Discriminant Analysis: Classifying and Predicting Sexual Offenders

Bhajneet Kaur, Laxmi Ahuja, Vinay Kumar

Abstract: Crimes or offences against women have their major contribution to spoil the physical and mental health of women. There are varieties of techniques and methods exists, through which crime against women can be analysed and predict for preventing and controlling it. In this paper a model has been proposed to classify and predict sexual offenders for two categories of victim namely major and minor using multiple discriminant analysis. IBM SPSS 21, statistical software has been used to build model with 85.3% of accuracy rate to classify and predict the cases. Dataset of sexual offenders has been extracted from the Chicago police portal system. After the compilation of the dataset four attributes have been used in the current analysis namely age, height, weight and victim_type. Three variables are taken as independent variables and one variable namely victim_type used as a target variable as it has two categories i.e. minor and major. All the three independent variables have their contribution in the model building but with different contribution rates. By using the standardized beta and Wilk’s lambda, ranking of contribution of independent variables has been generated. Age of the sexual offenders has contributed the most to predict sexual offender. The second ranked variable found as weight, its contribution towards prediction is higher than “Height” but lower than “Age”. The variable “Height” got third rank as low predictor to predict sexual offenders for victim types. The proposed model is very effective and would be used to implement or taken the different security measures in the various public or private places to prevent and control the various kind of sexual offences.

Keywords: sexual offenders, sexual offences, discriminant analysis, classification, prediction, SPSS.

I. INTRODUCTION

There are varieties of crimes against women. Sexual offences are one of them. Offences create menace in the society across the world and are the serious problems with destructive consequences. Sexual offenders are the persons who perpetrated towards sex crime. Basically sexual offenders are categorised as per various types of victims or offences [2]. Some Studies have been conducted on unique characteristics of sexual offenders to classify them [4], [9]. Variety of offences comes under the category of sexual offences like rape, sexual assault, sexual abuse etc. Various data mining and analytical techniques have been used earlier to classify or predict crimes or offences. This study emphasizes on the prediction of sexual offenders on the basis of three very important physical characteristics like Age, Height and Weight. In the proposed work discriminant analysis (DA) technique has been applied to discover the predictive equation or threshold values to predict sexual offenders. It used to identify a predictive equation based upon the predictors or explanatory variables or independent variables to classify the overall data into groups. Discriminant analysis is very similar to regression analysis with categorical dependent variable. The technique called multiple discriminant analysis because multiple independent variables have been taken for the analysis purpose. [18], [25], [27].

II. RELATED STUDY

There are various research’s works have been done in the field of crime and criminal activities against women. Sexual violence is very common to see in the previous literature to be condemned by state [6]. The understanding of sexual offender’s characteristics has been explained by [10]. There are various factors affecting the crime against women, a study has been conducted by [21] and few studies have been read about the understanding of risk factors of sexual offenders [7]. There are few similarities exist in the sexual offenders who commit activities related to sexual assault. Many rapists have negative image of women and shows hyper identification with the masculine role. 83% crime has been increased from 2007 to 2016 [22]. Most of the studies have been directed upon rape offence with elderly women by young adults [5]. Somehow, sexual offences are directly connected to the unsafe spaces. [30] studied about the sexual offences with the relation of private places but the study conducted by [33], the fear of sexual offences inhibits from the lack of safe public places. Most of the studies have been done with the relation with public and private unsafe spaces. A study has been conducted in Chicago by [11] to analyze the victim and offender mobility. According to study, 31.5% sexual assault occurs at offender’s place or its home address [11]. To predict the sexual offenders various predictive model have been developed by using various data mining techniques. A decision tree model has been developed on the basis of secondary dataset to classify and predict the sexual offenders on the basis of major /minor victims [26].

