Advanced Concepts in Data Analytics

Final Project: New York City Case Study using CRSIP-DM

Weight: 30% **Marks: /400**

Student Name:

Student ID:

Date:

This page was intentionally left blank.

Table of Contents

Final Project: New York City Case Study using CRISP-DM

Introduction

For the final project in this course, you'll use the skills you've developed in this course to analyze real-life data from the New York City Police Department. You'll focus on different aspects of data analysis and then create four reports to reveal your findings.

This project gives you the opportunity to:

- Apply your learning to a real-life data analysis problem
- Demonstrate your advanced data analytics skills
- Demonstrate your achievement of the course learning outcomes

Some class time will be provided for the four project sections, and your instructor will be available to answer questions and provide support. However, you will do much of the work on your own. Your first task is to read this student manual and be clear about the deliverables

Background

CRISP-DM

Cross-industry standard process for data mining, known as CRISP-DM, is a widely used, open standard process model that describes common approaches used by data mining experts. It breaks down the process of data mining into six major phases:

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

The sequence of the phases is not strict, and moving between different phases is usually required, as demonstrated in the figure below. You can familiarize yourself with CRISP-DM by reading the supplementary CRISP-DM 1.0 manual from IBM (ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf).

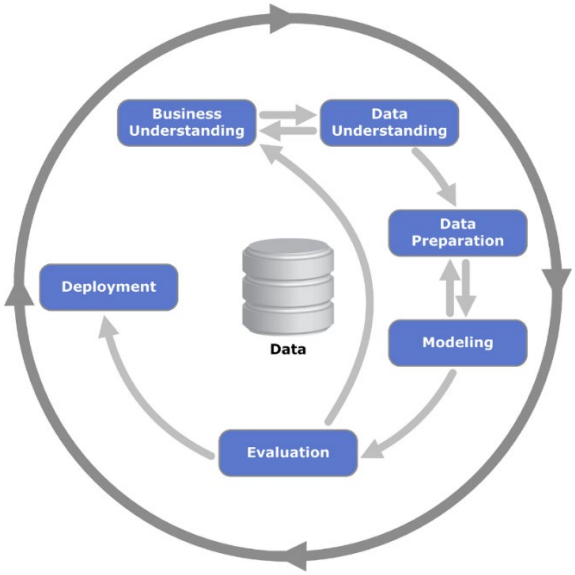

Figure 1: CRISP-DM Phases

Source: Figure 2 in CRISP-DM 1.0 Manual

New York City's Stop-Question-Frisk (SQF) Data

The Stop-Question-and-Frisk (SQF) program of New York City's Police Department is a policing strategy by which officers stopped and questioned hundreds of thousands of pedestrians annually and searched them for weapons and other contraband.

For this project you'll examine stop-and-frisk data for the year 2012. You'll following the CRISP-DM framework to generate findings and insights about the SQF program to include in your reports.

You can obtain the original data from Stop, Question and Frisk Data (https://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page).

Deliverables and Deadlines

There are four deliverables for this project:

1. Data and Visualization Report (7.5%)
2. Association Rule Mining Report (7.5%)
3. Cluster Analysis Report (7.5%)
4. Predictive Modelling Report (7.5%)

All report deliverables must be submitted by the end of the final day of class. It's highly recommended that you start early and improve your analysis along as the class progresses.

Report Guidelines

The data analytics project consists of four sections that combine to make a whole deep dive into the SQF data. You will create a report for each of the sections and summarize your findings with different analytical perspectives.

Think of each report as a professional document that a consultant would prepare for a client. The main goal is to provide value for the client. Use the following guidelines to prepare each report.

Structure

- Use a title page with the project title and your name.
- Include an abstract/executive summary (less than 1 page) that introduces the problem and highlights your key results.
- Organize the report using CRISP-DM, and then by issue (e.g., the impact of race)
- Number the pages.
- Include a table of contents for longer documents.
- End the report with a conclusion that summarizes the main findings and includes a list of recommendations.

Tables

- Do not copy tables from your analysis tool (i.e., don't use copy and paste or include screen shots). Create your tables in Word or Latex, and only add necessary information. Highlight important information.
- Tables need to be numbered and referenced in the text. Discuss what we can learn from the data in the tables.
- Use a table to describe features in a consistent and easy-to-read way.

