TERM PROJECT

PREDICTING CRIMES

11146315 SEOUNGWAN SEO
12146305 JEGYEONG KIM

Business Analytics
Prof. Kyoungok Kim

CONTENTS

01 BACKGROUND

02 ANALYSIS CONCEPT

03 ANALYSIS PROCESS

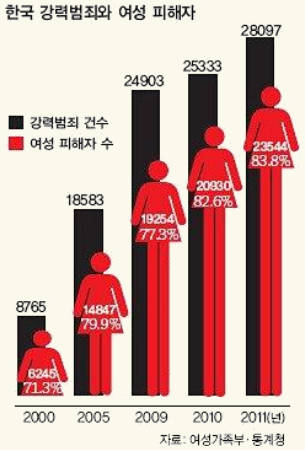
04 CONCLUSION

1. Background

Nowadays…

Random and violent crime is one of **biggest social problem.**
According to National Police Agency, the crime against women, children and the elderly is very serious and especially, the crime against woman accounts for 87 percent of crime.

The best way to handle crime is **prevention**, not after treatment.
We need the way to manage the **patrols** efficiently and effectively!

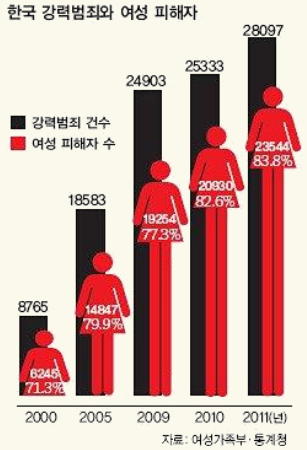


1 Background

Nowadays...

Random and violent crime is one of **biggest social problem.**
According to National Police Agency, the crime against women, children and the elderly is very serious and especially, the crime against woman accounts for 87 percent of crime.

The best way to handle crime is **prevention**, not after treatment.
We need the way to manage the **patrols** efficiently and effectively!



2 Concept

Purpose : Reduce crime rate

Prediction

When | Where | What

crime will be occurred?

Patrol

based on result of analysis

Prevent

the criminals from committing a crime

BACKGROUND
CONCEPT
PROCESS
CONCLUSION

1. Data Exploration

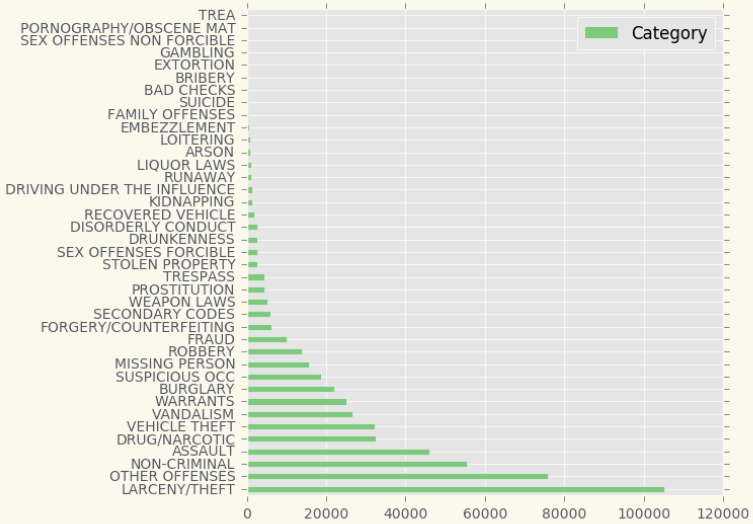
Type	Continuous Variables	Categorical Variables
Columns	Dates, X, Y	Category, Descript, DayOfWeek, PdDistrict, Resolution, Address

Dates	timestamp of the crime incident
Category	category of the crime incident.
Descript	detailed description of the crime incident.
DayOfWeek	the day of the week
PdDistrict	name of the Police Department District
Resolution	how the crime incident was resolved
Address	the approximate street address of the crime incident
X	Longitude
Y	Latitude

