Analysis and Prediction on Crime in Chicago City

Haoran Li, Oliver
Ziyue Peng

香港城市大學
City University of Hong Kong

Agenda

1. Identify problem (purpose of our study)
2. Data sources (where & how)
3. Prepare data (integrate, transform, clean, filter, aggregate)
4. Data Visualization
5. Build model & Evaluate model (applying evaluators)
6. Communicate results (final achievement)

1

Identify problem

- Crime issue
- Dangerous zones
- Low arrest rate
- Our purpose

2

Data Source

Data source was obtained from **Kaggle.com**

The original data was extracted from **Chicago Police Department's CLEAR** (Citizen Law Enforcement Analysis and Reporting) system

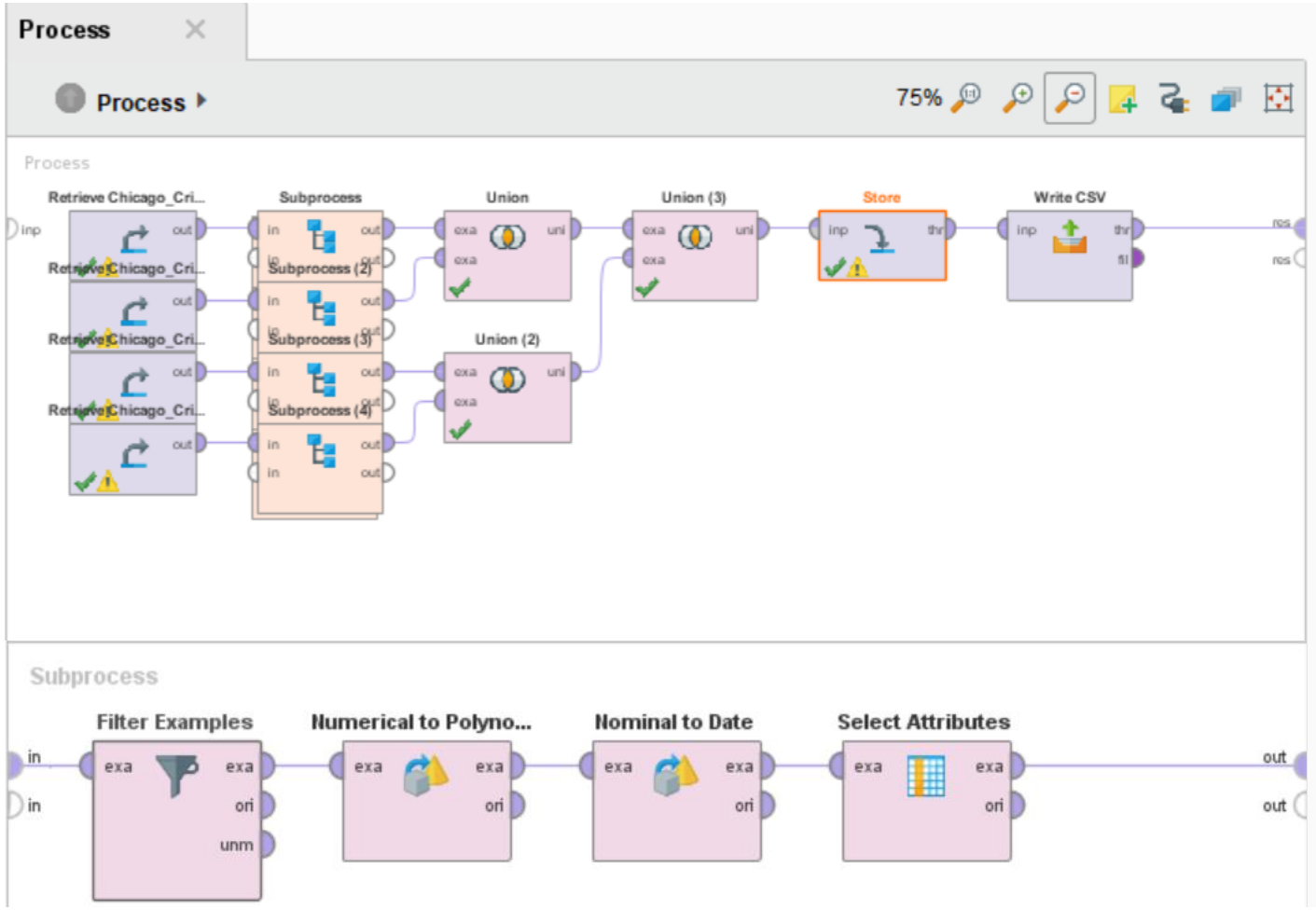
The dataset contains crime information from 2001 to 2017, and is consisted of nearly 8 million rows

kaggle

3

Data Preparation

- We have completed data cleaning on our own for machine learning part by using RapidMiner, a visualized data cleaning and data mining software.



