Collateral Consequences: Predicting Barriers to Employment Resulting from Incarceration Across United States

Manpreet Dhindsa
School of Data Science
University of Virginia
Charlottesville, VA
mkd8bb@virginia.edu

Joseph Eldredge
School of Data Science
University of Virginia
Charlottesville, VA
jwe2n@virginia.edu

Gretchen Larrick
School of Data Science
University of Virginia
Charlottesville, VA
jem3yb@virginia.edu

Karan Manwani
School of Data Science
University of Virginia
Charlottesville, VA
akp4he@virginia.edu

Heman Shakeri
School of Data Science
University of Virginia
Charlottesville, VA
hs9hd@virginia.edu

I. Abstract

Formerly incarcerated individuals face various societal, educational, housing, and employment challenges after being released from prison and re-entering society. Since employment impacts recidivism among formerly incarcerated individuals, this work aims to develop a system that will provide organizations with this data and a tool for formerly incarcerated individuals to understand what consequences are attached to a job description. In this paper, we used data from the National Inventory of Collateral Consequences of Conviction (NICCC). Moreover, we develop an interactive dashboard to empower job-seekers with knowledge on barrier crimes per industry and jurisdiction. Additionally, a model was developed to assist formerly incarcerated individuals and organizations with finding employment. We build a model that predicts the collateral consequences and NICCC industry keywords applied to a specific job posting using Natural Language Processing (NLP) techniques and the Logistic Regression model using a Bi-gram TFIDF matrix yield a 61% accuracy score. Additionally, we consider historically posted job listings on major job boards such as monster.com, indeed.com, dice.com, and careerbuilder.com with higher accuracy scores of 74% and 84% via Random Forest and XG Boost respectively using word-level matrices. Code is available on our GitHub page here.

II. Introduction

One difficulty that formerly incarcerated individuals face upon re-entering society is the aversion many employers have toward hiring them Holzer et al. [1]. Additional barriers to job finding can include access to data and information about where and how employment is possible. Related, online services which handle the task of matching people who have been incarcerated with eligible employers can take various complex forms. These include traditional Recommender Systems (RSs), Content-Based Recommenders (CBRs), which primarily consider only the preferences of the user, Knowledge-Based Recommenders (KBRs), which mainly rely on domain knowledge, and Reciprocal Recommender Systems (RRSs), which aim to recommend employers who are likely to reply positively to the recommendation receiver initiated interaction (see Almalis et al. [2]).

As employers seek to become fairer in their hiring practices, they may wish to use fair-by-design algorithms, like that described by Garca-Soriano and Bonchi [3], to ensure that all candidates receive equal consideration. However, for employees and applicants alike, existing well-known algorithms generally lack programmed domain expertise to consider regulatory constraints that should influence their decision-making in this area. Further, popular platforms do not often provide the job-seekers who have been incarcerated with interfaces to quickly understand how the consequences of their conviction might affect candidacy.

III. Problem statement and Background

A significant development that made consequences of conviction less uniform across States involves the U.S. Equal Employment Opportunity Commission (EEOC), which is designed to enforce civil rights laws dealing with workplace discrimination. Following a 2013 challenge by the State of Texas, which argued that it should be able to ban convicted felons from working in specific fields unilaterally, the Fifth Circuit Court found the EEOC had no authority to prohibit such bans [4]. This left States most of the discretion to determine the collateral consequences of convictions on employment in certain sectors, barring several nationwide (Federal) consequences that apply to all States. Ultimately, this ruling cemented radically different approaches to enforcing *barrier crimes*; each State has diverse laws and approaches to balancing workplace safety with opportunity for people convicted of such crimes. More research is needed, but we speculate that this lack of readily available data contributes to many employers' approach to default to using results of background checks and other screening methods without considering the barriers or lack thereof, which an applicant legally faces. Since employment impacts recidivism among formerly incarcerated individuals, this project aims to develop a system that will provide organizations with this data

and a tool for formerly incarcerated individuals to understand what consequences are attached to a job description.

We used data from the National Inventory of Collateral Consequences of Conviction (NICCC) to develop an interactive dashboard to showcase maps in which a user can see the number of various discretions across the different jurisdictions of the United States. The dashboard aims to educate and empower job-seekers with knowledge on barrier crimes per industry and per jurisdiction. Users can select their intended industry sector, and the map will adjust to show the consequences imposed per citation (conviction), jurisdiction (State), and particular industry. This way, different users can start to understand how difficult or easy it may be to live in a particular area or work in a specific industry, depending on the conviction. The dashboard can be hosted on platforms such as Heroku or Google Data Studio, which make it easy for end users to access this information.

IV. Data

Currently, the primary data source incorporated in the interactive dashboard is from the NICCC. This data includes barrier crimes listed by industry, discretion, and jurisdiction. The NICCC site states, "The legal content of this site is updated alongside state legislative sessions and is current as of the date specified on the details page of each consequence. Note that some content may not yet be updated through the most recently completed legislative session in a state."