BACKGROUND
CONCEPT
PROCESS
CONCLUSION

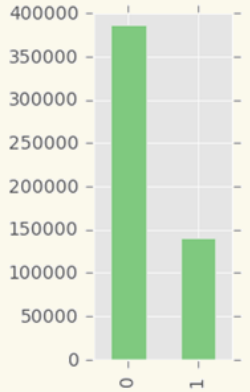
1 Data Exploration

The number of each crimes



1. Many trivial crimes and crimes that cannot be prevented by patrolling
2. Big gap between the number of categories

The number of serious crime & not-serious crime



1. Serious crime is minor class
2. Not-serious crime is major class

BACKGROUND
CONCEPT
PROCESS
CONCLUSION

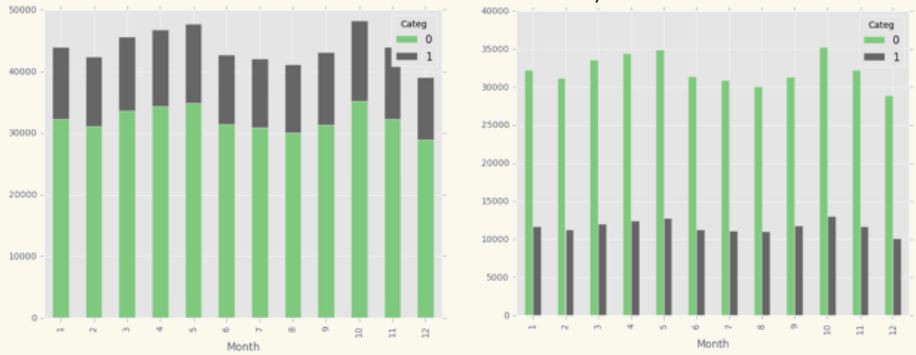
1 Data Exploration

The number of crimes by Year

1. Serious crime decrease slightly
2. Not-serious crime increase slightly

1. Both of serious and minor crime follows similar pattern and there are no big difference.

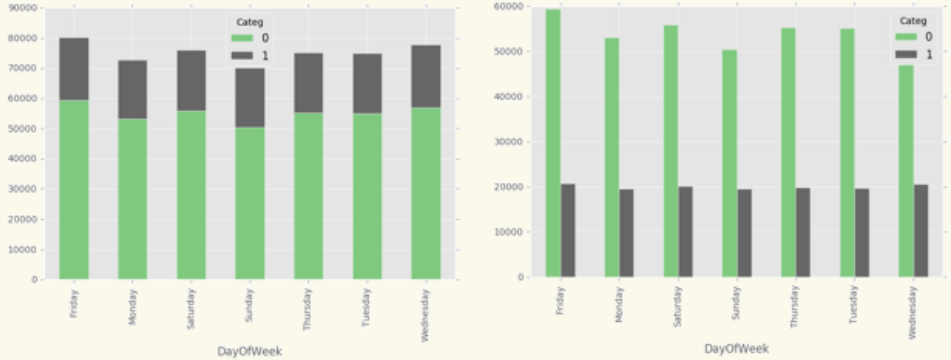
The number of crimes by Month



BACKGROUND
CONCEPT
PROCESS
CONCLUSION

1 Data Exploration

The number of crimes by Week of Day

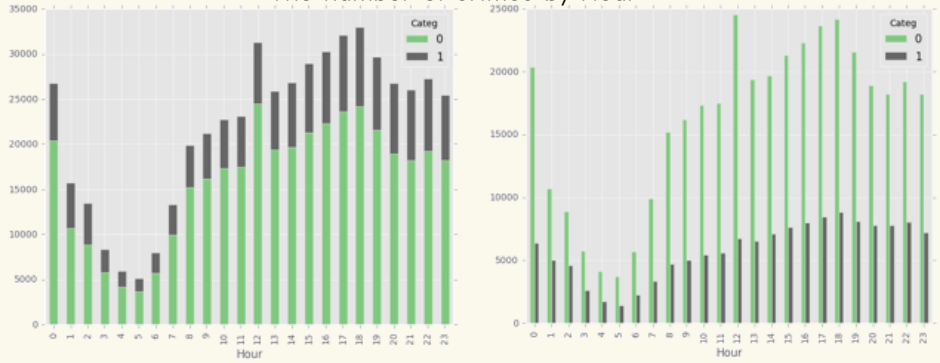


1. The number of not-serious crime is influenced by week of day
2. The number of serious crimes is same regardless of day of week

