Data Mining - Project 2
Predictive Modeling (Classification)

A classification and prediction based analysis of Dallas Crime Report Data

- Vivek Sanghvi Jain

Under Guidance of
Prof. Michael Hahsler

Table of Contents

List of Figures

1. Executive Summary & Business Understanding

There was a report published earlier last year, suggesting that crime rates in Dallas have increased. Dallas police are changing their tactics to combat the rise of all types of crimes. This is a matter of concern to anyone and everyone not only living in Dallas, but in areas surrounding Dallas.

This report is a continuation of an earlier report which dealt with Data and Visualisation based on the data collected from the Dallas Police Department, on online portal "*Dallas Open Data"*. Using this data, an in-depth analysis was performed using RStudio, which reveals some important statistics and sheds new light on rise of rise of crime in Dallas.

The primary objective of this report is to help answer 2 questions:

1. Can we predict if an incident will result in an arrest?
2. Can we predict some types of offences using the data?

To answer these questions, several different predictive models were attempted and different classification techniques were considered , some of them have been presented later. Thus, the predictive model helps us in predicting the probability of a future outcome based on certain parameters like existing data on top of which the model was built.

For the first question, we can use several parameters to figure out if an incident will result in an arrest or not. Once we use a classifier to figure this out, the police department can then effectively use this information to deploy an appropriate amount of force to investigate and eventually arrest the offender; this information also helps in police department and 911 operators prioritise first responder man power.

For the second question, I've taken a subset of the data to figure out if the Crime in consideration is against a child. So that, not only proper police force, but also child protective services can be involved in the investigation at the earliest.

Some popular predictive models include: Decision Trees, Rule Based Classifiers, Bayesian classifiers, Support Vector machines, Random forest,etc.

In this project Decision Trees have been extensively used. And the tree sheds some new information of effective handling of situations that might lead to an arrest and possible predictions of type of offences.

The report has been complied according to the guidelines of CRISP-DM Framework

2. Data Understanding & Data Preparation

Data Analytics is a very powerful process involving extensive analysis of a huge amount of data; which, if done effectively, can shed new insight on to known data yielding powerful inferences which can prove to be very useful.

For this report we have used a small part of the data from the Dallas Open Data, as explained in the earlier project. In addition to those, we added a few more attributes and the total set is shown below:

Col #	Column Name	Description, Example	Scale Type
1	IncidentNum	RMS generated number	Nominal
2	PCClass	Severity level of the Arrest Incident & Classification of Offence, according to the Penal Code	Ordinal
3	Premise	Location type where arrest incident took place For example, Apartment Parking, Residence	Nominal
4	ZipCode	Zipcode of the offense location	Nominal
	Date1	The date of the offense	Interval
6	Year1	Year of the offense	Interval
7	Month1	Month of the offense	Interval
8	Day1	Day of the offense	Interval
9	Time1	Time of the offense	Ratio
10	Date1DayOfYear	The calender number of the year 1?365	Interval
11	CallReceived	Date and time of call	Interval
12	CompRace	Complainant's Race(Asian, Black, White,NativeIndian,Italian, Latino)	Nominal
13	CompSex	Complainant's Sex(M or F)	Nominal
14	CompAgeAtOffenseTime	Complainant's Age at the time of the offense. Numeric	Ratio
15	UCROffDesc	UCR Offense description. A list of several descriptions	Nominal
16	Gang	Yes or no if offense is Gang realated	Nominal
17	Drug	Yes or no if offense is Drug realated	Nominal
18	**Status**	One of the following: Clear by Arrest, Clear by Exceptional Arrest, Closed/Cleared, Open,Returned for Correction or Suspended	Nominal
19	**arrest**	A Logical value true or false signifying if an arrest was made	Nominal
20	**childRelated**	A Logical value true or false signifying if a child was the victim	Nominal
21	**Division**	The Divisions in which Tx is divided. They can be Central, North-Central, North-East,North-West South-Central SouthEast, SouthWest	Nominal
22	**Watch**	Time for Policing is divided into 3- hour shifts or watches. They can be 1/2/3 signifying Midnight/Day/Evening Shift	Nominal

Table 2.1: Data Description for Selected Data Variables

i. Can we predict if an incident will result in an arrest?

