SESUG Paper 234-2019

Validating Classification and Recidivism in Corrections Through Outcome Analytics

Alan R. Mann, MS

ABSTRACT

A data-driven design for inmate classification in corrections is proposed as a fact-based, analytic, modeled alternative to the possibility of confirmation bias in risk assessment of prisoners entering incarceration, using retrospective assessment of recidivism risk measures of inmate data from New York State and Florida respectively. A framework utilizing the CRISP-DM (Cross-Industry Standard Process Model for Data Mining) from Abbott, D. (2014), was employed, deriving Key Performance Indicators (KPIs) of Class Severity and predicted Custody Risk groupings based upon Offense and Severity of New York State Offense Classification sentencing guidelines (Strazzullo Law Firm., n.d.). While initial findings from New York indicated a random and highly subjective pattern of classification, with class A misdemeanor offenders placed in Maximum custody, and class A-I felony offenders in Minimum, subsequent identification of objective markers in the New York and Florida data proved a more accurate, stronger scored, categorical and correlative relationship. This paper utilizes both visual and tabular illustrations, employing both Exploratory and Confirmatory Data Analysis, or both a-posteriori and a-priori results/conclusions in determining a data-driven case for inmate custody level classification. (Jebb, A. T., Parrigon, S., & Woo, S. E., 2017). The paper concludes with a discussion of deployment utilizing a dashboard, and iterative Exploratory Data Analysis.

INTRODUCTION

The major purpose of scoring a Custody Classification-based, data-driven design is to enable classifying an inmate as minimum, medium, or maximum risk through a dashboard or other analytic insight tools, as a predictive concept which diverts from the biographic approach of custody level classification requiring points per a stepwise process, such as the U.S. Bureau of Prisons' SENTRY system, relevant to management of adult and juvenile corrections in the United States. The business problem of classification should not solely depend upon a data-driven model, being guided toward a fairer approach to corrections that seeks to reduce the population in prisons through better mitigative practices, allowing the current subjective approach and data driven guidance which could avoid the mistake of placing an inmate in a differing grade of custody than needed (Cameron, B., 2019). The reality is that prisons are understaffed, violent, physically deteriorating repository facilities, rather than correctional tools. With understaffing at 20%, such as in Alabama, a cloud-based, secure classification system at the inmate record level could allow a cost-effective method for easing some personnel burden in prison management (Lockhart, P. R. , 2019).

The data-driven design would be utilizing supervised and unsupervised machine learning, basing the outcome though Categorical Data Analysis using Decision Tree, Logistic Regression, and Cluster Analysis models for predictive analytics. A Key Performance Indicator of predicted probability of correct classification based upon retrospective classification could be a Key Performance Indicator, (KPI), acting as a guide to calibrating a better outcome of offense-based classification. Mayhew, H., et al. (2016) state that advanced analytics needs to be a "purpose-driven" means to an end. It should be a means to offer options instead of a uniform path to a solution. The modeling methods could additionally utilize the purpose of corrections: to move toward an unbiased focus within the data through testing for bias, possibly breaking the cycle of recidivism, also known as "desistence," currently at 17% after 9 years in the U.S. (Alper, M., et al., 2018). Initially, descriptive statistics are needed to understand the bases for insight utilizing analytic model assessment.

ETHICAL CONSIDERATIONS

Any business or public service document needs to factor the matter of ethics to protect both the client and any research subjects' privacy, transactional information, identities, and operative practices to assure compliance with any and all protections afforded by the native country's regulations of business embodied in law and judicial precedent.

These principals could be ethical considerations for analytics professionals to follow. They are:

- Privacy, or protections of individuals afforded by legal and cultural boundaries.
- Confidentiality of Shared Private Information, extending to metadata shared in confidence with business institutions in which the owners of the data have an ongoing or historic relationship.
- Transparency, or holding others accountable to foster trust (Richards, N.M., et al., 2014). Legally, the Freedom of Information Act (Office of Information Policy (OIP), U.S. Department of Justice, 2019) of 1966 guarantees the right to transparency of government activities not deemed of serious (Secret), or critical (Top Secret) nature to damage the national security of the United States (Richards, N.M., et al., 2014).
- Identity Protection, or, protecting the least common denominator of the body politic, the 'we the people' of the Constitution. The determination of a person's reputation, economic and political qualities is dependent upon association of these qualitative elements with the individual (Richards, N. M., et al., 2014).
- The data used for this project is from the New York Department of Corrections and Community Supervision, limited to the 5 boroughs comprising New York City. (New York State Department of Corrections and

Community Supervision. (March 18, 2019). Broward County, FL COMPAS System extracts of 10,000 inmates have also been selected (Ofer, D., 2017). While New York State is anonymized at the name level, one could possibly obtain Personally Identifiable Information (PII) from the following variables:

-
    - Current_Age,
    - Snapshot_Year,
    - Authority (Offense Code Number, from the New York Criminal Code),
    - Gender,
    - Race,
    - Age

of the inmate for name, which could be written in a newspaper.

The Broward County, FL dataset has:

- Full names,
- Sex_Code_Text (Gender),
- Screening_Date,
- Marital_Status,
- Ethnic_Code_Text (Race)

as identification/demographic fields. This paper has only needed subjective and objective fields of Custody Level and Offense Severity/ Offense Class to describe either reasons for incarceration (New York), or Supervisory status (RecSupervisoryLevel) compared to the Recidivism Score Decile Code, giving a risk score from 1 to 10 for recidivism. No personally identifiable information, as defined by the National Institute of Standards and Technology (NIST, n.d.) is referenced in the analytic models, with the exception of variables Race and Age from New York in determining the degree of variability in the K-means clustering model. Both were dropped as analytic measures of comparison as risks of confirmation bias.

FRAMEWORK

Abbott, D. (2014) outlines the CRISP-DM (Cross-Industry Standard Process Model for Data Mining) sequence as a guide to a step-wise, document-friendly process of analysis. Abbott, D. (2014) cites the choice of CRISP-DM as “the most often used process model since its inception in the 1990’s.” Within the theoretical definitions of Tukey, (Exploratory Data Analysis), (Jebb, A. T., et al.,2017), which requires development of theories of data behavior,

validated by a-posteriori visualizations and discoveries, with Confirmatory Data Analysis (CDA), which takes a more structured or linear approach to data, or an a-priori approach, (Jebb, A. T., et al., 2017), this paper seeks to prove through CDA, along with visualizations that represent a behavioral discovery not initially apparent (EDA),(Ref: Figures 9a – 9d), is both an exploratory and confirmatory analysis through the CRISP-DM framework.

1. Following the CRISP-DM format, Phase 1 will cover the first 4 CRISP-DM steps of this Capstone paper: Business Understanding: Define the project;
2. Data Understanding: Examine the data; identify the problems in data;
3. Data Preparation: Fix problems in the data, and create derived variables
4. Modeling Build predictive or descriptive models: Build predictive and/or descriptive models

Phase 2 will discuss the Evaluation and Deployment steps of CRISP-DM as applied to a scoring of classification and recidivism.

PHASE 1:

BUSINESS UNDERSTANDING: Assumptions, Exploratory Analytics, and Initial Hypotheses

According to Alper, et al., (2018), violent crime comprises 1 in 4 of all offenses contributing to recidivism. Taking this a step further in data-driven classification, nearly 50% of our New York City resident cohort fell under Classes B, C, D. and E Violent Felonies. (NYS prison admissions, 2019).

Business Understanding of the aims of the NYS DOCCS Custody Classification system would be defined per Security (Custody) Levels as: “by severity of offense and/or other behavior,” and are usually assigned to prisons having a corresponding level of security. Using Federal guidelines for definition, (Understanding prison security levels, n.d.), our definitions of Minimum, Medium, and Maximum Security are:

- Minimum Security, (MIN), minimal history of violence, non-sex offenses, less than 10 years left in sentence. Inmates housed in dormitories;
- Medium Security, (MED), most inmates with history of violence. Less than 30 years left in sentence. Inmates housed in cells, high level of guarding, walled building surrounded by razor wire.

- Maximum Security, (MAX), also known as High Security. High concentration of violent prisoners, multiple walled facilities surrounded by razor wire and at most, gun towers. All classes of prisoners housed here, with danger to sex offenders and informants. Gang activity high.

For New York State only:

- Shock Incarcerations, new since March, 2019; mental fitness certified, inmate between ages 18 and 49, maximum security, eligible for release within 3 years, no violent felony offenses, including A-1 Felony classifications; No felony sex offenses, homicides, or escapes in record. No outstanding felony or immigration warrants in other states (New York State Department of Corrections and Community Supervision, March 18, 2019).

EXPLORATORY ANALYTICS FOR DATA UNDERSTANDING

Data Cleaning Techniques

Before any data understanding and analysis is performed, data cleaning and validation need to be performed. The difference between data quality and data cleansing hinges on one word, which is “validity” (Peterson, 2003). Validity defines the quest for quality data, which could be objectively incorrect, but still proves relevant in organizing enterprise events. Data hygiene is the task of making data objectively error-free and reliable. Important tools for data hygiene in Base SAS® include PROC UNIVARIATE for examining numeric data type variables for missing values and uneven quantile distributions of numeric values within the dataset. The corresponding tool for character data types would be PROC FREQ, testing for null values and uneven distributions of literal values within the character variables of the dataset (Cody, R., 1999).

Assumptions with Exploratory Analytics

The first assumption on the variables affecting classification is that socialization abnormalities could affect the classification of the inmate. Huebner, B. M., et al.,(2007) argue based upon their 2007 study of the effects on recidivistic behavior on young convicts list gang affiliation and drug use as two main effects of recidivism, which could help in risk assessment of inmates upon intake. While drug use is not a direct effect measured upon intake, mental health and gang affiliation are contributing factors. The data collected within NY daily inmates in custody, (2019) seem to confirm this categorical relationship.

The Null Hypothesis states that at a 95% confidence level, (alpha = 0.05), there is no statistical significance between either Gang Affiliation or Mental Health issues.

Using alpha = 0.05 (SAS will default to the 95% confidence limit:

Given: Mental or Gang_Aff = Probability of Custody Level [MIN, MED, MAX, or Shock Incarceration (SHO)]:

Null Hypothesis H_0 : There is no probabilistic effect / statistical difference between Mental or Gang_Aff (Y) by Custody Level (X_1);

Alternative Hypothesis H_1 : There is probabilistic effect / statistical difference between Mental or Gang_Aff by Custody Level;

$$E(Y),(\text{Response}) \text{ or } = \log(p/1-p) = ß_0\,(\text{Y-Intercept}) + ß_1X_1\,(\text{probability coefficient} * \text{predictor})\ ;$$

Figure 1: Logistic (Logarithmic) Regression Equation Example

A logistic regression model was run, using Gang Affiliation and Mental Health as dependent variables, with Custody Levels of Minimum, Medium, and Maximum (1, 2, and 3 respectively) as the independent variables. As Custody Level increases, the probability of Gang Affiliation increases twofold, possibly a geometric relationship. Note: Gang Affiliation is 14.5% of the overall Inmate Intake population in New York City.

