An Analysis of Canadian Prisoner Parole Data with Clustering and Statistical Methods

CSCI 4146

John Phillips
B00685175

Chaoran Zhou
B00551572

Introduction

The legal imprisonment of criminals has been a practiced concept in many cultures for centuries. Public support for imprisonment depends on popular notions of fair justice in regard to the deciding factors of imprisonment and release. While there are factors that should rightfully influence a prisoner's release such as the nature of the crime committed, there are also factors that should remain proportional to the population means, such as race and gender.

We have obtained a publicly available dataset that contains records of Canadian Parole Grants and Revocations. With this data in hand, we aim to find which factors, if any, are the most significant in whether a prisoner is allowed parole or not. In order to check this, multiple feature selection methods will be used to identify important features and then these results will used to construct predictive cluster models. Where applicable, this hypothesis is measured with a significance of 95%.

Dataset

For this project, the Conditional Release – Appeal Decisions datasets from 2009 through to 2017 obtained through the Government of Canada's Open Data Inventory were collected. All of the annual datasets of this type contained the same columns and type system, so the 8 most recent were used for this project. In total, there are 17 columns in each dataset. Besides columns such as the year and month of the parole appeal decisions, there are a few columns that need more in-depth explanation.

The Decision Type column describes what the outcome of the parole appeal meeting is intended as. For example, some of the values for this column are, "Day Parole – Denied", "Full Parole – Granted", "Full Parole – Revoked", among many others. There are 52 total Decision Types existing in the datasets that were collected for this project out of a total of 76 that are outlined by the source. There is another column called Decision Type Group that is a higher hierarchical description of the Decision Type column, with 10 different values. The Final Decision column notes if the decision specified in the Decision Type column was affirmed or if further review or hearings are required at a later date.

The Decision Purpose column indicates whether the prisoner meets the requirements for Accelerated Parole Review (APR). APR is a method introduced by the government to help free prisoners who were convicted of non-violent crimes quicker. There is another column called APR Qualifier which indicates if the prisoner who meets the criteria for APR is in the initial or final stages of applying for parole.

The Jurisdiction column specifies whether the prisoner is under a federal or provincial sentence, where provincial sentences are all sentences under two years and federal sentences are all sentences over two years. There is also a Sentence Type column, which specifies if the sentence

has a determined end date or if is indeterminate. Additionally, there is a Major Offence Group column, which describes the crime of the prisoner's imprisonment. Examples of the values in this column include "First Degree Murder", "Schedule 1 Offence of the Corrections and Conditional Release Act (CCRA) of a Sexual nature", and "Non-Schedule Offences of the CCRA".

Additionally, there is a Race column that contains the self-reported race of the prisoner and Race Group column which aggregates the similar values of the Race column into 5 overall categories.

Data Preparation

By aggregating the yearly datasets together, the resulting dataset contained 4927 rows of parole decisions.
Because our hypothesis is focused on finding attributes that bias a prisoner's ability to get parole granted, we created a Label column that contains either a Positive value, where the prisoner is granted some form of parole, or a Negative value, where the prisoner is either denied parole or has some form of privilege cancelled. The distribution of the data within this label field is 935 Positive and 3992 Negative, roughly approximating to 1/5 positive and 4/5 negative. With this knowledge, the tests in our project were conducted with two different datasets: a dataset with all valid 4287 records, and a dataset with 1000 records, containing 500 randomly selected positive records and 500 randomly selected negative records. The Decision Type column was then removed from both datasets.

The Review Type column was dropped altogether, as every record during this timeframe had the same value and was made as a paper-based decision. The Race Group column was favoured over the Race column as the Race column is based on self-reported values and thus includes some overlapping values (South Asian, S.E. Asian, Asiatic, Asi-E/Southeast, etc). All of the records that had a Final Decision value of anything other than Affirmed were removed, as it is not possible to trace the outcome of the decision. After this modification, the dataset contained 4287 records. The APR Qualifier column was then dropped, as only 252 records had a value of some sort for it, and the effect of APR on the decision could be seen in the Decision Purpose column regardless.

Following the preceding changes, we converted our dataset into a binary dataset via the use of one hot encoding. This was chosen over converting to label encoding due to the unranked nature of the dataset's categorical values.

Feature Selection

In order to find the attributes of the dataset that had the most impact on the record's label, we chose to employ multiple feature selection methods. Specifically, we selected Chi-square as a filter method, Forward Selection as a wrapper method, and XGBoost as an embedded method.

Chi-square

Chi-square was selected as a feature selection method due to its ease of implementation and ability to work with categorical data. The results of the Chi-square test with a p value of less than 0.05 on our sampled data can be seen below in table 1. The result of the Chi-square test on the total dataset can be seen in Table 3 on the following page.

