Crime Prediction in Swedish Municipalities

with machine learning algorithms

Bachelor's Thesis

15 ECTS

Author
Nils Dominguez Berndtsson

Supervisor
Hiba Nassar

Department of Statistics
Lund University
Sweden

2019-02-28

Contents

Abstract

In this thesis we use a number of common machine learning algorithms to predict crime rates in Swedish municipalities. As predictors we use a mix of municipal socioeconomic variables. For some years we are able to correctly classify up to 85% of the municipalities that have a high crime rate. The highest prediction accuracy rates are obtained from tree and clustering based models. Important factors for forecasting crime in Sweden seem to be divorce rates, male age, unemployment and unsuccessful high school education.

Acknowledgement

I would like to thank my supervisor Hiba Nassar for providing great ideas and feedback when writing this thesis.

1 Introduction

Crime is, needless to say, one of the big problems of society. Apart from causing pain and suffering on a community it also has vast negative economic effects on a society. Some of the negative effects are slowing economic growth, less market competition, less tourism and overall lower quality of life. To decrease the amount spent on crime prevention work it is important to have effective crime-prevention resources. One step in this direction is the evaluation of data-driven crime prevention methods. In the recent years an increase in crime data availability has been beneficial for the data-driven models of crime prediction. Often researchers seek to identify ”High crime areas” or so-called crime hot-spots where crime is more likely to occur. A lot of studies of crime prediction have been conducted at both city and state levels. As the relationships between features and crime are often nonlinear the research area is well suited for models that take this into account. Therefore in recent years, it has been attracting a lot of interest to use machine learning methods for crime prediction. This thesis builds on existing research and uses machine learning methods to predict crime in Swedish municipalities. As model features, we use a range of socioeconomic variables. The dependent variable is chosen as the number of reported crimes per 100.000 inhabitants and taken from the Swedish National Council for Crime Prevention (BRÅ). Findings include that variables such as divorces seem to have a strong lagged positive effect on a municipality's crime rate. Other factors that are important for predictions are unemployment, male age, successful high school graduation and wages. For prediction we obtain sensitivity values as high as 85%. They are however not very stable and the accuracy rate drops significantly for some years (to around 65%). The best predictions are given by the random forest model.

2 Theory

Numerous earlier studies have used the random forest algorithm when predicting crime as this method usually produces good results.

Using a similar framework as this thesis Alves et al. (2018) try to predict crime in Brazilian cities. They stress the importance of using models that account for non-linearity in the predicted relationships. Using a random forest algorithm they are able to explain 97% of the variance in homicide rates between different Brazilian cities. As feature variables they use 10 socioeconomic urban indicators for their predictions and also the number of traffic accidents and the lagged crime rate (10 years) to measure possible autocorrelated behaviour in crime rates. They compare the results from the random forest model with a linear model estimated with ordinary least squares. The predictor variables in the linear model proves to be inconsistent and have no effect on homicide rates.

Bogomolov et al. (2014) predict crime in London boroughs. They use 68 different metrics as predictor variables. To avoid correlation between the predictor variables they perform a principal component analysis.

As final classifier they use a random forest model which yield the best results. Their final model has an accuracy of 70% when explaining the crime rates in London boroughs. Ackerman (1998) examines the relationship between changes in mean crime rates and socioeconomic factors in smaller American cities. Findings include that poverty and weak family structure are the most prominent ones when explaining such changes.

2.1 Predictor variables

The thesis uses predictor variables with roots in socioeconomic theory. When dealing with machine learning methods predictor variables are often called features and we will also employ this term in the thesis. Below follows a description of seven variables that can be linked to crimes.

2.1.1 Divorce rates

Divorce rates have earlier been used in crime prediction (Bogomolov et al., 2014). In a study by Sampson et al. (2006) the marriage rate is shown to have a decreasing effect on crime rates.
An explanation to this is that marriage creates an interdependent system of obligation, mutual support and restraint that impose large cost on potential criminal activities (Sampson et al., 2006).

2.1.2 Successful High School graduation

Education has been shown to reduce the number of crimes as shown by Lochner and Moretti (2004). The authors here used data on high school graduation and also incarceration. They calculate the social benefits of a one percent increase in male high-school graduation to be as high as $1.4 billion in annual savings.

In a similar study from the UK Machin et al. (2011) estimate the social savings from an additional student qualification to be 10 000 pounds.

2.1.3 Unemployment

Regarding the effect of unemployment on crime, the effect has historically been divided into two schools (Melick, 2003). The first effect is due to that a person seeks to sustain the same standard of living it had before the unemployment phase. Secondly they point to the victims side of an unemployment increase. An increase in unemployment eventually leads to a decrease in wealth which would lead to less attractive targets for perpetrators. In this case unemployment is thus believed to have a negative effect on the number of crimes.

2.1.4 Relative income

The relation between relative income and crime levels has been studied but the relationship is not clear. An individual with a lower relative income will according to economic theory be likely to commit crime if

she lives next to wealthier individuals. (Becker, 1968). A lower relative income can also cause feelings of despair which further increases the propensity for individuals to commit criminal acts (Merton, 1938).

