***Canton Police Department Proposal***



***By Curtis Corazao***

***Phases***

● Business Understanding  
● Data Understanding  
● Data Preparation  
● Modeling  
● Deployment

Data Understanding

Business Understanding

Data Preparation

**Data**

Modeling

Deployment

***Tasks and Outputs Overview***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Business Understanding** | **Data Understanding** | **Data Preparation** | **Modeling** | **Deployment** |
| **Determined Business Objectives:**  Make Canton safer through effective police resources based on collected crime data related to violent crimes.  Success is to understand the causes using quantitative methods to determine the relationship to the violent crime data.  **Assess Situation:**  Personnel within the canton police department is available to achieve business objectives.  Data is based on subzones of the town of canton with calls of various crimes reported, male/female population, employment and wage income details of subzones.  Crime is assumed to be higher in lower income subzones.  Officer lives are put at risk as they enforce the laws in these high crime rated subzones with the benefit of creating a stable and lawful town.  **Determined Data Mining (DM) Goals:**  Identify Subzones within Canton that show high calls for violent crimes. Determine the average income and average age of the population within the subzones of focus.  **Produce Project Plan**  Initial project plan will include collecting data about violent crimes (e.g., number of rapes, assaults, burglaries), and the data related for the subzones, income levels, ages and additional demographic details that might relate to violent crimes. | **Collect Initial Data**  Existing data includes subzones of town, male/female population, income per capita, average household income, median age, education levels with addition to violent crime target variables.  **Describe Data**  Data includes quantitative measures of every valid crime incident such as date of the incident, age, income level, subzones, education level, and sex. In addition, it has individual call incidents about violent crimes such as rape, assaults and burglaries.  **Explore Data**  Data needs to be visually inspected and determine what to do with missing values. For example, what approach should be taken when income or age is missing from an individual violent crime call?  **Verify Data**  Data will be checked for null values and fixed in a meaningful way in order to perform a valid analysis. For example, if the income is missing, we could use the average income for the subzone. Same approach could be used for missing age. | **Select Data**  Determine which rows of data to include based on subzones with high calls of violent crimes. In addition, identify outliers and possibly exclude from the analysis to avoid erroneous conclusions.  **Clean Data**  This step is important to update any missing values with meaningful data. For example, use an average of income or age for null values in those columns.  **Construct Data**  This step involves codifying any categorical data into quantitative data so that it can be used in data science algorithms.  **Integrate Data**  Merging and appending data sets for analysis. For example, merging the data from one subzone to another subzone.  **Format Data**  This step involves which models to use and formatting the data to fit the data format requirements. Transforming data into vector data. | **Select Modeling Techniques**  This step involves which models to use to achieve desirable analysis outcome. For example, using the correlation analysis to determine which feature variables might be related to the target variables and validate statistical assumptions; For example, normality of the data.  **Generate Test Design**  Measure the goodness of the models that are built for analysis. For example, looking at the coefficient of correlations in a regression model and validating the statistical assumptions. For example, addressing multicollinearity issues.  **Build Model Parameter**  Importing the data into a model using python and deciding which portion for the base model and which portion for the testing.  **Assess Model**  Running the base model, using the test data, and comparing against other models to determine validity of the predictions as they explain target variables. | **Plan Deployment**  Create a detailed technical deployment plan which involves data collection from source systems and algorithms with existing systems. For example, using predictions from the models and displaying their results on existing systems.  **Plan Monitoring and Maintenance**  Determine which factors or influences need to be monitored for each model produced. This step involves reviewing the outcomes from the models and determine the needs for modifications for improvements of the data or model.  **Produce Final Report**  Prepare final report for presentation to top officials within the canton police department of the findings of what variables are most likely related to violent crimes so C.P.D. can take appropriate measures.  **Review Project**  The sheriff and other top ranking officers in the canton police department will review and provide feedback on the finished project. This involves reviewing the overall goals of what factors are related to violent crimes, cost and timing expectations of the project. |

***Business Understanding***

**Description** **Examples of Crime**

**Violent Crimes of interest:**

* Data consisting of calls for burglaries, assault, murders, fights, armed subjects, rapes etc.