To predict if an incident will result in an arrest, or not, we use the following class variable.

arrest

This attribute is the class attribute on which we use our multiple prediction models to determine if an arrest was made or not, or, for a new event, an arrest will be made or not?

This class variable predicts if an arrest will be made or not based on the parameters fed to the classifier, like information about watch, signal, PCClass, Premise, Time1, ZipCode, Division .

This class variable is created by the aggregation of the Status variables with values, Clear by Arrest" or "Clear by Exceptional Arrest" or "Returned for Correction".

Code used to aggregate different status into two nominal boolean true or false value signifying an event where an arrest was made is marked by arrest=True is:

#Code Snippet

```
data$arrest <- data$Status == "Clear by Arrest" | data$Status == "Clear by Exceptional Arrest" | data$Status == "Returned for Correction"
```

ii. Can we predict some type of Offence using the other data?

childRelated

Mentioned below are some statistics about Child abuse in Texas[ref]

Today, 185 Texas children will be victims of abuse.
In one year, more than 65,000 cases of child abuse were confirmed in Texas.
1 in 4 Girls is sexually abused before her 18th birthday.
1 in 6 boys is sexually abused before his 18th birthday.

And the more we can do to help them, the better. For this, childRelated class variable was created denoting of instances where a child was the victim of a crime. Although, the data shows only about a thousand records of offences against children, it would be helpful to have this information as early as possible so that appropriate authorities and child services can reach the child and provide proper care.

This attribute is the class attribute on which we use our multiple prediction models to determine if an a child is the victim of an arrest or not? If he/she is, then authorities can act accordingly.

This class variable is created by the checking CHILD (OFFENSES AGAINST) in the UCROffDesc variable with several other values.

Code used to create and update different offences as nominal boolean childRelated or not is given as:

#Code Snippet

```
data_red$childRelated <- data_red$UCROffDesc == "CHILD (OFFENSES AGAINST)"
data_red$childRelated <- as.factor(data_red$childRelated)
```

3. Modeling

Out of the several available classifiers, the classifiers that I have user are:

i. Decision Tree

ii. Conditional Inference Tree

iii. Naive Bayes

Accuracy is measured by the formula: accuracy <- function(truth, prediction) { tbl <- table(truth, prediction) sum(diag(tbl))/sum(tbl)

Q1 Model 1: Decision Tree (rPart)

