Patterns and Predictions of Crime in Chicago

PRINCETON UNIVERSITY

Matthew Yeh, May Jiang, Russell Slighton

COS 424: Fundamentals of Machine Learning, Professor Barbara Engelhardt

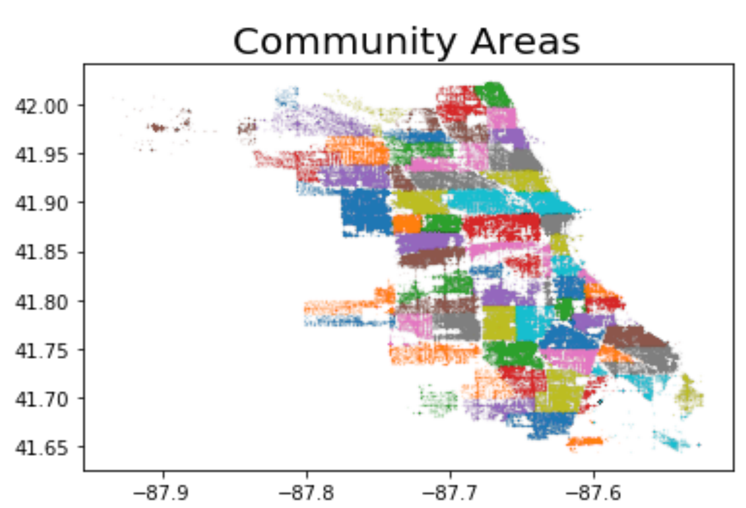
Abstract

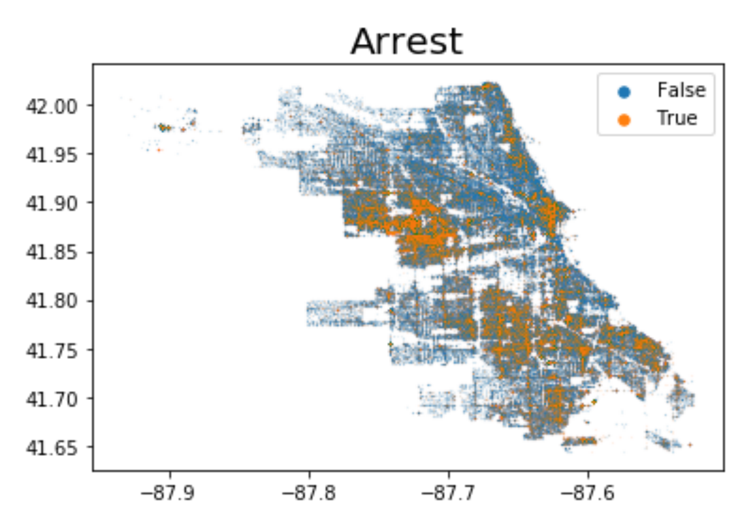
- Uncover patterns and latent structure in crime incidents in Chicago
  - Time
  - Location
  - Type of crime
- Utilize the features in the dataset and the latent structure discovered to predict and analyze criminal behavior
  - Violent and non-violent crimes
  - Whether arrest was made
  - Frequency of crimes
  - Type of crime
- Consider the consequences of our results in the context of policing and community