Response Profile		
Ordered Value	gang_aff	Total Frequency
1	1	987
2	0	5789

Figure 2: Probability modeled is gang_aff=1.

The Null Hypothesis of no significant difference is refuted by pr < Chi Square of less than 0.05.

Please refer to Figures 3, 4, and 5:

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1423.0719	1	<.0001
Score	1260.4166	1	<.0001
Wald	905.9922	1	<.0001

Figure 3: Gang Affiliation Affecting Custody Level

As Custody Level increases, the probability of Mental Health issues increases more gradually, possibly a geometric relationship. According to Understanding prison security levels, (n.d.), as custody levels move toward Maximum, gang activity is more prevalent, so this indicates a correct alignment with the data, in addition to gang activity as an effect upon higher custody levels. Note: Mental Health Issues is 44.8% of the overall Inmate Intake population in New York City.

Response Profile		
Ordered Value	mental_hlth	Total Frequency
1	1	3033
2	0	3743

Figure 4: Probability modeled is mental_hlth=1.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	52.8253	1	<.0001
Score	52.7456	1	<.0001
Wald	52.5054	1	<.0001

Figure 5: Mental Health Issues Affecting Custody Level

The Null Hypothesis of no significant difference is refuted by pr < Chi Square of less than 0.05:

The second assumption for exploratory consideration is that offenses affect Custody Level. Of interest would be the result of another Logistic Regression model, through SAS ® Enterprise Miner ™.

A predictive analysis to determine the strength of the second assumption of Custody Level effects was performed using New York State prison admissions, (2019), comparing a roll-up using SAS® PROC FREQ. Offenses. Listed by variable ‘Authority’, this variable holds categorical values of New York State Penal Code Offense Numbers. (Used in Statistical role of Predictor variable, or ‘X’), with a Target of categorical variable Custody Level, (Statistical role of Response variable, or ‘Y’) as Minimum, Medium, Maximum, and Shock Incarceration Custody Levels.

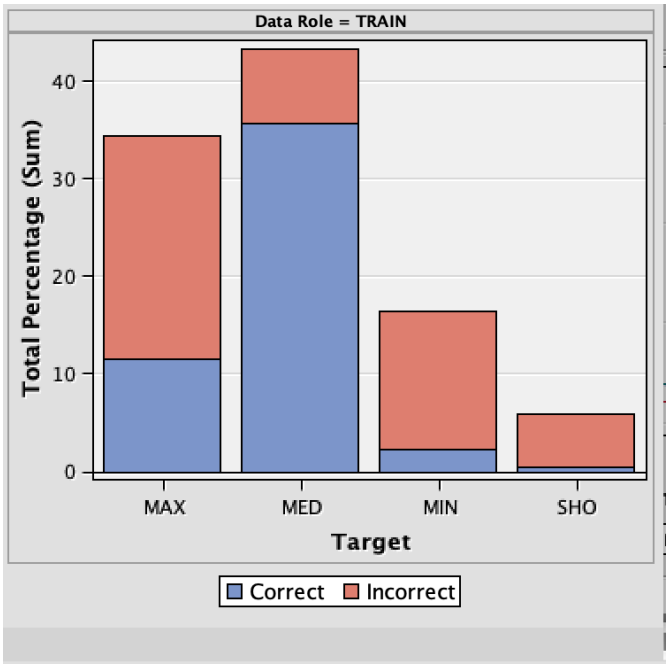

Figure 6a: Predictive and Actual Custody Levels vs Offenses per, NYC, 2018-2019

Regarding Target matching Outcome, Offenses vs Maximum Custody (Outcome), would predict at 33.33%, correct within the cohort at 11.46%. Medium Custody (Outcome), is predicted correct at 82.21%, correct within the cohort at 35.63% .

Minimum Custody (Outcome), is predicted correct at 13.92%, correct within the cohort at 2.29%.

Shock Incarceration (Outcome), is correct at 7.14%, correct within the cohort at 0.417%. Overall correct ratio is 50.529% ; incorrect is 49.471%.

Please refer to Figure 6b: “Correct and Incorrect Custody Levels.”

Outcome Type	Data Role	Target Variable	Target Label	Target	Outcome	Correct Text	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage	Correct
PREDICTION	TRAIN	Custody...		MAX	MAX	Correct	47.82609	33.333...	55	11.45833	0
PREDICTION	TRAIN	Custody...		MED	MAX	Incorrect	26.08696	14.423...	30	6.25	1
PREDICTION	TRAIN	Custody...		MIN	MAX	Incorrect	18.26087	26.582...	21	4.375	1
PREDICTION	TRAIN	Custody...		SHO	MAX	Incorrect	7.826087	32.142...	9	1.875	1
PREDICTION	TRAIN	Custody...		MAX	MED	Incorrect	31.06509	63.636...	105	21.875	1
PREDICTION	TRAIN	Custody...		MED	MED	Correct	50.59172	82.211...	171	35.625	0
PREDICTION	TRAIN	Custody...		MIN	MED	Incorrect	13.60947	58.227...	46	9.583333	1
PREDICTION	TRAIN	Custody...		SHO	MED	Incorrect	4.733728	57.142...	16	3.333333	1
PREDICTION	TRAIN	Custody...		MAX	MIN	Incorrect	19.04762	2.4242...	4	0.833333	1
PREDICTION	TRAIN	Custody...		MED	MIN	Incorrect	23.80952	2.4038...	5	1.041667	1
PREDICTION	TRAIN	Custody...		MIN	MIN	Correct	52.38095	13.924...	11	2.291667	0
PREDICTION	TRAIN	Custody...		SHO	MIN	Incorrect	4.761905	3.5714...	1	0.208333	1
PREDICTION	TRAIN	Custody...		MAX	SHO	Incorrect	16.66667	0.6060...	1	0.208333	1
PREDICTION	TRAIN	Custody...		MED	SHO	Incorrect	33.33333	0.9615...	2	0.416667	1
PREDICTION	TRAIN	Custody...		MIN	SHO	Incorrect	16.66667	1.2658...	1	0.208333	1
PREDICTION	TRAIN	Custody...		SHO	SHO	Correct	33.33333	7.1428...	2	0.416667	0

Figure 6b: Correct and Incorrect Predicted Custody Levels vs Offenses per, NYC, 2018-2019

There could be other effects, as judgements and possible protective actions, (those mentioned for sex offenders and informants, for example). Subjectively, New York may have underestimated those belonging in Maximum Security! A near 50:50 correct/incorrect ratio is similar to a coin toss to determine to path of an inmate's life. Therefore, a data driven classification system utilizing predictive analytics is possible, but the effects of subjectivity, and external effects of mental health and gang affiliation could affect custody level classification.

DESCRIPTIVE STATISTICS FOR DATA UNDERSTANDING

An assessment of Class of Offense for New York State inmates from New York City for the period January 1, 2018 December 31, 2018, exclusive of null values in Class of Offense provided the following descriptive statistics:

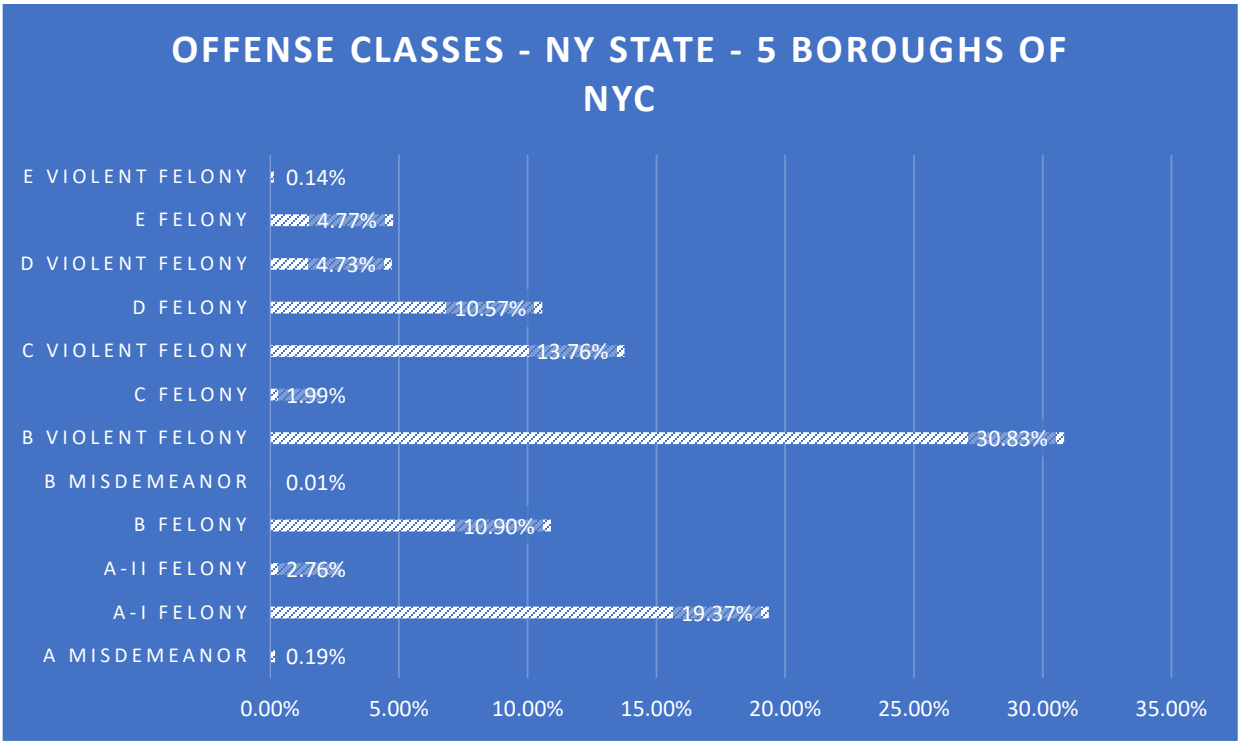

Figure 7: Classes of Offense

The classification of criminal offenses in New York State is arranged by most serious levels of misdemeanors and felonies, Misdemeanors covering petty and minor crimes, from petty theft to Driving while Intoxicated, First Offense, no injuries or property destruction. Classes A are the most serious, followed by Class B, with D and E being the least serious. Of particular interest is the occurrence of 50.2% of all classes of crimes in either Class B Violent Felony or A-I Felony (Murder or attempted murder). A possible concern would be that these 2 classes could bias classification outcome models, resulting in false positives regarding error testing. In the

above statistics for the 5 counties (boroughs) that comprise New York City, the most serious crimes committed by inmates admitted in 2018, were recorded, according to the New York State Department of Corrections and Community Supervision. (NYSDOCCS, n.d.). Conversely, the population distribution by risk expressed as Custody Level is (Please refer to Figure 8):

Custody_Level	COUNT	PERCENT
MAX	7645	49.5014245
MED	7293	47.22222222
MIN	319	2.065527066
SHO	187	1.210826211

Figure 8: Breakdown of Custody Level Classification, New York City, 2018-2019

Within each Custody Level classification, the following descriptive observations were made:

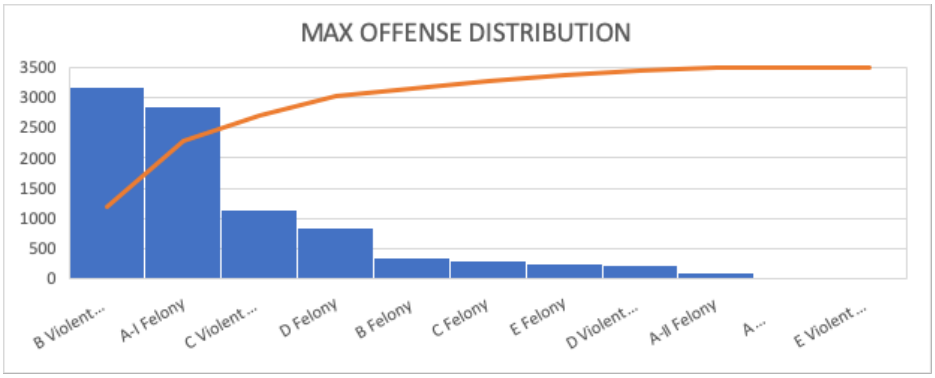

Figure 9a: Offense Class Representation in Custody Level MAX (Maximum)

However, when Class Severity is created and aligned objectively through predictive modeling toward data-driven classification: (See 9b, 9d, 9e, 9g)

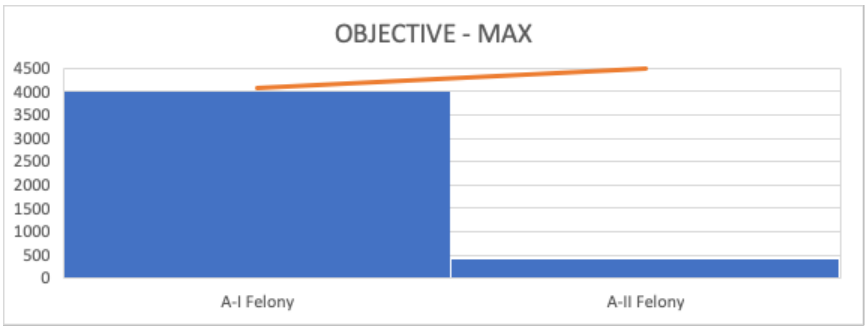

Figure 9b: Objective – Aligned Offense Class Representation in Custody Level MAX (Maximum)

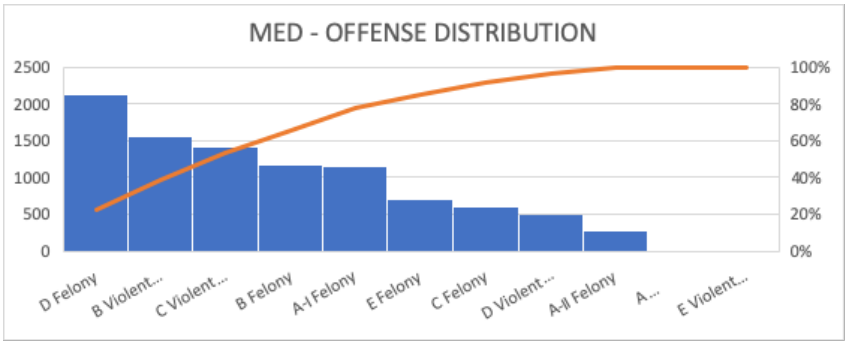

Figure 9c: Offense Class Representation in Custody Level MED (Medium)

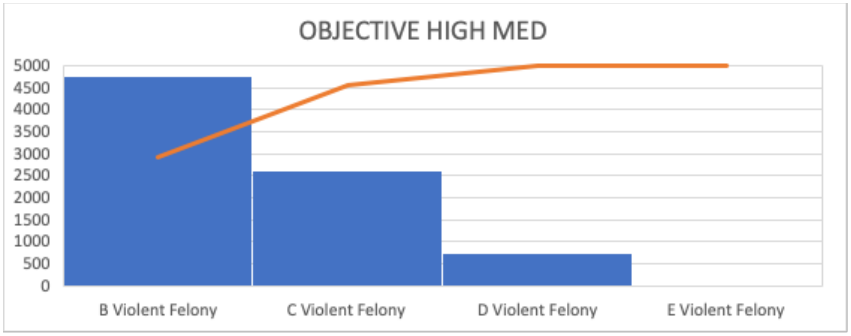

Figure 9d: Objective – Aligned Offense Class Representation in Custody Level HIGH MED (Medium)

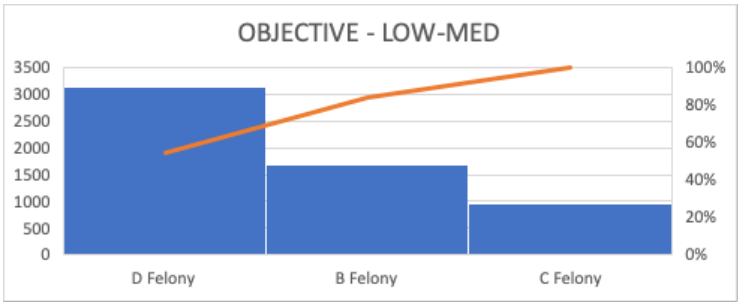

Figure 9e: Objective – Aligned Offense Class Representation in Custody Level LOW MED (Medium)

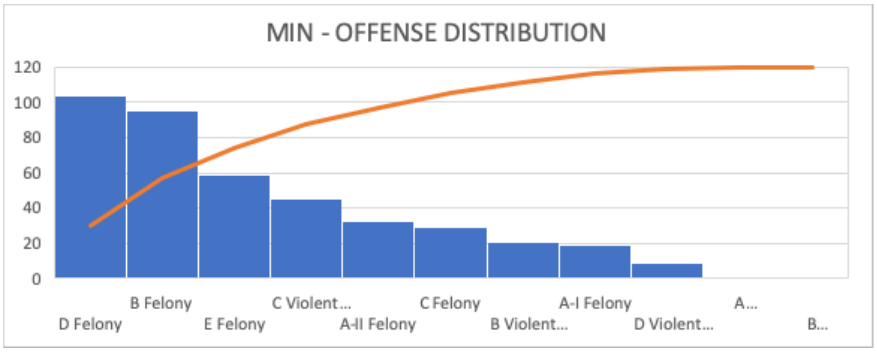

Figure 9f: Offense Class Representation in Custody Level MIN (Minimum)

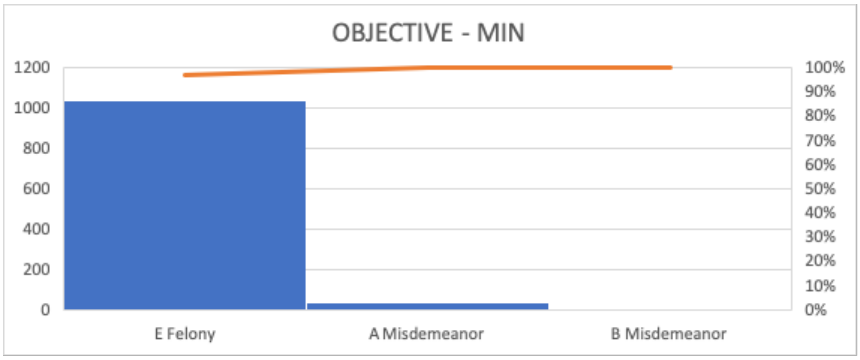

Figure 9g: Objective – Aligned Offense Class Representation in Custody Level MIN (Minimum)

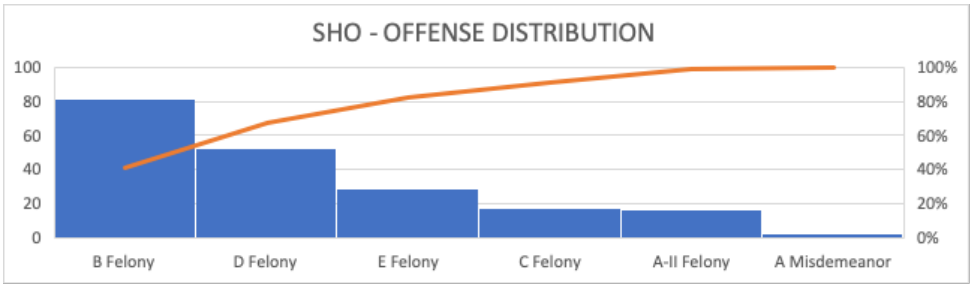

Figure 9h: Offense Class Representation in Custody Level SHO (Shock Incarceration) (No Objective-aligned Predictive Class, due to relationship with Maximum Custody Level)

Note: The above Objective levels were derived by a predictive model that is described from Pages 24 – 30. While this is partially out of synch in the report, a comparative set of Subjective and Objective Descriptive Statistics could better illustrate the story.

A list of New York Criminal Code Offense Classes (Strazzullo Law Firm, n.d.), other than

Class A-I and A-II felonies (Life imprisonment, or Death Penalty) follows (Ref: Figure 10):

No Priors	Non Violent Predicate	Violent Predicate	
B Violent Felony	Lowest: 5 years in prison	Lowest: 8 years in prison	Lowest: 10 years in prison
	Highest: 25 years in prison	Highest: 25 years in prison	Highest: 25 years in prison
B Non Violent Felony	Lowest: 1-3 years in prison	Lowest: 4 1/2 – 9 years in prison	Lowest: 4 1/2 – 9 years in prison
	Highest: 8 1/3 – 25 years in prison	Highest: 12 1/2 – 25 years in prison	Highest: 12 1/2 – 25 years in prison
C Violent Felony	Lowest: 3 1/2 years in prison	Lowest: 5 years in prison	Lowest: 7 years in prison
	Highest: 15 years in prison	Highest: 15 years in prison	Highest: 15 years in prison
C Non Violent Felony	Lowest: No Jail (Probation possible)	Lowest: 3-6 years in prison	Lowest: 3-6 years in prison
	Highest: 5 – 15 years in prison	Highest: 7 1/2 – 15 years in prison	Highest: 7 1/2 – 15 years in prison
D Violent Felony	Lowest: 2 years in prison	Lowest: 3 years in prison	Lowest: 5 years in prison
	Highest: 7 years in prison	Highest: 7 years in prison	Highest: 7 years in prison
D Non Violent Felony	Lowest: No Jail (Probation possible)	Lowest: 2 – 4 years in prison	Lowest: 2 – 4 years in prison
	Highest: 2 1/3 – 7 years in prison	Highest: 3 1/2 – 7 years in prison	Highest: 3 1/2 – 7 years in prison
E Violent Felony	Lowest: 1 1/2 years in prison	Lowest: 2 years in prison	Lowest: 3 years in prison
	Highest: 4 years in prison	Highest: 4 years in prison	Highest: 4 years in prison
E Non Violent Felony	Lowest: No Jail	Lowest: 1 1/2 – 3 years in prison	Lowest: 1 1/2 – 3 years in prison
	Highest: 1 1/3 – 4 years in prison	Highest: 2 – 4 years in prison	Highest: 2 – 4 years in prison
A Misdemeanor	Lowest: No Jail	Lowest: No Jail	Lowest: No Jail
	Highest: 1 year in prison	Highest: 1 year in prison	Highest: 1 year in prison
B Misdemeanor	Lowest: No Jail	Lowest: No Jail	Lowest: No Jail
	Highest: 90 days in prison	Highest: 90 days in prison	Highest: 90 days in prison
Violation	Lowest: No Jail	Lowest: No Jail	Lowest: No Jail
	Highest: 15 days in prison	Highest: 15 days in prison	Highest: 15 days in prison

Figure 10: List of New York State Offense Classes Other Than A-I and A-II felonies

Of interest in Figures 9a, 9c, 9f, and 9h is the absolute indication in descriptive form of harshly penalized offenses in Minimum security, while Maximum and Medium contain individuals having been convicted of Class A Misdemeanors! Shock Incarceration, as stated above is intended to divert first-time felons, other than those having committed Class A-I and A-II felonies ((New York State Department of Corrections and Community Supervision, March 18, 2019). Therefore, the determination of Custody Level is probably subjective, guided but not totally constrained by law.

