Analysis and Classification of Crimes in Chicago

Ye Zhou

September 22, 2017

1 Definition

1.1 Project Overview

Historically, Chicago saw a major rise in violent crime starting in the later 1960s. More recently, the crime situation in Chicago is getting even worse. Last year, Chicago has experienced a recent spike in homicides of 762 people, an increase of 58 percent over 2015. Compared with other largest cities in United States, Chicago has a significantly higher murder rate than New York or Los Angeles. There is no denying the fact that crimes have become a severe social concern for Chicago as prosperous communities in the long term.

1.2 Problem Statement

This project is aimed to analyze Chicago crimes in 2015 and 2016 and make prediction of crime category based on time and position information. The raw data from Chicago Data Portal is pre-processed to features (year, month, weekday, hour, location description, domestic, beat, district, ward, community area, latitude and longitude) and target variable (crime type). 20% of the data is randomly sampled and then split into train and test data sets (6:4). After an exploration in the visualization of the crime data, the logistic regression (benchmark model) is applied as a Benchmark model. At the end, a XGboost model is trained on train data set and hype-parameters are tuned using k-fold cross-validation. The performance of two models are evaluated and compared in the means of log loss function and accuracy on the test data set.

1.3 Metrics

This project use the accuracy, log loss, confusion matrix and classification report, from `sklearn.metrics` package, to train models and evaluate the performance. These are two metrics commonly used for classification problems.

- Log loss, is also known as logistic loss or cross-entropy loss, used in probabilistic classifiers, which is define as

$$-\log P(y_t/y_p) = -(y_t \log(y_p) + (1 - y_t) \log(1 - y_p)) \tag{1}$$

  where y_t is the true label in $\{0, 1\}$ and y_p is estimated probability that $y_t = 1$.

- Accuracy is defined by

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i) \tag{2}$$

  where y_i is the true label and $\hat{y}_i$ is the predicted class label for the ith observation. And $I(y_i \neq \hat{y}_i)$ is an indicator variable, denoting if the observation i is correctly classified.

- Precision is defined as the ratio tp / (tp + fp) where tp is the number of true positives and fp the number of false positives. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative.

- Recall is the ratio tp / (tp + fn) where tp is the number of true positives and fn the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples.

- F-beta score is the weighted harmonic mean of precision and recall.

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3}$$

2 Analysis

2.1 Data Exploration

The data is imported from the Chicago Data Portal website [4], which is a collection of city data related not only to crimes, but also to education, transportation, health, and so on.

The data set has 21 columns:

- ID - Unique identifier for the record.

- Case Number - The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.

- Date - Date when the incident occurred. this is sometimes a best estimate.

- Block - The partially redacted address where the incident occurred, placing it on the same block as the actual address.

- IUCR - The Illinois Unifrom Crime Reporting code. This is directly linked to the Primary Type and Description.

- Primary Type - The primary description of the IUCR code. It is target variable to predict in the project. The distribution of crime types is shown in Figure 1.

- Description - The secondary description of the IUCR code, a subcategory of the primary description.

- Location Description - Description of the location where the incident occurred.

- Arrest - Indicates whether an arrest was made.

- Domestic - Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.
- Beat - Indicates the beat where the incident occurred. A beat is the smallest police geographic area each beat has a dedicated police beat car. Three to five beats make up a police sector, and three sectors make up a police district. The Chicago Police Department has 22 police districts.
- District - Indicates the police district where the incident occurred.
- Ward - The ward (City Council district) where the incident occurred.
- Community Area - Indicates the community area where the incident occurred. Chicago has 77 community areas.
- FBI Code - Indicates the crime classification as outlined in the FBIs National Incident-Based Reporting System (NIBRS).
- X Coordinate - The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.
- Y Coordinate - The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.
- Year - Year the incident occurred.
- Updated On - Date and time the record was last updated.
- Latitude - The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.
- Longitude - The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.
- Location - The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block.

Entries with any missing data are discarded directly. In total, 510,070 non-null observations are imported for data visualization and 20% of which are randomly sampled for model training and testing. All the date values are between 2015 and 2016 and positional data are in Chicago, thus no outliers or abnormalities are detected in data set.

