Combining Machine Learning and Queueing Theory for Data-driven Incarceration-Diversion Program Management

Bingxuan Li[1], **Antonio Castellanos**[2], **Pengyi Shi**[1], **Amy Ward**[2]

[1]Purdue University, West Lafayette, IN, USA
[2]University of Chicago, Chicago, IL, USA
li3393@purdue.edu, antonio.castellanos@chicagobooth.edu, shi178@purdue.edu,
amy.ward@chicagobooth.edu

Abstract

Incarceration-diversion programs have proven effective in reducing recidivism. Accurate prediction of the number of individuals with different characteristics in the program and their program outcomes based on given eligibility criteria is crucial for successful implementation, because this prediction serves as the foundation for determining the appropriate program size and the consequent staffing requirements. However, this task poses challenges due to the complexities arising from varied outcomes and lengths-of-stay for the diverse individuals in incarceration-diversion programs. In collaboration with an Illinois government agency, we develop a framework to address these issues. Our framework combines ML and queueing model simulation, providing accurate predictions for the program census and interpretable insights into program dynamics and the impact of different decisions in counterfactual scenarios. Additionally, we deploy a user-friendly web app beta-version that allows program managers to visualize census data by counties and race groups. We showcase two decision support use cases: changing program admission criteria and launching similar programs in new counties.

Introduction

Recidivism, defined as a person's return to criminal activity after correctional interventions and sanctions, significantly contributes to the mass incarceration issue (Berk 2017; Leipold 2005). Traditional incarceration focuses on punishment rather than addressing the root causes of crime, such as substance use disorder and mental health issues. Consequently, the U.S. criminal system experiences high rates of recidivism. In an effort to break this revolving door of recidivism, incarceration diversion programs have emerged as a promising solution to tackle the root causes (Latessa, Johnson, and Koetzle 2020). These programs provide mental health and substance treatment, education, and community-based services as alternatives to traditional punishment methods. Various settings have shown that such programs can lead to significant reductions in recidivism rates, ranging from 10% to 30% (Peters and Murrin 2000; McNiel and Binder 2007; Lin et al. 2020).

Due to the complex nature of these programs and their emphasis on individualized interventions, successful implementation hinges on two crucial considerations: (i) program sizing and (ii) staffing. The optimal program size needs to strike a delicate balance between the societal cost arising from long-term recidivism reduction and in-program revocation, where "in-program revocation" means that individuals are removed from the diversion program and placed into conventional incarceration due to rule violations or other reasons, as these diversion programs are not strict incarcerations (Burrell 2006; DeMichele 2007). A too-small program may fail to provide sufficient treatment access to individuals who could benefit to improve their long-term success, while an excessively large program could result in numerous in-program revocations, raising potential safety concerns within the community and encountering public resistance. Once the program size is determined, appropriate staffing levels are necessary to cater to diverse needs from individuals in the program. Understaffing leads to overloaded case managers without sufficient bandwidth, which can significantly hamper the overall effectiveness of the program and reduce the likelihood of successful program completion. Furthermore, different program participants possess diverse needs; for instance, Hispanic individuals benefit from Spanish-speaking case managers (Mock 2022). Thus, accurate prediction of the census – the number of participants in the diversion program at any given time, considering their various characteristics – is needed for determining staffing.

Despite the increasing use of machine learning (ML) tools in recidivism prediction and probation decisions, there is a lack of decision support systems (DSS) capable of providing data-driven program sizing and staffing support. In particular, a DSS that aids in understanding the long-term impact of admission decisions on census, in-program revocations, and staffing needs is notably absent. For example, our community partner is considering expanding the eligibility criteria for current diversion programs and launching new programs in additional counties. However, the repercussions of these decisions on the in-program revocation and associated societal cost, existing participants, and potential changes in staffing needs remain unclear due to the complexities arising from varying program outcomes (completion or not) and length-of-stay (LOS) for individuals with distinct characteristics ("features"). Understanding the interplay of these factors requires a sophisticated model that can capture long-term dynamics and facilitate counterfactual evaluations.

In this work, we collaboratively develop and implement a

DSS to predict the census in diversion programs and support staffing decisions with our community partner, an Illinois government agency that runs a statewide program allowing diversion from state prisons through community-based services. Our key contributions are as follows:

1. We develop a DSS that integrates various prediction modules within a simulation-based framework. This DSS accurately predicts census, enabling informed program sizing and staffing recommendations via counterfactuals.
2. Our approach combines ML prediction with a queueing model that effectively abstracts the diversion process's complexities. Unlike conventional ML approaches that focus solely on prediction, our method offers interpretable insights into program entry and leaving dynamics as well as the impact of different decisions on these dynamics in counterfactual scenarios.
3. We deploy a beta-version of a user-friendly web app, enabling program managers to visualize the census data by counties and race groups. We demonstrate two use cases for program sizing and staffing support when initiating new programs.

Throughout this process, we closely involve our community partners, seeking their input during problem identification, model development, and result interpretation, and iteratively refining the model design with a focus on interpretability. We are currently on the path to deploying this technology.

Background and Related Work

Figure 1 illustrates the simplified process flow for individuals in the incarceration-diversion program offered by our community partner, which targets eligible probationers. After risk assessment and screening, eligible and willing individuals are admitted. The case managers then determine the specific program completion requirements that make-up the in-program activities, such as attending substance use treatment programs and cognitive-behavioral therapy sessions. Successful completion results in a *Completed* outcome. However, individuals may recidivate (commit new crimes while in program), leading to their probation being *Revoked*, or they may be unable to finish the program due to various reasons, marked as *Not Completed*. Some individuals also leave the program with "unknown" reasons in the dataset, which we refer to as *Other* in the outcome labeling. For admitted individuals, the initial screening date indicates the start of the program and is considered as their *arrival date* to the program; the *termination date* represents individuals leaving the program, accounting for all types of program outcomes. The difference between the two dates gives the individual's length-of-stay in the program. A more detailed explanation of the process flow is available in the Online Supplmenet.

Many existing research on incarceration-diversion program has focused on outcome prediction and the identification of predictors for successful completion (Loeb, Waung, and Sheeran 2015; Verhaaff and Scott 2015; Loong et al. 2021). Studies have also emphasized the role of probation/-case officers, showing that reduced caseloads can significantly reduce recidivism rates (Burrell 2006; Jalbert et al. 2010; Jalbert and Rhodes 2012), underscoring the benefits of proper staffing and manageable caseloads in enhancing program effectiveness and supervision. In comparison, we focus on a process-level view of incarceration-diversion programs, which incorporates ML prediction into a queueing-based simulation model.

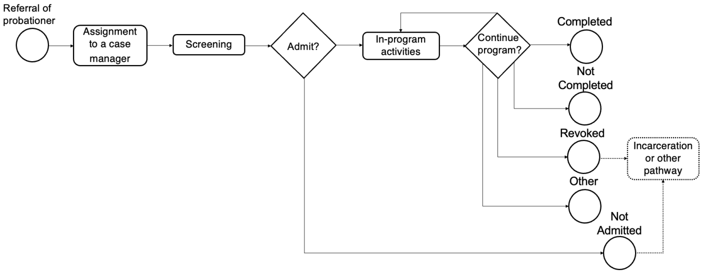


Figure 1: Incarceration-diversion program diagram.