A comparative view of recidivism patterns among classes of offenses shows a marked level within violent felonies (Refer to Figure 11):

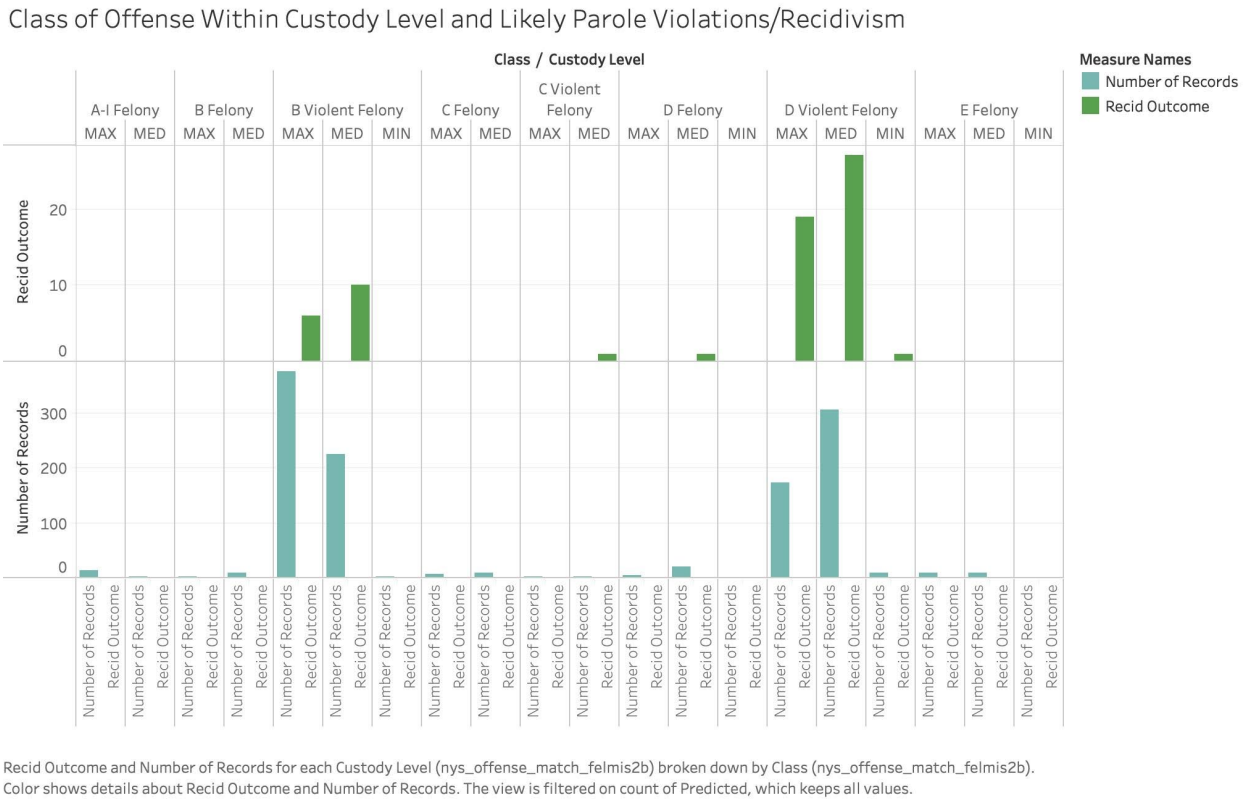

Figure 11: Breakdown of Custody Level, Offense Classification, and Recidivism Events, New York City, 2018-2019

DISCUSSION – PHASE 1

The above descriptive visualizations and frequency distributions could be illustrating subtle but apparent differences due to subjectivity in the relationship between overall offense, offense class, and risk control (custody level) events within the New York State inmate population from New York City, particularly influenced by violent felonies. Through the above descriptive and analytics-based visualizations, insight may be guided toward these effects:

- Gang Affiliation / Activity
- Mental Health Issues
- Violent Felonies Other Than Murder
- Offenses and Probabilities of Correct Classification

The insights displayed in Figures 6a through 6b would suggest a predictive bias toward custody level of Medium Security, which in fact is 47% of the 2018-2019 sample. The prediction is correct vs incorrect classification, based upon the occurrence of offenses as categories, with predicted outcomes of Maximum,

Medium, Minimum, and Shock Incarceration. Gang Affiliation and Mental Health could contain effects to influence more severe custody level classification. Figure 11 shows a descriptive trend toward custody level Maximum regarding Class B Violent Felonies. Given the above insights, where statistical software, machine learning software, and descriptive frequencies show a marked influence between Custody Level and offense classes, there could be a bias where human assignment is utilized more than offense guidance.

Therefore, our third assumption is that another effect, possibly subjective confirmation bias (Nickerson, R. S., 1998), defined as, "the seeking or interpreting of evidence in ways that are partial to existing beliefs, expectations, or a hypothesis in hand," could exist in classifying custody level. A guiding research question would now be, "How can subjective, confirmation bias be adjusted?" According to Ofer, D. (2017), subjective bias exists, such as racial bias in classification for the state of Florida.

PHASE 2: MODEL SELECTION

The CRISP-DM framework (Abbot, D., 2014), has thus far guided the above subject matter utilizing:

Stage	***Description***
Business Understanding .	Define the project
Data Understanding	Examine the data; identify problems in the data.
Data Preparation	Fix problems in the data; create derived variables.
Modeling Build predictive or descriptive models.	Build predictive or descriptive models.

Figure 12: Initial 4 steps CRISP-DM (Abbott, D., 2014)

Phase 2 is comprised of:

Evaluation	Assess models; report on the expected effects of models.
Deployment	Plan for use of models.

Figure 13: Final 2 steps, CRISP-DM (Abbott, D., 2014)

In Phase 1, Logistic Regression utilized in a simple bivariate analysis utilizing SAS® STAT ™, and a highly multivariate analysis utilizing SAS® Enterprise Miner ™. Subsequently, Decision Tree models were attempted to achieve a predictive view of the relationship between classes of offenses (CLASS) and offenses (MOST_SERIOUS_CRIME). An example follows in Figure 14:

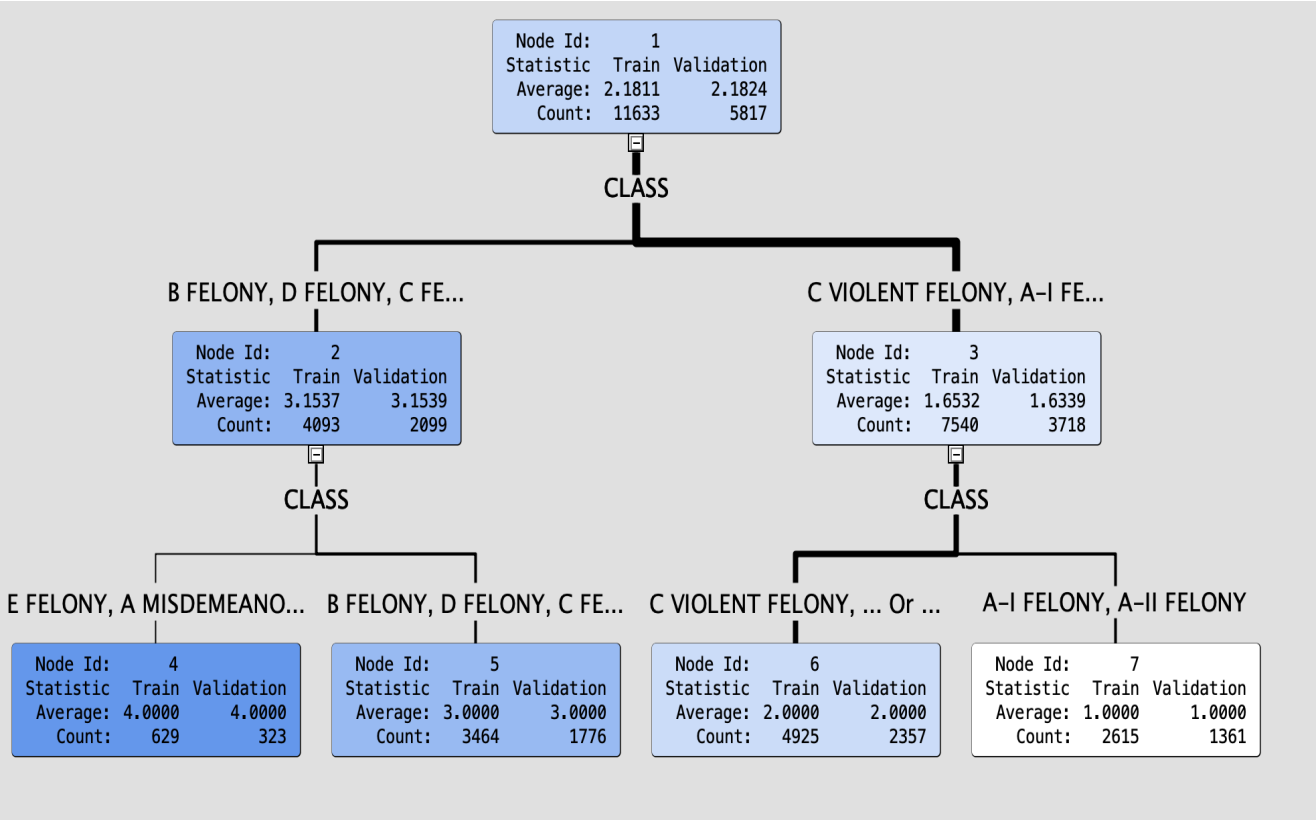

Figure 14: Decision Tree, Outcomes of Offenses per Offense Classes

Of interest in this analysis would be the trending effect of probabilities of felonies, especially Violent Offenses. This is a valuable probabilistic guide to trends of felonies and misdemeanors committed within the inmate population, and a guide to determine objective classification following the classification guidelines listed by Cameron, B. (2019). This would be a good planning tool for population-wide predictive analytics, it may also provide an alternative to the highly subjective category CUSTODY_LEVEL. The Decision Tree is the most commonly used model by data scientists, accurate to the seventh tier (Abbott, D., 2014). This will be discussed later in the paper.

