1 Linear regression

The regression problem we are going to solve is predicting the amount of autothefts per population in communities. We did a 5-fold outer crossvalidation with an inner loop 10-fold crossvalidation and feature selection.

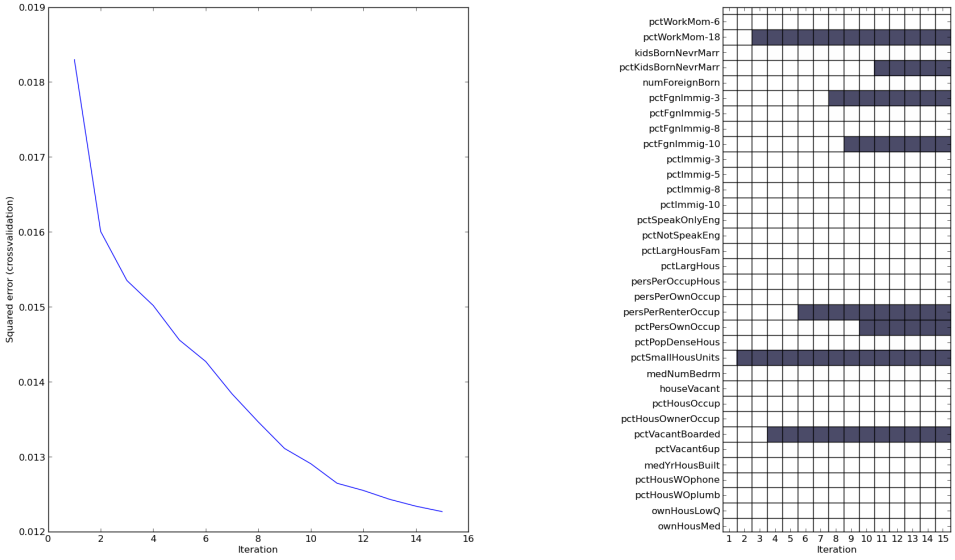

Figure 1: Training of the model in third fold of crossvalidation

The fitting parameters show that our model benefits from feature selection, as both the error and fit of the model to the test are better with feature selection. In figure 3 we see that the residuals look evenly distributed, and we asses that a transformation of variables will have little or no impact on the model.

```
Linear regression without feature selection:

- Training error: 0.00522342840258
- Test error:     0.0631319627131
- R^2 train:     0.785850567012
- R^2 test:     -1.59138161533
Linear regression with feature selection:

- Training error: 0.00990016946312
- Test error:     0.0151810436296
- R^2 train:     0.594114150015
- R^2 test:     0.37686275426
```

1.1 Model application

Our model is predicting the autotheft attribute fairly well with an average 1.5% error with feature selection of the attributes.