Simulation and queueing models have been utilized in similar criminal justice contexts. Taxman and Pattavina (2013) advocate for simulation modeling to explore recidivism reduction strategies, showcasing applications such as a web tool for diversion program targeting and discrete-event simulations to estimate the impact of risk-and-need-responsivity principles. Usta and Wein (2015) use a queueing network simulation to demonstrate that offering split sentences to low-level felons optimizes the trade-off between recidivism risk and jail congestion. Attari et al. (2021) leverage a queueing simulation model with recidivism prediction based on logistic regression and estimated program effects to make admission decisions in reintegration programs. Zhang, Shi, and Ward (2022) employ a theoretical queueing model to explore fairness and efficiency in routing customers in a two-stage service system motivated by incarceration diversion. Master et al. (2018) analyze a jail network queueing model with a continuum of classes capturing recidivism risk, and propose a two-threshold control policy to optimize crime rate and mean jail population. However, none of these studies have investigated the impact of admission volume on program census using an ML-based outcome prediction model, or explored the trade-off between in-program revocation and long-term recidivism rate to determine optimal program size. Furthermore, the aforementioned works did not emphasize technology deployment, a key focus of our work.

Data and Descriptive Statistics

We obtained de-identified data from our community partner for incarceration-diversion programs in four Illinois counties: DuPage, Cook, Peoria, and Will. The dataset spans from February 2011 to April 2022 for DuPage, from May 2012 to March 2022 for Cook, from November 2011 to March 2022 for Will, and from September 2013 to March 2022 for Peoria. We retrieve the raw data from our partner and consolidate multiple datasets based on the client ID. The consolidated dataset comprises records of adult participants admitted to the program, and it includes essential information such as the arrival date, termination date, program outcomes, and individual features including race, gender, education, marriage, housing, risk assessment scores, prior crime history, and referral sources. Using the arrival and termination dates, we also calculated the historical census each month (calculation details to be discussed in the next section).

Table 1 presents descriptive statistics on arrival, LOS, and outcomes for each county and race group: White (W), African-American (A), Hispanic (H), and Other (O). Additional descriptive statistics, such as age and gender distribution, are in the Online Supplement (Tables 1-3). Through the descriptive analysis, we identify two critical components that our DSS should incorporate. First, different counties exhibit distinct racial compositions and program outcomes, implying that the program sizing and staffing needs may vary across counties (recall the earlier example of the Hispanic group). Thus, it is imperative for the DSS to capture these demographic differences effectively. Second, while the average LOS is relatively consistent across counties, it significantly varies across program outcomes. Not-Completed and Other outcomes are associated with shorter average LOS compared to Completed and Revoked outcomes. Notably, county and race are known factors upon individual entry, whereas the outcome is unknown at that time. Hence, our DSS must incorporate a prediction function for the outcome at the entry point, enabling it to not only predict completion rates or in-program revocation rates but also predict the program census to provide effective staffing decision support.

Table 1: General Statistics for DuPage (Du), Cook (Co), Will (Wi), and Peoria (Pe).

Cy	R	%	Arvl. mly.	LOS yrs.	Outcome LOS ($\mu[\sigma]$) and %			
			$\mu[\sigma]$	$\mu[\sigma]$	Com	Not	Rev	Ot
Du		682	5.2[3.6]	1.6[1.0]	1.6[0.7]	1.7[1]	1.7[1]	1.1[0.7]
11					47%	19%	21%	13%
-	W	61%	3.2[2.7]	1.6[1.0]	50%	19%	19%	12%
22	A	21%	1.1[1.2]	1.7[1.1]	42%	18%	22%	18%
	H	14%	0.7[0.9]	1.7[1.1]	38%	24%	23%	15%
	O	3%	0.2[0.4]	1.3[0.8]	50%	5%	25%	20%
Co		826	9.1[12.3]	1.5[0.8]	1.5[0.5]	1.5[0.9]	1.5[0.8]	1.1[0.6]
12					51%	9%	28%	12%
-	W	8%	0.7[1.3]	1.6[1.1]	54%	7%	30%	9%
22	A	82%	7.4[10.2]	1.5[0.8]	52%	9%	27%	12%
	H	8%	0.7[1.1]	1.5[1.0]	40%	11%	33%	16%
	O	3%	0.2[0.6]	1.5[1.1]	78%	0%	22%	0%
Wi		600	7.0[3.5]	1.5[0.9]	1.9[0.6]	1.1[0.9]	1.1[0.7]	0.4[0.3]
14					45%	20%	33%	2 %
-	W	58%	4.0[2.5]	1.4[0.9]	46%	18%	33%	3%
22	A	31%	2.2[1.6]	1.6[1.0]	39%	24%	37%	0%
	H	9%	0.6[0.8]	1.6[0.8]	54%	18%	25%	3%
	O	2%	0.1[0.4]	1.7[1.2]	43%	29%	14%	14%
Pe		389	4.2[2.9]	1.7[0.9]	2.3[0.5]	1.1[0.7]	1.2[1.0]	0.8[0]
13					48%	46.7%	5%	0.3%
-	W	31%	1.3[1.4]	1.7[0.9]	42%	51%	6%	1%
22	A	66%	2.8[2.2]	1.7[0.9]	51%	45%	4%	0%
	O	3%	0.1[0.4]	1.5[0.8]	50%	50%	0%	0%

Methods

We begin by providing an overview of our simulation-based DSS and then proceed to elaborate on the main components in detail. Figure 2 outlines our method pipeline. Our DSS contains two primary steps:

1. **Data extraction and prediction module building:** We start by extracting historical data and performing necessary pre-processing. Subsequently, we develop an ML outcome classifier based on individual features, enabling us to predict the program outcome in the simulation. We also construct empirical LOS distributions for each possible outcome for LOS prediction.
2. **Census prediction based on simulation:** The arrival generation in the simulation follows a trace-based approach, where monthly arrivals follow historical rates, and each new arrival's features are sampled from the corresponding month. For each arrival, the ML classifier predicts the outcome, and their LOS is then sampled from the outcome-based empirical distributions. The overall census is predicted using Algorithm 1, outlined below. The census by subgroups, e.g., race groups, can be calculated in similar ways.

Next, we elaborate on the outcome prediction module and discuss how we address a critical issue when incorporating them in census prediction: the right-censoring issue. We end this section by discussing the LOS sampling method.

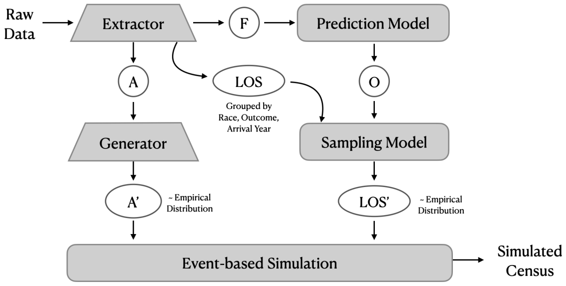


Figure 2: Overview of our framework and relevant parameters. Historical arrival date (**A**), clients' features (**F**), and LOS are extracted from the historical data through the extractor. These variables are used in "Generator" to generate new arrivals and in "Prediction Model" to predict program outcome (**O**); the latter will be used to sample LOS from the corresponding empirical distribution. The sampled LOS (**LOS'**) and generated new arrivals (**A'**) are used to generate census following Algorithm 1.

