DOC Assault Model Rebuild

PURDUE UNIVERSITY | The Data Mine

Jennifer Lu, Gabriel Clifton, Nathan Yoosang Cho, Sameer Dadoo, John Erick Talabucon TA: Aneesh Denduluri

MANAGEMENT PERFORMANCEHUB

ROADMAP TO CREATE A SUCCESSFUL ASSAULT MODEL

1 → MODEL SELECTION → 2 → MODEL IMPLEMENTATION → 3 → MODEL TESTING → 4 → MODEL EVALUATION → 5

INTRODUCTION & OVERVIEW

Management Performance Hub (MPH):
- Indiana's central hub for data-driven innovation, collaboration, and advanced analytics
- Unlocks data insights, integrates disparate systems, and drives informed decision-making

Indiana Department of Correction (DOC):
- Oversees 25,000 incarcerated individuals across 18 adult correctional facilities, 3 juvenile facilities, and multiple parole districts
- Deliver comprehensive support services aimed at reducing recidivism

BACKGROUND & MOTIVATION

DOC has transitioned to a data system with an updated structure. As part of this transition, DOC is rebuilding its 'Assault' predictive Random Forest model, reassessing key elements, and retraining it using current data and outcomes.

The MPH is hoping to achieve the following:
- Enhanced Predictive Accuracy: Incorporating better machine learning algorithms to improve assault risk assessment.
- Up-to-Date Model Training: Utilizing current prisoner data to ensure the model's relevance and accuracy.
- Improved Public Safety: make informed decisions and reduce recurring offenses.

MODEL BRAINSTORM & SELECTION

We want our model to...
- Make new discoveries with data
- Have a high accuracy

Poisson Regression
- Model that assumes counts follow a **Poisson distribution (**variance is equal to the mean)
- Suitable for modeling events over time.

Random Forest
- Model that takes **multiple decision trees** on different parts of the data
- Combines their predictions to improve accuracy and reduce overfitting

K-Nearest Neighbors (KNN)
- Model that classifies data points based on the majority label of their nearest neighbors
- Used to find new patterns and correlations within data and makes predictions based on similarity

Why KNN?

KNN best fits the three goals we want from a model

A KNN model...
- Works to find new relationships and correlations between data point features
- Adaptive to complex patterns within data that can be linear and non-linear
- Robust to small datasets, which fits the dataset

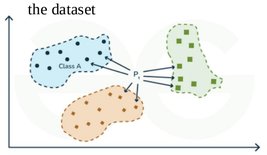

Figure 1: Visualization of KNN

LABEL ENCODING

To implement the KNN Model, we used resources and tools like the Scikit-learn, Seaborn, Matplotlib and Pandas library in Python to train, test, and visualize our data.

Why is label encoding necessary?

A KNN model calculates the difference between points on a plot and making predictions based on the nearest points, but you cannot plot categorical data without label encoding, or assigning the categorical points with respective numerical points.

Example: the "**Response**" column where we set "**Assault**" as 1 and "**NoAssault**" as 0.

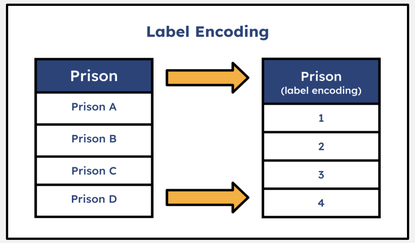

Figure 1: Visual representation of label encoding with Prison A, B, C, D to 1, 2, 3, 4

CORRELATION MATRIX

Why is a correlation matrix important?
- Find how strongly correlated one variable
- –1.0, a strong negative correlation, to 1.0 a strong positive correlation.
- A correlation matrix allows us to
- Find new trends in the data
- Reduce number of features used in our model

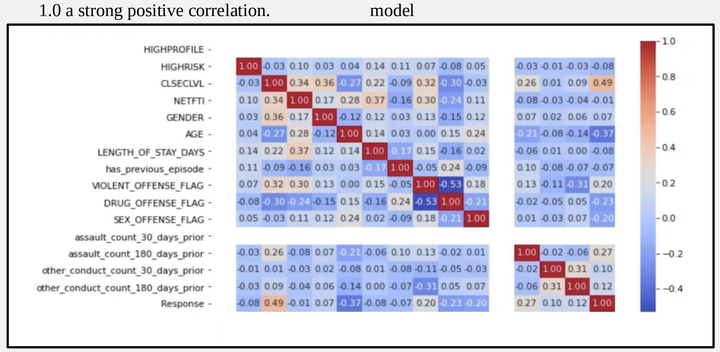

Figure 2: Correlation matrix of various categories of the columns represented with color as a heat map

Notable Trends for Model

1. Security classification level had a positive correlation with assault
2. Age had a strong negative correlation with assaults
3. Felons flagged for a drug related crime were likely also flagged for a violent offense crime
4. The felon's highest felony level did not have any correlation with assault
5. Felons flagged for a sex offense had a negative correlation with assault

MODEL IMPLEMENTATION & VISUALIZATION

Using correlation matrix patterns, we selected the following features to train, test, and model our data.
- Age
- Net days of stay (NETIFI)
- Classification level (CLSECLVL)
- Assaults in the past 180 days
- Violent offense flag
- Sex offense flag
- Previous episode in prison flag

K = 15 (calculated using √n, where n = size of the dataset)

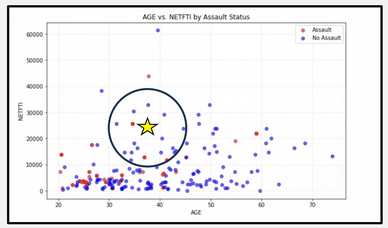

Figure 3: Visualization of KNN model comparing NETIFI vs AGE the star represents a new data point and the circle is the K nearest data points

ASSAULT MODEL METRIC REPORT

We used Sklearn's accuracy library to access the correctness of our model
The model accuracy was 84.404% on test data and 82% on training data

Precision values: (accuracy of positive predictions)	no assault 0.87	assault 0.79
Recall values: (coverage of actual positives)	no assault 0.90	assault 0.73
The F-1 score: (balance of precision and recall)	no assault 0.88	assault 0.76

What does this mean?
- Model is more effective when predicting a datapoint as **no assault** compared to **assault**.
- adjust parameters, or use a different model to increase accuracy when a data point is classified as an assault.

Our model is neither overfitting, nor underfitting.

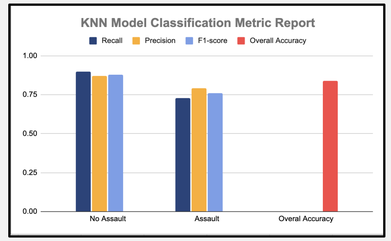

Figure 4: Visualization of the KNN model's precision, recall, f-1 score and overall accuracy for No Assault and Assault

FUTURE GOALS

Our goals are to...
- Implement and test more models such as **Possion Regression** and **Random Forest**
- Optimize our current model, fine tune hyperparameters, and find a way to encode all qualitative data

CONCLUSIONS

Our conclusions are based on the project's work over a semester regarding personal growth and learning lessons
- Used machine learning algorithms to create an assault model
- Through a correlation matrix, we realized that not all factors are as positively or negatively correlated to assaults as expected.
- Utilized qualitative and quantitative data analysis, machine learning techniques, and data visualizations to analyze trends in the data

ACKNOWLEDGEMENTS & REFERENCES

We would like to thank our TA **Aneesh Denduluri**, our mentor at the **MPH**, **Kelsey Chance**, and the Data Mine staff for guiding us throughout this project, especially **Kevin Amstutz**.