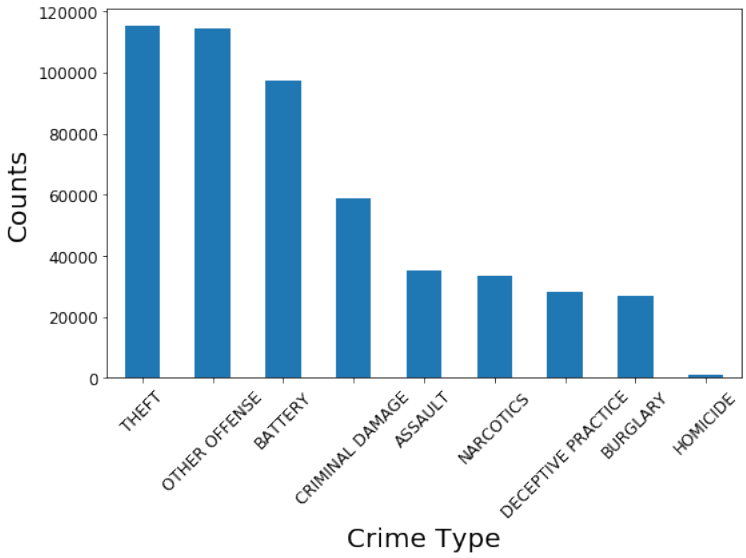

Figure 1: Total crimes in Chicago in 2015 and 2016

2.2 Exploratory Visualization

The features related to time are 'Year', 'Month' and 'Hour', extracted from the original feature 'Date'. As shown in figure 2a, in the past two years, Chicago experienced a high spike of crimes from May to October and has the lowest crime commits in February. Figure 2b represents the data, relying on the principal component analysis, where the first principal component corresponds to the direction with the largest variation. The biplot represents the axes of crime types and data point on the first two principal components (PCs). The first PC shows consistently that there are generally more crimes from May to October. The second PC indicates that BATTERY and NARCOTICS have a higher frequency in May and June, while THEFT occurs more from September to October. The biplot of the crime counts by hour and crime types is shown in Figure 2c. The first PC shows generally more crimes from noon to midnight. The second principal component shows that BATTERY and OTHER OFFENSE have a higher frequency at night (from 19 to 23), while THEFT and DECEPTIVE PRACTICE occurs more in the afternoon (from 12 to 15). Figure 2d shows a comparison of crimes between 2015 and 2016 for different crime types. A more detailed table of crime annual increase rate are shown in Table **??**, which shows that the count of Homicide crime increases by 60.60% from 2015 to 2016, while that of Narcotics is reduced by 53%.

The crime density on Chicago map is shown in Figure 3a, which shows that downtown area in Chicago has the highest crime density. Figure 3b shows the standardized crime count (divide by the maximum value among all communities) in each community for the past two years. The community shapefile is imported from the Chicago Data Portal website. Austin, Near North Side, Near West Side, South Shore are communities with standardized value greater than 0.5. Position related features are Location Description, Domestic, Beat, District, Ward, Community Area, Latitude, Longitude. The crime counts in different location for different types are shown in Figure 4a. The crime dependence on the Domestic feature

Table 1: Annual crime increase rate from 2015 to 2016 in Chicago

Primary Type	**2015**	**2016**	**Increase rate**
ASSAULT	16945	18076	6.67
BATTERY	48578	48666	0.18
BURGLARY	13084	13619	4.09
CRIMINAL DAMAGE	28527	30185	5.81
DECEPTIVE PRACTICE	14577	13507	-7.34
HOMICIDE	467	750	60.60
NARCOTICS	22839	10662	-53.32
OTHER OFFENSE	57744	56690	-1.83
THEFT	56827	58327	2.64

is shown in Figure 4c. Beat, District, Ward and Community Area are different ways of Chicago city district assignment. In this report, Community is used as an example for analysis. From Figure 4b, the second PC shows that BATTERY and NARCOTICS have a higher frequency in AUSTIN, while THEFT occurs more near Chicago downtown (LOOP and NEAR NORTH SIDE). A more careful look at the district distribution of different crime types are shown in Figure 5. It shows that Austin is the community with high crime rate of BATTERY, CRIMINAL DAMAGE, ASSAULT, NARCOTICS and OTHER OFFENSE; THEFT and DECEPTIVE PRACTICE occurred more near downtown; and BURGLARY has high outbreak in southern (e.g. South Shore) and western communities.