Outcome Prediction

The outcome prediction is a multi-classification problem with label $k \in \mathcal{K}$. We consider the following ML models as candidates to handle this multi-classification problem: logistic regression (LR) with Lasso (L1) regularization, decision tree (DT), gradient boosting tree (GBT), and multi-layer perceptron (MLP).

We perform data pre-processing and feature selection, which involves data transformation (log and square root), features grouping, outlier elimination, etc. These steps aim to enhance the model's understanding of data relationships and identify relevant features for better predictive capabilities. An automated data pre-processing pipeline was developed to streamline these essential steps. Hyper-parameter

Algorithm 1: Census Prediction Algorithm

Input: Outcome and LOS prediction models, and the number of arrivals in each month t, $\{A_t\}$.

Generate Arrivals: Sample arrival time for the ith individual who arrived in month ℓ as $a_{\ell,i}$, predicted outcome as $\hat{k}$, and predicted LOS as $\hat{L}_{\hat{k},i}$.

Generate Departures: The number of departures in month t, $D_t = \sum_{k\in\mathcal{K}} D_{k,t}$, for outcome $k \in \mathcal{K} = \{\text{complete, not complete, revoke, other}\}$, with

$$D_{k,t} = \sum_{\ell\leq t}\sum_{i=1}^{A_\ell} \mathbf{1}(t-1 < a_{\ell,i} + \hat{L}_{\hat{k},i} \leq t)\mathbf{1}(\hat{k}=k), \qquad (1)$$

and $\mathbf{1}(\cdot)$ is the indicator function.

Generate Census: The census for month t can be calculated recursively as

$$X_t = X_{t-1} + A_{t-1} - D_{t-1}. \qquad (2)$$

optimization frameworks (Optuna for GBT, Keras Tuner for MLP, GridSearchCV for LR and DT) were used to automate the tuning process with 200 trials to test hyper-parameter combinations. We have tuned the parameters such as number of hidden layers, activation functions, learning rate, where the tuning range of these parameters and the final selected values for each model are available in the Online Supplement. Stratified k-folds cross-validation was implemented to avoid overfitting.

Addressing Right-censoring

While incorporating outcome and LOS predictions into the queueing simulation, we encountered a significant challenge related to calibrating the census using historical data, primarily due to right-censoring. Specifically, a portion of individuals had missing termination dates when calculating LOS, with the percentage ranging from 11% to 38% across different counties. Upon investigation, we discovered that these individuals were primarily still in the program at the time of data cutoff, resulting in *right-censored* termination dates. Dismissing these individuals was not an option since they represented a substantial portion of the system's load, and ignoring them would lead to significant under-prediction of the census. Meanwhile, attempting to fill the missing LOS using data from those with termination dates (as in conventional missing data imputation methods) led to miscalibration issues. This was because the non-right-censored individuals are primarily those who were revoked or did not complete the program and had shorter LOS (Table 1).

To address the right-censoring issue, we employed a new two-stage sampling approach. Specifically, at each arrival instance, we flip a coin with the right censoring probability to predict if this individual will be right-censored. If yes, we assign an extended LOS value so as to make sure that this individual is still in-program by the cutoff time. Otherwise, the person is predicted as not right-censored and we sample LOS values from the corresponding empirical distributions (from those who were not right censored in the data). This two-stage sampling method turns out to be critical in calibrating our simulation prediction results to the observed data, as we demonstrate in the next section. Through this two-stage approach, we effectively address the limitations posed by the right-censoring issue and significantly enhances the simulation's accuracy in replicating the historical census data.

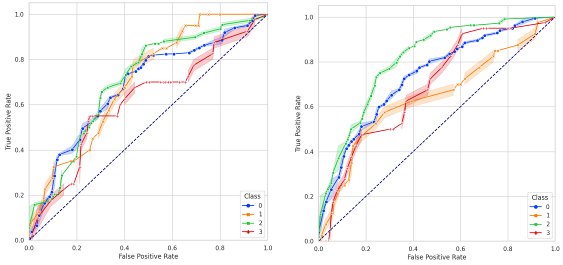
LOS Sampling Discussion

Recall that LOS is defined as the days between the arrival date and the termination date. Initially, we attempted to predict length of stay using various ML prediction models, including MLP, LASSO regression, regression tree, and XGB. However, all of them yielded unsatisfactory results with large Mean Squared Error (MSE). A closer examination revealed that most of our features are categorical, and even within the same combination of features, there were significant variations in LOS. We also tried to fit distributional models (e.g., assuming Gaussian distribution and learning the mean and variance as a function for different feature combinations) but this approach was also unsuccessful as there was no suitable distributional fit for the LOS data (Kolmogorov-Smirnov tests for typical distributions like normal, log-normal, or exponential all failed). Histograms of the LOS by counties demonstrate this challenge (see Online Supplement, Figure 2). Therefore, we resort to directly sampling from the empirical distribution based on the combination of county, year, race, and outcome, which are the most predictive features for LOS.

Results

Outcome Prediction

We implemented LR and DT using sklearn, GBT using LightGBM, and MLP using Keras. For model performance comparison, we choose the One-vs-Rest (OVR) ROC AUC score and weighted F1 score as metrics to compare their prediction performance, with LR serving as the baseline model. The OVR strategy compares each class against all the others simultaneously, making the ROC AUC score applicable for multiclass classifiers. We split the dataset into training and testing sets for evaluation, using a stratified strategy to maintain class balance.



(a) MLP

(b) GBT

Figure 3: OVR AUC results for each ML model. Class 0-3 corresponds to complete, not complete, revoke, other, respectively.

Table 2 shows the out-sample performance of each model after fine-tuning. Figure 3 further illustrates the four OVR ROC Curve for GBT and MLP. From both the table and

Table 2: Program outcome prediction results.

Model	ROC AUC score	Weighted F1 score
LR	0.590	0.389
DT	0.703	0.497
GBT	**0.718**	**0.526**
MLP	0.691	0.454

the plots we can can observe that GBT demonstrates the best performance on the testing set. Consequently, we selected GBT as the outcome predictor and integrated it into our framework. The feature importance plot from GBT is in the Online Supplement (Figure 1).