Table 1: Chi-square Feature Selection on Balanced Dataset

Feature	P Value	Score
Sentence Type: Indeterminate	1.06768e-24	105.266
Jurisdiction: Provincial	5.29656e-14	56.6165
Major Offence Group: Second Degree Murder	3.93729e-11	43.6448
Major Offence Group: Schedule 1 Without Sex	4.04812e-7	25.6713
Major Offence Group: First Degree Murder	4.55663e-7	25.4429
Sentence Type: Determinate	1.52224e-6	23.1197
Race Group: Other/Unknown	1.23817e-3	10.4325
Race Group: Asian	1.01118e-2	6.61509
Jurisdiction: Federal	4.75445e-2	3.92601

We selected the three features from this list with the highest Chi-square score and lowest p value for our analysis. These three values corresponded to the top three features of the Chi-square test from the whole dataset, whereas the following attributes had some variance. There was also a concern that using all of the values above 95% significance would lead to the model being over-trained on the training dataset. We predicted that these three would likely be reflected as significant features in most samples.

Table 2: Chi-square Feature Selection on Total Dataset

Feature	P Value	Score
Sentence Type: Indeterminate	5.03794e-27	115.885
Jurisdiction: Provincial	8.04197e-15	60.325
Major Offence Group: Second Degree Murder	7.16979e-12	46.9805
Major Offence Group: First Degree Murder	1.7769e-7	27.2618
Sentence Type: Determinate	1.99816e-7	27.0349
Major Offence Group: Schedule 1 Without Sex	1.39052e-6	23.2937
Race Group: Other/Unknown	5.44519e-4	11.9567
Race Group: Caucasian	1.14739e-2	6.3904
Race Group: Aboriginal	1.33436e-2	6.12298
Race Group: Asian	1.66497e-2	5.73293
Jurisdiction: Federal	4.48961e-2	4.02254

Forward Selection

Forward selection was selected as a feature selection method due to its easy approach and its ability to generalize and avoid overfitting, which was suspected to be a potential problem with the dimensionality of the dataset. We implemented a separate Forward Selection method for each of the clustering algorithms that were to be used. Each algorithm would be run with each attribute individually, and then the attribute that produced the highest harmonic mean between the homogeneity and completeness would be selected as a feature. This feature would then be carried over into the next iteration and the process would be repeated until the maximal harmonic mean was reached. The results can be seen listed below in Table 2.

Table 3: Forward Selection for Each Cluster Algorithm

K-Means	DBSCAN	Affinity Propagation
Sentence Type: Indeterminate	Sentence Type: Determinate	Sentence Type: Determinate
Jurisdiction: Federal	Jurisdiction: Federal	Jurisdiction: Federal
Decision Purpose: By Exception	Decision Purpose: By Exception	Race Group: Caucasian

XGBoost

XGBoost is a gradient boosting algorithm for feature selection. It was chosen as one of the tools for this project due to being one of the fastest running gradient boosting algorithms and its strong performance on modeling datasets. The results of the XGBoost algorithm for a sampled dataset can be seen below in Figure 1 with the results on the entire dataset in Figure 2 below it.