A second model of interest would be logistic regression with high multiple regression predictor occurrences of felony and misdemeanor criteria (criterion variables) as CLASS, and responses in MINIMUM, MEDIUM, MAXIMUM, or SHOCK INCARCERATION Custody Levels, as seen in Figures 6a and 6b. While this model is also applicable as a population planning tool, it is not relevant to inmate-level processing, but it introduced the **probability of placement** in MINIMUM, MEDIUM, MAXIMUM, or SHOCK INCARCERATION. This model is informative in our study, as it indicates Incorrect and Correct proportions within the 3 outcomes of MINIMUM, MEDIUM, and MAXIMUM.

Bias in subjective selection by human beings could have caused the Incorrect readings. The need to control for bias is apparent, examined for the third analytic model, with the Firth bias-adjustment estimates option as a candidate. (Bar, H., 2012). Firth, D. (March, 1993) develops a method of adjusting the maximum likelihood estimate to reduce inherent bias, such the degree of inaccurate classification, caused in part by prejudicial subjective scoring (Ofer, D., 2017). However, Firth would only be utilized for binary/bivariate (2 x 2) models only; this is not a classic binary model, so invoking the Firth option is ignored in SAS®. A propensity scoring method could be a possibility for mitigating error from bias. (D'Agostino, R. B., 1998), or, a weight option for adjustment. A third bias mitigation method could be PROC STDIZE, which standardizes a distribution.

A comparative modeling exercise was done on a training dataset in SAS® Enterprise Miner ™. Using the Training and Validation subsets of the complete New York State / 5 Borough Prison intake population, (NYC), using New York State prison admissions, (2019). A summary comparison of data mining derived models of Random Forest, Decision Tree, High Performance (HP) Regression (Logistic), and conventional Logistic Regression (Reg) approaches follows. (Ref. Figure 15):

Fit Statistics				
Model Selection based on Valid: Average Squared Error (_VASE_)				
			Valid:	Train:
			Average	Average
Selected			Squared	Squared
Model	Model Node	Model Description	Error	Error
Y	Tree3	Decision Tree (3)	0	0
	Reg3	Regression (3)	2.86E-31	7.72E-31
	HPDMForest	HP Forest	4.15E-31	0.000343849
	HPReg	HP Regression	1.00E-29	6.12E-30
	Neural2	Neural Network (2)	0.000000441	0.000000414

Figure 15: Model Selection Results Summary

Using the Abbott, D. (2014) argument of the “best accuracy approach” relies upon the best representation of reality in the model, irrespective of statistical coefficients such as R^2 or average squared error, although in

determining model selection, the degree of error in a machine-learning environment could help in justification. To be conclusive, to repeat Abbott, 2014, the assessment should match the business objective(s) concluded within Business Understanding.

Another option would be scoring and testing for fit, or percentage of error in the data based upon the observed and expected scored results of the logistic model. A dataset derived from NY daily inmates in custody: From New York City open data. (2019)., (2019), 'nys_offense_match_FelMis2b' ('Felonies & Misdemeanors v 2b) was used. Given the alternative option, the third analytic model for evaluation would therefore be a scored predictive logistic regression model, programmed in SAS® STAT ™, using the following procedural code:

```
proc logistic data=d.nys_offense_match_FelMis2b descending plots=EFFECT;
   model CUST_LEV_NUM = offense_num  /  alpha=0.05 cl link=logit rsquare;
     oddsratio offense_num;
     score out=d.score1 fitstat;
    output out=d.out4o p=predicted l=lower u=upper reschi=resid ;
run;
```

Initially, accuracy of matches, or 'fit' would be of interest in cross-checking the model expressed by the "Correct/Incorrect " distribution visualized in Figure 6b. The Error Rate calculated by the Logistic model shows a 41.27% error rate, or, custody level classification is only 58.73% correct. A check on the dataset shows 8002 errors, where the observed score value of Custody Level (CUSTODY_LEVEL_NUM), where:

- 3 = Maximum;
- 2 = Medium;
- 1 = Minimum;
- 0 = Shock Incarceration

differs from the predicted score value, which is derived from the highest probability per Custody Level. As there are 19,389 observations, with 8,002 in error, the error percentage is 8,002/19,389, or 0.41270824.

Fit Statistics for SCORE Data

Data Set	Total Frequency	Log Likelihood	Error Rate	AIC	AICC	BIC	SC	R-Square	Max-Rescaled R-Square	AUC	Brier Score
D.NYS_OFFENSE_MATCH_FELMIS2B	19389	-15807.7	0.4127	31623.49	31623.49	31654.98	31654.98	0.031568	0.038955	.	0.515937

Figure 16: Fit Statistics for Subjective Score Data

Using Hatcher, L. (2003) matrix of R^2 strength:

R-Square	Description
+-1.00	Perfect
+-0.80	Strong
+-0.50	Moderate
+-0.20	Weak
'0.00	None

Figure 17: R^2 Score Strength Matrix

With the resulting SCORE table as the source for empirical probabilities and accuracy scores, the highest Custody Level probability corresponding with the CUSTODY_LEVEL_NUM value (P3, P2, P1, P0) would be the predicted Custody Level for classification. This could be the best way to leverage predictive insight of whether or not the assigned Custody Level Classification outcome (the Dependent, or Y variable) followed the predicted Custody Level based upon the predictors, or offenses (The Independent, or X variable), in this model, being 'offense_num," or a dummy numeric variable assigned for each offense, which would be an objective variable .

SUBJECTIVITY VS. OBJECTIVITY

A subjective approach causing error rates as stated above could be an effect of confirmation bias (Nickerson, R. S., 1998). Minimization of subjectively created responses to offense events would need scores of objective variables, such as the degree or ranking of criminal offenses, as illustrated above in Figure 10. For the present, a Key Performance Indicator as Custody Level is reflecting too much variance as expressed in the R^2 coefficients in Figure 15. Therefore, in order to apply a more correlative predictive model of classification, a nearest-neighbor approach to establishing justification for a "new metric" (Abbott, 2014) should be evaluated.

A fourth analytic method, K-means clustering, could provide insights to randomly clustered probabilities among categories such as offense classes (CLASS), but would guide toward the outliers and most uniform cluster categories, allowing analysis similar to the Decision Tree (Figure 14), and Multivariate Logistic Regression predictive error outcome plot (Figure 6b). A clustering model, initially standardizing the more objective KPIs could allow a more robust assessment.

Two offense-related variables were added, **Class_Severity** and **CS_Grp, (**Ref: Figure 18), replacing 'nys_offense_match_FelMis2b,' creating dataset "nys_offense_groups_clean2:"

Class_Severity	CS_Grp	CLASS	COUNT	PERCENT
1	1	A-I Felony	4021	20.7385631
2	1	A-II Felony	426	2.19712208
3	2	B Violent Felony	4761	24.55516014
4	2	C Violent Felony	2602	13.4199804
5	2	D Violent Felony	730	3.765021404
6	2	E Violent Felony	21	0.108308835
7	3	B Felony	1685	8.69049461
8	3	C Felony	944	4.868740007
9	3	D Felony	3131	16.14833153
10	4	E Felony	1034	5.332920728

11	4	A Misdemeanor	33	0.170199598
12	4	B Misdemeanor	1	0.005157564

Figure 18a: Objective Variables, derived from 'Authority' (Offense): 'Class_Severity,' 'CS_Grp,' and 'Class'

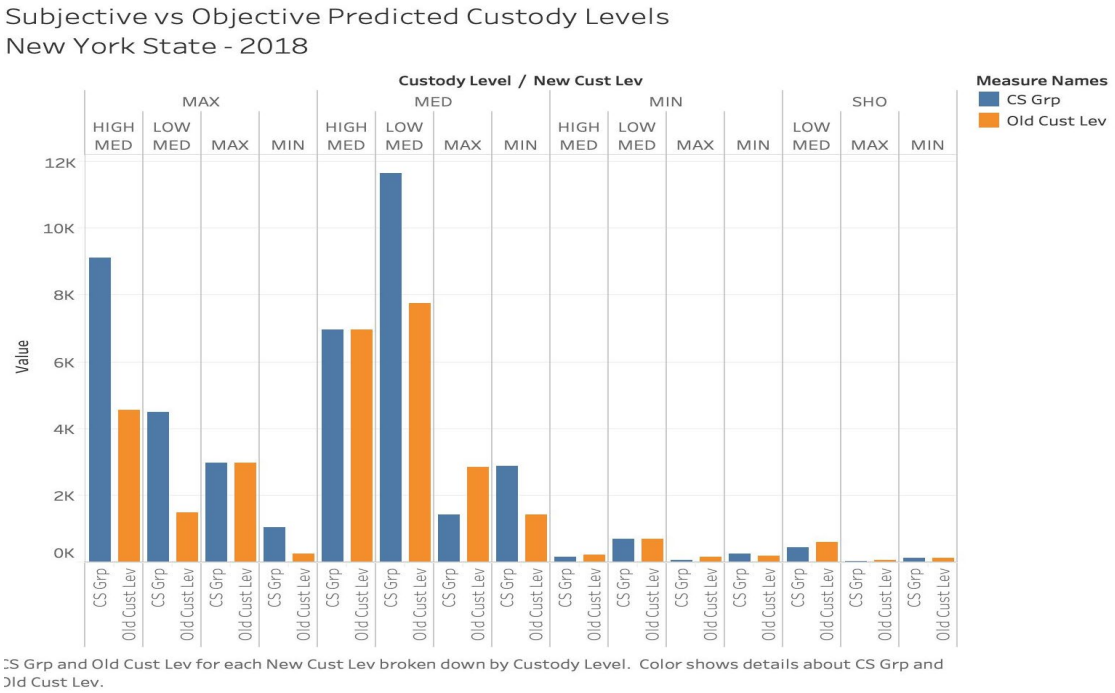

Figure 18b: Predicted Objective Levels (CS Grp), compared with Subjective Custody Levels

Note that variable CS_Grp aligns with the Decision Tree in Figure 14, with Custody Levels MAX , MED, MIN respectively replaced by CS_Grp '1', '2'&'3', and '4.'

Therefore, after careful evaluation, both a logistic regression derived, inmate-level generated probability ranged list of errors, a K-Means derived cluster model to determine the best Predictor should coexist as analytic tools, with the new Response variable based in the Decision Tree illustrated above in Figure 14.