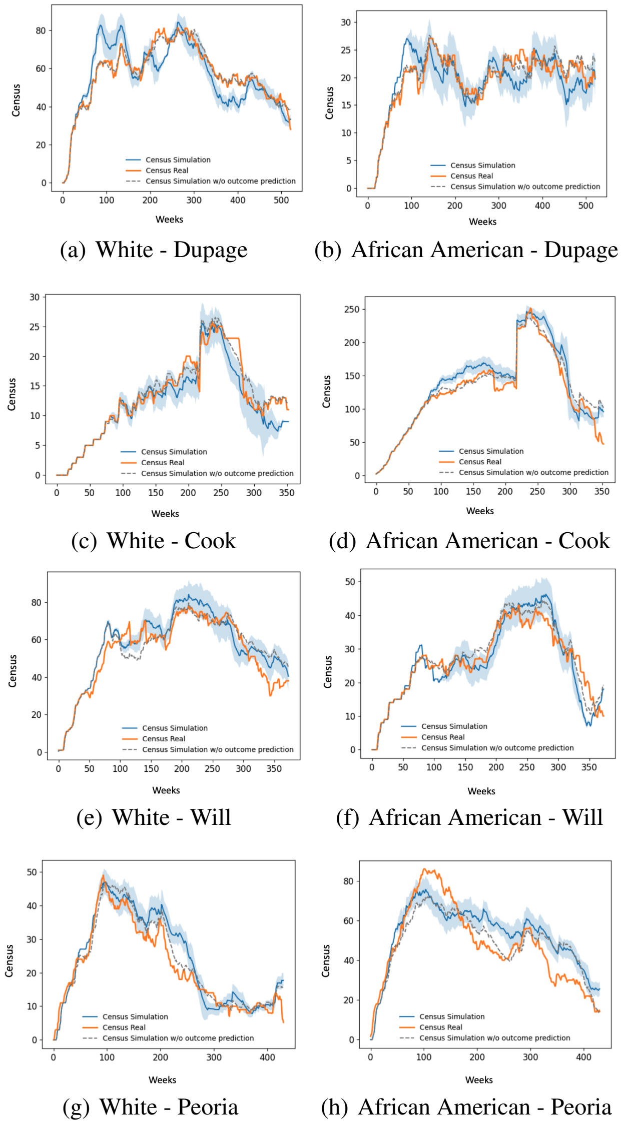
Census Prediction

Figure 4 shows the census prediction results for White and African American groups in each the four counties. We omit the comparison for Hispanic and Other as their typical census is below 10. Note that the simulation errors come from two sources: (a) the error from the outcome prediction; (b) the error from LOS sampling. Therefore, we show two simulation curves in each plot of Figure 4: one using the actual outcome (eliminating error from source (a)) and another using the predicted outcome. We can observe that the curve using the actual outcome closely aligns with the real (observed) census, suggesting that the error from LOS sampling is minimal and our approach to addressing right-censoring is effective. On the other hand, the curve using the predicted outcome (the final output) exhibits more deviation but remains reasonably close and captures the general trend. The mean absolute percentage errors (MAPE) when using the actual outcome are 3.01, 4.95, 5.51, and 15.78 for DuPage, Cook, Will, and Peoria counties, respectively; while the MAPE when using the prediction outcome are 9.40, 10.05, 7.03, 21.21, respectively. These findings suggest that the primary source of error lies in outcome prediction, especially for later years. This is expected since the prediction accuracy tends to be lower for years with fewer observations (similarly for counties with fewer participants such as Peoria); also see the last section for our ongoing efforts to improve outcome prediction accuracy. Nevertheless, the simulation outcome replicates the trends and patterns observed in the actual data, which validates the capability of our DSS to reasonably capture the real-world dynamics. This sets the stage for utilizing the model output to perform counterfactual analyses and informed decision-making. These simulation calibration results have been shared with the management team of the incarceration-diversion program and have garnered favorable feedback.

Counterfactual Analyses

We present two use cases for our DSS: program sizing for Will county and staffing estimation for launching similar programs in a new county.

For the first use case, as we previously discussed, there is a tradeoff between a larger program size with increased admissions, which may increase in-program revocations and raise potential community safety concerns, and a smaller program, which restricts the program's potential benefit on



(a) White - Dupage

(b) African American - Dupage

(c) White - Cook

(d) African American - Cook

(e) White - Will

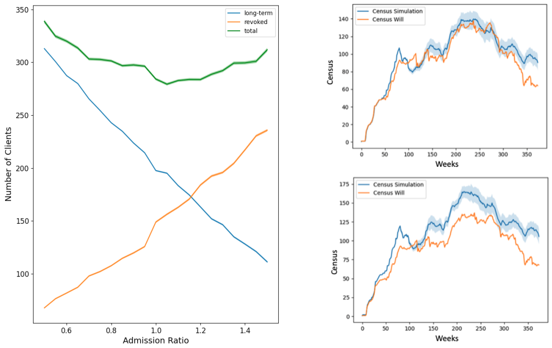
(f) African American - Will

(g) White - Peoria

(h) African American - Peoria

Figure 4: Census Prediction Results for DuPage, Cook, Will, and Peoria. Shaded areas are 95% prediction interval for the final output using predicted outcomes.

a wider range of individuals. To illustrate, Figure 5(a) depicts the relationship between the proportion of individuals admitted to the program and the following: (i) in-program revocation rates, predicted by the outcome prediction model; (ii) long-term recidivism rates, with current admission rate as the baseline (1.0) and each additional admitted individual leading to a 30% reduction in recidivism mentioned in the introduction; (iii) the weighted sum of (i) and (ii) with weights 0.8 and 1.2, with more emphasis on the latter as it represents the long-term societal benefits. We can observe that as admissions increase, in-program revocations in (i) also increase, while long-term recidivism in (ii) decreases. Consequently, the sum yields a U-shaped curve, suggesting that the optimal program size is slightly larger than the existing program size (baseline 1.0). This finding supports the expansion of admission criteria that our community partner is currently considering.



(a) Program sizing (b) Census change

Figure 5: Counterfactual analysis: left plot for program sizing, and right plot for census prediction for staffing support (top one for Will, and bottom one for new county with larger Hispanic population).

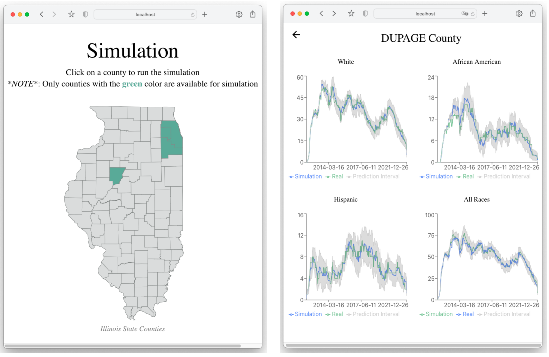
In the second use case, we explore the scenario of establishing a new program with demographics and admission criteria akin to Will County, but with a higher proportion of Hispanic participants and fewer White and African-American individuals. The resulting census, as shown in the bottom panel of Figure 5(b), is observed to be higher than that in Will. This is attributed to the fact that Hispanics exhibit higher completion rates, leading to longer LOS and hence, higher census. This analysis emphasizes the importance of ensuring adequate case manager resources to support Hispanic individuals, not only due to the increased arrival rates but also the elevated census resulting from the longer LOS.

Ongoing Deployment

To facilitate the adoption and deployment of our DSS for our community partner, we have developed a user-friendly web-based platform; see Figure 6 for screenshots. The platform consists of a web interface created with React and Tailwind CSS, as well as a back-end using Flask to run the ML model and simulation. This intuitive web-based platform enables users to interact with the DSS seamlessly. The homepage displays a map of Illinois, allowing users to select a specific county (highlighted in green for the current four counties in Figure 6(a)) and initiate the simulation. Once the simulation is complete, an interactive plot displaying the census prediction results is generated (implemented with using D3.js and rCharts) to provide an intuitive understanding of the outcomes. See sample results in Figure 6(b).

One significant challenge we faced during implementation was optimizing response and simulation times to provide results within seconds. Lengthy waiting times for results to be shown could deter users from using the app. To address this, we first improved the speed of the ML prediction module by preloading the model and warm-start. Then we optimized the data-transfer in user workflows. By implementing a Persisting Redux state to store fetched data in local storage, we reduced both the back-and-forth communication between client and server. Moreover, we engineered the parallel data fetching function to reduce client-server waterfalls and the total time to load data. We also processed and aggregated data on the server-side, offloading the processing burden from the client-side.These changes significantly reduced the average run time from 8 minutes to 15 seconds.