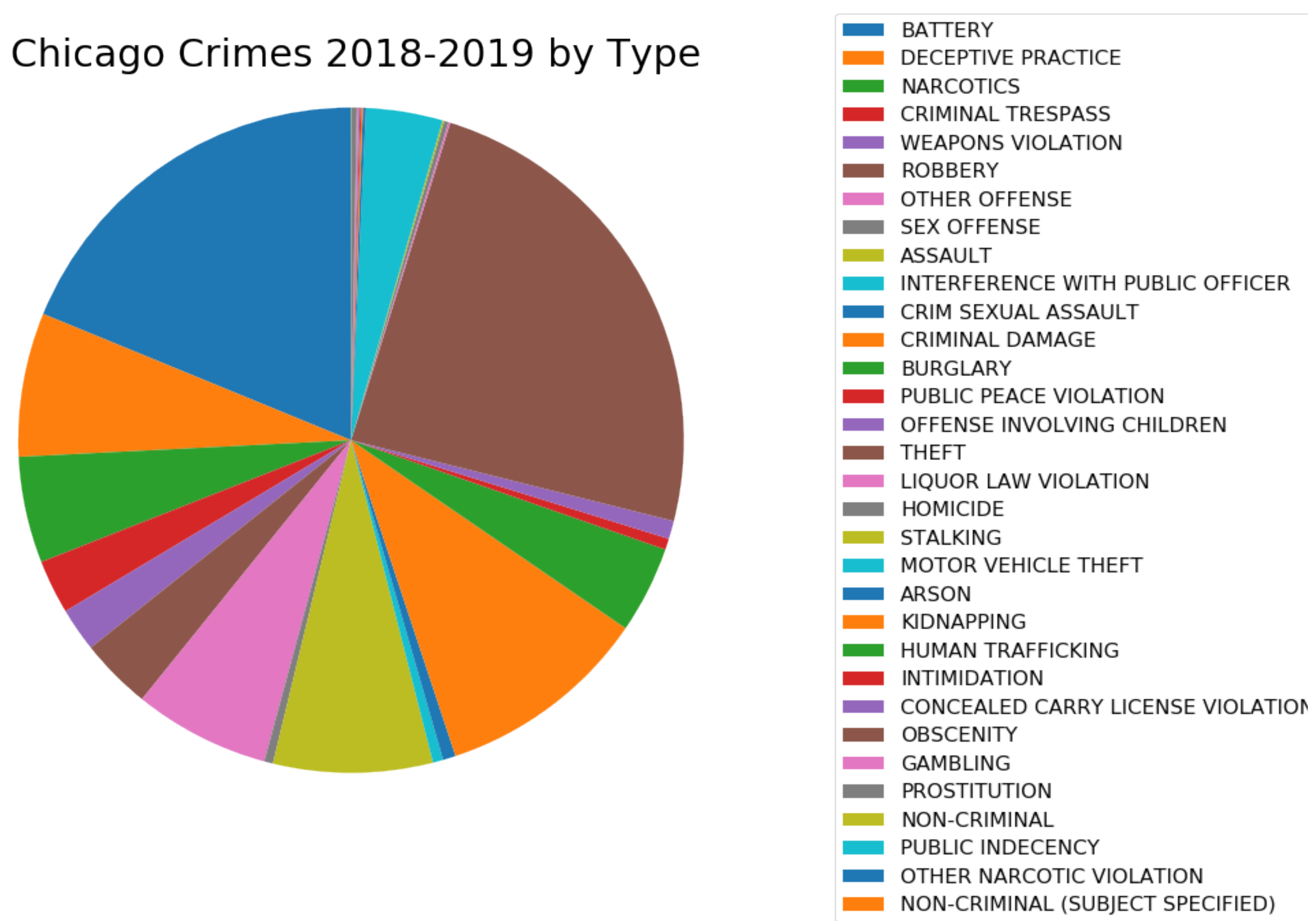
Background

Dataset
- Continuously updated (2001 – Present)
- City-proper Chicago
- Exact location of crime
- 1.8 GB

General Information
- Crime disproportionately allocated
- City proper Chicago
- 10th highest for murder, 3rd for violent
- Most gang-infested city
- Police underreporting crime