2.1.5 Inequality

An empirical causal relation between inequality and crime has been shown by Fajnzylber et al. (2002). As a measure of inequality we use the Gini coefficient which is calculated for every municipality and every year. It is one of the most common measures of income inequality, developed by Gini (1912) and shows how income is distributed in a population. The coefficient ranges between 0 and 1 where 0 represents perfect equality (everybody has the same income) and 1 perfect inequality (one person has all the income). In this thesis the index is calculated with data taken from Statistics Sweden (SCB) on the number of municipal inhabitants in different income classes.

2.1.6 Age

Agewise crime has been claimed to peak when individuals enter adulthood and then decline with age (Quetelet, 1984). In the ”General theory of crime” developed by Gottfredson and Hirschi (1990), crime is linked to the theory of self control. This theory states that ineffective parenting before the age of 10 can cause a lack of self control in children that later make them more prone to committing criminal acts. Documentation of the connection can be found in Hirschi and Gottfredson (1983).

2.1.7 Public health

Among our model features we also include an index for public health. A motivation for this is that people with low incomes could have more physically and mentally demanding and stressful jobs. As people with these jobs often live in areas with more socioeconomic problems that have higher crime rates one could also link public health to crime.

3 Data

One of the most time consuming parts when writing this thesis has been to find and structure the data set. Most of the data is taken from Statistics Sweden (SCB). We employ yearly data for Swedens 290 municipalities for the time period 2000-2016. This time period was chosen as it for some variables was hard to find reliable data for earlier and later dates. In the small number of cases when we have dealt with missing observations linear interpolation has been used.
When predicting we use different lags for the features. This is done to optimize the number of right predictions. The final lag structure chosen was to lag the variable for divorces 5 years and the rest of the features 1 year. As men tend to be overrepresented in crime statistics (Bäckman et al., 2018) the variable age used in the thesis is the male age. Regarding unemployment it would have been optimal to use data on real

unemployment rates. This was however not available on a municipal level. Instead we have resorted to using data on the municipal ratio of people who receive unemployment benefits. The sickness number for a municipality is based on the number of days with paid sickness benefits.
Table 1 shows each predictor variable (or feature), it's lag and where the data was obtained from.

Table 1: **Variables used**

Variable	**Lag**	**Source**
Crimes per 100.000 inhabitants (response variable)		National Swedish Center for Crime Prevention
Divorces per number of marriages	(-5)	SCB
Mean Wage	(-1)	SCB
% of pop on Unemployment benefits	(-1)	SCB
Male mean age	(-1)	SCB
% of high school students that graduate with complete grades	(-1)	National Swedish Agency for Education
Gini coefficient	(-1)	SCB
Sickness number	(-1)	Swedish Insurance Agency

The models use a binary dependent or response variable which is based on the number of reported crimes per 100 000 inhabitants in each municipality per year. The variable is one if the rate of reported crimes in the municipality is higher than 12000 per 100.000 persons per year and zero otherwise. A ratio like this corresponds to the municipalities being above or equal to the sample 93rd percentile.
For structuring the dataset and estimating the models we have used the statistical programming language R. Code for this is available upon request.

4 Method and Results

One motivation for the thesis is to see if machine learning methods are at all applicable for this sort of problems. Therefore, we have in the thesis used a range of different machine learning methods for predictions (displayed in Table 2), and compare them to see which of them gives the best results. We begin with a description of performance measures for the models used. Following this we introduce the principal component method which is used in some of the models. The following subsections will begin with a short presentation of each model and end with the results for that particular model.

Table 2:

Models used
Binary classification tree
Random Forest
AdaBoost
K-Nearest Neighbour
Naive Bayes classifier

4.1 Sensitivity, Specificity and MSE

To evaluate the models used in the thesis, three different performance measures have been used.
These are the Sensitivity, Specificity and Mean squared error (MSE). We will now give a brief explanation of each one of these measures.

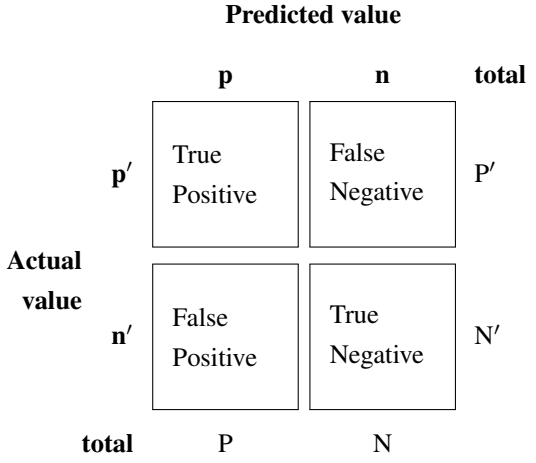

Figure 1: Confusion matrix

To accurately assess what our models get right and which types of problems they have we use the Confusion matrix displayed in Figure 1. When a municipality with a high crime rate is correctly classified we say that it is a True Positive (TP). According to the same reasoning a correctly classified municipality with a low/normal crime rate is a True Negative (TN). The last two options are False Positives (FP) and False Negatives (FN). Formulas for the performance measures Specificity and Sensitivity are defined in formula (1):

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{1}$$

The Sensitivity is the ratio of the correctly classified high crime rate municipalities and the Specificity is the ratio of correctly classified low crime rate municipalities. The last measure we use is the MSE which in our case measures the ratio between the number of correctly classified municipalities and the total number of classifications. The name Mean squared error can in a classification setting seem misleading. With a continuous response variable (Y_i) where $i = 1, ..., N$ the measure is defined as:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i - Y_i)^2 \tag{2}$$

With a binary response variable the squared difference between the two terms in the sum of (2) will be 0 for correctly classified observations and 1 for misclassified observations. Thus we can in this case define MSE as:

$$\text{MSE} = \frac{MC}{N} \tag{3}$$

where MC is the number of misclassified observations. Naturally we want the MSE to be as small as possible.