After multiple rounds of testing, we deployed the webapp to Google Cloud Platform using Ray, a distributed compute framework to scale AI application to production. We supported multi-node severing using Ray Serve, and built auto resource configuration function to auto scale-up or down according to the resources requested by application using Ray Jobs and Ray Cluster. We have already presented the webapp to the management team of our community partner, receiving strong support, and are actively incorporating their feedback to improve the design iteratively.


Simulation

Click on a county to run the simulation

NOTE: Only counties with the green color are available for simulation

(a) Homepage (b) Sample result page

Figure 6: Demo of the web-based app for ongoing deployment of our DSS.

While the current version covers specific counties in Illinois, our plan is to progressively extend its coverage to include more counties and jurisdictions, focusing on generalization and scalability in refining our DSS. Other future deployment efforts include (i) refining the DSS through more advanced ML techniques (like transfer learning and ensemble methods) to improve the program outcome prediction, especially for years and counties with small samples, and collaborating with our community partner to identify/collect more predictive features such as socio-determinants; (ii) allows users to upload their own data into the web app, providing the flexibility to run simulations and counterfactual to increase the applicability of our DSS to a broader range of users and settings. On the implementation side, we have adopted a tiered approach, starting with leadership buy-in (completed first phase) and actively gathering feedback from end-users (ongoing second phase). We understand the potential negative effects and are committed to continuous improvement, including assessing the algorithm's real-world impact, evaluating and addressing any unintended consequences (e.g., bias), and ensuring practicality within various budgets and political constraints. By effectively bridging AI technology and decision-making, our DSS demonstrates great potential to address a significant societal problem. With a clear path to deployment, our tool holds the promise of positively impacting society by fostering safer communities through more effective diversion programs.

Disclaimer

Views expressed in this work are those of the authors and do not reflect the official policy or position of the State of Illinois or the Illinois Criminal Justice Information Authority.

We commit to an author presenting the paper in-person at AAAI-24. If exceptional circumstances arise that prevent the authors from presenting the paper in person, we will notify AAAI immediately. We acknowledge that if we do not notify AAAI about exceptional circumstances, then the paper will not be included in the proceedings.

References

Attari, I.; Crain, P. A.; Shi, P.; Helm, J. E.; and Adams, N. 2021. A simulation analysis of analytics-driven community-based re-integration programs. In *Proceedings of the Winter Simulation Conference*, 1–12.

Berk, R. 2017. An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *Journal of Experimental Criminology*, 13: 193–216.

Burrell, B. 2006. Caseload standards for probation and parole (September 2006). *American Probation and Parole Association (APPA)*. Accessed: 2023-07-28.

DeMichele, M. T. 2007. Probation and parole's growing caseloads and workload allocation: Strategies for managerial decision making. *The American Probation & Parole Association (APPA)*.

Jalbert, S. K.; and Rhodes, W. 2012. Reduced caseloads improve probation outcomes. *Journal of Crime and Justice*, 35(2): 221–238.

Jalbert, S. K.; Rhodes, W.; Flygare, C.; and Kane, M. 2010. Testing probation outcomes in an evidence-based practice setting: Reduced caseload size and intensive supervision effectiveness. *Journal of Offender Rehabilitation*, 49(4): 233–253.

Latessa, E. J.; Johnson, S. L.; and Koetzle, D. 2020. *What works (and doesn't) in reducing recidivism*. Anderson.

Leipold, A. D. 2005. Recidivism, Incapacitation, and criminal sentencing policy. *U. St. Thomas LJ*, 3: 536.

Lin, Z. J.; Jung, J.; Goel, S.; and Skeem, J. 2020. The limits of human predictions of recidivism. *Science advances*, 6(7): eaaz0652.

Loeb, R. C.; Waung, M.; and Sheeran, M. 2015. Individual and familial variables for predicting successful completion of a juvenile justice diversion program. *Journal of Offender Rehabilitation*, 54(3): 212–237.

Loong, D.; Barnsley, J.; Aubry, T.; and Dewa, C. S. 2021. Individual factors associated with recidivism among mental health court program clients. *International Journal of Law and Psychiatry*, 74: 101651.

Master, N.; Reiman, M. I.; Wang, C.; and Wein, L. M. 2018. A continuous-class queueing model with proportional hazards-based routing. *Available at SSRN 3390476*.

McNiel, D. E.; and Binder, R. L. 2007. Effectiveness of a mental health court in reducing criminal recidivism and violence. *American Journal of Psychiatry*, 164(9): 1395–1403.

Mock, L. 2022. A profile of Latinx participants in Adult Redeploy Illinois. *Illinois Criminal Justice Information Authority*. https://icjia.illinois.gov/researchhub/articles/a-profile-of-latinx-participants-in-adult-redeploy-illinois.

Peters, R. H.; and Murrin, M. R. 2000. Effectiveness of treatment-based drug courts in reducing criminal recidivism. *Criminal justice and behavior*, 27(1): 72–96.

Taxman, F. S.; and Pattavina, A. 2013. *Simulation strategies to reduce recidivism*. Springer.

Usta, M.; and Wein, L. M. 2015. Assessing risk-based policies for pretrial release and split sentencing in los angeles county jails. *PloS one*, 10(12): e0144967.

Verhaaff, A.; and Scott, H. 2015. Individual factors predicting mental health court diversion outcome. *Research on Social Work Practice*, 25(2): 213–228.

Zhang, Z.; Shi, P.; and Ward, A. R. 2022. Routing for fairness and efficiency in a queueing model with reentry and continuous customer classes. In *2022 American Control Conference (ACC)*, 4882–4887. IEEE.

Online Supplement

1 Detailed Program Process Flow

Figure 1 provides a more detailed version of the one in the main paper. The incarceration-diversion process begins by an eligible probationer being referred to the program, which can be initiated by a judge, probation officer, or public defender. Upon referral the case is assigned to a expert case manager who conducts a comprehensive screening of the individual's demographic information, offense details, criminal history, drug use, and other pertinent factors. Additionally, a risk assessment score is assigned during this screening process. Our current partner employs the Level of Service Inventory - Revised (LSI-R) score.

Based on the screening results, if the individual is deemed suitable for program admission and expresses willingness to participate, the diversion program starts; otherwise, the case may follow alternative pathways, potentially leading to incarceration or going through other correctional routes. For an admitted individual, the case manager then determines the specific program completion requirements that make-up the in-program activities, such as attending substance use treatment programs, cognitive-behavioral therapy sessions, and vocational training. The case manager regularly asseses the individual's progress within the program, which includes risk assessments, drug tests, and more. These assessments guide decisions regarding the individual's program status – whether they should continue in the program, if any adjustments to the requirements are necessary, or if program termination is warranted.

An individual who successfully fulfills all program requirements receives a *Completed* outcome. However, in some cases, individuals may recidivate, committing new crimes while in the program, resulting in their participation in the incarceration-diversion program being *Revoked*. Consequently, they are terminated from the program. Therefore, in the main paper "in-program revocation" refers to the process of removing an individual from an incarceration-diversion program and transferring them back to traditional incarceration or another correctional settings due to non-compliance or other program-related reasons. Additionally, individuals may fail to complete the program for various reasons, such as being transferred to another county or going missing. These cases are labeled as *Not Completed* in the program outcome. Some individuals may exit the program for "unknown" reasons, which we classify as *Other* during the machine-learning training process.