III. PURPOSE OF THE STUDY

Sexual violence is the serious cause for any country, state or society which affects the public health and produces the disturbance. Sexual offender is the person who commits the sexual offence. The purpose of the study is to build an effective model to classify and predict the sexual offenders based upon minor and major victims. Multiple discriminant analysis technique has been chosen for the data analysis to generate a threshold values because the target variable is the categorical variable with the data values of minor and major victims. First of all the pre-testing must be done on the data to check whether data
Discriminant Analysis: Classifying and Predicting Sexual Offenders

IV. CONCEPTUAL MODEL DEVELOPMENT

A. Dataset Specification

The secondary dataset has been used to propose the predictive model sourced from the online portal of Chicago city police department. Dataset has been categorized into various attributes of sexual offenders. Entire information of sexual offenders updates on daily basis. After the compilation of dataset few attributes are taken for the analysis purpose to build a predictive model. Three attributes of the dataset are taken as independent variables i.e. Age, height and weight of sexual offenders. One attribute is taken as a target variable from the dataset namely victim_type. This variable is a categorical or nominal variable defined by two values i.e. N and Y. Y means the sexual offender offences with minor victim and N means offender offences with major victim. To build the predictive model the victim_type has been used as a dependent variable or target variable and age, weight and height are used as independent ones. Multiple discriminant analysis has been applied in this research because more than one variable are used to build a model as independent category. All the independent variables are continuous in nature or they can be said in the metric scale of dataset.

B. Tools and Techniques

Discriminant analysis technique of regression family has been applied in this research to predict the sexual offenders for major and minor categories using SPSS tool. In this technique Wilk’s lambda test gives the significance output for the predictive model. The significance testing has been done for both variable wise and model wise. There are some other techniques of regression series like binary regression or logistic regression analysis which can be used to predict the crime and criminal acts as one of the study conducted to find the disbandment of the data (Upadhyaya et al., 2018). Based on unstandardized coefficient values a predictive equation has been formed and based upon standardized beta values the ranking of the independent variables has been generated.

C. Pre-Testing: Assumption check for Discriminant Analysis

- **Dependent variable must be in Nominal Scale**
  To apply Discriminant analysis (DA), the dependent variable must be categorical variable or nominal scale. Here, in the dataset dependent and target variable has two values 1 (Y=Minor) and 0 (N=Major). Assumption of discriminant analysis has been met with this particular attribute i.e. victim_type.

- **Independent Variable must be in metric data**
  In order to run discriminant analysis the independent variables (IDVs) or explanatory variables (EVs) must be in metric data. There are three IDVs are taken from the data for analysis namely Age, Height and Weight. All three variables have been checked; therefore all the values in the physical attributes are continuous in nature.

- **Testing of Outlier**
  Outliers are the extremely high or extremely low values of the dataset. In order to execute Discriminant analysis (DA), all the values in dataset must be free from outliers. So, boxplot method has been chosen to test the outlier detection in the data. Outlier detection and correction is the part of data cleaning process. Therefore, all the possible outliers have been corrected from the dataset.

D. Significance Testing and Independent Variables

Table 1 defines the results of the assumption i.e. no multicollinearity. No Multicollinearity exists when there is no correlation between independent variables. As shown by Table 1, the correlation values between the explanatory or independent variables are -.182, -.226, .257 between Age and height, Age and weight, and weight and height. Hence, very low correlation values have been found through the table. Therefore, the multicollinearity assumption has been full-filled with the data of independent variables. Multicollinearity only checked within the independent variables.

<table>
<thead>
<tr>
<th>Table-1: Correlation result between Variables</th>
<th>Table Column Head</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation</td>
<td>Age</td>
</tr>
<tr>
<td>Age</td>
<td>1.000</td>
</tr>
<tr>
<td>Height</td>
<td>-.182</td>
</tr>
<tr>
<td>Weight</td>
<td>-.226</td>
</tr>
</tbody>
</table>

Table- II: Wilk’s Lambda

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Table Column Head</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilk’s Lambda</td>
<td>.766</td>
</tr>
<tr>
<td>Significance</td>
<td>.000</td>
</tr>
</tbody>
</table>

As per the values of Wilk’s lambda the significance values are also produced. For all the variables the significance values are .000 which are less than .05. So, all the explanatory variables have their some contribution to predict the sexual offenders for minor and major victims. There is one more interpretation found through above table that all the values are significant, means a relation has been exist between independent and dependent variable. Therefore, age, height and weight of sexual offenders are related with target variable victim_type.

Retrieval Number: D9851118419/2019©BEIESP
DOI:10.35940/ijrte.D9851.118419

Published By:
Blue Eyes Intelligence Engineering & Sciences Publication
V. MODEL ANALYSIS AND RESULTS

After checking and fulfilling all the assumptions, the technique discriminant analysis has been applied on the dataset using IBM SPSS 21.