2.3 Algorithms and Techniques

In this section, you should focus on describing how your algorithms work, in terms of how they train and predict. What's the procedure behind the scenes? What's the theory behind it? This discussion doesn't have to be mathematically rigorous, unless that suits you better; essentially, however you can best explain these concepts, this is how you should approach it As an example, if you were describing the SVM, you might discuss such concepts as "maximizing the margin" and the "kernel trick"

In this project, the tree-based Gradient Boosting will be used with the package `xgboost`, which is the abbreviation of 'extreme gradient boost'. This algorithm combines sequential weak learners into one strong powerful learner. Each weak learner in the model corresponds to a shallow classification decision tree, where the most commonly occurring class is assigned at each leaf node. The classification tree is grown based on two common measures of purity. The Gini index is defined by

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{4}$$

a measure of total variance across the K classes. An alternative to Gini index is cross-entropy,

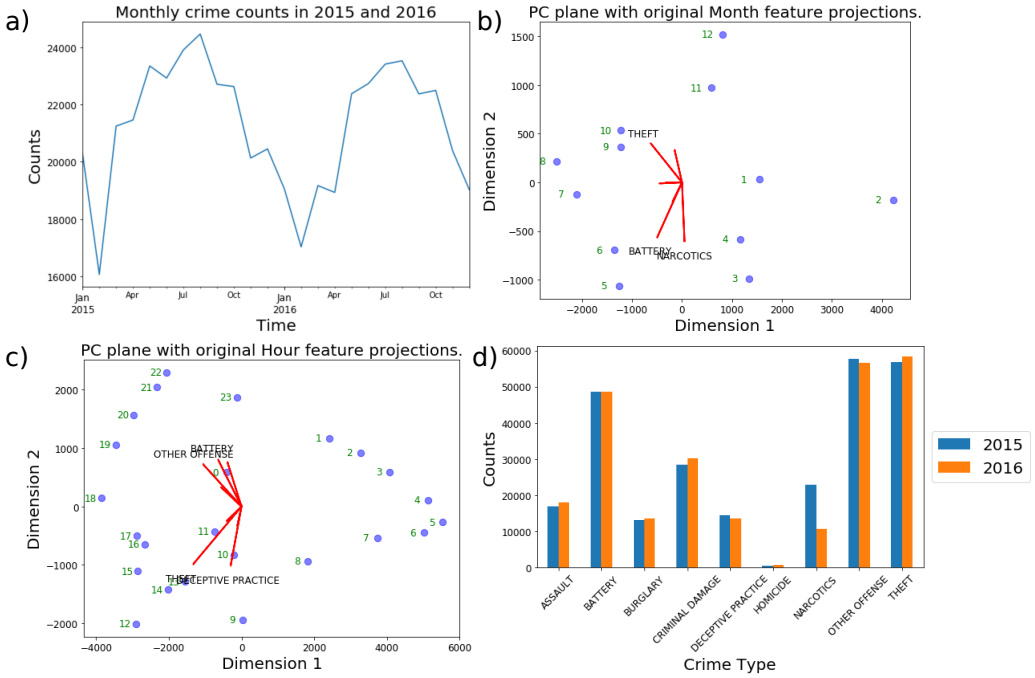

Figure 2: a) The monthly crime counts in Chicago from 2015 to 2016. b) Principal component analysis of crimes by month and type. c) Principal component analysis of crimes by hour and type. d) Comparison of crime count between 2015 and 2016.

given by

$$D = \sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk} \tag{5}$$

Both of these two measures have value between 0 and 1. When G or D is close to 0 or 1, then the node is purer, aka one class is more dominant in this node. Split with largest drop of Gini index or cross-entropy is made during tree growth.

The hyper-parameters wait to be tuned are:

- `learning_rate`: Boosting learning rate.
- `n_estimator`: number of boosted trees to fit.
- `min_child_weight`, which defines the minimum sum of weights of all observations required in a chiral.
- `max_depth`: the maximum depth of each weaker learner. Deep trees lead to over-fitting more easily.
- `gamma`: the minimum loss reduction required to make a split. Higher value prevents the model from over-fitting.
- `subsample`: the fraction of observations to be randomly samples for each tree. Lower value prevents the algorithm from over-fitting.