4.2 Principal component analysis (PCA)

Some models used (K-nearest neighbour, Naive Bayes classifier, AdaBoost) employ a principal component analysis. There are two main reasons for doing this. The first one is to deal with possible multicollinearity between features. The second one is to deal with some peculiarities of the used models,
These being that the K-NN method is sensitive to scaling and that the Naive Bayes model assumes that the features are independent.
The PCA takes the variance of the dataset and decomposes it into different factors. This is done by decomposing the covariance matrix into eigenvectors and eigenvalues. Together they explain how much variance the dataset has in the eigenvectors directions. The eigenvector with the highest eigenvalue becomes the first principal component for the covariance matrix. To choose an appropriate number of factors a scree plot is often used. This plot shows the fraction of total variance explained by each principal component. After observing such a plot, in this thesis 3 components were chosen to represent the variance in the dataset.

4.3 Decision trees

We next present the first method used in the thesis. Decision trees is a collective name for tree based statistical methods. With a continuous response variable they are called Regression trees and with a discrete

one Classification trees. In our case the response variable is binary so we are dealing with classification trees.

Classification trees split the feature space into different regions. The splits are done on the different features. As the number of features tend to be numerous the splits are typically done in high dimensional space.

We will now provide a short example on how Classification trees work. For ease of explanation we will consider a feature space in two dimensions.

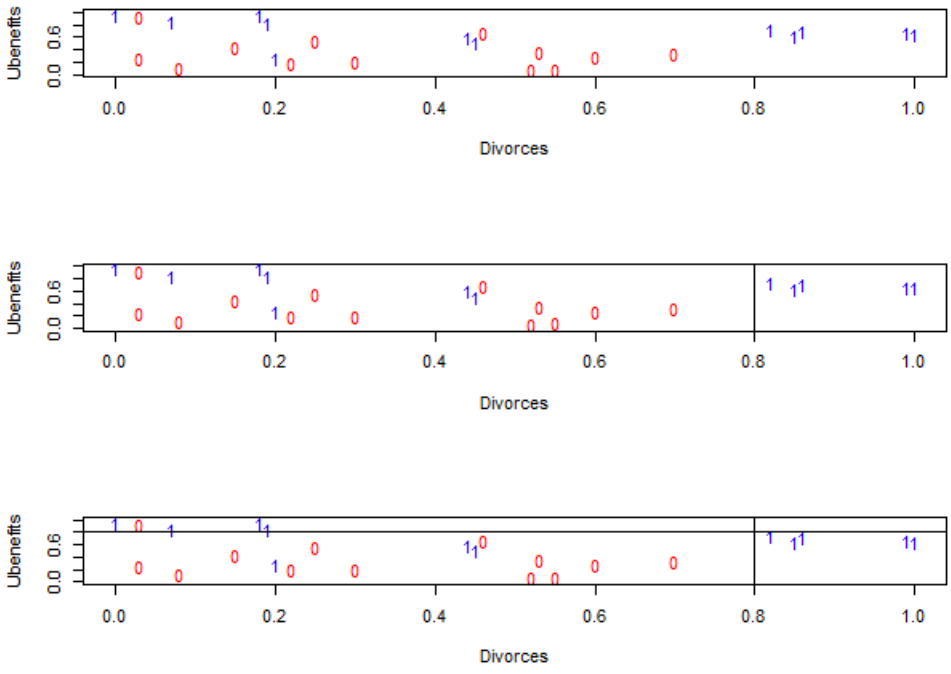

Figure 2: Splitting the feature space into different regions

In the two dimensional case (Figure 2) we have two model features, Divorces and Unemployment benefits and a binary response variable which is high/low crime municipality. The different colours of the dots indicate the different values of the response variable. Here blue ones indicate a municipality that has a high rate of crime and the red zeros are municipalities with lower crime rates. When classifying we measure the Mean squared error (MSE). This is the percentage of region observations that do not belong to the majority class in their region. Before we perform the first split the unit square that the features cover can be seen as a single region and the error is thus $12/25$.

We now perform the first split on the divorces feature (Divorces> 0.8). The new MSE is calculated as the new number of misclassified observations (considering we now have 2 regions) divided by the total number

of observations. That is we have by performing the split diminished our MSE to: $(7+0)/25 = 7/25$. We continue by making the next split on Ubenefits > 0.8. The new MSE then diminishes further and becomes: $(1+3+0)/25 = 4/25$.