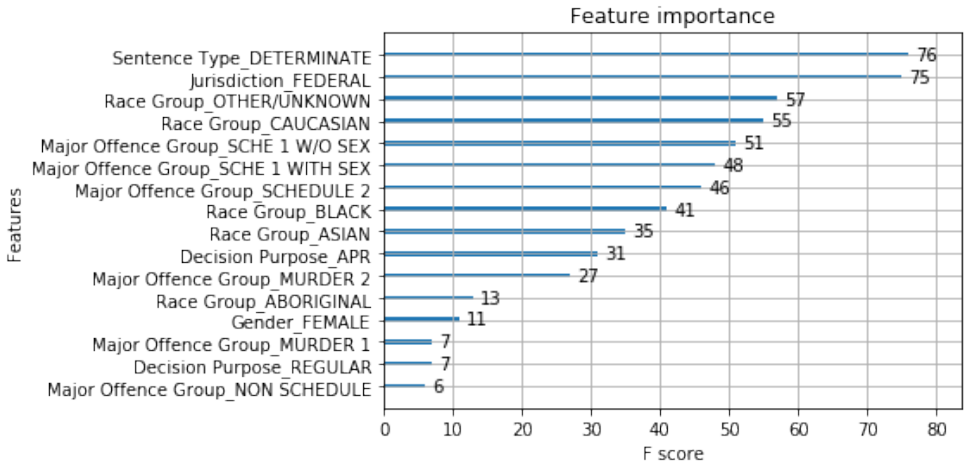

Figure 1: XGBoost on Balanced Dataset

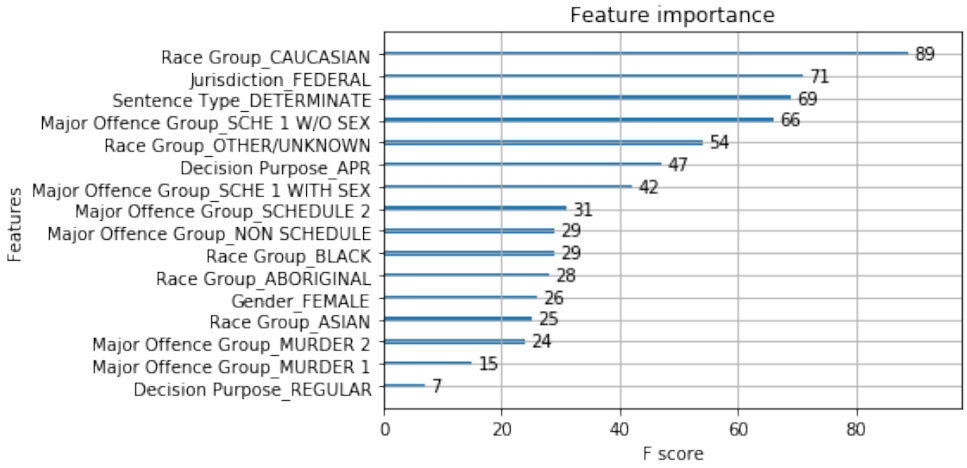

Figure 2: XGBoost on Total Dataset

Clustering Algorithms

For our task of building a cluster model, we selected three different algorithms: K-Means, DBSCAN, and Affinity Propagation. We selected K-Means and DBSCAN particularly because they are quick to run algorithms. Affinity Propagation was selected because it works well for cases where there are several clusters of varying sizes. In any case where more than two clusters were created, they would be condensed based on the label of majority until only two remained.

Each clustering algorithm was run with three randomly sampled datasets containing 500 positively labelled records and 500 negatively labelled records based on the parameters of the model. The results are described in Figure 3 on the following page. As can be seen, all three feature selection algorithms performed fairly comparatively. However, the clustering algorithm itself made a noticeable difference. K-Means and DBSCAN performed on par with one another, but the Affinity Propagation algorithm resulted in clusters that had a lower purity by 3 percentage points.