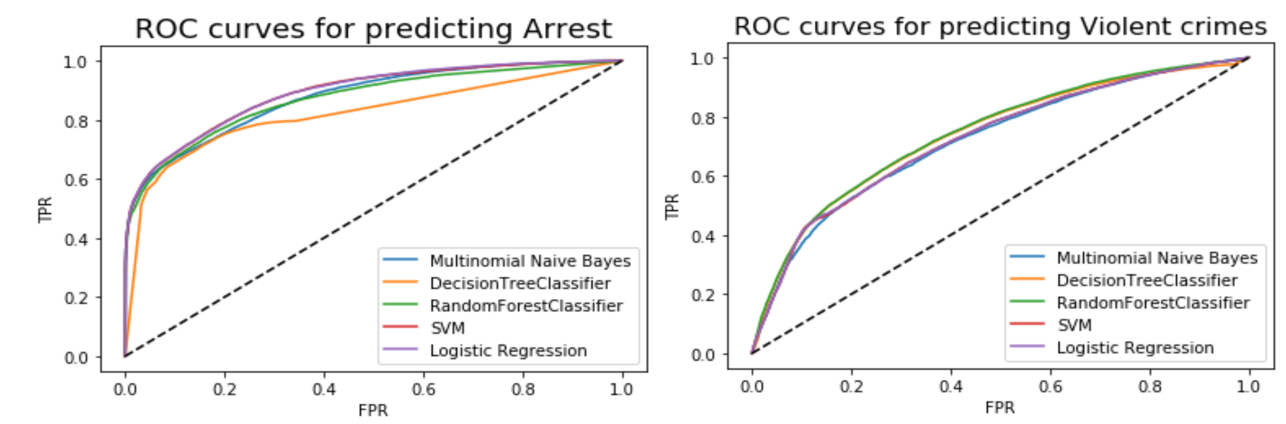




Classification

Question

Can we predict if a crime was violent or not?



Model	Arrest		Violent	
	Accuracy	AUC	Accuracy	AUC
Multinomial Naïve Bayes	0.889	0.877	0.711	0.720
DecisionTreeClassifier	0.877	0.821	0.723	0.739
RandomForestClassifier	0.883	0.891	0.724	0.745
SVM (Linear Kernel)	0.891	0.890	0.725	0.726
Logistic Regression	0.892	0.891	0.724	0.726

Latent Structure

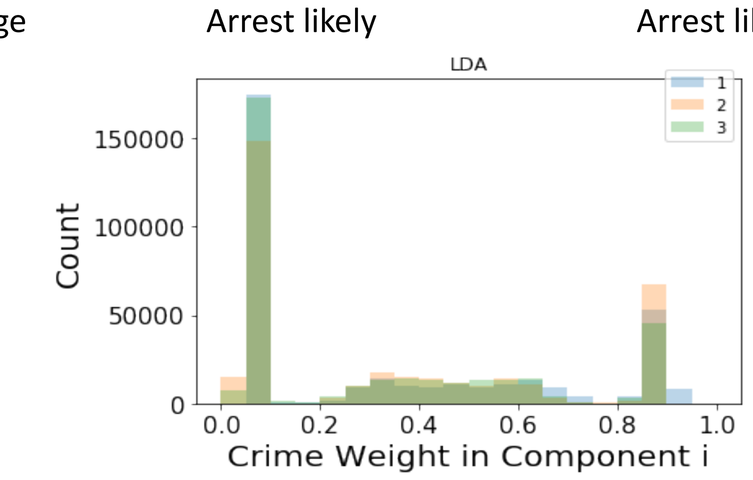
Question

Can we predict crimes by understanding key components?

Method
1. Make data multinomial
2. Train Latent Variable Models
3. Examine ability to predict and latent components