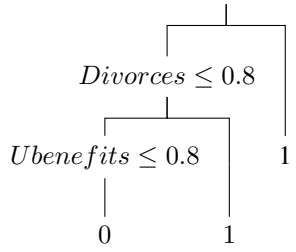

Figure 3: Classification tree example

A more common way to show this procedure is to draw classification trees as the one in Figure 3 which is the tree plot way of showing what we just did.

As we usually have more than two features the splits are commonly done on coordinates in high dimensional planes. The splitting process is usually continued until we reach a set stopping criterion. When this stopping criterion is reached we have a number of terminal nodes that have a value of either 1 or 0. The values of the terminal nodes show the predictions for the municipalities that have the feature values that lead down to the node. Each terminal node is a region and the predictions are simply the majority vote (majority class) in each region.

When splitting the nodes we are interested in which splits that give the highest information gain or most homogeneous tree nodes. As a measure of homogeneity one could of course use the MSE. It is however more common to use so called information criteria that are based on the MSE as measures of gains in homogeneity. There are several possibilities when choosing information criteria. In this thesis we use one called the Gini Index. To understand the formula for this index we first define

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{\mathbf{x_i} \in R_m} \mathbf{I}(y_i = k), \tag{4}$$

the proportion of class k observations in region R_m. The notation in this formula should be understood as follows: First m is an index for different regions or disjoint sets in which we have divided the feature space where $m = 1, ..., M$. The maximal value that M (and hence m) can take is N, in which case we have a unique region for every observation. Secondly N_m is the number of observations in region R_m. Then we let $\mathbf{x_i}$ be the explanatory variable or feature vector for observation i where $i = 1, ..., N$. In our case $\mathbf{x_i}$ is thus a vector with 7 components describing the different features in our model. (Note however that in our simplified example in Figure 2, $\mathbf{x_i}$ has been replaced by a 2 dimensional vector.) Finally y_i (which equals 0 or 1) is the response to the explanatory variable $\mathbf{x_i}$, and therefore k is either 0 or 1.

Ideally we would want $\hat{p}_{mk}$ to be one for one k and zero for the others. This would imply that all $\mathbf{x_i} \in R_m$ have the same response. In practice we try to come as close to this ideal as possible so that R_m is as homogeneous or pure as possible.
The Gini index for region m is computed as:

$$\sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}). \tag{5}$$

We then compute the normalized measure for a state. This becomes the weighted average of the Gini measures for all the different regions:

$$\frac{1}{N} \sum_{m=1}^{M} N_m \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}), \tag{6}$$

(since $N = N_1 + N_2 + ...N_m$). Here K is the number of outcome classes for the response variable (in our case $K = 2$). In the ideal case the Gini index is zero and we want it to be as small as possible.
We split our tree by computing a Gini-index for a state and compare it before and after different splits. Finally we choose to make the splits that give the highest reduction in Gini index.

4.3.1 Results from Classification trees

Table 3:
Classification tree predictions

Predictions (CR_t)	Specificity	Sensitivity	MSE
CR_{2016}	0.97	0.75	0.07
CR_{2015}	0.98	0.68	0.05
CR_{2014}	0.98	0.64	0.05
CR_{2013}	0.96	0.65	0.06

Table 3 shows the result table for the prediction of high or low crime rates with the classification trees.
Here CR stands for the variable high/low crime rate as described in the Data part. When predicting the different municipalities crime rates for a given year we use data on the earlier years to train on. For example when predicting crime rates in 2016 (CR_{2016}) we train on data from 2000-2015, for CR_{2015} we train on data from 2000-2014 and so on. We can see that the model sensitivity is highest for CR_{2016} and lower and fairly stable for earlier years.
The classification tree in Figure 4 show that divorces is an important variable to split on. If we move on from

this variable, the variables age, unemployment benefits, wage and successful high school education seem to matter.

To reduce overfitting classification trees are often pruned. Pruning means reducing the trees size by removing parts of it that are insignificant for classification.

If we prune the tree using cross validation the chosen variables to split on are divorces, age and wage. The way we split on the feature variables are all according to theory. Figure 4 shows some different combinations that can lead to high municipal crime rates. If the divorce rate is lower than 0.016 we see that it is a combination of high unemployment benefits rates and high divorce rates or low middle wage that lead to a high crime rate. If the divorce rate is higher from the beginning the probability of a high crime rate is higher. The attributes of high crime rate municipalities are then high divorces, low male age, high unemployment benefits and unsuccessful high school education.

The pruned tree in Figure 5 pick out high divorce rates, low male age and low wages as the three most contributing factors for municipalities with high crime rates.