class variable: arrest

parameters used: PCClass, Premise, Day1, Division

Accuracy: 89.24%

Kappa value: 0.16

Kappa: 0.16

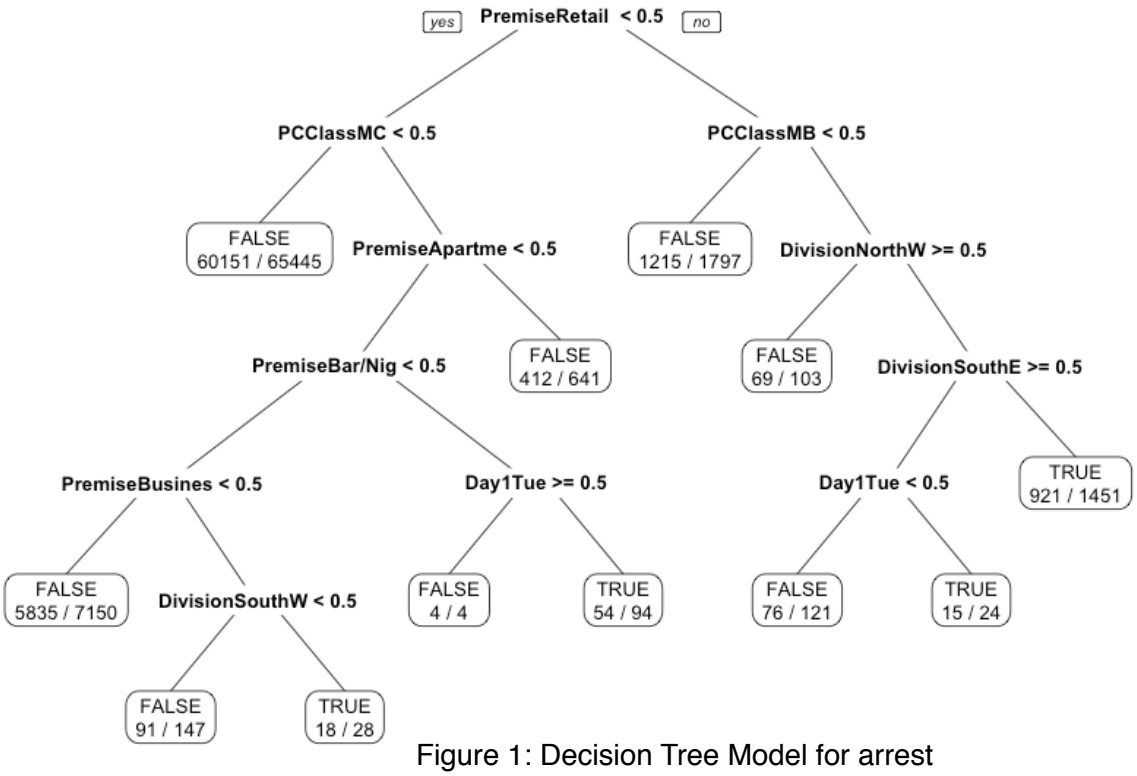

Figure 1: Decision Tree Model for arrest

Code Snippet:

```
datanew <- data_red[,c("arrest","PCClass","Premise","Day1","Division")]
#With training and testing, this code gives:
inTrain <- createDataPartition(y=datanew$arrest, p = .75, list=FALSE)
training <- datanew[ inTrain,]
testing <- datanew[-inTrain,]


RPdnfit1_2 <- train(arrest ~ ., data = training, method = "rpart",
                    control=rpart.control(minsplit=10),
                    trControl = trainControl(method = "cv", number = 10),
                    tuneLength=20)
RPdnfit1_2
rpart.plot(RPdnfit1_2$finalModel, extra = 2)
> pred <- predict(RPdnfit1_2, newdata = testing,na.action = na.pass)
> head(pred)
confusionMatrix(data = pred, testing$arrest)
Confusion Matrix and Statistics
```

From the Base Method to determine base random accuracy: False/(True+False): Probility: 88.8%

```
Accuracy : 0.8924
95% CI : (0.8888, 0.8961)
No Information Rate : 0.8885
P-Value [Acc > NIR] : 0.03572
Kappa : 0.1618
Mcnemar's Test P-Value : < 2e-16
Sensitivity : 0.9895
Specificity : 0.1156
```

This tree predicts if an incident can result in an arrest or noticing several of the aforementioned parameters.

In the beginning, you look if the probability of the Premise being Retail is less than 0.5? If true,traverse to the left subtree and check if probability of PCClass being MC is less than 0.5. If Yes, traverse lets subtree, and come to the conclusion Not Arrested. If Premise isn't Retail, Travel Right Subtree of Root Node, Check if probability of PCClass being MB is less than half and proceed so and and so forth till you reach any of the leaf nodes. For eg. 921 out of 1451 times an incident results in an arrest if probability of Division being SouthEast is more than 0.5, being NorthWest is more than 0.5, PCClass being MB>0.5

The model performs well and has a high accuracy, but not a very high Kappa. It uses the training set effectively. But still doesn't yield an accuracy much higher than the base accuracy

