CSE583 - Big Data & Policing

HW 3

Deadline - 8 March, 2019
Total Marks - 140

Instructions

1. Preferably use Python3 for this assignment.
2. All datasets pertaining to this HW are available in this Google Drive folder
3. The final submission must consist of all your source code and a report containing all your findings.
4. In case of any doubts/issues, please post on the Moodle forum.
5. Do not copy. This could result in an 'F' grade in the course.

Part 1 (70)

- Use the `Isis Admin's Hacked Phones.zip` dataset to answer the following questions -
  1. Which company's phone is the most popular amongst ISIS operatives across all countries? Use a bar plot to show the distribution of phone manufacturers. (10)
  2. What are the top 5 most popular OUTGOING numbers? What are their names, and which country do they belong to? (10)
  3. What are the top 5 most popular INCOMING numbers? What are their names, and which country do they belong to? (10)
  4. Show radial plots representing the frequency of calls made in Yemen, Syria and India. Is there a difference in call pattern? Or do all three countries make calls at similar timings? Note: Remember to convert all times to local time before plotting the radial plots. (20)
  5. Using a bar plot, compare and contrast the average call duration for ISIS operatives across all the countries. Which country has the highest average? Does it also have the highest number of calls? Similarly, which country has the lowest average, and does it also have the lowest number of calls? (20)
  6. [BONUS] Across all the images and videos, find the face of the ISIS operative which appears most number of times. Which country does he/she belong to? (15)

Part 2 (70)

- Use the `savry_2017.csv` dataset. Refer to the `model.ipynb` file provided to understand the existing prediction model, as explained by Chato sir in a guest lecture. Go through the dataset carefully, and answer the following -
  1. Determine the relationship between human risk assessment (`professional_risk_evaluation`) and SAVRY total score (`sum_all_risk_items`). Is there a correlation? (10)
  2. What are the top 3 offenses predicted least accurately by the current model? (10)
  3. Instead of trying to predict recidivism, using the same model predict recidivism_severity. Using an ROC curve, explain how the prediction is better/worse. (10)
  4. Using the same set of features, build two models to predict violent and non violent recidivism separately. Are predictions more accurate for these two models as compared to the existing model? (10)
  5. Use two other learning models of your choice to predict recidivism (e.g. SVM). Explain the difference/similarity in prediction results. Compare and contrast the ROC curves of all three models. (30)