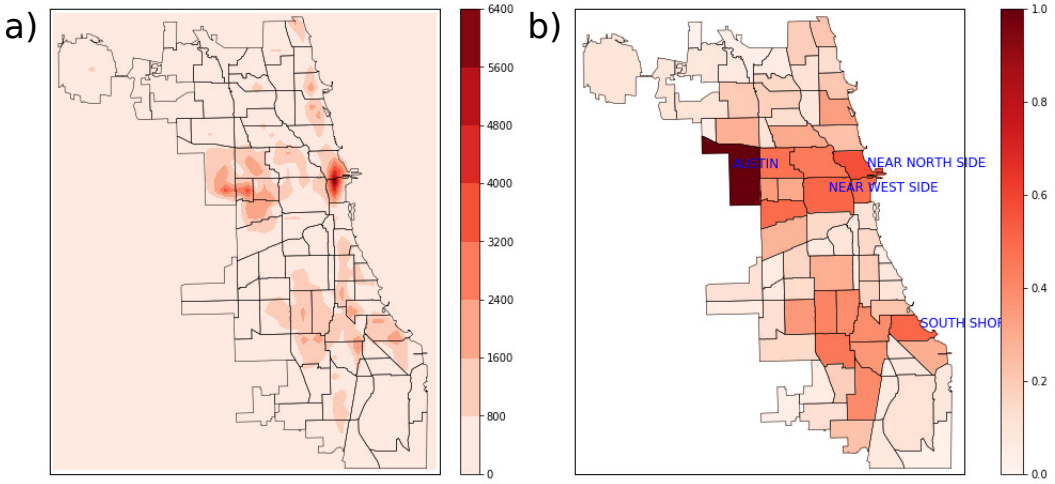

Figure 3: a) Crime density in Chicago in 2015 and 2016. b) Crimes in Chicago communities.

2.4 Benchmark Model

As a classification problem, the logistic regression model is used as a benchmark model. Logistic regression is a linear model, which could be used for the multi-class categorization.

The prediction probability for observation i and class j is defined as:

$$P(Y_i = j) = \frac{e^{\beta_j \cdot X_i}}{1 + \sum_{k=1}^{K-1} e^{\beta_k \cdot X_i}} \tag{6}$$

where β_k is the set of regression coefficient associated with class k and X_i is the set of explanatory variables associated with observation i.

Two different regularization method, Ridge and Lasso will be tested with a range of regularization parameter values.

- `penalty`: the norm used in the penalization.
- `C`: inverse of regularization strength .

3 Methodology

3.1 Data Preprocessing

The rows with NA values are dropped. New features, Month, Weekday (from Monday to Sunday) and Hour are created from original feature Date. The information of IUCR and FBI Code are excluded, because they contain the crime classification information. Despite the original positional features, new feature named 'Cluster' is created by clustering the latitude and longitude columns, using Gaussian Mixture Algorithm.

The Primary Type of crimes is the target variable for prediction, which has 33 categories in the 2016 crime data. To make the model more robust and efficient, the Primary Type is transformed to major crime types (theft, battery, criminal damage, assault, deceptive