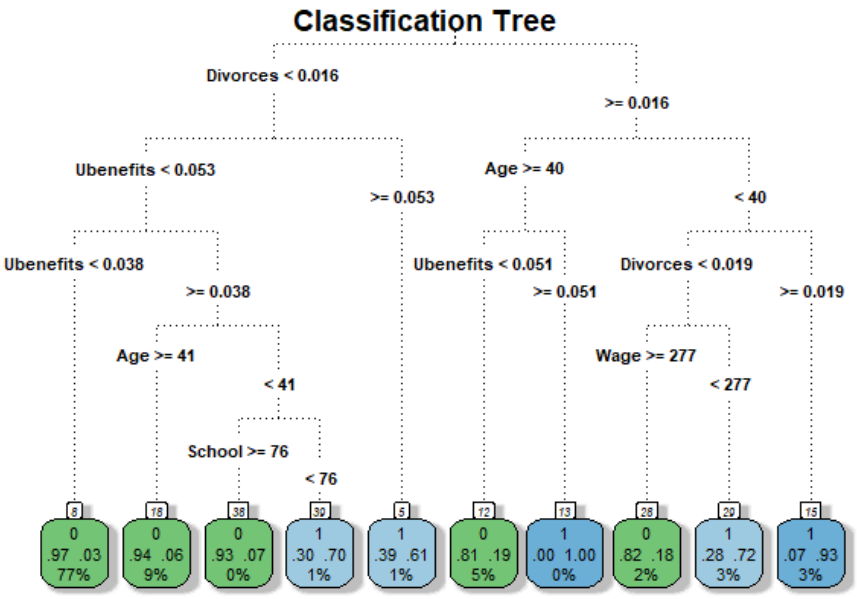

Figure 4: CR_{2016}, classification tree from the rpart package

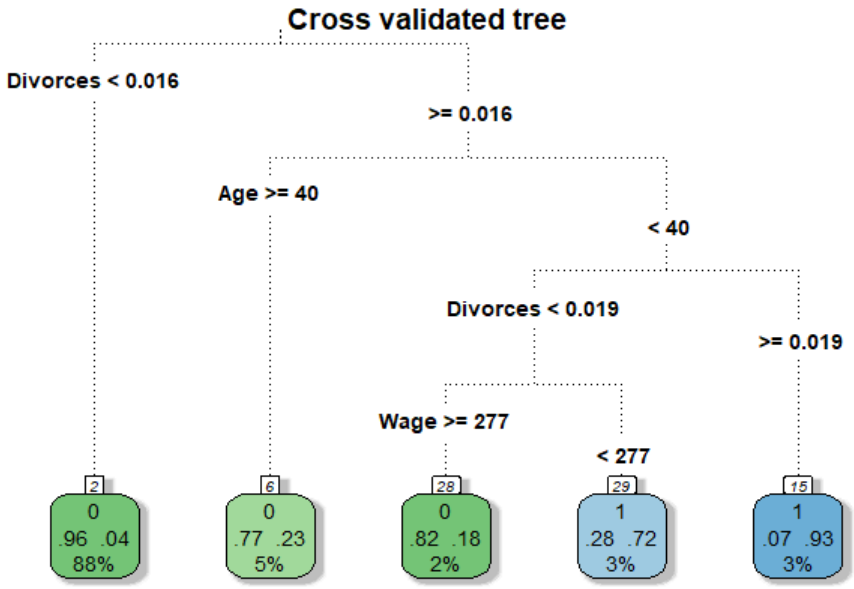

Figure 5: CR_{2016}, classification tree from the rpart package pruned with cross validation

4.4 Random Forest

Classification trees often tend to overfit the data. Apart from pruning a popular way to deal with this problem is to use a method known as random forest. The random forest method uses a combination of bagging and random feature selection to grow a group of trees (forest). Bagging draws bootstrap samples from the training data and then grows trees on these subsets. At each node when growing the trees a number of randomly selected features l that is less than the total number of features n is considered when choosing how to make the optimal splits. The predictions become the average decision from all the grown trees. The random forest method is a so called ensemble learning method. It uses the mean prediction of many weak classifiers as its prediction. Below we display pseudocode for the random forest algorithm based on how it is presented in Friedman et al. (2001).

Algorithm 1 Random Forest

1. For b=1:B

1. Draw a bootstrap sample Z from your data.

   (a) Randomly select l from your total n features where $l < n$

   (b) Make the optimal split of the node based on the l features.

   (c) Repeat a) and b) at each of the resulting nodes until a minimum node size is reached

2. This will result in an ensemble of 1,...,B trees

3. The classification is finally decided by a majority vote of the 1,...,B grown trees.

4.4.1 Results from Random forest

Results from the random forest predictions are shown in Table 4.

Table 4:
Random Forest predictions

Predictions (CR_t)	Specificity	Sensitivity	MSE
CR_{2016}	0.98	0.85	0.05
CR_{2015}	0.98	0.68	0.05
CR_{2014}	0.99	0.68	0.03
CR_{2013}	0.97	0.88	0.02

We can see that we with our features are able to get a fairly high accuracy in predicting whether the municipality will have a high or low crime rate. The predictions are best for the years 2016 and 2013 and drop in accuracy for 2015 and 2014. The predictions yield surprisingly good results for some years but do not seem very stable between the chosen years. The drop in specificity for 2014 leads one the believe that something important happened in the data that year and that it is important to find out more about this peculiarity. Regarding the choice of parameters models trained with cross validation or bootstrapping from the caret package give the same results as when using the default choices for the randomForest package in R.