3

Data Preparation

- We have also used built-in methods in spark to do some other kinds of data type transformation, since RapidMiner is not supporting all types of transformation

```
root
 |-- id: integer (nullable = true)
 |-- beat: string (nullable = true)
 |-- type: string (nullable = true)
 |-- location: string (nullable = true)
 |-- arrest: boolean (nullable = true)
 |-- datetime: timestamp (nullable = true)

Total number of crime records: 7939294
```

PART 4

Data Visualization

5

Build Models

In general, we have chosen 3 machine learning models to apply in our project: Naïve Bayes, Logistic Regression, Linear Support Vector Machine.

Naïve Bayes

NB is one of the simplest model in ML. It converges faster than other models, providing fast speed as well as relatively good performance

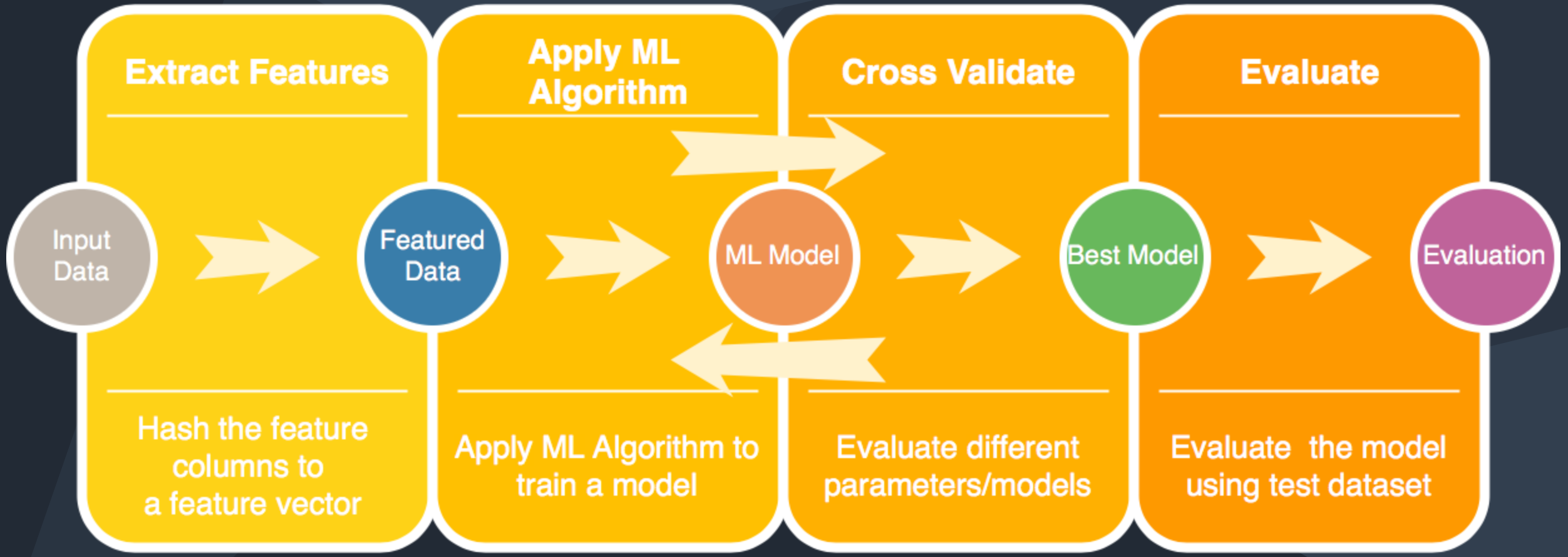
Logistic Regression

LR has a nice probabilistic interpretation. And when you don’t have too many features, it has very high accuracy.

Linear Support Vector Machine

If we define features as vectors, then our data can be located in a high dimensional space. SVM is a hyperplane that divide the hyperplane to help doing classification.

Workflow


Extract Features
Apply ML Algorithm
Cross Validate
Evaluate
Input Data
Featured Data
ML Model
Best Model
Evaluation
Hash the feature columns to a feature vector
Apply ML Algorithm to train a model
Evaluate different parameters/models
Evaluate the model using test dataset