2 More Descriptive Analysis

In Tables 1 and 2 we provide descriptive statistics for other features in our prediction models. Specifically, Table 2 includes the categorical variables in the data set. These variables are for each individual: (i) Risk: their recidivism risk scaled into 3 levels. (ii) AdOffense: name of the offense committed that lead to them being admitted into the program. (iii) OffenseClass: class of the offense. (iv) Pdrug: Primary drug that the individual stated at admission to consume. (v) ReferralReason: reason for referral to the program. (vi) WhoReferred: Role of the person who referred the individual to the program. (vii) Gender. (viii) EmploymentS: employment type at admission . (ix) MaritalS: marital status at admission. (x) HousingS: states the living situation at admission. (xi) MedicaidSt: Medicaid enrollment status at admission. (xii) UniqueAgents: Number of unique agents that saw the person during his LOS. (xiii) FinalProgramPhase: Final program level that the person achieved. (xiv) RewardedBehavior: indicator to whether the persons received any reward during his LOS in the program, and (xv) Sanctions: indicator to whether the person received any sanction during his LOS in the program. Table 1 provides mean and standard deviation for continuous variables in the dataset. These variables are (i) AgeAtEnroll: Age of the individual at enrollment. (ii) CriminalHistScore: ranges from 0 to 30 and its the estimated score of the individual's past criminal history. (iii) and (iv) AvgMVistis (AvgReqMVisits) Average number of actual (required) monthly visits to the individual during their LOS. (v) and (vi) TotalMVistis (TotalReqMVisits): Total number of actual (required) visits to the individual during their LOS.

Table 1: Continuous Covariates Summary Statistics.

Variable	County			
	DuPage $\mu(\sigma)$	Cook $\mu(\sigma)$	Will $\mu(\sigma)$	Peoria $\mu(\sigma)$
AgeAtEnroll	30.5 (9.9)	41.8 (13.1)	33.6 (9.2)	35.5 (11.0)
CriminalHistScore	5.1 (1.6)	17.9 (11.3)	5.2 (1.7)	5.6 (1.4)
AvgMVisits	1.7 (1.0)	0.5 (0.5)	1.5 (0.8)	4.3 (1.9)
TotalMVisits	39.8 (30.5)	11.2 (11.0)	29.6 (20.0)	93.7 (56.2)
AvgReqMVisits	1.6 (1.2)	–	0.0 (0.0)	–
TotalReqMVisits	38.2 (35.3)	–	0.3 (0.6)	–

3 Details of the Outcome Prediction

Pre-processing. We apply the following pre-processing and feature selection process: (*i*) Data transformation: log transformation and square root transformation, were applied to enable the model to better understand the relative relationships between data points. (*ii*) Features grouping and outlier elimination: To enhance model performance, we visualized the distribution of unique values for each feature. Features with less than 3 unique values were dropped. For categorical features, we eliminated unique values that represented less than 10 percent of all unique values. For numerical features, we removed outliers using the Interquartile Range (IQR) method. (*iii*) Feature selection: we filtered features based on feature importance score for GBT model. We develop an automated data pre-processing pipeline to perform these three steps.

Hyper-parameters tuning. After data pre-processing, we split the processed dataset into a training set and a testing set in a 9:1 ratio. We applied a stratified splitting strategy to ensure that each class appears in both the training and testing sets in the same proportion. All models are trained using

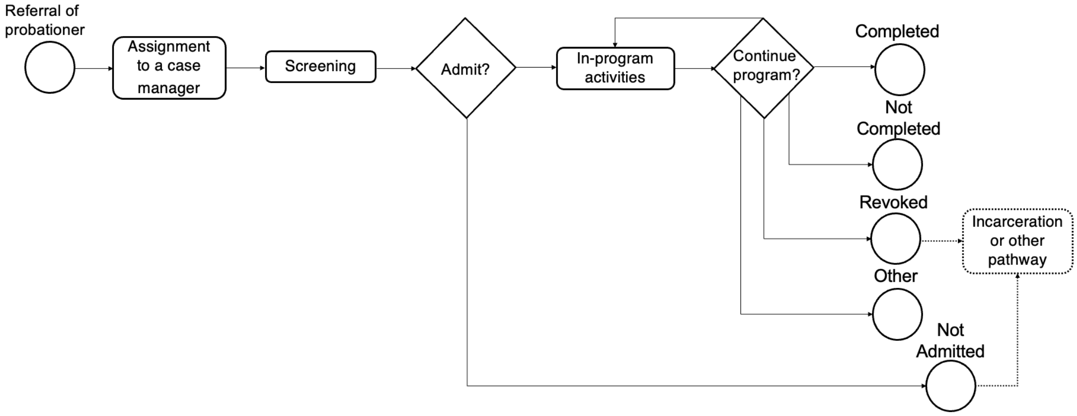


Figure 1: Incarceration-diversion program diagram.

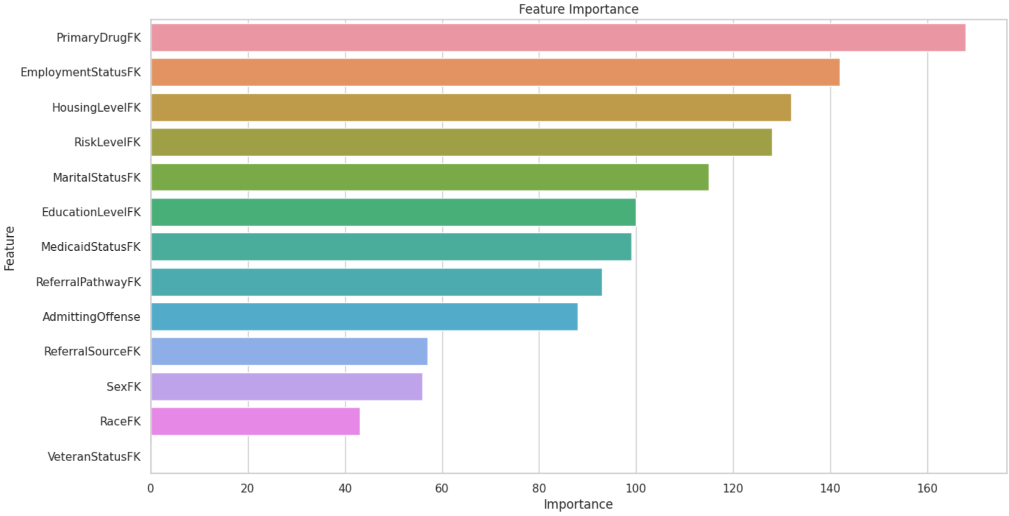


Figure 2: Feature importance plot.

the training set and tested on the same testing set. For hyper-parameter tuning, we implemented stratified k-folds cross-validation to ensure the models were not overfitting, where each fold must contain the same percentage of samples of each class as the complete set. We tested different values of k with best performance achieved at $k = 5$ for all models. We fine tune hyper-parameters for each ML model to improve the model performance. We automated this tuning process using hyper-parameter optimization frameworks: in our implementation, we used Optuna to optimize hyper-parameters for GBT, Keras Tunner for MLP, and GridSearchCV for LR and DT. We created trials to test different combinations of hyper-parameters, and selected the trial with the best model performance as the final set of hyper-parameters. Specifically, for the MLP, we first tested several possible configurations to narrow down the search space. We then evaluated various hypter parameters (including the number of nodes, layers, optimizers, etc.) to identify the top 50 structures with the best performance using Random Search. We further tuned the learning rates, number of epochs for each of these structures to gain the best out-of-sample performance. We followed the same procedure for other models. Tables 3 to 6 demonstrate parameters of each model, along with tested value and best value.