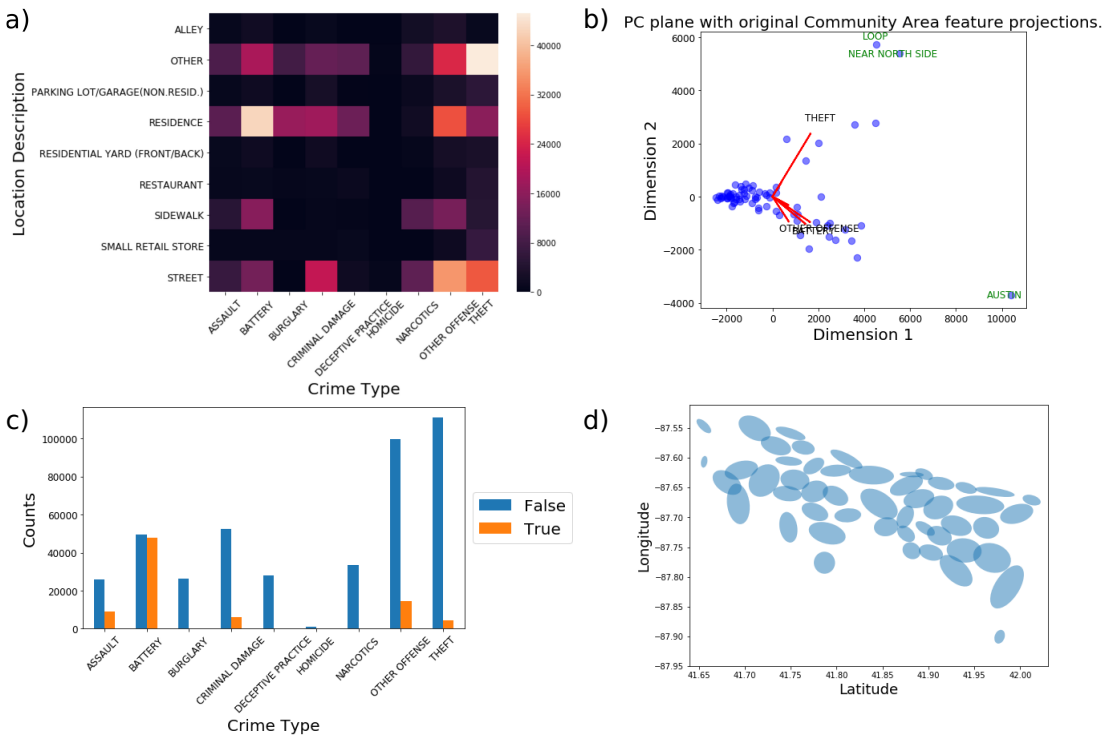

Figure 4: a) Relation between location description and crime type. b) Principal component analysis of crimes by community and type. c) Domestic related and unrelated crimes for different crime types in Chicago. d) Clusters of crimes by latitude and longitude using Gaussian Mixture Model.

practice, burglary, narcotics, robbery), and severe crime type (homicide). Figure 1 shows the total count of different crimes in Chicago from 2015 to 2016.

3.2 Implementation

- **Data analysis**
  - `Pandas`
  - `Numpy`
- **Data visualization**
  - `Matplotlib`
  - `seaborn`
  - `mpl_toolkits.basemap`

In the data visualization, the principal component analysis is used for constructing biplot to interpret the relationship of hour, month and community to crime types (imported from `sklearn.decomposition.PCA`). Before training the model, the Location Description feature is transformed to labels using `sklearn.preprocessing.LabelEncoder`. New feature,

named Cluster, is created by Gaussian Mixture Model based on Latitude and Longitude (`sklearn.mixture.GaussianMixture`). Due to the large amount of data, 20% of data are randomly selected for model training and testing with `random`. Then the data set is randomly split to train set and test set by `sklearn.model_selection.train_test_split`. The logistic regression model is imported from `sklearn.linear_model.LogisticRegression` and XGboost modeli is installed from `xgboost` packages. The parameter tuning and model evaluation use the accuracy and log loss function, replying on `accuracy_score`, `log_loss`, `confusion_matrix` and `classification_report` from `sklearn.metrics`.

3.3 Refinement

Hyper parameter is tuned using `sklearn.model_selection.GridSearchCV, KFold`. For logistic regression, the `penalty` ('l1' or 'l2') and `C` ($[0.01, 0.1, 1]$) are tuned. The accuracy for each grid is shown in Figure 6.

The parameters tuned in XGboost algorithm are as follows:

- `learning_rate: [0.01, 0.05, 0.1, 0.15, 0.2]`
- `n_estimator: [200, 300, 400, 500, 550, 600, 650, 700]`
- `max_depth: [2, 3, 4, 5, 6]`
- `gamma: [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.1, 1.2]`
- `subsample: [1, 0.95, 0.9, 0.8, 0.7]`

The grid search result for XGboost model is too large to be included in the report (please find in log folder). The parameters corresponding to the optimal model result is:

- `learning_rate = 0.1`
- `max_depth = 4`
- `gamma = 1`
- `n_estimators = 650`
- `subsample = 1`

4 Results

4.1 Model Evaluation

Two hyper-parameters, `penalty` and `C`, are tuned in the logistic regression, where `penalty` denotes the type of regularization (Ridge or Lasso) and `C` controls the regularization strength. The best parameters are chosen as ridge regression with `C = 0.1`, which yields the optimal accuracy 35.5% of training data. The evaluation performance (log-loss function and accuracy) on the test data-set are 1.667 and 35.8%.