---

1. In the morning, the number of crimes is small, otherwise, the amount of crimes is very high in dinner time.

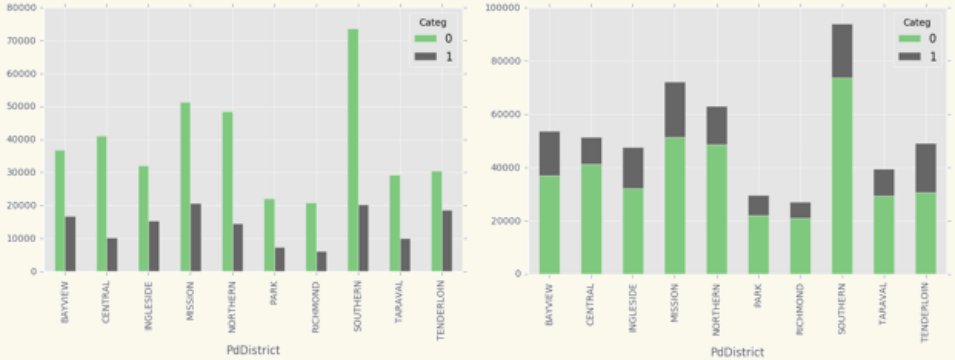
The number of crimes by Hour



BACKGROUND
CONCEPT
PROCESS
CONCLUSION

1 Data Exploration

The number of crimes by Police Department District

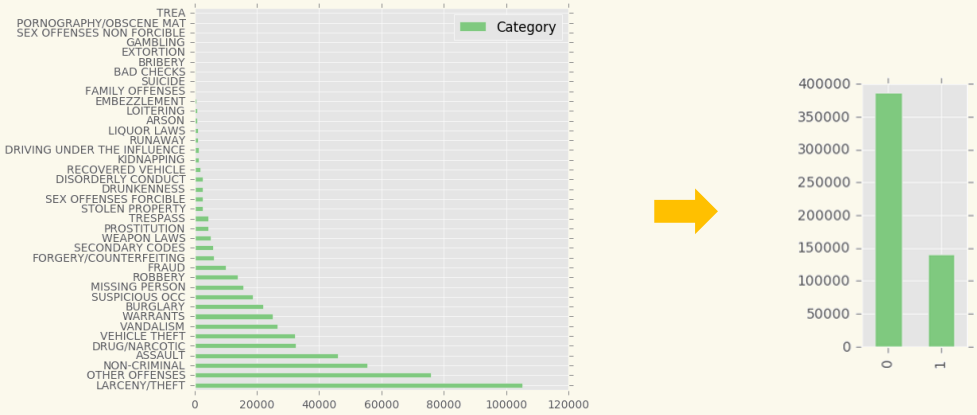


1. Each district has own characteristics
2. The number of crime is highly influenced by PdDistrict.

As a result of Data exploration,
Almost of variables look like important.
However, we will add and delete some variables
by comparing score later for better accuracy.

2. Preprocessing

2.1. Convert 'Category' Column to Binary Target Variable



We want to find only serious crime and the crime that we can prevent in advance.

Convert 'Category' Column to binary target variable by serious/not serious crime.

Serious crime is class 1 that we want to find.

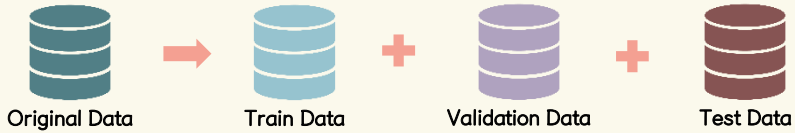
Not-serious crime is class 2 that is major class.

※The list of target class 1 is as followings:
KIDNAPPING, ARSON, ASSAULT, VEHICLE THEFT, ROBBERY, WEAPON LAWS, DRUG_NARCOTIC, TRESPASS, RUNAWAY, SEX OFFENSES FORCIBLE.

BACKGROUND
CONCEPT
PROCESS
CONCLUSION

2. Preprocessing

2.2. Drop Columns
2.3. Convert 'Dates' Column
2.4. Split Data
2.5. Make dummy variables



Continuous Variables	Categorical Variables
Dates, X, Y	Category, ~~Descript~~, DayOfWeek, PdDistrict, ~~Resolution~~, ~~Address~~

Make dummy variables

2016-06-09 14:30

Year 2016
Month 6
Hour 14
Minute 30

Drop 'Descript', 'Resolution' column because it is not influence to output.