The data from the NICCC has 19411 rows, and the features include 'Citation', 'Citation URL', 'Title', 'Number of Consequences', 'Relevant Subsections', 'Related Statutes', 'Notes', 'Current Through', 'Jurisdiction', 'Consequences', 'Keywords', 'Offense Type', 'Discretion', and 'Duration.'

Employment data used for the model development is collected through an API aggregation site, www.Demyst.com which produced historical jobs data from four different sources, Monster.com, Indeed.com, careerbuilder.com, and dice.com. Data was collected using the follwing keywords that related to the jobs listed in the NICCC database: 'Adult Care', 'Banking', 'Education', 'Finance', 'Healthcare', 'Public Employment', 'Teacher', and 'Transportation'.

V. Methods

The proposed dashboard provides an interface for users

- to understand barriers to employment based on specific convictions.
- to enter a job description in order to predict consequences.

A. Designing Collateral Consequences Dashboard

The goal was to utilize the data from the NICCC and convert it into a form where someone with a particular conviction can visually understand the implication on employment throughout the United States. A primary feature includes a map representing all 50 states showing the total number of citations (or convictions) that impact employment. A user can click on the map, and all data visuals within the map filter to that state. Additional visualizations will be included that allow users to sort by Offense Type, Industry Keywords, and State.

B. Predicting NICCC Consequence Keywords

To assist formerly incarcerated individuals and organizations in finding employment, we determined it would be useful to provide a tool to show them what collateral consequences apply to a specific job posting. Our approach was to use Natural Language Processing (NLP) techniques to build a model that could predict the NICCC industry keywords from a given job description AND predict the same keyword using a description of the consequence. This provides a common keyword for job-seekers who also know a description of their conviction.

1) Predicting NICCC Keyword from Consequence Descriptions: We built the model starting with the NICCC database using the title for each collateral consequence as the predictor variable and the keywords as the outcome variable. The title field provides a brief description of each collateral consequence, and the keyword indicates the industry in which the corresponding title is applicable.

The dataset had 98 unique classes with some classes having approximately 5000 records and others having less than 10 records. Tables I and II list the top 5 and bottom 5 classes based on the number of corresponding records.

TABLE I: Top 5 Classes

Class	Count
Health care	4,965
Public employment	3,775
Public Office	2,371
Education & schools	1,828
Teachers & instructors	1,115

TABLE II: Bottom 5 Classes

Class	Count
Jury service	7
Deferred adjudication & diversion	3
Judicial evidence & witnesses	3
Pardon & executive relief	1
Civil commitment	1

The pre-processing of the data involved removing stop words and special characters from the predictor variable. The TFIDF matrix was then created using the data from the predictor variable. We experimented with two different TFIDF matrices; one was constructed on a word level, and the other was constructed on a bi-gram level. The response variable was then encoded for it to be used in the models. A 70/30 train-test split was used to build and test the model.

We built the models using the scikit-learn package in Python. The initial results were not promising with low accuracy scores. Based on our analysis, the main reason for the low accuracy scores was limited data for many classes. Therefore, we oversampled for classes with less than 3,000 records. This would give us more data to train our model, but the drawback is it is prone to overfitting.

2) Predicting NICCC Keyword with data from Job postings: Another approach utilized was to build a model with data sourced from historically posted job listings. Job postings were sourced from four major job boards: monster.com, indeed.com,

dice.com, and careerbuilder.com. The NICCC industry keywords were used to find jobs from the data. The top 9 industry keywords in Table III in the NICCC data set accounted for most of the collateral consequences, allowing a focused approach to build the model.

TABLE III: Industry Keywords for sourcing Jobs Data

Adult Care	Banking	Education
Corrections	Finance	Health care
Public Employment	Teacher	Transportation

A total of 500 to 600 jobs were pulled from the various job boards for each of these keywords. The job description and job title were used as the predictor variables, and the keyword/industry that we used to pull the jobs was used as the outcome variable. This method of sourcing the data retrieves various job categories within an industry. For instance, using the healthcare keyword, our pull would have jobs such as Customer Service Rep and Physical Therapist in the same category as long as both jobs are in the healthcare industry.

The pre-processing of the data involved removing stop words and special characters from both the predictor variables. Then, the TFIDF matrix was constructed after vectorizing both predictor variables on a word level.