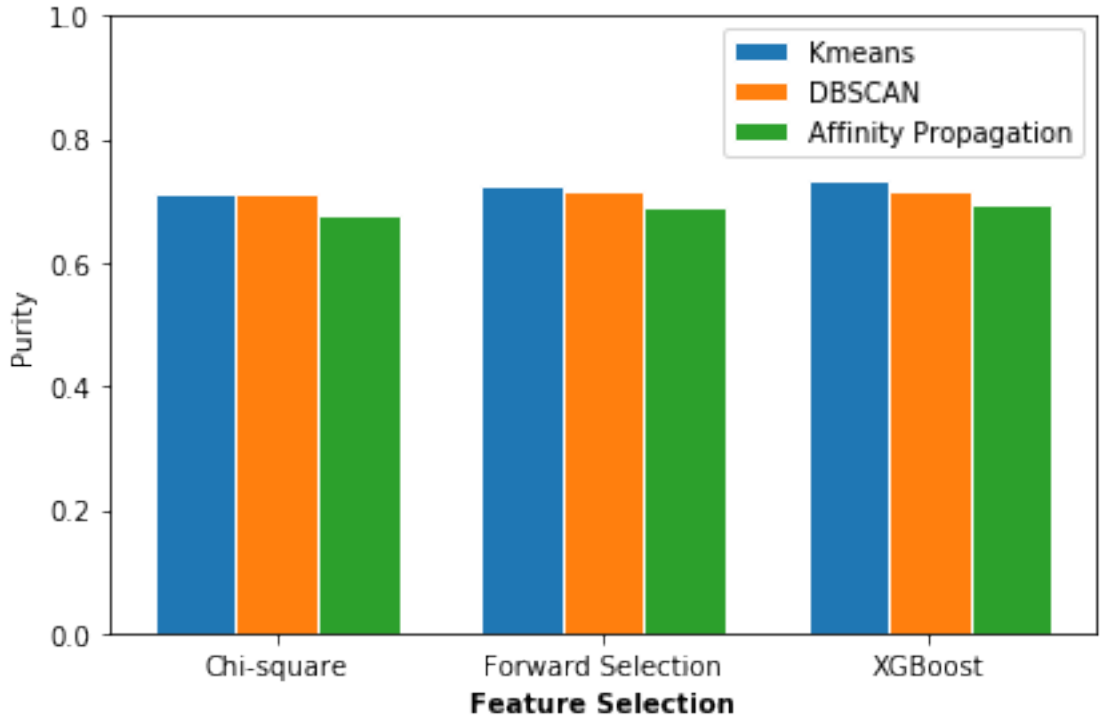

Figure 3: Average Purity of Clustering Models

Conclusions

Based on the results of the cluster models and their positive performance, we can conclude that there are indeed many features that impact a prisoner's likelihood of being granted parole or having parole revoked. Even with only three attributes each, the cluster models consistently predicted a majority of the records. It can also be concluded that the K-means and DBSCAN algorithms perform a sufficiently good job in regard to building a model.

Future Work

Future projects should focus on applying a more statistical analysis on the relationship between the significant features that have been identified in this report and the final outcome of the parole decision. For example, only the three most significant features from the Chi-square tests were used for modeling. This should be re-done using all of the features that were found to be above the significance threshold and compared to the methods used in this project. Overfitting may indeed become a problem, but the models produced from this project were very general.

References

Brownlee, J. (2016, August 31). *Feature Importance and Feature Selection With XGBoost in Python*. Retrieved from machine learning mastery: https://machinelearningmastery.com/feature-importance-and-feature-selection-with-xgboost-in-python/

Kaushik, S. (2016, December 1). *Introduction to Feature Selection methods with an example (or how to select the right variables?)*. Retrieved from analytics vidhya: https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/

Kodžoman, V. (2017, April 20). *Feature importance and why it's important*. Retrieved from Data, what now?: https://datawhatnow.com/feature-importance/

Moffitt, C. (2017, February 06). *Guide to Encoding Categorical Values in Python*. Retrieved from Practical Business Python: http://pbpython.com/categorical-encoding.html

Raschka, S. (2016, March 2). *What is the difference between filter, wrapper, and embedded methods for feature selection?* Retrieved from github: https://github.com/rasbt/python-machine-learning-book/blob/master/faq/feature_sele_categories.md

Parole Board of Canada. (2017, Novenmber 23). *Conditional Release - Appeal Decisions*. Retrieved from Government of Canada: https://open.canada.ca/data/en/dataset/097dfb3f-65d8-4b78-8442-a961052bc168?=undefined&wbdisable=true

Parole Board of Canada. (2017, Novenmber 23). *Conditional Release - Appeal Decisions 2016-2017*. Retrieved from Government of Canada: https://open.canada.ca/en/external-entity/solr_inventory-2aab42488173235b9daaaa53bd1cd588

Parole Board of Canada. (2017, Novenmber 23). *Conditional Release - Appeal Decisions 2015-2016*. Retrieved from Government of Canada: https://open.canada.ca/en/external-entity/solr_inventory-096a7527f406df1eb5ebf6f45049e6ba

Parole Board of Canada. (2017, Novenmber 23). *Conditional Release - Appeal Decisions 2014-2015*. Retrieved from Government of Canada: https://open.canada.ca/en/external-entity/solr_inventory-a85292345516164c0e8ab7249dbfc730

Parole Board of Canada. (2017, Novenmber 23). *Conditional Release - Appeal Decisions 2013-2014*. Retrieved from Government of Canada: https://open.canada.ca/en/external-entity/solr_inventory-7f9ad3cdeaf09efb702a659dfa895d97

Parole Board of Canada. (2017, Novenmber 23). *Conditional Release - Appeal Decisions 2012-2013*. Retrieved from Government of Canada: https://open.canada.ca/en/external-entity/solr_inventory-4f87112adcc21fc04471f4312ef050ac

Parole Board of Canada. (2017, Novenmber 23). *Conditional Release - Appeal Decisions 2011-2012*. Retrieved from Government of Canada: https://open.canada.ca/en/external-entity/solr_inventory-37d7e5eeb20253366ea535c4d87aae80

Parole Board of Canada. (2017, Novenmber 23). *Conditional Release - Appeal Decisions 2010-2011*. Retrieved from Government of Canada: https://open.canada.ca/en/external-entity/solr_inventory-666a86a267679075b07b12442b547e39

Parole Board of Canada. (2017, Novenmber 23). *Conditional Release - Appeal Decisions 2009-2010*. Retrieved from Government of Canada: https://open.canada.ca/en/external-entity/solr_inventory-3b37d72fe92df8ce92e2decb522f72c8