A. Ranking of Variables

As resulted from the Table 2, Age variable has been found the very good predictor to predict or classify offenders and got Rank-1 in prediction equation. The wilk’s lambda data value of the variable “Age” is 0.76 Weight is the second highest predictor among the all variables with its .86 score through wilk’s lambda. Third variable, height scored .93 score, means height is the low predictor in the predictive equation. Variables ranks are generated through their wilk’s lambda score, are mentioned in Table 4.

Table- III: Standardized beta values of Explanatory Variables

<table>
<thead>
<tr>
<th>Variables</th>
<th>Beta Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>.880</td>
</tr>
<tr>
<td>Height</td>
<td>.341</td>
</tr>
<tr>
<td>Weight</td>
<td>.592</td>
</tr>
</tbody>
</table>

B. Eigen value with canonical correlation

In this analysis canonical discriminant function has been used which has been provided the canonical correlation. Canonical correlation is the output of the association between discriminant scores and two groups (minor and major). Value of the canonical correlation close to 1 indicates strong correlation between the predicted score and two groups. In the above table the total canonical score found as 0.629 which indicates a good correlation. The table also gives the eigenvalue i.e. 0.653. Eigen value has been generated by explained part of regression equation by divided unexplained part of regression equation.

Table- V: Eigen values with Canonical Correlation

<table>
<thead>
<tr>
<th>Eigenvalue</th>
<th>% of Variance</th>
<th>Cumulative %</th>
<th>Canonical Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>.653</td>
<td>100.0</td>
<td>100.0</td>
<td>.629</td>
</tr>
</tbody>
</table>

Here, eigenvalue is good for the regression model so our model is good to predict. Two more values have been shown by the Table 5, cumulative percentage and percentage of variance. Percentage of variance helps to evaluate the most spread canonical variable. Cumulative percentage is responsible for the total dispersion by the canonical variable. Therefore, a good prediction of sexual offenders can be done on the basis of their physical traits age, height and weight.

C. Overall Model’s Significance Testing: Wilk’s Lambda

Table- VI: Eigen values with Canonical Correlation

<table>
<thead>
<tr>
<th>Wilks’ Lambda</th>
<th>Chi-square</th>
<th>Significance value</th>
</tr>
</thead>
<tbody>
<tr>
<td>.605</td>
<td>621.933</td>
<td>.000</td>
</tr>
</tbody>
</table>

Table 6 denotes the overall evaluation of Wilk’s lambda through three independent variables age, weight and height. There is un-explanatory part of the dependent variable by the independent ones. In the above table value of the wilk’s lambda is significance because in the last column i.e. significance column the value displays .000 which is less than .05.

D. Predictive Equation

A predictive discriminant regression equation has been formed through the un-standardized beta values of age, weight, height from Table 7. These values are described as explanatory part of each variable in the predictive equation. The discriminant equation has been mentioned below:

Sexual offender Category (DS) = -11.456 + .079 * Age + .008 * Height + .016 * weight

The above equation calculates the discriminant score (DS) for each variable for each record. Here in the above equation the value of the constant is -11.456. For every record of the physical characteristics of sexual offender, predicted value has been calculated for target variable i.e. victim type by putting the values of Age, Height and Weight. As per the discriminant score the predictive value of the target variable has been defined on the basis of discriminant function.

Table- VII: Variable Equation

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Un-standardized Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>.079</td>
</tr>
<tr>
<td>Height</td>
<td>.008</td>
</tr>
<tr>
<td>Weight</td>
<td>.016</td>
</tr>
<tr>
<td>(Constant)</td>
<td>-11.456</td>
</tr>
</tbody>
</table>

The discriminant equation has been formed through unstandardized coefficient scores only, not from the standardized coefficient scores. In the unstandardized score units are not same, even they are different for each individual independent variable. Standardized and normalized regression scores have been taken for assigning the ranks of IDVs and unstandardized coefficient scores are used to form the discriminant equation to find the discriminant scores.