Figures

- Include a numbered caption with every figure or plot. Discuss it in the text and reference it by number. What does it reveal and how is it helpful? Figures without a discussion are useless.
- If you have several options to plot the same thing, then choose the best visualization. Highlight important aspects using graphs.
- All graphs must include proper names for both axes.
- Make sure your graphs are readable (i.e., text/numbers in graphs should be about the same size as the regular text in your document). Graphs should be appealing (e.g., colors, legend). Figures should not waste space!

Code

- If it is necessary to include code, upload it in a separate file.
- You can include short pieces of code in the main report if they are especially interesting.

Report 1: Data and Visualization

Weight: 7.5% **Marks: /100**

Marking Criteria

Follow the CRISP-DM framework and answer the following questions as part of your report.

Business Understanding [20 marks]

- What is the purpose of the SQF program?
- How would you define and measure the effectiveness of such a program?
- What data would you need be able to judge its effectiveness?

Data Understanding [80 marks]

- Describe the meaning and type of data (e.g., scale, values) for each attribute in the data file. [10 marks]
- Verify data quality. Are there missing values? Duplicate data? Outliers? Are those mistakes? How do you deal with these problems? [20 marks]
- Give simple, appropriate statistics (e.g., range, mode, mean, median, variance, counts) for the most important attributes in these files, and then describe what they mean or whether you found something interesting. [10 marks]

  Note: You can also use data from other sources for comparison.
- Visualize the most important attributes appropriately (at least 5 attributes). [15 marks]

  Important: Provide an interpretation for each chart, explaining each attribute and why you chose the visualization you did.
- Explore relationships between attributes. Look at the attributes and then scatter plots, correlation, cross-tabulation, group-wise averages, etc., as appropriate. [15 marks]
- Compare the reasons for an SQF and what type of force was used by the officer. [10 marks]

Report 2: Association Rule Mining

Weight: 7.5% **Marks: /100**

Marking Criteria

Follow the CRISP-DM framework and answer the following questions as part of your report.

Data Preparation [30 marks]

- Construct the required transaction data set for frequent itemset and association rule mining.

Modelling [60 marks]

- Create frequent itemsets and association rules. [50 marks]
- Use tables and visualizations to help explain your results. [10 marks]

Evaluation [10 marks]

- What findings are the most interesting? Why?

Report 3: Cluster Analysis

Weight: 7.5% **Marks: /100**

Marking Criteria

Follow the CRISP-DM framework and answer the following questions as part of your report.

Data Preparation [20 marks]

- Define and prepare your class variables. [10 marks]

  Note: You may have to combine different columns.
- Remove variables that are not needed/useful for the analysis. [5 marks]
- Describe the final dataset that is used for classification and include the scale/range for the new combined variables. [5 marks]

Modelling [75 marks]

- Perform cluster analysis.
  - Cluster the location for a crime of your choice. [20 marks]

    Note: Found clusters might be different depending on the time of day.
  - Cluster stopped people by reasons for stop. [20 marks]
  - What else can you use cluster analysis for in the data set? [20 marks]
- How did you determine a suitable number of clusters for each method? [10 marks]
- Use internal validation measures to describe and compare the clusters (some visual methods would be good). [5 marks]

Evaluation [5 marks]

- Describe your results. What findings are the most interesting? How can these findings be used?

Report 4: Predictive Modelling

Weight: 7.5% **Marks: /100**

Some examples of classification tasks are (you can use others):

- Can we predict if a person is armed using the other variables?
- Can we predict if an arrest will be made?
- Can we predict what force will be used by an officer given the other variables?

Marking Criteria

Follow the CRISP-DM framework and answer the following questions as part of your report.

Data Preparation [30 marks]

- Define and prepare your class variables. [20 marks]

  Note: You may have to combine different columns.
- Remove variables that are not needed/useful for the analysis. [5 marks]
- Describe the final dataset that is used for classification and include the scale/range for the new combined variables. [5 marks]

Modelling [60 marks]

- Create at least three different classification models (different techniques) for each of the classification tasks. [30 marks]
- Discuss the advantages of each model for this classification task. [5 marks]
- What are the most important variables found by each model? [5 marks]
- Assess how well each model performs (use training/test data, cross validation, etc., as appropriate). [20 marks]

Evaluation [10 marks]

- How useful is your model for the police? How would you measure the model's value if it were used? [5 marks]
- How would your implement your model to improve policing? What other data should be collected? How often would your model need to be updated? [5 marks]