**Problem Area:**

* Violent crime is active in several subzones within the town of Canton.
* The data mining within this project is intended to identify subzones within Canton that are most effected by violent crime and respond using all available personnel/resources within C.P.D.

**Determined Business Objectives:**

* Make Canton safer through effective police resources based on collected crime data.
* Success is to understand the causes using quantitative methods to determine the relationship to the violent crime data.
* It is assumed lower income subzones have the most volume of violent crime due to poor economic circumstances for the population.
* Business requirements will include patrol officers assigned to subzones with surveillance cameras set up at key areas of interest.
* Nonviolent crimes such as minor drug possession should be considered not a priority. Officers should issue ticket fines to offenders of minor drug possession if they find them instead of formal prosecution.

**Assess Situation:**

* Personnel within the canton police department is available to achieve business objectives.
* Data is based on subzones of the town of canton with calls of various crimes reported, male/female population, employment and wage incomes details of subzones.
* Crime is assumed to be higher in lower income subzones.
* Officer lives are put at risk as they enforce the laws in these high crime rated subzones with the benefit of creating a stable and lawful town.
* Officers are expected to stay vigilant and aware of their surroundings while performing their duty. Secured radio channels are available and open for additional police reinforcement or medical assistance.

**Determined Data Mining (DM) Goals:**

* Identify Subzones within Canton that show high calls for violent crimes. Determine the average income and average age of the population within the subzones of focus.
* Build models using available crime data to predict likelihood of subzones with the most violent crimes.
* Assign each subzone a rank to help determine the volume of violent crime activity.

**Produce Project Plan**

* Initial project plan will include collecting data about violent crimes (e.g., number of rapes, assaults, burglaries), and the data related for the subzones, income levels, ages and additional demographic details that might relate to violent crimes.

***Project Plan Timeline***

|  |  |  |  |
| --- | --- | --- | --- |
| ***Phase*** | ***Time*** | ***Resources*** | ***Risks*** |
| Business Understanding | Week 1 | All Analysts | Shift in C.P.D. resources against crime. |
| Data Understanding | Week 2 | All Analysts | Data problems, technology problems. |
| Data Preparation | Week 3 | Data mining consultant, some database analyst time | Data problems, technology problems; absence of data in electronic media. |
| Modeling | Week 4 | All Analysts | Technology problems, inability to find adequate model |
| Deployment | Week 5 | Data mining consultant, some database analyst time | Shift in C.P.D. resources against crime. Inability to implement results |

***Data***

***Understanding***

**Description Canton Police Examples**

**Data availability:**

* Data is assumed to be in electronic media, if the data is not available then the project will have to include tasks to convert the data from paper sources to electronic media. The data should include demographics of age, gender of the incidents within subzones in addition to income level details. Race should be included to the data if possible.
* More data is likely to be available as the project progresses.
* More details about the demographic details of subzones is desirable for analysis if it’s not already present.

**Collect Initial Data**

* Existing data includes subzones of town, male/female population, income per capita, average household income, median age, education levels with addition to violent crime target variables.
* Data shows demographics of age and sex of the population of subzones. Race should be an addition to the data if possible.

**Describe Data**

* Data includes quantitative measures of every valid crime incident such as date of the incident, age, income level, subzones, education level, and sex. In addition, it has individual call incidents about violent crimes such as rape, assaults and burglaries.
* The amount of data is assumed to be enough for analysis; the more data the better.
* Coding schemes columns in like Subzones, household earnings, median age, income per capita, male/female population, education level etc.

**Explore Data**

* Data needs to be visually inspected and determine what to do with missing values. For example, what approach should be taken when income or age is missing from an individual violent crime call?
* Hypothesis may change as we dive further into what the data shows.
* Subsets of the data need to be identified earlier on for analysis.

**Verify Data**

* Data will be checked for null values and fixed in a meaningful way in order to perform a valid analysis. For example, if the income is missing, we could use the average income for the subzone. Same approach could be used for missing age.