The best feature about this Model is the fact that it is humanly understandable, just a series of true/false questions. One Inference from this model is the fact that Business Premises in SouthWest Division are more likely t translate an incident into an arrest

Q1 Model 2: Condition Inference Tree (CTree)

class variable: arrest

parameters used: PCClass, Premise, Day1, Division

Accuracy: 89.33%

Kappa value: 0.17

#CodeSnippet

```
datanew <- data_red[,c("arrest","PCClass","Premise","Day1","Division")]
C45Fit <- train(arrest ~ ., method = "ctree", data = datanew,
 tuneLength = 5,
trControl = trainControl(
method = "cv"))
```

The tree also predicts if an incident can result in an arrest or noticing several of the aforementioned parameters.

Its uncommon, but both the CTree and the part decision tree have very similar accuracy values and kappa values. Thereby, their results would also be similar.

The model performs has a high accuracy, but not a very high Kappa. and still doesn't yield an accuracy much higher than the base accuracy

Q1 Model 3: Naive Bayes Classifier

class variable: arrest

parameters used: PCClass, Premise, Day1, Division

Accuracy: 88.89%

Kappa value: 0.01

#CodeSnippet

```
datanew <- data_red[,c("arrest","PCClass","Premise","Day1","Division")]
library(RWeka)
C45FitNB <- train(arrest ~ ., method = "nb", data = datanew,
            tuneLength = 5,
            trControl = trainControl(
              method = "cv"))
C45FitNB
```

The tree also predicts if an incident can result in an arrest or noticing several of the aforementioned parameters.

In this case, the accuracy isn't much higher than the base accuracy, but the kappa value is very close to 0. So this classifier with the current data doesn't yield any additional information. With new data, predictions might still be accurate.

Q2

Because we have a very small data set in for answering the Question if a child is a victim of crime, we solve the class imbalance by training the model for a smaller dataset and then use it for predection. In this case, all the 1028 cases of the cleaned data set from Project 1which had child victims, are sampled with another 1028 cases which dint have the said information, thus making the base accuracy to 50%

```
library(sampling)
datachildnew <- data_red[,c("childRelated","PCClass","Premise","Day1","Division","CompRace")]
data_red_1_arresttrue <- datachildnew
data_red_1_arresttrue$childRelated <- factor(datachildnew$childRelated == "TRUE",
                                levels = c(FALSE, TRUE), labels =c("False", “True”))
fit4<- train(childRelated ~ ., data=data_red_1_arresttrue, method = "rpart",
          trControl = trainControl(method = “cv"))
id <- strata(data_red_1_arresttrue, stratanames="childRelated", size=c(1000,1000), method="srswr")
data_red_1_balanced <- data_red_1_arresttrue[id$ID_unit, ]
table(data_red_1_balanced$childRelated)

fit <- train(childRelated ~ ., data = data_red_1_balanced, method = "rpart",
          trControl = trainControl(method = "cv"),
          control = rpart.control(minsplit = 10))
data_red_1_balanced <- na.omit(data_red_1_balanced)
datachildnew<-data_red_1_balanced            #For Simplicity, datachildnew will be used as balanced dataset
```

The above piece of code creates a balanced dataset on which the following models are executed.

Model 1: Decision Tree (rPart)

class variable: childRelated

parameters used: PCClass, Premise, Day1, Division, CompRace

Accuracy: 85.40%

Kappa value: 0.708

#CodeSnippet

```
fitnewchild <- train(childRelated ~ ., data = datachildnew , method = "rpart",
                control=rpart.control(minsplit=10),
                trControl = trainControl(method = "cv", number = 10),
                tuneLength=5)
fitnewchild
rpart.plot(fitnewchild$finalModel, extra=2)
```