To determine the best Predictor, given the near complete randomness of subjectivity using the logistic model, a K-Means analysis was run using SAS® STAT ™, comparing the independent predictive variables' interaction as nearest-neighbors, using R^2 as a guide, along with the overall R^2 to determine percentage of acceptable variance, or "fit," given as >= 0.70 (Bafna, J., March 15, 2017). The method used was initial standardization of the KPIs 'Class_Severity' and 'CS_Grp' , using PROC STDIZE, then processing the resulting dataset using PROC FASTCLUS with the following variables: Race, Class_Severity, offense_num, CS_Grp. After testing, the most optimal level of clustering was found to be 24, where variation was found to be at or greater than an acceptable level of 0.70 (0.79636), as defined in Bafna, J. (March 15, 2017). The results are (Ref: Figure 19):

Statistics for Variables				
Variable	**Total STD**	**Within STD**	**R-Square**	**RSQ/(1-RSQ)**
RACE	1	0.28733	0.91754	**11.127032**
Class_Severity	1	0.18631	0.96533	**27.843306**
Offense_Num	1	0.25049	0.937331	**14.956846**
CS_Grp	1	0.06216	0.996141	**258.133436**
OVER-ALL	1	0.2144	0.954085	**20.779548**
Approximate Expected Over-All R-Squared =	0.79636			

Figure 19: Objective Variables Correlation Statistics K-Means Nearest Neighbor Model

Given the high level of correlation among clusters, illustrated in Figure 18, there is confidence in measuring the utilization of Class_Severity as the Predictor, and CS_Grp as the new classification parameter in a logistic regression model.

The new predictive model, free of confirmation bias as defined by Nickerson, R. S., 1998, shows a nearly uniform, error-free classification method. Note the logarithmic probability pattern (Abbott, 2014), indicative of a strong logarithmic relationship between 0 and 1. (Ref: Figure 20):

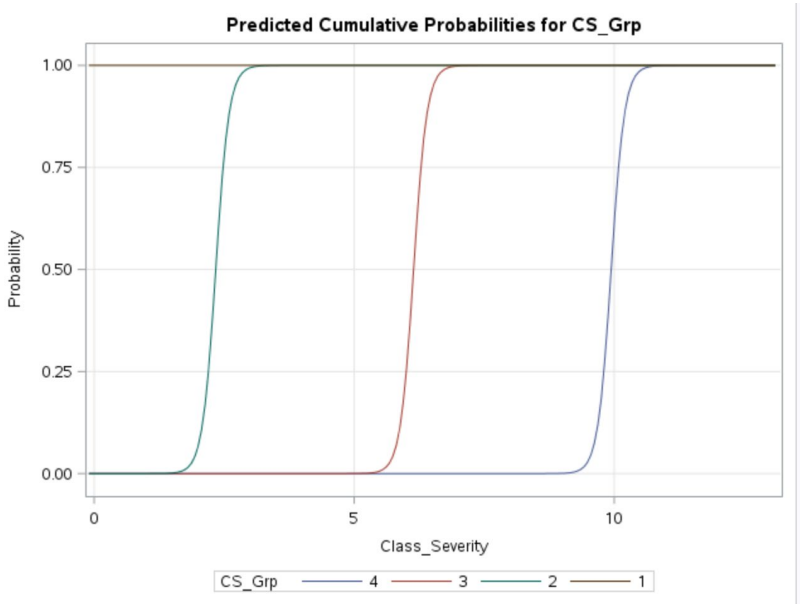

Figure 20: Objective Variable CS_Grp vs Class_Severity – Logistic Predictive Cumulative Probabilities

Fit Statistics would now be:

Fit Statistics for SCORE Data

Data Set	Total Frequency	Log Likelihood	Error Rate	AIC	AICC	BIC	SC	R-Square	Max-Rescaled R-Square	AUC	Brier Score
D.NYS_OFFENSE_GROUPS_CLEAN2	19389	-67.2332	0	142.4665	142.4685	173.9563	173.9563	0.912676	0.999339	.	0.000348

Figure 21: Fit Statistics for Objective Score Data

With the redefined SCORE table as the source for empirical probabilities and accuracy scores, the highest Custody Level probability corresponding with the CS_Grp value (P3, P2, P1, P0) could be the objective, data-driven predicted Custody Level for classification as defined in the above Business Objectives.

ASSESSING RECIDIVISM WITH THE LOGISTIC MODEL

The New York State Inmate dataset (NY daily inmates in custody: From New York City open data, 2019) contains parole violation intake tracking in labeling the inmate's status, deriving the binary dummy variable 'recid_outcome,' comprising roughly 4.6% of the cohort. As Alper, M., et al., 2018, illustrate am 83% recidivism rate within a 9 year period in 30 states presents a nationwide problem in corrections. Ofer, D. (2017) argues that a subjective confirmation bias effect exists in Florida (N=60843), pertaining to recidivism, tested in the categorical (logistic) relationship between Supervision Level (similar to CS_grp above), as Response, and Decile Score of Recidivism Risk as Predictor, reflecting a low to high score of 1 through 10 respectively. Using the following SAS® STAT ™, logistic regression model, the results could refute Ofer, D. (2017). (Ref: Figures 22 and 23):

```
proc logistic data=d.florida_recid_only descending plots=EFFECT;
   model  RecSupervisionLevel  = DecileScore /alpha=0.05 cl link=logit
   rsquare;
      oddsratio RecSupervisionLevel; score out=d.score6 fitstat;
   output out=d.RecidFL_grp p=predicted l=lower u=upper reschi=resid ;
run;
```

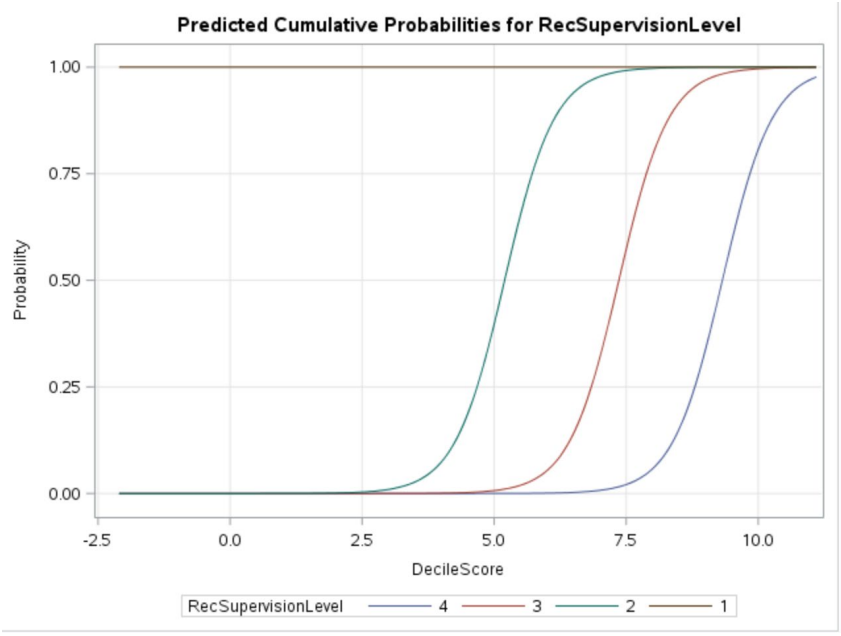

Figure 22: Objective Variable RecSupervisionLevel vs DecileScore – Logistic Predictive Cumulative

Fit Statistics for SCORE Data

Data Set	Total Frequency	Log Likelihood	Error Rate	AIC	AICC	BIC	SC	R-Square	Max-Rescaled R-Square	AUC	Brier Score
D.FLORIDA_RECID_ONLY	20281	-6714.2	0.1005	13436.41	13436.42	13468.08	13468.08	0.756516	0.865162	.	0.182786

Figure 23: Fit Statistics for Objective Recidivism Score Data

Using the Bafna, J. (March 15, 2017) standard of 0.70 as an acceptable level of R^2 variability, we may infer with an R^2 of 0.756516, and an error rate of 0.1005, a moderate level of categorical correlation exists, indicating some subjective effect of risk assessment of recidivism in Florida. Note that the Supervision Level classification of 1 has no variability. According to Abbott, D. (2014), the R^2 could be an excellent indicator, given that social science correlations could return relatively good relationships with R^2 coefficients of 0.30. As this could qualify as a social science exercise, we would be confident, given the standard of 0.70 in Bafna, J. (March 15, 2017). A success rate of nearly 90% would suggest a relatively accurate prediction of recidivism risk in Florida's prison population, where an 83% risk (Alper, et al, May 23, 2018) exists countrywide.

DISCUSSION – PHASE 2

The classification model utilizing objective statistical data in analyzing the New York State dataset (NYS prison admissions: Beginning 2008: From New York State open data, 2019) could have value in determining the risk of custody and recidivism. While effects of gang affiliation and mental health are interesting measures, the overwhelming discovery in this study has determined a large confirmation bias effect, rendering an initial data-driven solution with an error rate of 41.27% nothing more than a "coin toss! " To manage confirmation bias within a subjective classification culture, the assistance and analytic alternative of an objective, criminal-code based Predictor(KPI = Class_Severity), against a new Response of a Classification Group (KPI = CS_GHrp), based upon criminal-code offense severity could provide a true, data-driven solution for Inmate Risk Assessment.

DEPLOYMENT – STEPPING BACK TO MOVE FORWARD

With the above model of Logistic Regression, supported by variable validation through Decision Tree and K-Means clustering models, plus descriptive statistical visualizations, a Business Intelligence / Reporting tool could be useful at the workstation of an inmate intake interviewer, a corrections planner, or a prison warden. However, before visualization/reporting is designed, a data warehouse infrastructure to support a Business Intelligence suite of

performance and administrative-level tools would need to be designed, tested, and implemented to support a data-driven change, which would deliver the above analytic solutions. To tell the story, a foundation needs to be built. The following overview utilizes the experience of private enterprise in developing a Business Intelligence framework which would support a holistic enterprise/agency rather than a siloed, or office-specific solution.

A convincing and sponsored business need to build or sustain Business Intelligence in the enterprise would be based upon a mature but archaic business reporting system, limited to 2 dimensional tabular reports . The business requirements, such as business rules driving data structures and analytic aggregation need to drive the continued development of the solution, from conception to completion. Otherwise, the enterprise, from the experience of this writer, is merely chasing a software/hardware panacea presented by a convincing sales team, an aggressive management team, biased consultants, or politically motivated IT department, resulting in a difficult bridge between the promise of the purchased products and the actual business need of the enterprise departments needing a better insight into the enterprise's Organizational Memory, developed through Information Integration, which delivers Insight Creation, and allows effective Presentation of summarized and detailed analytic data (Sabherwal and Becerra-Fernandez, 2011), through a system that can evolve as the needs of the enterprise change through continued learning and improvement (Teich, 2018).

INSIGHT CREATION THROUGH DATA DEPLOYMENT

Assuming a scalable and well-tested data infrastructure delivers:

- Transactional data of events in the time dimension, such as entry into the prison system;
- Master data, which defines and describes business-relevant people, places, and things;
- Reference data, stored in tables containing lists of demographic, geographic, account, and other data acting in concordance between other Transactional and Master data tables and schemas. (Kimball, R., & Ross, M., 2013).