5 Evaluate Models

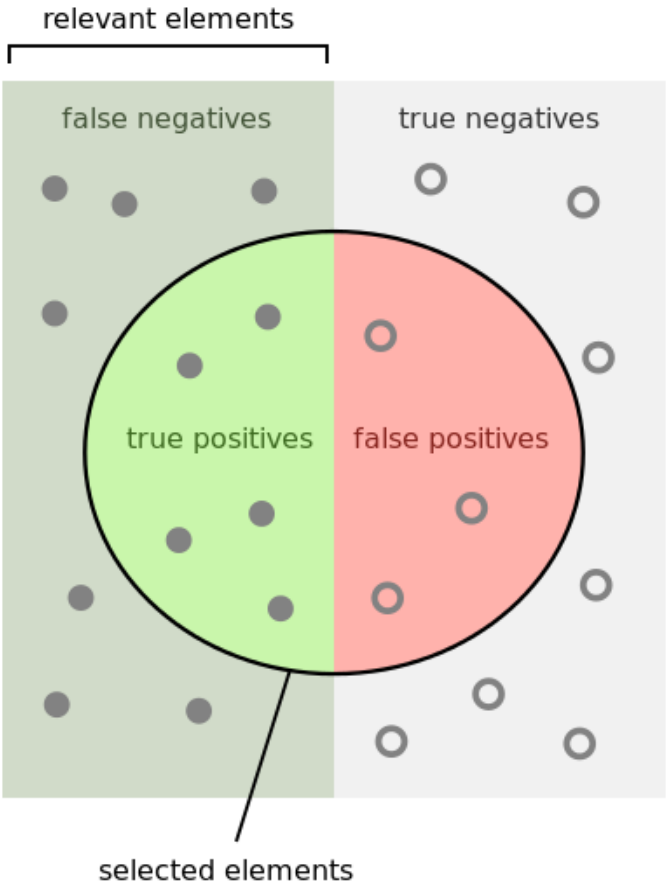
$$Precision = \frac{True\ Positive}{Labelled\ Positive} = \frac{True\ Positive}{True\ Positive + False\ Postive}$$

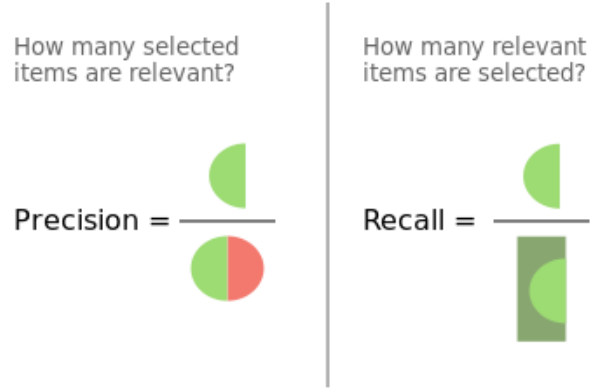
How correctly label?

$$Recall = \frac{True\ Positive}{Fact\ Positive} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

How well distinguish?

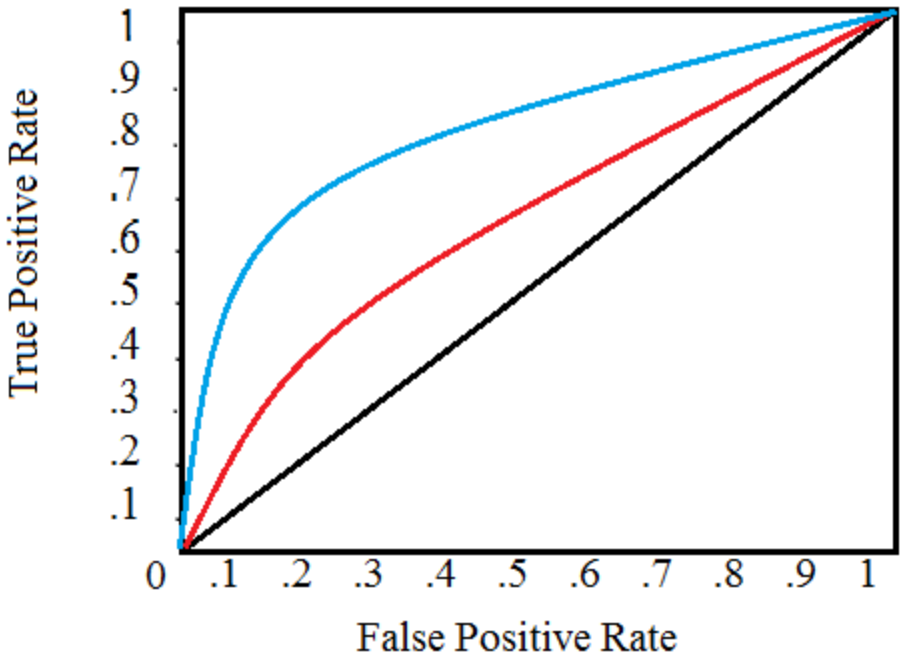
$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$





5 Evaluate Models

AUC (Area Under ROC Curve)



$$False\ Positive\ Rate = \frac{False\ Positive}{Fact\ Negative} = \frac{False\ Positive}{False\ Positive + True\ Negative}$$

$$True\ Positive\ Rate = Recall$$

Draw for each threshold

(0,0) all samples labelled negative

(1,1) all samples labelled positive

(0,1) all samples labelled correctly

(1,0) all samples labelled incorrectly

AUC=1 -> perfect model

Results

- Information in data
- Methodology of machine learning
- Prediction of arresting results
- Real world effects

Q&A

2018

THANK YOU!

Our greatest appreciation to the general support of Prof. Yi Shang and members of TA group!