Table 2: Categorical Covariates Summary Statistics (N/A or Other Categories are Omitted).

Variable	Categories	County			
		DuPage	Cook	Will	Peoria
Risk	Highest	24.3	32.0	2.3	1.0
	High	60.7	26.2	35.1	24.7
	Medium	11.0	15.6	42.1	47.0
AdOffense	Drugs	43.0	67.8	31.7	37.0
	Property	31.1	17.6	52.5	46.3
	DUI	11.1	2.3	3.8	1.0
OffenseClass	Class 4	42.5	–	11.5	20.6
	Class 3	13.5	–	5.7	5.7
	Class 2	16.0	–	5.7	5.1
Pdrug	Heroin	27.0	43.6	32.3	9.5
	THC	18.6	18.5	17.5	21.6
	Coc.Crack	7.8	10.9	21.0	11.6
ReferralReason	Tech Violation	31.2	0.0	12.8	0.0
	3/4 Felon	20.5	70.5	59.2	80.0
	1/2 Felon	9.8	16.5	23.7	14.7
WhoReferred	Prob Officer	64.7	97.3	1.8	0.0
	Judge	32.0	1.3	0.7	91.3
	Pub. Defender	0.6	0.0	75.3	2.8
Gender	Female	25.2	21.3	21.7	19.8
	Male	74.8	77.5	78.2	80.0
EmplymntS	Full Time	49.7	85.7	38.2	6.7
	None	32.3	4.8	59.2	92.0
	Part Time	18.0	9.4	2.7	1.3
MaritalS	Single	86.4	85.6	15.0	22.9
	Married	5.9	7.1	1.8	5.7
	Divorced	4.7	2.3	0.2	1.8
EducationS	HighSchool	40.3	37.2	34.3	13.6
	No HighSchool	32.6	52.4	10.8	12.3
	Some College or Graduated	19.4	3.5	11.8	4.4
HousingS	Friend or Family	62.3	27.9	6.2	17.7
	Own/Rent	29.0	15.5	2.7	11.1
	No Home Reported	5.9	23.9	16.5	70.2
MedicaidS	Yes	23.8	48.4	8.3	3.3
UniqueAgents	4	11.6	2.2	8.6	–
	3	27.9	31.9	22.3	2.3
	2	60.6	65.9	69.1	97.7
FinalProgPhase	Level 3/4	11.1	15.7	32.3	0.3
	Level 1/2	56.5	14.4	22.7	3.1
	Level 0	2.9	35.5	7.0	27.0
RewardedBehv	Yes	4.0	29.1	2.5	1.5
Sanctions	Yes	91.8	99.3	89.8	41.1

Table 3: Gradient Boosting Tree Hyper-parameter Tuning Results.

Parameter Name	Parameter Value	
	Tested	Best
learning_rate	[0.01, 0.5]	0.44
max_depth	[5, 20]	10
lambda_l1	[1e-8, 10]	0.106
lambda_l2	[1e-8, 10]	1.031
num_leaves	[10, 60]	42
bagging_fraction	[0.6, 1.0]	0.612
feature_fraction	[0.6, 1.0]	0.701

Table 4: Decision Tree Hyper-parameter Tuning Results.

Parameter Name	Parameter Value	
	Tested	Best
criterion	gini, entropy	gini
max_depth	[1, 15]	6
min_samples_split	[1e-8, 10]	9
num_leaves	[10, 60]	42
min_samples_leaf	[0.6, 1.0]	2

Table 5: MLP Hyper-parameter Tuning Results.

Parameter Name	Parameter Value	
	Tested	Best
dense activation function	relu, tanh, sigmoid	relu
number of hidden layers	[1,5]	2
num of neurons per hidden laye r	[12, 512]	45, 20
layer activation function	relu, tanh, sigmoid	relu, sigmoid
dropout rate	[0.0, 1.0]	0.1, 0.1
optimizer	adam, sgd, rmsprop	adam
learning rate	[0.001, 0.1]	0.00399

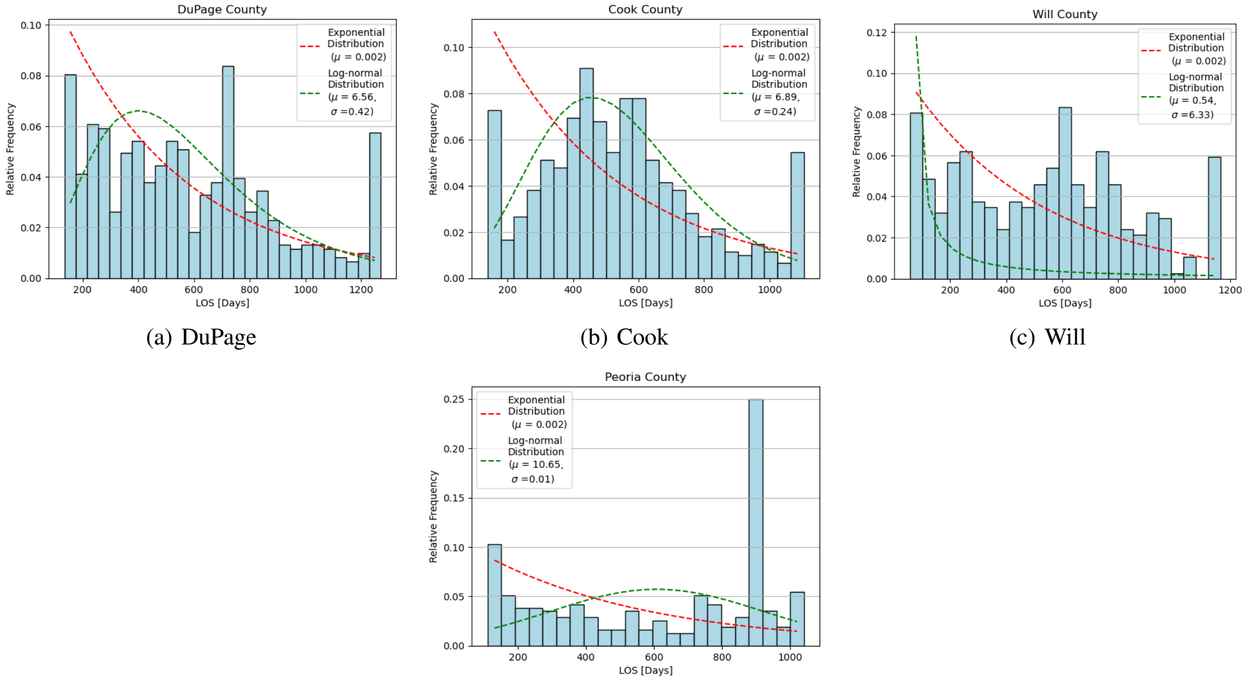
Table 6: Logistic Regression Hyper-parameter Tuning Results.