***Data Preparation***

**Description**  **Canton Police Examples**

**Select Data**

* Determine which rows of data to include based on subzones with high calls of violent crimes. In addition, identify outliers and possibly exclude from the analysis to avoid erroneous conclusions.
* Determine what data will be used for the base model construction and which will be used for testing the model.

**Clean Data**

* This step is important to update any missing values with meaningful data. For example, use an average of income or age for null values in those columns.
* This involves filling in missing data, addressing data errors, coding inconsistencies and bad metadata.

**Construct Data**

* This step involves codifying any categorical data into quantitative data so that it can be used in data science algorithms.

**Integrate Data**

* Merging and appending data sets for analysis. For example, merging the data from one subzone to another subzone.
* Merging and appending data sets will be considered depending on what the data offers.

**Format Data**

* This step involves which models to use and formatting the data to fit the data format requirements. Transforming data into vector data.
* These may include regression models and cluster analysis.

**Data preferences:**

* Data involving demographic details, income and subzones should be the main focus in the analysis. For example, incident data of violent crime that includes date of the incident and demographic specifics described previously of the subject involved in the incident. In addition to the type of violent crime.

***Modeling***

**Description** **Canton Police Examples**

**Select Modeling Techniques**

* This step involves which models to use to achieve desirable analysis outcome. For example, using the correlation analysis to determine which feature variables might be related to the target variables and validate statistical assumptions; for example, normality of the data.
* Models will include regression, cluster, vector, SVD and histogram analysis.

**Generate Test Design**

* Measure the goodness of the models that are built for analysis. For example, looking at the coefficient of correlations in a regression model and validating the statistical assumptions. For example, addressing multicollinearity issues.
* Goodness will include R squared values, cluster nets, etc.

**Build Model Parameter**

* Importing the data into a model using python and deciding which portion for the base model and which portion for the testing.
* Parameter settings may involve only including subzones and their details that exceed above a certain volume of violent crime data.

**Assess Model**

* Running the base model, using the test data, and comparing against other models to determine validity of the predictions as they explain target variables.
* Each model should be considered how difficult it is to implement in deployment.
* Subzones with the lowest household income are expected to have the highest volume of violent crime in the area.
* All variables are considered when keeping or dropping when creating the most accurate R-squared values.
* Using SVD analysis when we have a lot of missing values.

***Deployment***

**Description**  **Canton Police Examples**

**Plan Deployment**

* Create a detailed technical deployment plan which involves data collection from source systems and algorithms with existing systems. For example, using predictions from the models and displaying their results on existing systems.

**Plan Monitoring and Maintenance**

* Determine which factors or influences need to be monitored for each model produced. This step involves reviewing the outcomes from the models and determine the needs for modifications for improvements of the data or model.
* Models should be considered expired after a three years.
* Years of data should be collected to show the level of progress of lowering crime in certain subzones.

**Produce Final Report**

* Prepare final report for presentation to top officials within the canton police department of the findings of what variables are most likely related to violent crimes so C.P.D. can take appropriate measures.
* Summary of data mining, results both models and findings
* Recommendations for further data mining work, including interesting leads discovered.

**Review Project**

* The sheriff and other top ranking officers in the canton police department will review and provide feedback on the finished project. This involves reviewing the overall goals of what factors are related to violent crimes, cost and timing expectations of the project.
* Models like the regression model should be deployed immediately with their findings and recommendations to leadership of C.P.D. to explain the relationship between feature variables to the target variables of violent crime.

***Summary of Proposal***

* Regression models, cluster analysis and other models will help determine what subzones to focus police resources on in order to lower crime in the town of Canton.
* The project is intended to help identify subzones within Canton with the highest volume of violent crime so C.P.D. can commit its resources and personnel to those areas in order to lower the crime rate.
* Enables the creation of clean and transformed data  
  prior to modeling with the goal of addressing issues of missing values and codifying the demographic data for analysis.
* Established procedures and best practices for  
  modeling; for example, addressing issues of data normality, evaluation of the base models and deployment and integration into existing systems.
* The cost of implementing this proposal will involve estimating the cost C.P.D. resources and the data science consulting resource cost which could be finalized at a later stage.