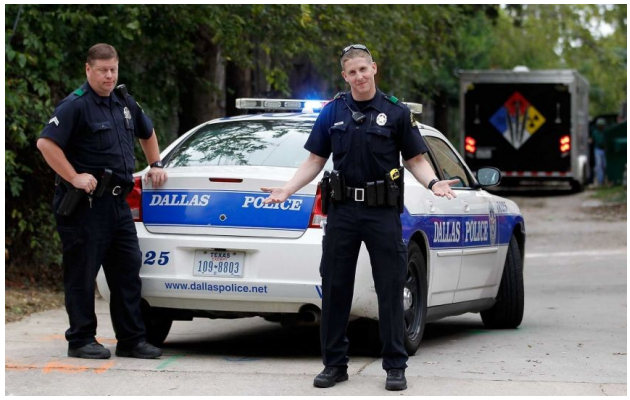
Data Mining Spring 2016

Project 2: Predictive Modeling (Classification)

Assigned: 2/17/2016
Due: 3/16/2016 (via Canvas)
Points: 100

Please submit your report in **PDF format.**



You will work again with crime report data for Dallas published by Dallas Open Data, and you can use the cleaned data from Project 1.

Write a report covering in detail all steps of the project. The results have to be reproducible using your report. Carefully describe every assumption and every step in your report. Also, mention any program/code/additional data that you are using for your analysis.

Here are two questions you should answer (you can come up with more questions counting towards exceptional work):

1. Can we predict if an incident will result in an arrest?
2. Can we predict some types of offense using the other data?

Note: Be careful what features you use for prediction. Not all information might be available at the time you need to make the prediction. For example, the Penalty Code Class (e.g, Misdemeanor A, Felony 1) might only be added after the investigation is concluded. However, you would like to predict if an arrest will follow before deciding how to optimally allocate scarce policing resources between investigations. You need to argue why you are using certain features.

Follow the CRISP-DM framework

Steps 1 and 2 have already been performed in Project 1.

3. **Data Preparation [30]**

- Define and prepare your class variables used for the different questions. You may decide to discretize or aggregate (i.e., combine values) for your class and/or other features. [20]
- Select features that might be useful for modeling. Create features if necessary (e.g., transformation to rates, time differences, etc.). [5]
- Describe the final dataset that is used for classification (include the scale/range of new features) [5]

4. **Modeling [50 points]**

- Create at least 5 different classification models (different techniques and different parameters). [20]
- Discuss the advantages of each model for this classification task. [5]
- What are the most important features found by each model. Are they the same. Discuss what this means. [5]
- Assess how well each model performs (use training/test data, cross validation, etc. as appropriate). [20]

5. **Evaluation and Deployment [10 points]**

- How useful is your model for the stake holders (e.g., law enforcement, general public, politicians)? How could stake holders act on the model. [5]
- How would you measure the model's value if it was used. [5]

Exceptional Work [10 points]