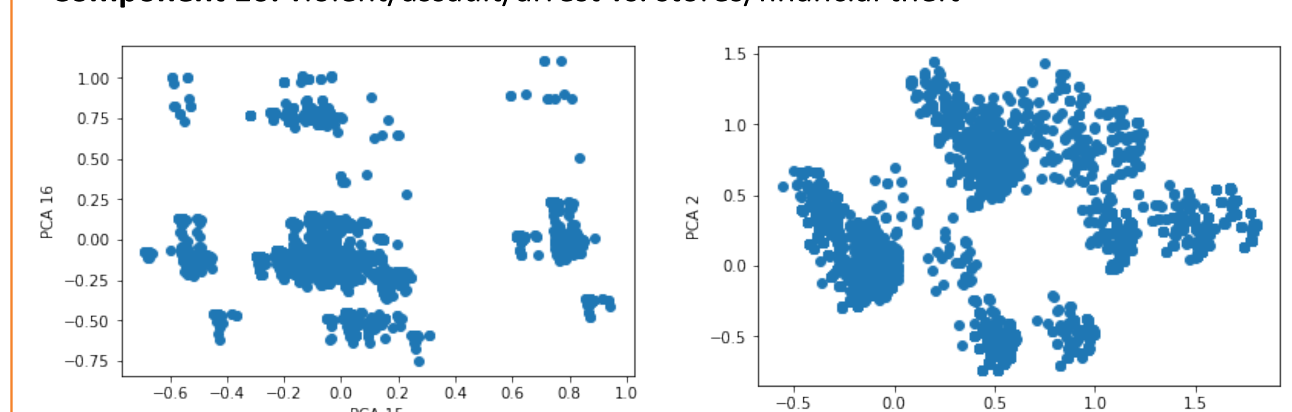
LDA

Component 1	Component 2	Component 3
Residential	Street	Store
Violent	transportation	Sidewalk
Domestic	location_sketchy	Violent
Domestic battery	Theft	Battery
Hours 6am-12pm	Night time	Night time
Property damage	Arrest likely	Arrest likely



PCA

Component 1: residential/domestic abuse vs. street/transportation/vehicle trespass
Component 15: drugs/weapons, night vs. fraud/electronic harassment, morning
Component 16: violent/assault/arrest vs. stores/financial theft



KMEANS
Centroid 3: store crimes
Centroid 4: arrest, street, late night, drugs
Centroid 5: violent, residential, non-domestic
Centroid 7: domestic, residential, late night

GMMs:
manually determined no latent structure found

Time Series Regression: Crime Rate

Question

Can we predict the number of crimes in the next week?

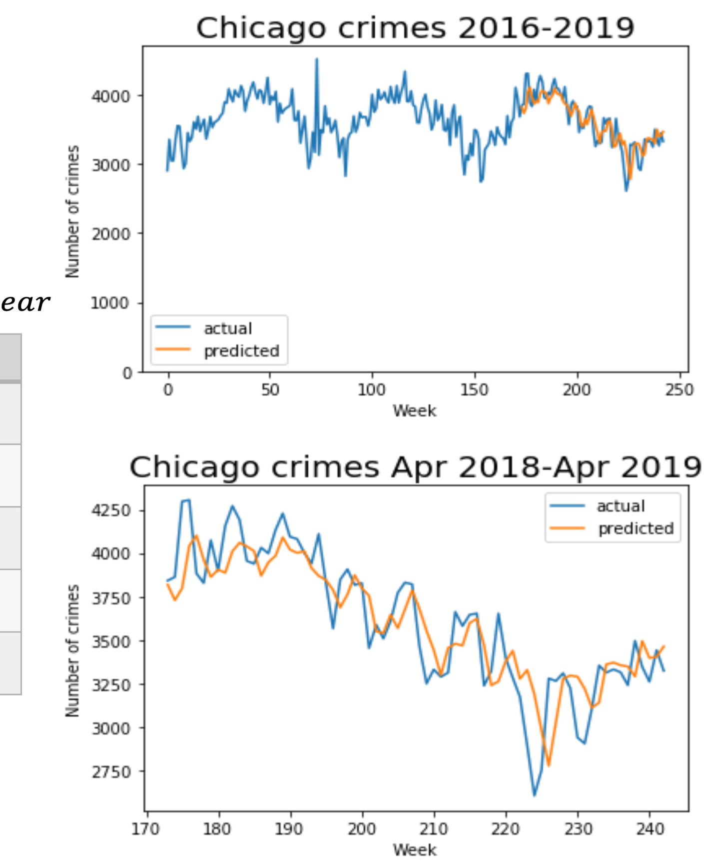
Autoregressive Models

1. $Y_t = \beta_0 + \beta_1 Y_{t-1}$
2. $Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2}$
3. $Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-year}$
4. $Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-year} + \beta_4 Y_{t+1-year}$

Model	R^2	RMSFE
Mean	0.000	393.91
AR 1	0.703	214.56
AR 2	0.714	210.56
AR 3	0.726	206.26
AR 4	0.745	198.76

