Paichana Kularb
57090015

Data Mining Project 2

Data Preparation

Attributes that will be used in all predictions are as following:

Attributes	Description	Derived from
day	Day of the week(Monday-Sunday)	Datestop
hours	Time of the day (hours)	timestop
age	Age of the suspect	
sex	Sex of suspect	
race	Race of suspect	
frisked	Was suspect frisked	
searched	Was suspect searched	
perobs	Period of obersvation(minutes)	
perstop	Period of stop(Minutes)	
typeofid	Suspect's identification type	
weight	Suspect's weight(Kg)	
height	Suspect's height(Cm)	ht_feet, ht_inch
bmi	Suspect's BMI	height,weight
stopReason	Reason for stop	cs_objcs,cs_descr, etc
crimsup	Crime suspected	

- BMI is used instead of build because BMI is a more accurate description and is less prone to human errors. Since different people can have different opinion on a person's build.
- Crimsup is cleaned so that every record that means the same thing a grouped. For example, 'FEL' and 'FELONY'. Factor's level that are less frequent is also grouped into 'RARE'.
- stopReason is created by combing all of the cs_xxxx attributes.

Machine Learning libraries

Libraries	**Model**
e1071	SVM and Naïve Bayes
rpart	CART
C5.0	C5.0 tree
randomForest	randomforest
nnet	Neural Network

Predict if a person is armed?

Columns that indicate which type of weapon is found is combined into one column. The column will indicate if any type of weapon is found on the suspect. Combine: *pistol, riflshot,asltweap,knifcuti,machgun,othrweap* into *weaponfound.* This will be the y variables to predict whether a person is armed.

Additional attributes are:

Attributes	**Description**	**Derived from**
isforceused	was force is used by the officers	pf_hands,pf_wall,etc
arstmade	Was arrest made	

- Attributes that are numeric is normalized into having a mean of 0 and standard deviation of 1.
- The data is very unbalanced:

Dataset	**Unarmed**	**Armed**
Train	10523(94.3%)	641(5.7%)
Test	1170(94.4%)	70(5.6%)
Total	11693(94.3%)	711(5.7%)

This might cause the model to be biased towards the suspect is unarmed.

Classification models:

- **Support Vector Machine model**

1. Parameters

.1.1.Kernel = Radial Basis

.1.2.Gamma = 1/features

.1.3.Epsilon = 0.01

.1.4.Constant = 1

2. Results of not armed

.2.1.Accuracy = 94.4%

.2.2.Recall = 100%

.2.3.Precision = 94.4%

3. Results of armed

.3.1.Recall = 0%

.3.2.Precision = 0%

	N Actual	Y Actual
N Fitted	1170	70
Y Fitted	0	0

Support Vector Machine predicted all the suspect as not armed. This is because most of the data in this dataset contains a lot of unarmed suspect.

- **Classification and Regression Tree**

1. Results of not armed

.1.1.Accuracy = 94.4%

.1.2.Precision = 95.5%

.1.3.Recall = 98.8%

2. Results of armed

.2.1.Precision = 51.7%

.2.2.Recall = 21.4%

	N Actual	Y Actual
N Fitted	1156	55
Y Fitted	14	15

- **Neural Network**

1. Parameters

.1.1.Weight of each sample = 1

.1.2.Hidden layers = 5

.1.3.150 Iterations

2. Results of not armed

.2.1.Accuracy = 94.7%

.2.2.Recall = 98.4%

.2.3.Precision = 96.1%

3. Results of armed

.3.1.Precision = 54.8%

.3.2.Recall = 32.8%

	N Actual	Y Actual
N Fitted	1151	47
Y Fitted	19	23

Model Comparison

The accuracy of neural network is the highest amongst the 3. The precision and recall for the 3 models is very high if the positive class is unarmed suspected. This is because of the imbalanced data of the dataset.

The interesting part is the result when the class armed suspect is positive. The Neural Network’s and CART’s precision is 54.8% and 51.7% respectively which is quite high comparing to the number of armed samples (5.7%). However, SVM only predicted that every suspect is unarmed.

Evaluation and Deployment

The model is not that use full to the police, the precision and recall of armed suspect is too low. Some data that are used is also not possible to get before stopping, is weapon found for example.

The model could be improved by collecting more data about armed suspect. The precision and recall will likely increase as more data is collected.

Predict if a person is arrested?

Additional attributes are:

Attributes	Description	Derived from
Isforceused	was force is used by the officers	pf_hands,pf_wall,etc
weaponFound	Was weapon found	pistol, riflshot,asltweap,knifcuti,machgun,othrweap

- Attributes that are numeric is normalized into having a mean of 0 and standard deviation of 1.
- The data is unbalanced:

Dataset	Unarmed	Armed
Train	8792(78.8%)	2372(21.2%)
Test	969(78.1%)	271(21.9%)
Total	9761(78.7%)	2643(21.3%)

Classification models:

Random Forest

1. Parameters

   .1.1.Number if trees = 500

2. Results of not arrested

.2.1. Accuracy = 86.6%

.2.2. Precision = 88.9%

.2.3. Recall = 94.7%

3. Results of arrested

.3.1. Precision = 75.4%

.3.2. Recall = 57.6%

	N Actual	Y Actual
N Fitted	918	115
Y Fitted	51	156

Neural Network

1. Parameters

.1.1. Weight of each sample = 1

.1.2. Hidden layers = 5

.1.3. 100 Iterations

2. Results of not arrested

.2.1. Accuracy = 85.2%

.2.2. Precision = 90.5%

.2.3. Recall = 93.2%

3. Results of arrested

.3.1. Precision = 64.9%

.3.2. Recall = 72.7%

	N Actual	Y Actual
N Fitted	903	95
Y Fitted	66	176

Stack by voting (Random Forest, SVM, Naïve Bayes, Neural Network and CART)

1. Results of not arrested

    .1.1.Accuracy = 86.7%

    .1.2.Precision = 91.2%

    .1.3.Recall = 91.8%

2. Results of arrested

    .2.1.Precision = 70.1%

    .2.2.Recall = 68.3%

	N Actual	Y Actual
N Fitted	890	86
Y Fitted	79	185

Model Comparison

The accuracy of stacking 5 models is the highest at 86.7%, follow with Random forest (86.6%) and Neural Network(85.2%). All of the precision and recall of suspect not getting arrest is high because of the sample and biasness.

The important part is the precision and recall when suspect is arrested. Random Forest precision is 75.4% but only 57.6% are recalled. While Neural Network recalled 72.7% but with 64.9% precision. And for the Stacking model precision is 70.1% with 68.3% recall.

Evaluation and Deployment

The model is not useful to the police because the data it requires might not be possible to obtain beforehand.

The model could be improved by collecting more data about arrested suspect. The precision and recall will likely increase as more data is collected. Attributes such as birth country, years in New York, salary could increase the model's performance.

Predict type of force used by officer?

The force type is separated into many columns each column only indicates whether a type of force is used. There is a problem in which more than one force could be used on a single suspect. I have created new class for different combinations. Only the top 15 combinations in terms of frequency will be a class. The rest will be combined in to a class called rare.

- Additional attributes are:

Attributes	**Description**	**Derived from**
arstmade	Was arrest made	
weaponFound	Was weapon found	pistol, riflshot,asltweap,knifcuti,machgun,othrweap
pct	Precinct of stop	

- Attributes that are numeric is normalized into having a mean of 0 and standard deviation of 1.

- The data is also unbalanced.

Class	**Frequency**
NoForce	8531
Hancuff	1326
Hand	737
Other	650
Hand Hancuff	259
Rare	249
Wall	162
Hand Wall	116
Hand Hancuff Wall	76
Hancuff Other	75
OnGround Hand Hancuff	62
WeaponDrawn	57
Hancuff Wall	36
WeaponPointed	28
WeaponDrawn Hancuff	20
Hand Other	20

Classification Models

Naive Bayes

1. Results

   1.1. Accuracy = 54.6%

Class	**Precision(%)**	**Recall(%)**
NoForce	75.4	71.3
Hancuff	19.6	28.2
Hand	12.8	19
Other	32.3	31.3
Hand Hancuff	6.3	3.7
Rare	20	4
Wall	7.1	7.7
Hand Wall	0	0
Hand Hancuff Wall	0	0
Hancuff Other	0	0
OnGround Hand Hancuff	0	0
WeaponDrawn	40	40
Hancuff Wall	0	0
WeaponPointed	11.1	50
WeaponDrawn Hancuff	0	0
Hand Other	0	0

Support Vector Machine

1. Parameters

   1.1.Kernel = Radial Basis

   1.2.Gamma = 1/features

   1.3.Epsilon = 0.01

   1.4.Constant = 1

2. Results:

   2.1.Accuracy = 67.9%

   2.2.Precision = 67.9%

   2.3.Recall = 100%

   2.4.SVM predicted all of the testing datapoints as NoForce.

C5.0 tree

1. Results:

    1.1. Accuracy = 67.2%

Class	Precision(%)	Recall(%)
NoForce	74.6	90.1
Hancuff	30.7	21.8
Hand	32.6	19
Other	54.8	25.4
Hand Hancuff	0	0
Rare	33.3	8
Wall	33.3	15.4
Hand Wall	83.3	45.5
Hand Hancuff Wall	33.3	10
Hancuff Other	0	0
OnGround Hand Hancuff	0	0
WeaponDrawn	60	60
Hancuff Wall	0	0
WeaponPointed	50	50
WeaponDrawn Hancuff	0	0
Hand Other	0	0

Model Comparison

The majority of the force use is not using any force, which takes up to 67.9% of the dataset. SVM only predicted that no force is used and have the highest accuracy. But it failed to predict any other force type. The C5.0 Tree have the highest recall and precision of force type excluding no force while the accuracy is about the same as SVM at 67.2%. Naïve Bayes accuracy is only at 54.6%.

Evaluation and Deployment

This model is also not useful for the police. The precision and recall of each force type is too low. Collecting some of these data is also seems to be impossible actually stop and frisked the suspect.

Collecting more data points that include more numbers of force type will increase the model's effectiveness. Attributes such as birth country, years in New York, salary could increase the model's performance.