Table 2: Classification report

Crime Type	Precision	Recall	F1-score	support
ASSAULT	0.13	0.00	0.00	2834
BATTERY	0.55	0.60	0.57	7838
BURGLARY	0.36	0.50	0.42	2116
CRIMINAL DAMAGE	0.27	0.10	0.14	4700
DECEPTIVE PRACTICE	0.41	0.34	0.37	2253
HOMICIDE	1.00	0.11	0.20	117
NARCOTICS	0.55	0.70	0.62	2587
OTHER OFFENSE	0.37	0.36	0.36	9071
THEFT	0.47	0.68	0.55	9290
avg / total	0.41	0.45	0.42	40806

Five parameters in XGboost model has been tuned, the optimal model result on training data is 44.6%, with `learning_rate = 0.1`, `max_depth = 4`, `gamma = 1`, `n_estimators = 650` and `subsample = 1`. The model performance, log-loss function and accuracy, on the test data set are 1.431 and 44.5%, respectively. A more detailed classification report is shown in Table 2. The crime types that have F1-score smaller than 20% are ASSAULT and CRIMINAL DAMAGE.

The accuracy of predictions in train and test set in both models are very close, implying the regularization parameters are well chosen and there is no significant overfitting in this problem. As explained in the Data Exploration section, neither outliers nor abnormalities are big concerns in this project.

4.2 Justification

Compared with the accuracy of benchmark model (35.8%), XGboost algorithm obtained an enhanced accuracy rate near 44.5%. The final accuracy is still better than the accuracy of the guess of the majority class THEFT out of 9-category output (22.8%).

5 Conclusion

5.1 Free Form Visualization

The analysis of feature importance from the XGboost algorithm is shown in Figure 7. For the positional information, latitude and longitude are the most important features. The artificial feature, Cluster, created from these two features also plays certain role. Among all time dependent features, hour is the most related to the crime type in the model.

The confusion matrix of final prediction on the test set is shown in Figure 8. The diagonal values in the normalized confusion matrix correspond to the precisions for each category, the value of which are listed in Table 2. We see consistently that CRIMINAL DAMAGE and ASSAULT are difficult to predict. We also notice that most HOMICIDE crimes are classified

as OTHER DEFENSE. We also notice that even though the precision of HOMICIDE equals 1. However, recall is as low as 0.11, leading to a low F1-score.

5.2 Reflection

To sum up, this project is aimed to predict the Chicago crime type based on time and position information. The raw data from Chicago Data Portal is pre-processed to new features and target variable, and then split into train, test data set. After the exploration in the visualization of the crime data, we trained the logistic regression (benchmark model) and XGboost models to obtain finely tuned parameters using k-fold cross-validation on training data set. The performance of two algorithms is evaluated with log loss function and accuracy and finally reported on the test data set. With XGboost algorithm, the optimal accuracy rate is obtained at 44.5%, much better than the accuracy of random guess with 9-category output (11.1%). Despite all this, the accuracy still seems to be low, which is probably due to the nature of this problem. For instance, as shown in Figure 5, Austin has high crime rate for most of the crime types, which indicates the difficulty of successfully categorizing the crime types in that community.

The most challenging and interesting part is to exploit new feature from original ones. As shown in Figure 7, Latitude and Longitude are most important features for model prediction. Thus the collection of information from these two features is very promising to further enhance the model performance. The decision tree algorithm divides regions in rectangular shapes by nature. Therefore, in this project, I used Gaussian Mixture Model to cluster the crime positions to create a new feature, in order to provide extra information to the XGboost model. From the Figure 7, this new feature bears certain importance in the model.

5.3 Improvement

- Increase the data set. For instance, earlier crimes data might be included.
- Feature engineering. The import of new features, such as a binary feature holiday, may further improve the model performance.
- Parameter tuning. Further examination of hyperparameters in the XGboost model or cluster number in the Gaussian Mixture model may further enhance the model performance.
- Model selection. Other classifiers, such as Random Forest, K Nearest Neighbors, SVM, Neural Network, may be tested.
- Ensemble modeling. The results of different models may be synthesized to improve the accuracy of prediction.