Table- VIII: Group Probabilities

<table>
<thead>
<tr>
<th>Target Variable: Type</th>
<th>Prior Cases Used in Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unweighted</td>
</tr>
<tr>
<td>Major</td>
<td>.502</td>
</tr>
<tr>
<td>Minor</td>
<td>.498</td>
</tr>
<tr>
<td>Total</td>
<td>1.000</td>
</tr>
</tbody>
</table>

The above Table 8 displays the group probabilities of the dependent variable. As discussed in the data description the target variable has 2 categories major and minor types. In the first column probability of the victim minor and major has been defined as per the dataset.
Discriminant Analysis: Classifying and Predicting Sexual Offenders

From the total probability i.e. 1, 0.502 or 50.2% cases of major type have been used and 0.498 or 49.8% minor cases have been included to build discriminant model. In the next two columns of the above table number of cases is defined taken for the analysis i.e. 623 cases are of major category and 618 cases are of minor category type have been included.

VI. MODEL’S ACCURACY AND CLASSIFICATION SUMMARY

Table- IX: LDA Function at Group Centroids

<table>
<thead>
<tr>
<th>Target Variable</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Major</td>
<td>.804</td>
</tr>
<tr>
<td>Minor</td>
<td>-0.811</td>
</tr>
</tbody>
</table>

The above Table 9 discusses about discriminant function. The prediction of this target variable namely victim type has been done on the basis of discriminant function after evaluating score for each variable. For the prediction of sexual offenders of victim type minor and major, -0.811 & .804 boundary values have been found through the discriminant function. On the basis of these two threshold values the prediction of sexual offenders can be done on the basis of major and minor victims. The prediction range can be defined on the X-axis on the basis of two threshold values, which has been derived below:

(-1) (0) (+1)

(-0.811) (0) (.804)

(Minor Type) (Either Minor or Major) (Major Type)

Fig.1. Model Prediction

The above Figure 1 shows the threshold values have been segregated into 3 categories explained below:

A. Category-1

In this category, model predicts all the correct values of victim type minor. A threshold value has been defined here i.e. -0.811, which means that all predictive values of sexual offenders are of victim minor below the value -0.811. Below this number all the values are predicted for minor type and the actual values were also minor.

B. Category-2

This category belongs to the model prediction of major type and a threshold value has been found in this i.e .804. The major type prediction is found in the positive direction. If, discriminant score is greater than .804 for any type of cases then the particular case will surely come under the category of major. After checking the real data all the cases of sexual offenders were of major type whose discriminant score have found above .804.

C. Category-3

Between the range -0.811 to .804 there are some cases which can be either of minor or major. There are two probabilities would be possible. First, offenders were of minor victim but prediction found as major. Second, offenders were of major victim but prediction found as minor. The actual prediction has not done in between this range. The reason is that the model prediction accuracy is not 100%, it has been found less than 100%. So there might be some cases, can be wrongly predicted. So this zone is known as zone of confusion.

D. Zone of Confusion

As per out analysis the predictive score of below 0 (origin) must be of minor type but actual data do not match with it. Same, the discriminant score above than 0 must be come under the target category major but the real data do not say the same thing. Real data values are different from the predicted values in this zone. Hence, this particular zone between -0.811 to +.0.804 known as zone of confusion.

Table 10: Classification Results Summary

<table>
<thead>
<tr>
<th>Target variable: Type</th>
<th>Predicted Group Membership</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Major</td>
<td>Minor</td>
</tr>
</tbody>
</table>

Original

<table>
<thead>
<tr>
<th>Count %</th>
<th>Major</th>
<th>Minor</th>
</tr>
</thead>
<tbody>
<tr>
<td>586</td>
<td>37</td>
<td>623</td>
</tr>
<tr>
<td>147</td>
<td>471</td>
<td>618</td>
</tr>
<tr>
<td>94.1</td>
<td>5.9</td>
<td>100.0</td>
</tr>
<tr>
<td>23.8</td>
<td>76.2</td>
<td>100.0</td>
</tr>
</tbody>
</table>

As mentioned in Table 10, the accuracy rates have been defined both for the major and minor types. Classification and prediction results for sexual offenders have been clearly mentioned in the above table. The total column defines the total number of cases of both the category of target variable. Also defines the total percentages of cases used in the analysis. The column predicted group membership discusses about the total number of predicted and unpredicted cases. In case of first category namely “major type” total 623 cases involved in the analysis to build the predictive model. Out of 623, 586 cases are accurately predicted by the model and 37 are uncertainly predicted. The total percentage of the right prediction is 94.1% in case of sexual offender of major victim type and only 5.9% cases are wrongly predicted by the model. In case of minor victim type, the overall 618 cases are involved in the analysis. Out of 618, 471 cases of second category of i.e. minor type have been correctly predicted and 147 are left as uncertain situation. Overall, 76.2% prediction has been done correctly and 23.8% not done correctly in cases of minor type. The overall accuracy rate of the prediction is 85.2%, which has been generated by the discriminant analysis algorithm for both the categories i.e. victim minor and major. The accuracy rate generated by the proposed model is good. Therefore, the proposed model can predict 85.2% of the sexual offenders correctly. The proposed model can be used for the further prediction of the sexual offenders on the basis of their physical characteristics.
VII. CONCLUSION AND FUTURE SCOPE