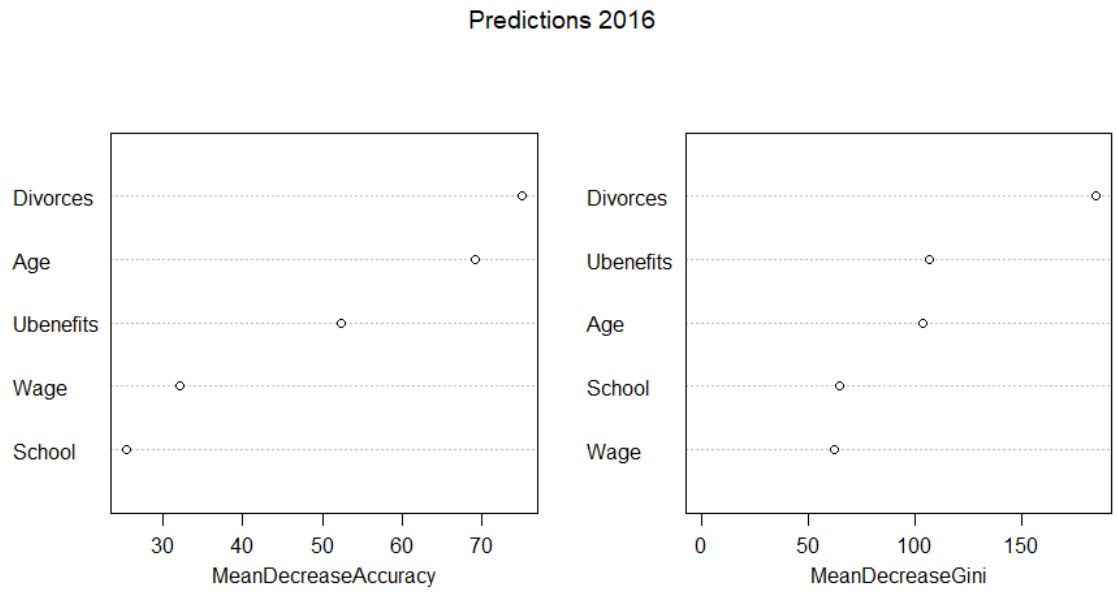

Figure 6: Random forest variable importance plots for predictions of CR_{2016}

Figure 6 shows variable importance plots obtained from the randomForest package in R. The mean decrease in accuracy shows how much more observations on average we would misclassify if we removed the specific variable. For example the left plot in Figure 6 says that if we remove the divorce variable from the model we would misclassify around 75 more observations on average. The Mean decrease in Gini measures how much we reduce node impurity when spitting on the variables. A definition of node impurity can be the gains in homogeneity we obtain by splitting at the node.
Judging from the variable importance plots divorces seem to be the biggest factor influencing crime. This variable is followed by Ubenefits, Age, School and Wage. A drawback with the random forest method is that it does not provide any information about the relationships between the features and crime rates apart from the variance importance plots. The closest we get to studying such relationships are with the binary classification trees.

4.5 AdaBoost

The AdaBoost classifier, first proposed by Freund and Schapire (1997), is an early boosting ensemble algorithm. Like the random forest algorithm it combines many base models to produce one optimal predictive model. As weak classifiers we choose classification trees which are often employed due to their computational efficiency. The idea with AdaBoost is that it assigns weights to each observation in the dataset. By continuously updating these weights it can subsequently assign more weight to observations that are harder

to classify. Pseudocode for the AdaBoost algorithm as displayed in Friedman et al. (2001) is shown below.

Algorithm 2 AdaBoost

1. Initialize the observation weights $w_i = 1/N$, $i = 1, 2, \ldots, N$
2. For a number of trees $t = 1$,..., T:
   (a) Fit a classifier $G_t(x)$ to the training data using weights w_i.
   (b) Compute
   $$\text{err}_t = \frac{\sum_{i=1}^{N} w_i I(y_i \neq G_t(x_i))}{\sum_{i=1}^{N} w_i}.$$
   (c) Compute $\alpha_t = \log((1 - \text{err}_t)/\text{err}_t)$.
   (d) Set $w_i \leftarrow w_i \cdot \exp[\alpha_t \cdot I(y_i \neq G_t(x_i))]$, $i = 1, 2, \ldots, N$.
3. Output $G(x) = \text{sign}\big[\sum_{t=1}^{T} \alpha_t G_t(x)\big]$.

Initially we give all observations in the training sample equal weights ($w_i = 1/N$) where N is the total number of observations. Further we have that $y_i \in \{-1, 1\}$. The error rate for the first classifier, G_1 is computed as in b) using the initial weights. We then use the error rate to compute the next generation of weights. The classification algorithm is then reapplied to the reweighted observations to fit a second classifier and we continue this way. The probably most important step is in d) where the weights are updated.
Here misclassified observations have their weights scaled by a factor $\exp(\alpha_t)$. This will increase their influence when fitting the next classifier in the sequence.
The final output in 3 is the decision based on the combined weighted decisions of the weak classifiers where higher weight is given to classifiers with a lower error rate.

4.5.1 Results from AdaBoost

Table 5:
AdaBoost predictions

Predictions (CR_t)	Specificity	Sensitivity	MSE
CR_{2016}	0.94	0.7	0.08
CR_{2015}	0.98	0.68	0.04
CR_{2014}	0.97	0.72	0.05
CR_{2013}	0.96	0.71	0.05

The parameters for the AdaBoost model are chosen with cross validation from the caret package. We can see that the predictions are somewhat under the ones from the random forest model. The results are similar to the ones obtained when using single binary classification trees.

