS 2015. In *National Crime Records Bureau*. http://ncrb.gov.in/StatPublications/CII/CII2015/FILES/Snapshots-11.11.16.pdf

Edwardes, S. M. (2008). Crime in India. In *Economic and Political Weekly* (Vol. 43, Issue 3, pp. 6–7). READ BOOKS.

National Crime Statistics (page 196). (2013). National Crime Records Bureau, India. http://ncrb.nic.in/CD-CII2012/Statistics2012.pdf

IBM. (2001). *Random Forest Node.* Retrieved from IBM: https://www.ibm.com/support/knowledgecenter/en/SS3RA7_sub/modeler_mainhelp_client_ddita/clementine/python_nodes_rf.html

IBM. (2012). *Clustering Models*. Retrieved from IBM: https://www.ibm.com/support/knowledgecenter/SS3RA7_15.0.0/com.ibm.spss.modeler.help/nodes_clusteringmodels.htm?view=kc

IBM. (2012a). Retrieved from TwoStep Cluster Node: https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/clusternode_general.htm

IBM. (2012b). *Clustering Models*. Retrieved from IBM: https://www.ibm.com/support/knowledgecenter/SS3RA7_15.0.0/com.ibm.spss.modeler.help/nodes_clusteringmodels.htm?view=kc

IBM. (2012c). *Auto Classifier Node*. Retrieved from IBM: https://www.ibm.com/support/knowledgecenter/SS3RA7_15.0.0/com.ibm.spss.modeler.help/binary_classifier_node.htm

IBM. (2012d). *Types of Models*. Retrieved from IBM: https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/understanding_modeltypes.htm

IBM. (2012e). *About SVM*. Retrieved from IBM: https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/svm_about.htm

IBM. (2012f). *C5.0 node*. Retrieved from IBM: https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/c50node_general.htm

IBM. (2012g). *C5.0 Node Model Options*. Retrieved from IBM: https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/c50_modeltab.htm

IBM. (2012h). *Misclassification Costs*. Retrieved from IBM: https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/oracle_coststab.htm

IBM. (2014). *IBM SPSS Decision Trees.* Retrieved from IBM: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/23.0/en/client/Manuals/IBM_SPSS_Decision_Trees.pdf

IBM. (2018). *QUEST and C&RT ignoring nominal input variables with more than 25 categories*. Retrieved from IBM Support: https://www.ibm.com/support/pages/quest-and-crt-ignoring-nominal-input-variables-more-25-categories

IBM Analytics. (n.d.). *What is predictive analytics?* Retrieved from IBM: https://www.ibm.com/analytics/predictive-analytics