Various research works have been done to analyse and predict the crime against women. Various policies have been drawn and various initiatives have been taken by the public and private organization to control it but the crime against women is increasing day by day. Sexual offences are the major crimes against women. To prevent and control these major sexual offences a classification and prediction discriminant model has been proposed. By using the discriminant analysis a first equation has been formed to predict the sexual offenders. Four attributes are taken to analyse the model namely age, height, weight and victim_type. One attribute type has two categories minor and major. This is known as a target variable of the dataset. The other three variables age, height and weight are the explanatory variables. The proposed model is 85.2% accurate to classify and predict the sexual offenders with the accuracy level of major and minor types are 94.1% and 76.2%. This model has been built on the basis of very important physical characteristics of sexual offenders like age, height and weight. Through the discriminant function a threshold values have been generated for two groups of target variable i.e. minor and major. The sexual offenders for minor type of victims would be predicted through this model when discriminant score will be below -0.811 and correct prediction will be done for major type when discriminant score will be calculated above 0.804. The model can be used further to classify and predict the sexual offenders on the basis of two categories major and minor. The proposed model can be used for the security implementation for minor and major separately in schools, colleges, workplaces, parks etc. Through this model various crime and criminal organization can take the measures of both kinds of victims. This model can be enhanced by adding more attributes of sexual offenders or offences.

ACKNOWLEDGMENT

We would like to show our gratitude to Ms. Akanksha Upadhyaya, Assistant Professor, Rukmini Devi Institute of Advanced Studies, New Delhi for sharing her pearls of wisdom about the technique applied in this research work.

REFERENCES

Discriminant Analysis: Classifying and Predicting Sexual Offenders


AUTHORS PROFILE

Ms. Bhajneet Kaur is Research Scholar at Amity Institute of Information Technology, Amity University, Noida. She is also working as an Assistant Professor in Rukmini Devi Institute of Advanced Studies, New Delhi affiliated to GGSIPU. She has teaching and research experience in the field of Computer science, Information Technology and Management. She has presented research papers in various National and International conferences, also received the best paper award. Her research work got published in National and International journals including scopus enlisted journal. She has also been invited as a resource person in research field in various faculty development programs.

Dr. Laxmi Ahuja M.Phil(CS) and Ph.D(CSE) working as Professor in Amity University with over 16 years of experience in academics. Published more than 40 research papers in International and National Journals and Conferences. Her area of expertise includes Search Engine, Data Mining and Cloud computing She has organized various Guest Lectures and Workshops and participated in many workshops and Faculty development Programmes. She is actively involved in research activities She has currently Eight research scholar under her guidance. She has also published number of Books related to operating system and Information Technology. She is Member of Computer Society of India (CSI) vice chairman of IETE The Institute of Electronics and Telecommunications Engineers and senior member of IACSIT. She is a Member of Technical Programme Committee (TPC) in various conferences across globe.

Dr. Vinay Kumar did his Ph. D. from University of Delhi, MCA from Jawaharlal Nehru University, Delhi and Master of Science in Mathematics from Bihar University. He has extensively contributed to the knowledge arena through his books and research publications. He worked as scientist in National Informatics Centre (NIC), Department of Communication & IT, Government of India for approximately 21 years. He worked, then, with Vivekananda Institute of Professional Studies (VIPS), affiliated to Guru Gobind Singh Indraprastha University, as Professor of Information Technology. He worked as Dean of the IT Department and Department of Research and Publication till 5th July, 2018. He also worked as chief editor of bi-annual journal named “Vivekananda Journal of Research” during the period. He has Published over 100 research papers in refereed journals. He has authored a book of Discrete Mathematics and a memoir Killed Instinct. He also contributes to Times of India blog.