Following 3-4 panel subject matter dashboard design methods in Wexler, S. (2017), population and individual detail analytics, descriptive and non-parametric could be constructed from our predictive analysis. An interactive structure could be: (Ref: Figure 24):

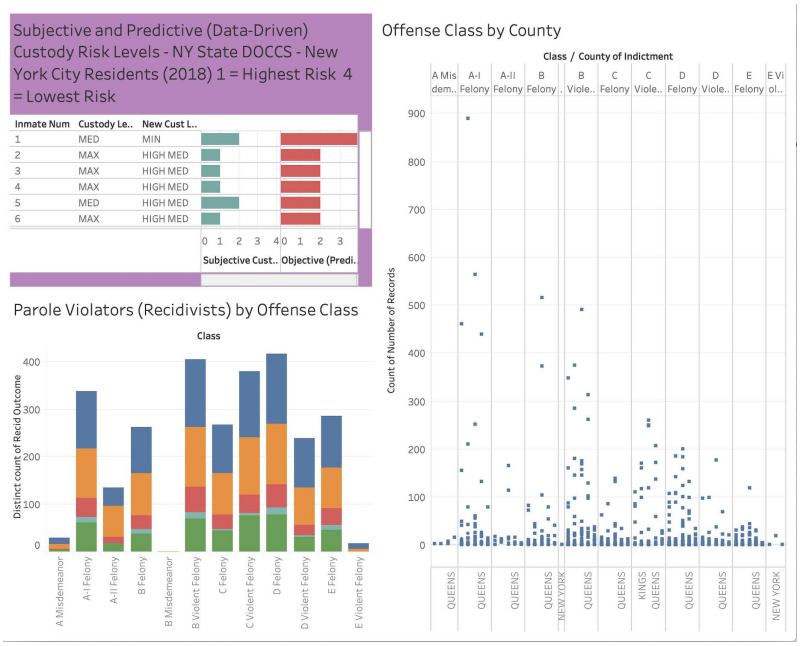

Figure 24: Prototype Analytic Dashboard

The story in the above dashboard starts with the Comparative Prediction panel, where Subjective and Predicted levels appear side-by-side, to allow a transaction-level, inmate-detail comparison of the subjective Assigned Custody Level, and the objective Data-Driven Predicted Custody Level. The lower left hand quadrant summarizes Parole Violators, or confirmed recidivists per Offense Class. The right 2 quadrants display Offense Classes per County (Borough) within New York City . More panels could be created, from Demographics of Race and Age per Offense Class. The distribution of Custody Levels within all New York State prison facilities could assist management in strategically planning for growth or contraction in prison housing based upon Custody Level, expanded to show Offense Class. The above would be more of an Executive dashboard, concerned with mostly summary KPI (Predicted Custody Level) information (James, 2010).

SAS (2016), details the Deployment Stage of the Analytics Life Cycle as:

- “Implement the models through testing and vetting;” Models are representations of reality, or what could be expected to occur based upon data.

- “Act on new information;” Once an answer is found, or a prediction delivered based upon facts, act iteratively.

- “Ask again.” Interrogatories change with data and changing effects of external and internal conditions.

Nothing is static.

An example of, “ask again,” utilizing Exploratory Data Analysis (Jebb, A. T., et al.,2017), concerning deployment of Gang Affiliation analytics follows.

Another analytic solution to classification of inmates upon intake would need to be the guide in both exploratory analysis (initial visualization of the data) and presentation of findings (scoring solution):

- Can security level classification of daily inmate intake data from New York City’s Department of Corrections be used in a predictive model to predict classification outcome, and scored to assist in managing the best secure inmate intake ?

Repeating Abbott, D. (2014), the most popular visualization analysis of unsupervised data is the Decision Tree. Upon initial exploration, a decision tree could be applied to the NY daily inmates in custody: From New York City open data. (2019) data sample regarding pre- Shock Incarceration Custody Level: (Minimum, Medium, or Maximum Security) for a sample from the first half of 2018:

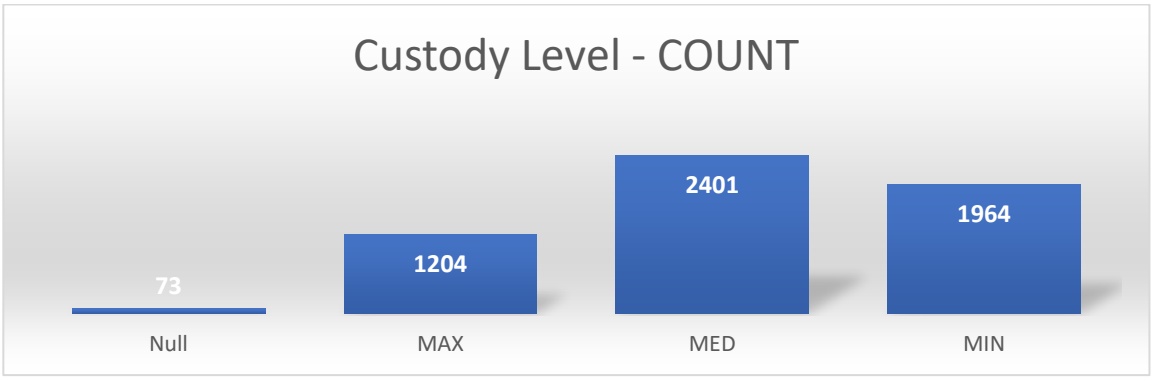

Figure 25: Descriptive Custody Level Distribution

Percentages would be:

Medium	0.40478
Minimum	0.344635
Max	0.250585

Figure 26: Descriptive Custody Level Distribution - Percentages

It is essential that data preparation and exploration are complete, facilitating a sound methodology of visualization and insight. A Decision Tree analysis of probabilistic occurrence (Ref: Fig. 26) seems to stem from Gang Affiliation, then Infractions, followed by demographic predictors GENDER and AGE!

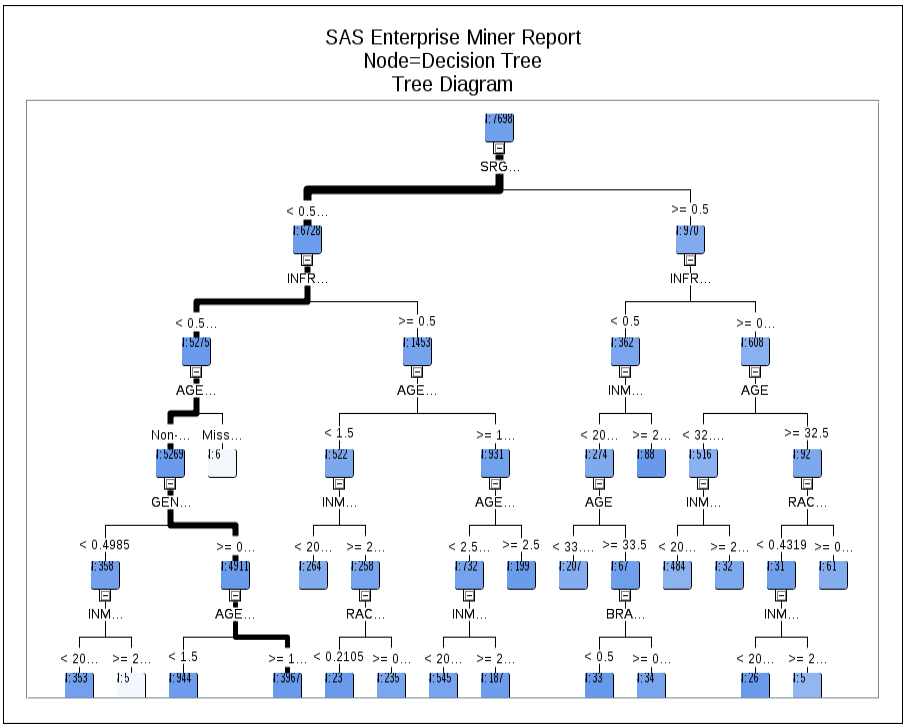

Figure 27: Decision Tree Analysis – Gang Affiliation

Response Rate of graphic point of inflection, or sectional cutoffs would be best met by the method, as a scale of predictive scores:

Range for Predicted	Mean Predicted	Max Predicted	Min Predicted	Mean Target	Max Target	Min Target
0.353 - 0.372	0.35935	0.37176	0.35327	0.35935	0.40478	0.00000
0.335 - 0.353	0.34808	0.34826	0.34783	0.34808	0.40478	0.00000
0.316 - 0.335	0.32422	0.32784	0.31961	0.32422	0.40478	0.00000
0.297 - 0.316	0.31162	0.31486	0.30359	0.31162	0.40478	0.00000
0.279 - 0.297	0.28877	0.29107	0.28619	0.28877	0.40478	0.00000
0.260 - 0.279	0.26627	0.26627	0.26627	0.26627	0.40478	0.25059
0.242 - 0.260	0.25059	0.25059	0.25059	0.25059	0.25059	0.25059
0.000 - 0.019	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

Figure 28: Decision Tree Analysis – Predictive Scores

With the target as CUSTODY_LEVEL_ALT: (Custody Level, equal to the New York State data field for CUSTODY_LEVEL), confirmation of the Decision Tree primary predictors of Gang Activity ("SRG_FLG") and Infractions (INFRACTION). The Model Assessment reaches a Point of Inflection at 0.353-0.372, (Ref: Figure 33), which leads us to the scoring solution where Gang Activity and additional Infractions are on the charges of the entering inmate, the minimal custody assignment should be MEDIUM regardless of whether or not the charges are minimal, given gang affiliation and / or accompanying infractions.

In brief, this probably doesn't make much sense, as it points toward Gang Affiliation and Infractions guiding toward the MED (Medium) Security Level, heading toward MAX. Perhaps Infractions were in addition to the Top Charge, also known as Most Serious Offense.

Scoring the model follows.

The comparative visual for presentation of the 2 most likely contributing factors, Gang Affiliation and Infractions, as indicated in the exploratory Decision Tree would be:

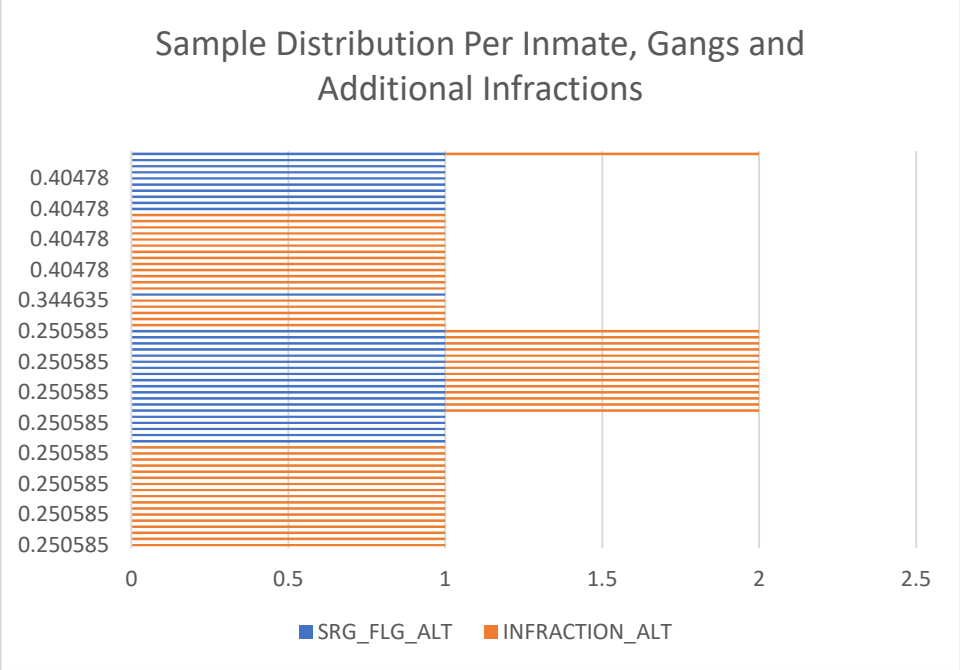

Figure 29: Decision Tree Analysis – Model Assessment using

Population Percentage Scores:

Medium (0.40478) Minimum (0.34635) and Maximum (0.250585)

If either a gang affiliation or additional infractions appear on the arrest warrant or ticket during processing, there could be a ratcheting of the guidance up by 1 security factor if initially classified at MIN or MED. There appear to be definite interactions when the population in Maximum contains Gang affiliation or additional infractions.