VI. Results

A. Dashboard

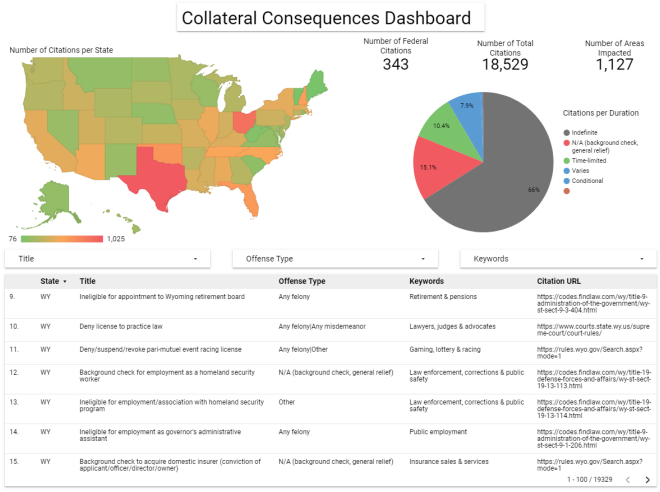

Fig. 1: Google Data Studio Dashboard representation of NICCC data (version 1)

Dashboard generation was completed using Google Data Studio (GDS) [5], shown in Figure 1. This dashboard is a user-friendly dynamical visualization tool that allows users to filter the categories to obtain answers to their individualized questions. The highest priority visualization is a map that totals all citations based on the user's choice filters. These filters are Title (a summary of the citation), State, Offense Type, and Keywords (Industry or subject of implication). In addition to the map, the duration a citation is active is represented by a pie graph. Finally, a table that details the values represented in the visualization is found at the bottom. The table contains all data for the filtered information by adding the Citation URL, which gives a user the direct link to the citation from its corresponding state.

B. Models

1) Predicting Keyword Using NICCC Database after oversampling: As shown in tables IV and V, the results of the model using the NICCC database for training reached accuracy scores of 61%. The most accurate model was the Logistic regression model using a Bi-gram TFIDF matrix. The F1 score of this model was slightly higher at 65%.

TABLE IV: Accuracy Score of Models using NICC data

TFIDF Matrix	Logistic Regression	Naive Bayes	Random Forest	XG Boost
Word Level	60%	56%	62%	61%
Bi-Gram Level	61%	60%	62%	59%

TABLE V: F1 Score of Models using NICC data

TFIDF Matrix	Logistic Regression	Naive Bayes	Random Forest	XG Boost
Word Level	64%	59%	65%	65%
Bi-Gram Level	65%	65%	65%	62%

2) Results Using Job Postings: The accuracy scores of most of the models improved using this approach, with the most significant improvement observed in Random Forest and Gradient Boosting models. Table VI shows these models had accuracy scores of 74% and 84%, respectively. Given the balanced dataset, the F1 scores were also very similar. However, the models with Bi-gram TFIDF matrices fail to produce better results than the word level matrices in this approach; hence we recommend using the latter.

TABLE VI: Accuracy Score of Models Built from Job Postings

Model	Mean CV Accuracy	Test Accuracy
Naive Bayes	61%	62%
Logistic Regression	64%	66%
Random Forest	73%	74%
Gradient Boost	84%	86%
XG Boost	85%	86%

VII. Conclusions and Future Research

The developed tool provides a job search interface to assist formerly incarcerated people in finding employment. We suggest building the keyword prediction process into the application's front-end, e.g. where the dashboard is embedded. Therefore, we recommend selecting an algorithm that balances high accuracy (and F1 score) with a more reasonable (within seconds) runtime, achieved by the Random Forest classifier. Even though the achieved 86% accuracy and the corresponding F1 score in this project, the target score must be set by a subject

matter expert with knowledge of the user base, the methods of data collection by the NICCC, and the desired computational performance of the application.

As mentioned, when training the model, The lack of data across different sectors posed a challenge to producing high accuracy and reliably predicting industries, so we explored different ways to increase our data sets and make them more balanced. For future research, one method to improve this approach would be to use NLP with synonyms or embeddings. We found the package spaCy extremely useful for synonym replacement and variable of Title from the NICCC data set to produce synonyms within a given domain/sector. For example, including keywords within the financial domain will improve the model's robustness against overfitting and helps to handle out of sample job descriptions in that domain.

Finally, the dashboard should ideally encapsulate all data needed to find appropriate information for potential jobs. As more information is obtained, it can be seamlessly added to a backend application like Google Data Studio. The integration of the model will not interfere with the end-user tasks of filtering and querying the data based on job specifications to obtain accurate information. As individuals use this dashboard, the model can be improved and target scores refined because SMEs may better understand the end user's needs, the consequences of conviction, and their needs in terms of information retrieval.

REFERENCES

[1] H. J. Holzer, S. Raphael, and M. A. Stoll, *Can employers play a more positive role in prisoner reentry?* Urban Institute Washington, DC, 2002.

[2] N. D. Almalis, G. A. Tsihrintzis, N. Karagiannis, and A. D. Strati, "Fodra—a new content-based job recommendation algorithm for job seeking and recruiting," in *2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA)*. IEEE, 2015, pp. 1–7.

[3] D. Garca-Soriano and F. Bonchi, "Maxmin-fair ranking: Individual fairness under group-fairness constraints," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery Data Mining*, ser. KDD '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 436–446. [Online]. Available: https://doi.org/10.1145/3447548.3467349

[4] W. P. BARR, U. A. General *et al.*, "In the united states court of appeals for the fifth circuit," 2019.

[5] Google. (2022) Google data studio. [Online]. Available: https://datastudio.google.com/