4.6 K-Nearest Neighbours (K-NN)

The K-NN algorithm works similar to the tree. The methods have in common that they create clusters of observations which are then used for predictions. The method however uses a different methodology than the tree when creating the clusters. Clusters are now formed by connecting a data point to it's nearest neighbour. The number of nearest neighbours for each point are decided by a number (K).
The distance between the points is usually measured by Euclidean distance in feature space.

$$d(p,q) = \sqrt{\sum_{j=1}^{n} (q_j - p_j)^2} \tag{7}$$

Here d is the distance between the points $p = (p_1, .., p_n)$ and $q = (q_1, ..., q_n)$ in Euclidean n-space (in our case $n = 7$). For the algorithm to work well the distance between the observations points in feature space need to be normalized/have the same scale. Below we present pseudocode for the K-NN algorithm.

Algorithm 3 K-NN

1. Initialize K to your chosen number of neighbours
2. For each observation in the feature space
   (a) calculate the Euclidean distance to the other observations
   (b) Add the computed distances together with their indices to an ordered list
3. Sort the lists with distances and indices by distances from smallest to largest in ascending order
4. Pick the first K entries from the sorted lists
5. Obtain the labels of the selected K entries

Each observation is finally classified to the most frequent class amongst it's K-nearest neighbours.

4.6.1 Results from K-Nearest Neighbours

Table 6:
K- Nearest Neighbours predictions

Predictions (CR_t)	Specificity	Sensitivity	MSE
CR_{2016}	0.95	0.75	0.06
CR_{2015}	0.97	0.6	0.06
CR_{2014}	0.97	0.64	0.06
CR_{2013}	0.97	0.82	0.03

The number of neighbours (K) is chosen to be 5 from cross validation. These results are more unstable than the ones from boosting. The predictions are similar to the ones from the random forest model in that they give high sensitivity to the first and last prediction year.

4.7 Naive Bayes classifier

With the naive Bayes classifier we use the Bayes rule (8). Two assumptions the model makes is that the feature variables are conditionally independent and normally distributed.

$$P(Y \mid X) = \frac{P(Y)P(X \mid Y)}{P(X)} \tag{8}$$

The denominator on the RHS of (8) is in the Naive bayes classifier setting traditionally dropped as it is independent of Y and will not affect the final outcome of the classifier. This gives us (9). In a model with features X_j where $j = 1, ..., n$ we can rewrite (9) as (10). Y in formula (10) and (11) is a binary variable that takes the values 0 or 1. Assuming that the features are conditionally independent of each other we arrive at (11).

$$P(Y \mid X) \propto P(Y)P(X \mid Y) \tag{9}$$

$$= P(Y)P(X_1, X_2, .., X_n \mid Y) \tag{10}$$

$$= P(Y) \prod_{j=1}^{n} P(X_j \mid Y) \tag{11}$$

$P(Y = 1)$ can be estimated from our training data. In the case with a binary response variable this is the number of $Y = 1$ divided by the total number of Y:s.
$P(X_1, X_2, .., X_n \mid Y = 1)$ can also be estimated from the training data with maximum likelihood estimation. We then make the assumptions that our features are conditionally independent and normally distributed.

The maximum likelihood estimation gives us estimates of the Gaussian distribution parameters μ_j and σ_j^2 for the model features. When predicting with the test data we use the estimated parameters to compute probabilities of observations belonging to a certain class.

4.7.1 Results from Naive Bayes classifier

Table 7:
Naive Bayes classifier predictions

Predictions (CR_t)	Specificity	Sensitivity	MSE
CR_{2016}	0.97	0.5	0.06
CR_{2015}	0.97	0.36	0.08
CR_{2014}	0.96	0.55	0.07
CR_{2013}	0.95	0.59	0.07

The Naive Bayes classifier has the overall lowest sensitivity values of all the models used. This is combined with relatively high values in MSE.
When using this model the principal component analysis was thought to be of great importance due to the strong assumptions the model makes. It is possible that the grouping together of feature variables that the principal component method makes is not very effective for making predictions with the Naive Bayes classifier. There could also be a problem with non normality of the three extracted principal components used for prediction.

5 Discussion

We have shown that it is possible to predict crime rates in Swedish municipalities with high accuracy. The most contributing crime factor seems to be weak family structures measured in form of divorces. As this variable is lagged 5 years it seems that the negative consequences of divorces has a non immediate but rather lagged effect on crime. Other factors that play a role in municipality crime rates are high rates of unemployment benefits, success in high school and low wages. We have measured model performance using three measures. These are MSE, Specificity and Sensitivity. Considering these measures we attach the highest importance to the Sensitivity. Sensitivity is the models ability to correctly classify municipalities that have a high crime rate. As the ”high crime rate” municipalities make up 7% of our sample it goes without saying that it is more impressive to correctly classify a municipality that has a high crime rate than to correctly classify a municipality that has a low crime rate. In our case the Specificity and MSE should be used to make sure that a high Sensitivity has not come at a too high cost in any of these other measures.