Confusion Matrix and Statistics
Reference
Prediction False True

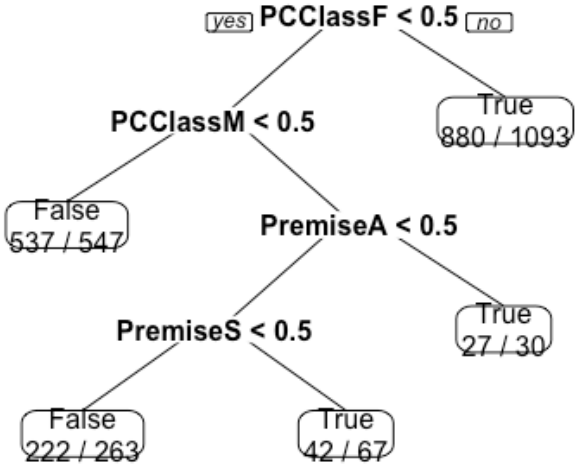

Figure 2: Decision Tree Model for childRelated

False	759	51
True	241	949

Accuracy : 0.854
95% CI : (0.8378, 0.8692)
No Information Rate : 0.5
P-Value [Acc > NIR] : < 2.2e-16
Kappa : 0.708
Mcnemar's Test P-Value : < 2.2e-16
Sensitivity : 0.9490
Specificity : 0.7590

This tree predicts if an incident involve a child victim.

In the beginning, you look if the probability of the PCClass being F is less than 0.5? If True, then the incident does involve an offence towards a child victim 880 out of 1093 times. If False,traverse to the left subtree and check for PCClass M probability, if less than 0.5, then the incident doesn't have a child victim, if false, go to right subtree and check probability of premise being A if less than 0.5, child victim there, if not, check probability for premise S is less than 0.5, 42 out of 67 times, it involves child victims and 222 out of 263 times, it doesnt..

The model performs well and has a high accuracy of 85.40% as compared to the base accuracy of 50%, and a very high Kappa value at 0.708 as well. It uses the training set effectively.

The best feature about this Model is the fact that it is humanly understandable, just a series of true/false questions. The Premises and PCClasses have been abbreviated with their first letter for simplicity and representation purposes in this model. Premise A standing for Airport and so on and so forth.

Q2 Model 2: Condition Inference Tree (CTree)

class variable: childRelated

parameters used: PCClass, Premise, Day1, Division, CompRace

Accuracy: 87.40%

Accuracy SD: 0.02

Kappa Value: 0.748

Kappa SD: 0.05

#CodeSnippet

```
C45FitNBchildctr <- train(childRelated ~ ., method = "ctree", data = datachildnew,
                  tuneLength = 5,
                  trControl = trainControl(
                    method = "cv"))
```

The ctree model has a high accuracy rate and a high kappa rate , even higher than the part decision tree. It also predicts if an incident involves offence to a child victim.

Q2 Model 3: Naive Bayes Classifier

class variable: childRelated

parameters used: PCClass, Premise, Day1, Division, CompRace

Accuracy: 45.80%

Kappa Value: 0.07

#CodeSnippet

```
library(RWeka)
C45FitNBchildnb <- train(childRelated ~ ., method = "nb", data = datachildnew,
                  tuneLength = 5,
                  trControl = trainControl(
                    method = "cv"))
C45FitNBchildnb
```

For this instance the Naive Bayes Classifier has poor accuracy and poor kappa value.

4. Evaluation and Deployment

The models that have been built can be effective in some of the following cases:

1. Prioritisation: Law Enforcement can effectively prioritise and send appropriate number of officers for incidents that may result in arrest.
2. Once deployed after refinement with more training data, child victims can be helped faster.

To perform an evaluation on the model, two metrics can be used and both can be attributed to a cost measure. One is the resources needed to run the model and the other is the gain achieved by proper predictive classification over standard misclassification error.

Stakeholders who can benefit from the models:

1. Law Enforcement Department: Proper classification of possible crime at early stages can be very helpful to law enforcement department.
2. Citizens. Any kind of predictive modelling gives law enforcement an edge to tackle crime which is helpful for regular Citizens who are the probable victims of the crime.

5. References:

[1] http://www.cactx.org/child-abuse-in-texas