MATH 4939 Project
Team 3

Introduction

"Arrests" dataset is based on the police treatment of individuals arrested in Toronto for possession of small amounts of marijuana from 1997 to 2002. The dataset is just a part of a large data set mentioned in a series of articles in the Toronto star. The dataset contains 5226 observations with 8 variables as below.

y=released: whether or not the person who is arrested is released with a summon (Yes or No)
x_1=colour: The arrested persons race (Black or White)
x_2=year: 1997 - 2002
x_3=age: The age of the arrested person in years
x_4=sex: Gender of the arrested person (Male or Female)
x_5=employed: Is the arrested person employed (Yes or No)
x_6=citizen: Is the person a citizen of toronto (Yes or No)
x_7=checks: Number obtained from the police databases (of previous arrests, previous conviction, parole status, etc.) the arrested persons name appeared upon labeled from 1 to 6

According to the dataset, the variable "released" is the independent variable y, and the rest are dependent variables. In this project, we will build two models using logistic regression method, compare the two models in different ways, and find out the factors which can influence the independent variable "released" significantly in order to explore the patterns of discrimination in the dataset.

Model

We will build two models for the dataset. The first model is a simple model, which is : $\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7$. Another model is more complicated, which contains the interaction terms and quadratic terms.

We are going to compare these two models in different ways and use the better one to find out the potential patterns of discrimination.

Compare and Analyze Models in Different Ways

1.Wald test and likelihood test
2.Forward or Backward Selection
3.Anova analysis
4.95% C.I. data ellipses for coefficients β_i
5.xy plot
6.Residual plots
7.Added Variable Plot
8.Box-Cox plot
9.Data ellipses
10.Influence and leverage
11.Prediction

Conclusion

According to all the analysis we have done, we find there exists discrimination as for if a person who is arrested will be released or not.