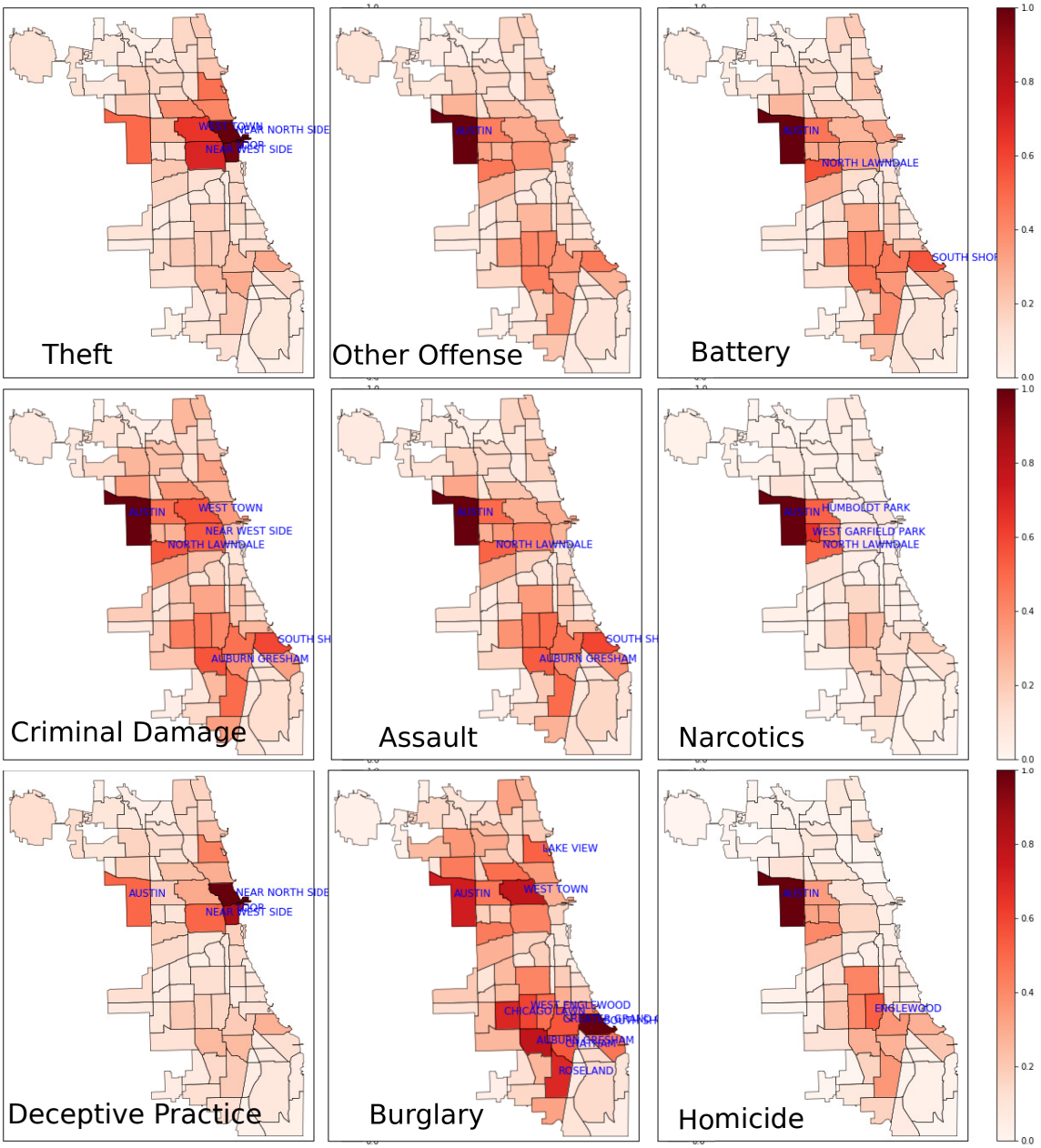

Figure 5: Standardized crime count in Chicago communities from 2015 to 2016 for different crime types.