Parameter Name	Parameter Value	
	Tested	Best
penalty	l1, l2, elasticnet	l1
l1_ratio	[0, 1]	(Not Used)
C (Inverse of regularization strength)	[0.0, 1.0]	0.018
solver	liblinear,lbfgs,saga	saga

4 Figures for Arrivals and LOS

Table 7: LOS by Outcome Summary Statistics

County	Race	LOS yrs. by Outcome Mean (SD)		
	grp.	Com.	Not Com.	Revok.
DuPage	All	1.6 (0.7)	1.4 (0.9)	1.7 (1.0)
(2011-	W	1.7 (0.7)	1.3 (0.9)	1.7 (0.9)
2022)	AA	1.6 (0.8)	1.6 (1.0)	1.8 (1.1)
	H	1.5 (0.7)	1.6 (1.0)	1.7 (1.0)
	OT	1.3 (0.6)	0.5 (0.2)	1.9 (1.1)
Cook	All	1.5 (0.5)	1.3 (0.8)	1.5 (0.8)
(2012-	W	1.8 (0.7)	1.3 (0.4)	1.1 (0.8)
2022)	AA	1.5 (0.5)	1.3 (1.0)	1.5 (0.8)
	H	1.6 (0.4)	1.1 (1.0)	1.6 (0.8)
	OT	1.6 (0.7)	–	0.6 (0.3)
Will	All	1.9 (0.6)	1.0 (0.9)	1.1 (0.8)
(2014-	W	1.9 (0.7)	0.9 (0.8)	1.1 (0.7)
2022)	AA	2.0 (0.9)	1.3 (0.9)	1.2 (0.9)
	H	2.1 (0.7)	0.6(0.7)	1.3 (0.7)
	OT	2.2 (0.8)	1.0(1.4)	2.4 (–)
Peoria	All	2.3 (0.5)	1.1 (0.7)	1.2 (1.0)
(2013-	W	2.2 (0.6)	1.3 (0.8)	1.1 (1.0)
2022)	AA	2.3 (0.5)	1.1 (0.7)	1.3 (1.1)
	OT	2.3 (0.5)	0.8 (0.3)	–



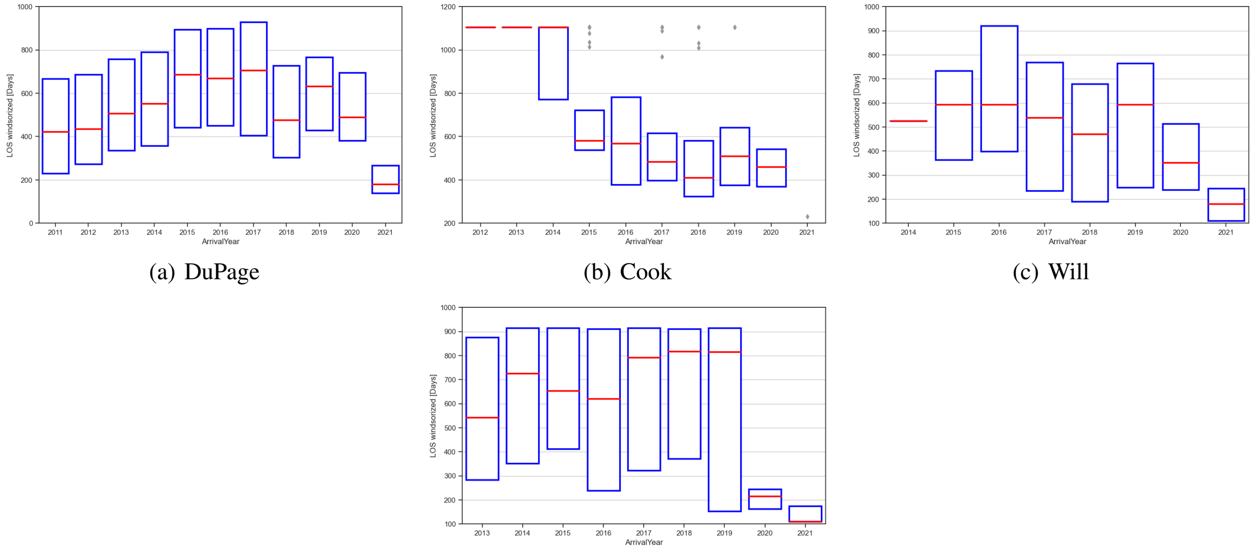
(a) DuPage

(b) Cook

(c) Will

(d) Peoria

Figure 3: Length of Stay Distributions (Complete Data, Windsorized)



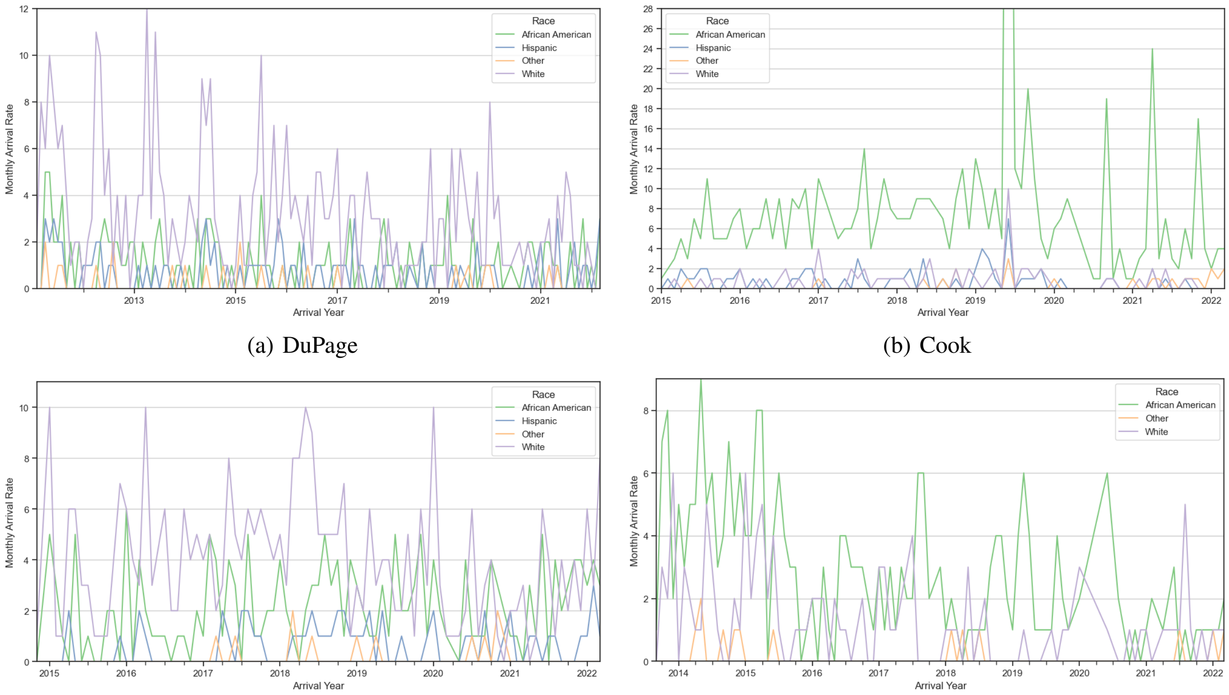
(a) DuPage

(b) Cook

(c) Will

(d) Peoria

Figure 4: Length of Stay per Year (Complete Data, Windsorized)



(a) DuPage

(b) Cook

(c) Will

(d) Peoria

Figure 5: Monthly Arrival Rate by Race