Project Proposal

Predicting the Probability of Recidivism

By: Kyle Hanson, Christine Orosco, Myra Rust

Background: It should be of no surprise to anyone that recidivism, the tendency of a criminal to relapse into criminal behavior, is a significant problem in our society. Again, and again, you read news articles or watch television programs about persons who were previously incarcerated, committing additional criminal acts after being released from prison. The question that we ask ourselves is, could these additional crimes have been prevented? Ideally, if we could determine who is likely to reoffend, better decisions can be made on how long of a sentence they should receive in the first place or if they should be allowed to be released on parole.

Problem Statement: According to a 2005 study conducted by the Bureau of Justice Statistics, "67.8% of released prisoners were arrested for a new crime within 3 years." (Durose et al., 2014). The reality is recidivism rates are significant and that must weigh heavily on the minds of judges and parole board officials while making decisions about a person's fate. As it stands today, decision makers are provided with high-level statistics on recidivism rates, but can we do better? Can we provide them with more accurate information about the likelihood that an individual would reoffend? By doing this, it would allow judges and parole board officials to make better decisions, not based on generalizations, but on individualism and hopefully mitigate or prevent additional crime. Our project seeks to leverage historical data, machine learning, and predictive analytics to provide an accurate predictive model that can be used to calculate the probability that an individual will reoffend based off that individual's demographics and offense related information.

Scope: The scope of this project will include the project team working collaboratively to accomplish five milestones.

- Milestone 1: Communication Plan, due March 21, 2021.
- Milestone 2: Project Proposal, due March 28, 2021.
- Milestone 3: Preliminary Analysis Report, due April 18, 2021.
- Milestone 4: Intermediate Results PP Presentation, due May 16, 2021.
- Milestone 5:
    - Project Presentation, due June 3, 2021.
    - Final Paper and Technical Report, due June 5, 2021.

Document Overview: The remainder of this document contains a discussion of the preliminary requirements including, technical approach, data source, analysis, requirement development, model deployment, testing, model evaluation, expected results, and project management topics including, project plan and project risk.

Preliminary Requirements: Preliminary requirements are a summation of what are the necessary elements involved for the lifecycle of project; with the emphasis being on the genesis of the project, but also extend to the culmination and result.

Technical Approach: Throughout the course of the project, the overlying aim is to be able to provide more accurate information about whether a convicted individual is more likely to reoffend based on past known factors. After the team identifies relevant data sources necessary to complete the assessment, the next step is to import and clean the dataset(s). After considering the project objectives, the team will determine if supplemental data is required to continue. If necessary, the team will conduct a search of additional data sources or insights. Once we have enough data to proceed, we will conduct exploratory data analysis seeking to subset the variables that hold more intrinsic value to our original question set. Upon completion of this step, the project can proceed to the next phase which is model selection and execution. Basing our decision on the problem statement, the team has identified that a regression model is best suited for the project. Regression modeling provides a quantitative metric such as a probability estimate, and this would provide more value to the decision makers over a binary prediction result of will offend/will not offend. Upon conclusion of the model execution, we will analyze the results and construct a deployable model.

Data Sources: 3-Year Recidivism for Offenders Released from Prison in Iowa dataset. This dataset includes data collected from 2010-2018 and provides a comprehensive set of features.

- *https://data.iowa.gov/Correctional-System/3-Year-Recidivism-for-Offenders-Released-from-Pris/mw8r-vqy4*

Analysis:

- Conduct Exploratory Data Analysis on the initial dataset, including any supplemental datasets as necessary.
- Initial selection of a Regression model for model execution with the possibility of additional models depending upon project analysis and model results.
- Define model testing methodology and results success criteria by analyzing the final dataset and identifying model input parameters.
- Conduct an analysis of model testing results for current model input parameter and hyperparameter refinement.

Development Requirements:

- Exploratory data analysis and modeling will take place on local machines.
- If a neural network is necessary to execute the algorithm, a cloud computing service is available.

Model(s) Deployment: As stated previously, Regression algorithms are the most suitable methods with the possibility of employing neural network configurations and the use of Tensor flow libraries.

Testing and Evaluation: For testing and evaluation, we will split the data into testing, training, and validation subsets. The typical ratios for test and training splits are a 60/40 or 70/30. The team will use existing recidivism rates as criteria for model success.

Expected Results: The objective of the project is to predict the probability that an individual will commit additional criminal acts within three years of being released from prison. We expect to produce an accurate model that provides the probability rate for recidivism for individual persons based on demographic and crime related factors.

Project Management: Project Management will follow the CRISP-DM model. The steps include Business Understanding, which is understanding the requirements, selecting the topic and ensuring the topic meets course requirements. The next phase is Data Understanding, the team will conduct an evaluation of the data to ensure the data is sufficient to meet the project objectives and goals. Additionally, the team will begin an exploration of the data to identify data relationships and any issues with data quality. In the third phase Data Preparation, the team will clean the data, create new attributes if necessary, and prepare the data for input to the model algorithms. In the fourth phase Data Modeling, we will refine model selection if necessary, build and test the model, and assess model results. While CRISP-DM appears to depict a linear process, it does not restrict the repetition of previous phases. The team will be flexible should there be a need to refine any phase of project or modify portions of the project. This is most likely during the model development and execution phases. During this phase it may become apparent that the team needs to consider another algorithm based upon the results at the time. Project management anticipates this recurrent activity and will adjust tasking to ensure the team remains on schedule and meets the required milestones. Upon completion of model execution, the team will prepare a final report that will provide a summary of the project, identify the results, address additional findings, and identify any issues or concerns.

Execution of the Plan: Microsoft teams seems to be the de facto platform for team interaction. As such, most of the project management execution will be done through Teams. Teams facilitates document collaboration and review. Because Teams does not facilitate a Kanban board type function, the Team will use a spreadsheet to list and track tasks assignments and due dates. Additionally, the spreadsheet will identify the course milestones. The spreadsheet will be posted to the Team Group on Blackboard.

Project Risks: All team members have a monetary, time, and effort investment in the successful completion of the project. Although there are no anticipated risks to the project at this time, the team must consider some possibilities. The biggest risk is the loss of a team member. If this should occur at any time during the project, the other two team members are ready and able to compensate for the loss. Another project risk is that during the Data Understanding Phase, it may become apparent that the datasets in-hand does not satisfy the objectives. To overcome this risk, the team has already identified other sources of datasets. The third major risk to the project is the model execution platform. The team will ensure that more than one platform is available with enough compute power to run the model's algorithms. If necessary, team members have no cost cloud compute services available to run their model's algorithms.

References:

Durose, M., Cooper, A., Snyder, H. (2014, April 22). Recidivism of Prisoners released in 30 states in 2005: Patterns From 2005 – 2010 – Update. Bureau of Justice Statistics. Retrieved from https://www.bjs.gov/index.cfm?ty=pbdetail&iid=4986.

Iowa Department of Correction. (2020, November 16). 3-Year Recidivism for Offenders Released from Prison in Iowa. Iowa Data. Retrieved from https://data.iowa.gov/Correctional-System/3-Year-Recidivism-for-Offenders-Released-from-Pris/mw8r-vqy4.