Drop 'Address' columns because of too many categories over 20,000

BACKGROUND
CONCEPT
PROCESS
CONCLUSION

3. Analyze

For us, **Accuracy** and **Recall** is most important score factors.
Assume that if polices patrol, the crime is prevented.

Precision means how many times we prevent serious crime among patrols we go to.
Recall means how many times we prevent serious crime among really occurred crimes.

Precision

Recall

3. Analyze
3.1. Selecting Variables
3.2. Selecting Sampling Method
3.3. Selecting Analysis Method

Selecting Variable(Naïve Bayes)

	Original	Without 'Year'	Without 'Month'	Without 'DayOfWeek'
Accuracy	0.6522	0.6586	0.6516	0.6458
Precision	0.3517	0.3428	0.3502	0.3487
Recall	0.3643	0.3087	0.3615	0.3618
F Score	0.3579	0.3249	0.3558	0.3645

Selecting Sampling Method(Decision Tree)

	Accuracy	Precision	Recall	F-Score
Original	0.8417	0.7197	0.6636	0.6905
Under-sampling	0.5786	0.5766	0.5840	0.5802
One-class Learning	0.7355	0.5782	0.0223	0.0431

Selecting Analysis Method

	Naïve Bayes (Gaussian)	Decision Tree	SVM	KNN	Logistic Regression
Accuracy	0.6518	0.8417	0.6243		0.7335
Precision	0.3506	0.7197	0.2777		0.4316
Recall	0.3622	0.6636	0.2574		0.0050
F Score	0.3563	0.6905	0.2672		0.0100

1. Conclusion

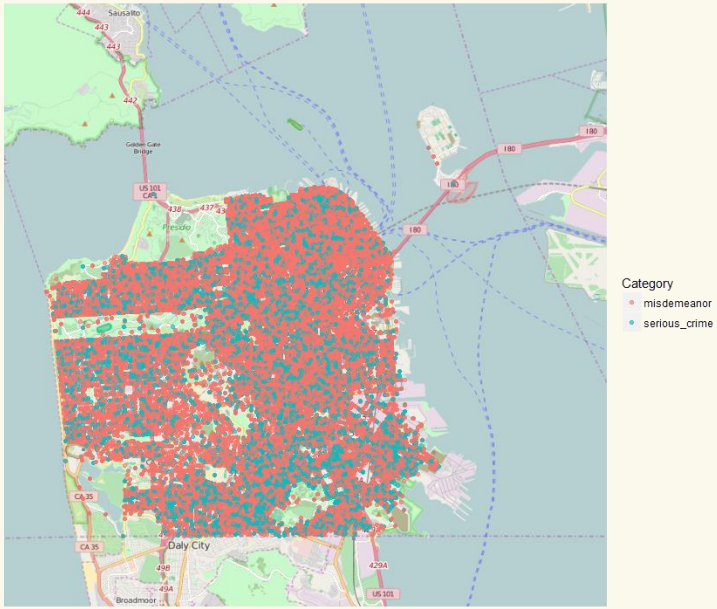
1.1. Final Performance of test set

1.2. Visualization

Final Performance of test set

Decision Tree	Accuracy	Precision	Recall	F-score
Train	0.8417	0.7197	0.6636	0.6905
Test	0.8396	0.7154	0.6611	0.6872

Visualization predicted crime of test set



BACKGROUND
CONCEPT
PROCESS
CONCLUSION

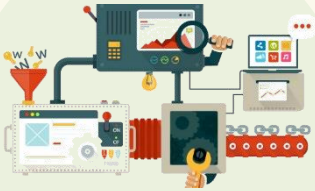
Conclusion

We think performance of our model is good.

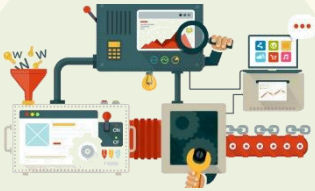
If patrol can prevent serious crime,
crime ratio would decrease to 70% based on our model.

If we have data of Korea,
we will be able to predict crime
and get similar performance with this data.

We hope that this study and analysis exercise
would be helpful to potential victims.



Q & A



THANK YOU