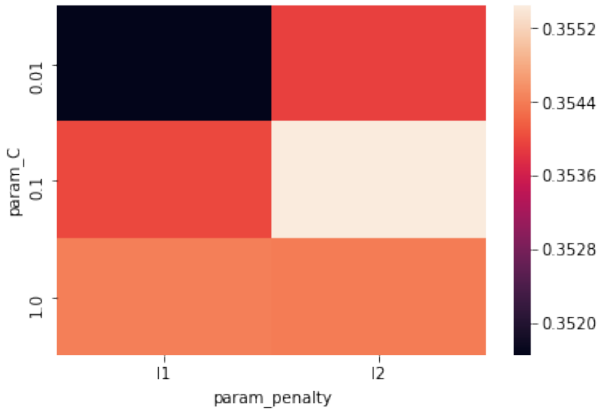

Figure 6: Grid search result of parameters for logistic regression.

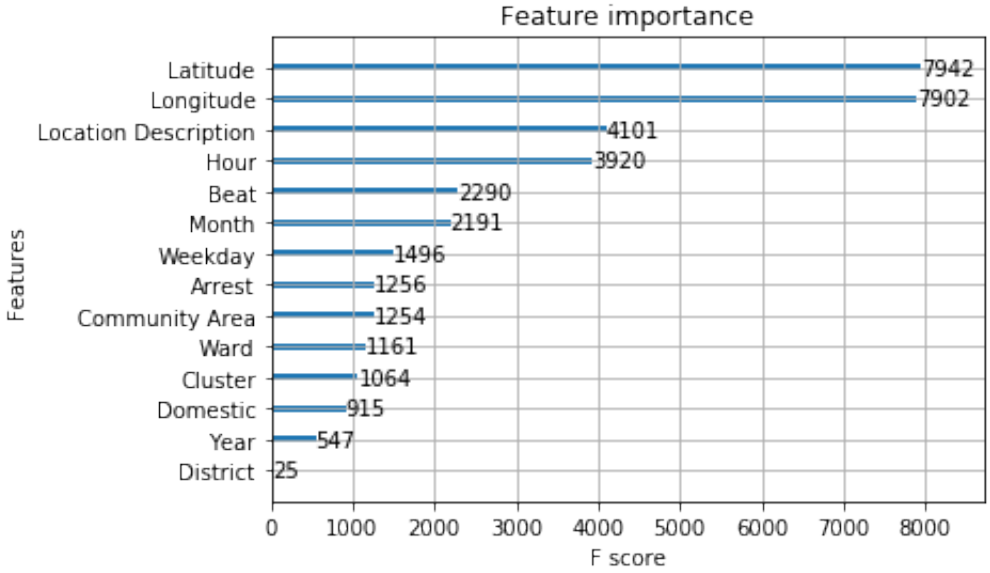

Figure 7: Feature importance in the XGboost model for Chicago crime categorization problem.

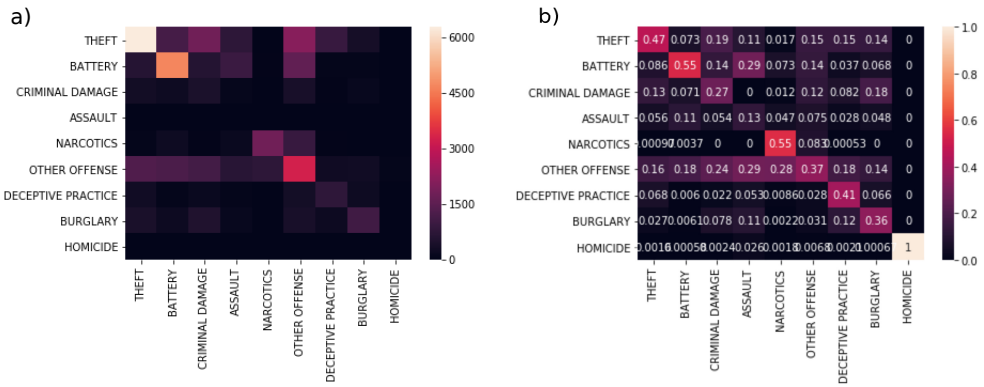

Figure 8: Confusion matrix of crime classification with (b) and without (a) normalization