Predicting by location (district, community area, ward) performed worse

Predicting only violent crimes performed slightly worse



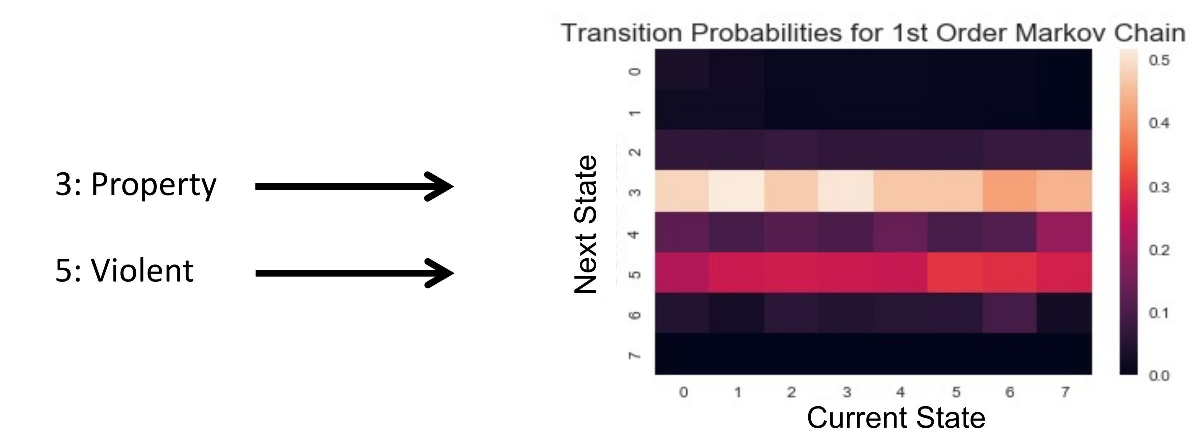
Temporal Structure of Crime Types

Question

Do some types of crimes tend to follow others?

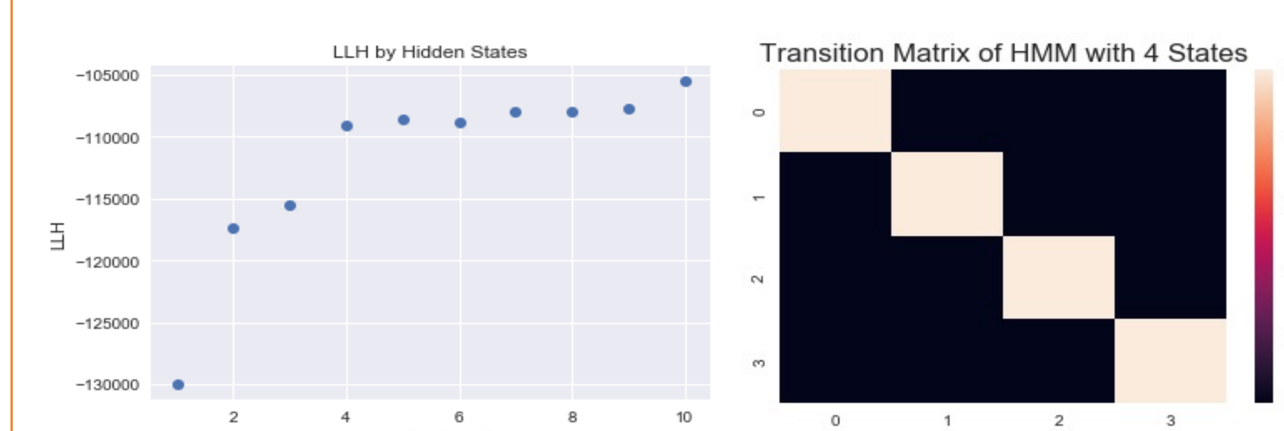
Method
1. Data grouped by *community area* (maximizes IGR) to produce training sequences.
2. Ordered data by time (ascending).
3. Crimes binned into eight natural categories.

Manifest Structure: Markov chain



Structure due to dataset imbalance.

Latent Structure: HMM



No improvement to performance above four hidden states.

No obvious latent structure.
As predictive model:
f1 = 0.31

References

City of Chicago. Crimes - 2001 to present. https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2.

Criminal Justice Information Services Division. 2017 Crime in the United States. https://ucr.fbi.gov/crime-in-the-u.s/2017/crime-in-the-u.s.-2017, 2017.

Acknowledgements

Kaggle's data collection was critical to our research.
We would like to thank the course staff and Professor Engelhardt for helping us and teaching us all throughout the semester.