A very small proportion have been classified for MIN (Minimum) Security; most have been placed in either MAX (0.250585) or MED (0.40478). MIN (Minimum) has the fewest occurrences in the sample! It would appear that for the most part, the classifications could fit the crime, which means the model could be a self-fulfilling prophecy of what was termed a “coin toss” above! If gang activity and/or Infractions accompanying the Top charge are present, then the classification must be Medium or Maximum; if normally Minimum given the offense, the security level should be increased to Medium or if Medium, Maximum.

However, when an objective predictive model is applied as in New York State to New York City Class Severity, and bucketed in Class Severity Group, or CS_Grp, the score of predictive Custody Level would be:

Response Profile				
Ordered Value	Custody Level	CS_Grp	Total Frequency	Percentage
1	MIN	4	2220	0.36441234
2	Low-MED	3	1292	0.21208142
3	High-MED	2	1949	0.31992777
4	MAX	1	631	0.10357846

Figure 30: Objective Custody Level

When applied to a Logistic regression model, factoring Gang affiliation (GANG_AFF) and CLASS_SEVERITY as Predictors of CS_GRP, the correlation is extremely weak, confirmed by only 14.5% of all inmates indicated as having gang affiliation. With a R^2 of just 0.019857, and a low predictive MAX Custody Level, an unsupervised method of a

Decision Tree could be a preferred modeling method of analytic insight (Ref: Figure 27).

Fit Statistics for SCORE Data											
Data Set	Total Frequency	Log Likelihood	Error Rate	AIC	AICC	BIC	SC	R-Square	Max-Rescaled R-Square	AUC	Brier Score
D.NYC_OFFENSE_MAIN2	6092	-7835.5	0.6356	15678.97	15678.98	15705.83	15705.83	0.019857	0.021463	.	0.704198

Figure 31: Gang Affiliation – Predictive Custody Level Target

CONCLUSION

Utilizing public domain data from New York City, New York State and Florida, a categorical model of Custody Level-based was compared with Decision Tree and K-Means Clustering in determining the best fit of predictors and response variables for inmate classification. The analysis was based upon the business objective of a data-driven model, rather than solely depending upon a subjective assessment by a human examiner. With an error

rate of 41.27%, and an extremely low R^2 coefficient of 0.031586, the subjective assessment method, using response variable Custody_Level against predictor Offense_Number , a dummy variable representing Offense from the New York State Criminal Code indicated subjectivity, possibly confirmation bias (Nickerson, R. S., 1998). An alternative method, using Decision Tree modeling to determine an alternative Key Performance Indicator of Class Severity Grouped Risk as Response, and a scored Risk Level of Class Severity, ranged with 1 as most severe, 12 as least severe, as predictor were developed. Predictor validation was accomplished through a K-Means clustering model, indicating a very strong, non-random, bias minimal R^2 coefficients of 0.96533 for Class_Severity, and 0.996141 for CS_Grp, the objective data-driven Custody Level response. When re-analyzed in a logistic regression model, the R^2 coefficient was calculated at 0.912676, with an error rate of 0. This would prove that a data-driven model using objectively derived parameters could be practical in determining non-subjective inmate classification. A follow-up investigation factoring sexual offenses, violence, and other behavioral markers, including recidivism would contribute a more accurate, reality-based classification, keeping data driven analytics a reality, while factoring subjective assessments.

An argument for a Business Intelligence platform, encompassing Organizational Memory, Information Integration, Insight Creation (Analytics), and Presentation Delivery Systems (Sabherwal, et al., 2011), using experience from private enterprise would be necessary to sustain a data-driven culture in a government agency tasked with adult corrections. An example of Deployment of a Decision Tree when checked with a Logistic Regression model proves the value of Exploratory Data Analysis at any stage of analytic assessment.

Finally, NEVER tie yourself down with just one model, and be prepared to test, retest, and compare the results among models, and select whichever is the best representation of reality, not the best justification of your thesis statement, or theory. Science is not ideology, but an objective process dependent on facts, not suppositions.

Data Science: It's not for the faint of heart!

ACKNOWLEGEMENTS

Nobody accomplishes an academic project alone.

- Adam Collis, MS: advised on conditions and cultural attitudes within juvenile and adult corrections.
- Reuben Hine, MS: advised on analytic modeling techniques.
- Ben Kohn, MIT, MS: master storyteller through visualization.
- Louise Mann, MA, this writer's favorite editor and Chief of the Grammar Police. years journey as an adult learner.
- Stefanie Reay, MS, for technical advice.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Alan R. Mann

amann370@icloud.com

REFERENCES

Abbott, D. (2014). *Applied predictive analytics: Principles and techniques for the professional data analyst.* Indianapolis, IN. John Wiley & Sons. p 19-23.

Alper, Mariel, Durose, Matthew R., & Markham, Joshua. (May 23, 2018). "Update on prisoner recidivism: A 9-year follow-up period (2005-2014)." *US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics*, NCJ 250975. Retrieved from: https://www.bjs.gov/index.cfm?ty=pbdetail&iid=6266.

Bafna, J. (2017). K-Means clustering with SAS. Retrieved from https://dzone.com/articles/k-means-clustering-with-sas

Bar, H. (2012). Bias adjustment in logistic regression models. Retrieved from https://www.cscu.cornell.edu/news/statnews/stnews82.pdf

Cameron, B. (2019). Federal prison consultant - Inmate classification - Prisoner classification and BOP designation. Retrieved from https://www.federalprisonauthority.com/inmate-classification-bop-designation/

Cody, R. (1999). *Cody's Data Cleaning Techniques Using SAS Software*. Cary, NC. SAS Institute.

D'Agostino, R. B. (1998). Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Statistics in Medicine, 17, 2265-2281. Retrieved from http://www.uvm.edu/~rsingle/stat380/F04/papers/Dagostino-StatMed-1998_PropensityScores.pdf

Firth, D. (March, 1993). Bias reduction of maximum lkelihood estimates. Biometrika, 80(1), 27-38. Retrieved from https://www.jstor.org/stable/2336755?origin=JSTOR-pdf&seq=1#page_scan_tab_contents

Hatcher, L. (2003). *Step-By-Step Basic Statistics Using SAS*. SAS Institute, Cary, NC, p 290-296.

Huebner, B. M., Varano, S. P., & Bynum, T. S. (2007). Gangs, guns, and drugs: Recidivism among serious, young offenders. Criminology & Public Policy, 6(2), 187-221. doi:10.1111/j.1745-9133.2007.00429.x

James, L. (2010). 6 key features of any business intelligence solution. Retrieved from https://www.yellowfinbi.com/blog/2010/11/yfcommunitynews-6-key-features-of-any-business-intelligence-solution-100207

Jebb, A. T., Parrigon, S., & Woo, S. E. (2017). Exploratory data analysis as a foundation of inductive research. Human Resource Management Review, 27(2), 265-276. doi:10.1016/j.hrmr.2016.08.003

Kimball, R., Ross, M. (2013). The data warehouse toolkit: The definitive guide to dimensional modeling (Third ed.). Indianapolis, IN: John Wiley & Sons, Inc.

Lockhart, P. R. (2019). America is finally being exposed to the devastating reality of prison violence. Retrieved from https://www.vox.com/policy-and-politics/2019/4/5/18297326/prison-violence-ohio-alabama-justice-department-lawsuit

Mayhew, H., Saleh, T., & Williams, S. (2016). Making data analytics work for you instead of the other way around. Retrieved from: https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/making-data-analytics-work-for-you-instead-of-the-other-way-around

National Institute of Standards and Technology. (2010). Guide to protecting the confidentiality of Personally Identifiable Information (PII) Special Publication 800-122. Retrieved from https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-122.pdf

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology, 2(2). p 175-220.

NYSDOCCS, n.d. “New York state law: penal law.” Retrieved from: http://ypdcrime.com/penal.law/offense_level.htm

NY daily inmates in custody: From New York City open data. (2019). Retrieved from:https://www.kaggle.com/new-york-city/ny-daily-inmates-in-custody.

New York State Department of Corrections and Community Supervision. (March 18, 2019). “Shock Incarceration Facilities.” Retrieved from: http://www.doccs.ny.gov/Directives/0086.pdf

NYS prison admissions: Beginning 2008: From New York State open data. (2019). Retrieved from: https://www.kaggle.com/new-york-state/nys-prison-admissions-beginning-2008.

Ofer, D. (2017). COMPAS recidivism racial bias: Racial bias in inmate COMPAS re-offense risk scores for Florida (ProPublica). Retrieved from: https://www.kaggle.com/danofer/compass#propublicaCompassRecividism_data_fairml.csv

Office of Information Policy (OIP), U.S. Department of Justice (2019). "What is FOIA?" Retrieved from: https://www.foia.gov/about.html.

Peterson, T. (2003) Data Scrubbing. *Computerworld*, 37(6), 32.

Richards, N. M., & King, J. H. (2014). BIG DATA ETHICS. Wake Forest Law Review, 49(2), 393-432.

Sabherwal, R., & Becerra-Fernandez, I. (2011). Business intelligence: Practices, technologies, and management. Hoboken, NJ: John Wiley & Sons, Inc.

SAS Institute. (2016). "Manage the analytical life cycle for continuous innovation: From Data to Decision." Retrieved from: http://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/manage-analytical-life-cycle-continuous-innovation-106179.pdf

Strazzullo Law Firm. (n.d.). New York State sentencing guidelines. Retrieved from https://www.strazzullolaw.com/table/

Teich, D. A. (2018, December 26, 2018). Machine learning and artificial intelligence in business: Year in review, 2018. Forbes. Retrieved from: https://www.forbes.com/sites/davidteich/2018/12/26/machine-learning-and-artificial-intelligence-in-business-year-in-review-2018/#5723f6b42041

Understanding prison security levels. (n.d.). Retrieved from https://www.prisonerresource.com/security-levels/

Wexler, Steve, Shaffer, Jeffrey, Cotgreave, Andy. (2017). The Big Book of Dashboards: Visualizing Your Data Using Real-World Business Scenarios. Wiley. Hoboken, NJ.