Our conclusion is that when evaluating the different models the Random forest model gives the best predictions. This is because this method gives the highest overall Sensitivity combined with high values on the Specificity and a low value on the MSE.
Apart from that the Naive Bayes model has the poorest results no clear distinction can be made about which one of the remaining models that is superior. The single classification tree and AdaBoost have high and stable Sensitivity values combined with reasonable values on Specificity and MSE. The KNN has high Specificity and Sensitivity values for CR_{2016} and CR_{2013} but lower for the remaining years.
A generalized additive model (GAM) was also tested on the dataset but had to be removed due to problems with autocorrelated residuals. Similarly a logistic regression model that initially was tried on the data set also had to be discarded due to autocorrelated residuals. The problems with the GAM and logistic regression models highlight the benefits of using non-parametric tree based and clustering methods on this type of data.
A problem faced when working with the thesis has been the quality of data. Even though data on 290 municipalities over 12 years may seem relatively big the machine learning methods used are normally employed on vastly bigger datasets. One could thus in future similar studies look for data with higher frequencies or simply wait until more data on the subject is available. Another critique could be the choice of dependent variable. It is not certain that number of reported crimes is a good proxy for actual crimes. It would for example be problematic if people in some municipalities where especially prone on reporting crime. We have also made the rather strong assumption of that people only commit crime in the municipality where they live. A more rigorous study would have checked for spillover effects across municipal borders.
The two variables that least affect crime are public health and inequality. In the case of public health The hypothesis for using the health variable was that high values on this variable would be correlated with physically and mentally stressful jobs. The despair of having these types of jobs or having them as the only option to resort to could increase the propensity to conduct criminal activities. As this hypothesised relationship did not seem to be significant it could be that the low wage or unemployment variable is a better measure of this proposed "despair" factor. A possible explanation for the absent effect of the Gini coefficient (as a measure of income inequality) is that the intra-municipality effect of inequality is non significant. It is possible that the relationship between inequality and crime is only measurable on a larger scale (for example on a regional or national level). The predictions become somewhat better when lagging the model features which shows that the effect they have on crime rates is not an immediate one.

6 Conclusion

Concluding we have shown that socioeconomic factors have large explanatory power in the crime rates of Swedish municipalities. Training on data before the predicted year we are able to get sensitivity rates around 80%. They are however unfortunately not very stable and vary a lot when predicting for different years. This means that sudden shocks are important to take into account when predicting crime rates. The features that this study shows are useful for predicting crime rates are of interest to researchers in associated

fields. They also provide important information to police and policy makers engaged in crime preventive work. For further research one could try to model the autocorrelated crime rates for municipalities using f.ex multivariate time series models. The relatively high accuracy rates of the models used give positive indications for further studies of forecasting crime rates in Sweden.

References

Ackerman, W. V. (1998). Socioeconomic correlates of increasing crime rates in smaller communities. *The Professional Geographer 50*(3), 372–387.

Alves, L. G., H. V. Ribeiro, and F. A. Rodrigues (2018). Crime prediction through urban metrics and statistical learning. *Physica A: Statistical Mechanics and its Applications 505*, 435–443.

Bäckman, O., R. Hjalmarsson, M. J. Lindquist, and T. Pettersson (2018). Könsskkillander i brottslighet-hur kan de förklaras? *Ekonomisk Debatt 46*(4), 67–78.

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy 76*(2), 169–217.

Bogomolov, A., B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A. Pentland (2014). Once upon a crime: towards crime prediction from demographics and mobile data. In *Proceedings of the 16th international conference on multimodal interaction*, pp. 427–434. ACM.

Fajnzylber, P., D. Lederman, and N. Loayza (2002). Inequality and violent crime. *The Journal of Law and Economics 45*(1), 1–40.

Freund, Y. and R. E. Schapire (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences 55*(1), 119–139.

Friedman, J., T. Hastie, and R. Tibshirani (2001). *The elements of statistical learning*, Volume 1. Springer series in statistics New York, NY, USA.

Gini, C. (1912). Variabilità e mutabilità (variability and mutability). *Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi (1955) ed. Bologna*.

Gottfredson, M. R. and T. Hirschi (1990). *A general theory of crime.* Stanford University Press, Redwood City, CA, USA.

Hirschi, T. and M. Gottfredson (1983). Age and the explanation of crime. *American Journal of Sociology 89*(3), 552–584.

Lochner, L. and E. Moretti (2004). The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *American economic review 94*(1), 155–189.

Machin, S., O. Marie, and S. Vujić (2011). The crime reducing effect of education. *The Economic Journal 121*(552), 463–484.

Melick, M. D. (2003). The relationship between crime and unemployment. *The Park Place Economist 11*(1), 30–36.

Merton, R. K. (1938). Social structure and anomie. *American sociological review 3*(5), 672–682.

Quetelet, A. (1984). *Adolphe Quetelet's research on the propensity for crime at different ages*. Anderson Publishing Company New York.

Sampson, R. J., J. H. Laub, and C. Wimer (2006). Does marriage reduce crime? a counterfactual approach to within-individual causal effects. *Criminology 44*(3), 465–508.