LightGBM is a gradient boosting framework that uses tree-based learning algorithms. It is designed to be distributed and efficient with the following advantages: Faster training speed and higher efficiency, Lower memory usage, and Better accuracy, Support of parallel, distributed, and GPU learning, and Capable of handling large-scale data. (4)

Catboost

CatBoost is an algorithm for gradient boosting on decision trees. It is developed by Yandex researchers and engineers, and is used for search, recommendation systems, personal assistant, self-driving cars, weather prediction and many other tasks at Yandex and in other companies, including CERN, Cloudflare, Careem taxi. It is in open-source and can be used by anyone. (5)

Model Building

In this section, I will use the prediction of the third year as a case study to explain how I approached this project. So, the outcome variable is the recidivism of the respondent after three years post-incarceration. I built several machine learning models and analyzed their results.

I start with the most basic, a logistic regression, which predicts the likelihood of recidivism. Logistic regression would serve as our baseline. Following that, I have performed Random Forest Classifier, Xgboost, LightGBM, and Catboost algorithms. To evaluate training set performance, I have implemented Stratified K-fold Cross-Validation Method. This technique is a variation of KFold that returns stratified folds. Since there are many categorical variables in our data, the folds preserve the percentage of samples for each categorical variable.

To boost the performance of the algorithms, I have implemented a hyperparameter tuning exercise. I applied the grid search method, which is a process that searches exhaustively through a manually specified subset of the hyperparameter space of the given algorithm. I used the "pruning" technique to stop training earlier when the learning curve was much worse