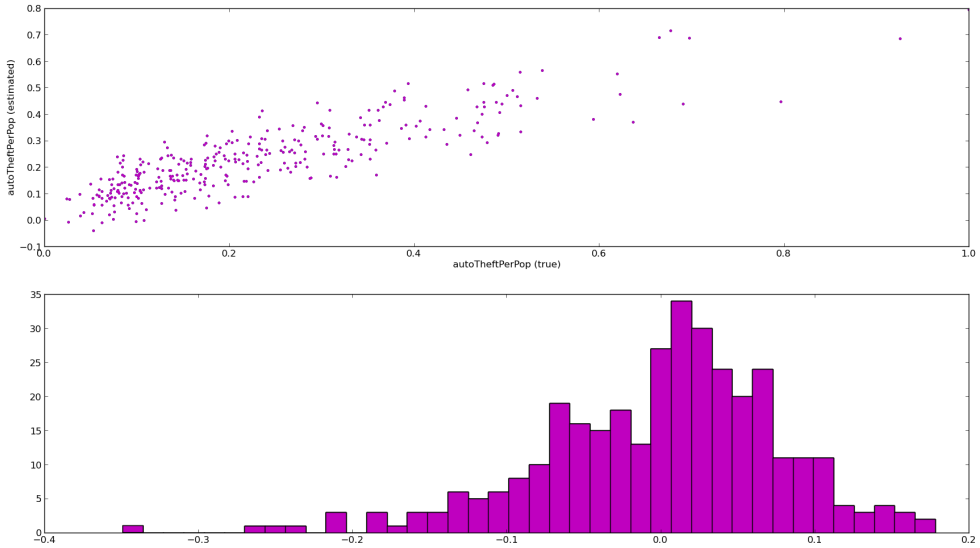

Figure 2: Predicted attribute with the true values and histogram of errors

The model coefficients of the linear model are shown below, in ordered list of the attributes and their coefficient value. policePerPop2 (unexplained in the dataset) and the number of persons living in urban areas are the most positively contributing attributes to the model.

```
['pctImmig-5' 'persPoverty' 'pctMaleDivorc' 'pctFemDivorc' 'persUrban'
 'policePerPop2']
[   1.88989563    2.27254948    2.88873715    3.5109234    97.79176162
  875.63649576]
```

Below are the most negatively contributing attributes where the number of policeofficers per population in the community is contraindicating of autothefts, which makes a lot of sense.

```
['policePerPop' 'pop' 'pctAllDivorc' 'policeField' 'pctImmig-8'
 'pctPersOwnOccup']
[-875.43290039  -98.96072598   -6.54728995   -2.31676831   -1.81957907
   -1.36791983]
```

1.2 Artificial neural network

We could not fit an Artificial neural network due to errors in our code we couldn't fix.

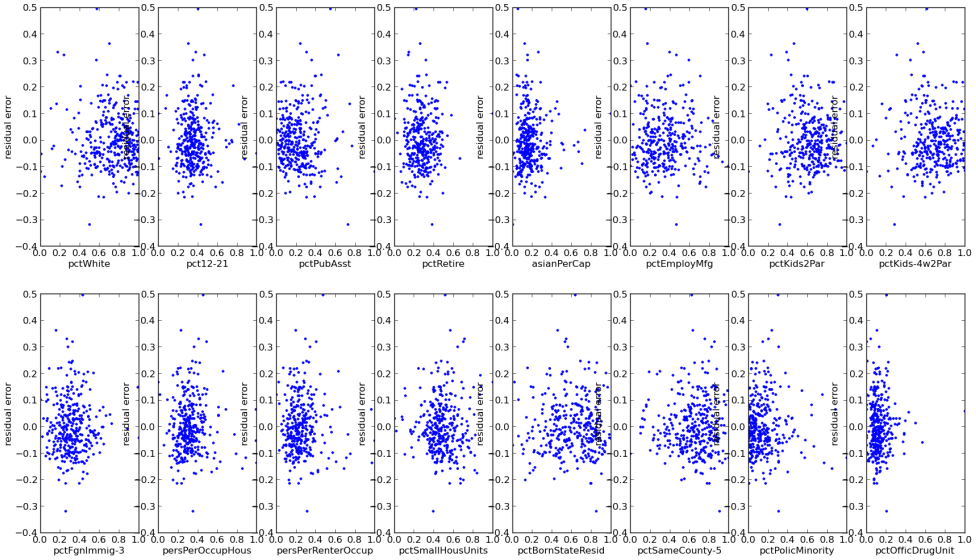

Figure 3: Residual error plots of some of the attributes

1.3 Comparison of models

Comparing the regression model with and without feature selection shows that the predicted true values are significantly different.

```
Regressions are significantly different. (p=0.00179354397592)
```

2 Classification

2.1 The problem

We chose to look at classifying by state, as that was one of the few nominal attributes in our data set. We also considered computing a new attribute corresponding to the dominant ethnicity of the community population. However, we are sceptical that any meaningful classification can be made.

2.2 Model application

Our naive bayes implementation yielded an error rate of 73.17% for our dataset using k=10 for the k-fold crossvalidation.

Our K-nearest neighbours implementation (run with maximum neigbours 30 and only on a subset of our attributes due to time constraints) gave the following error rates

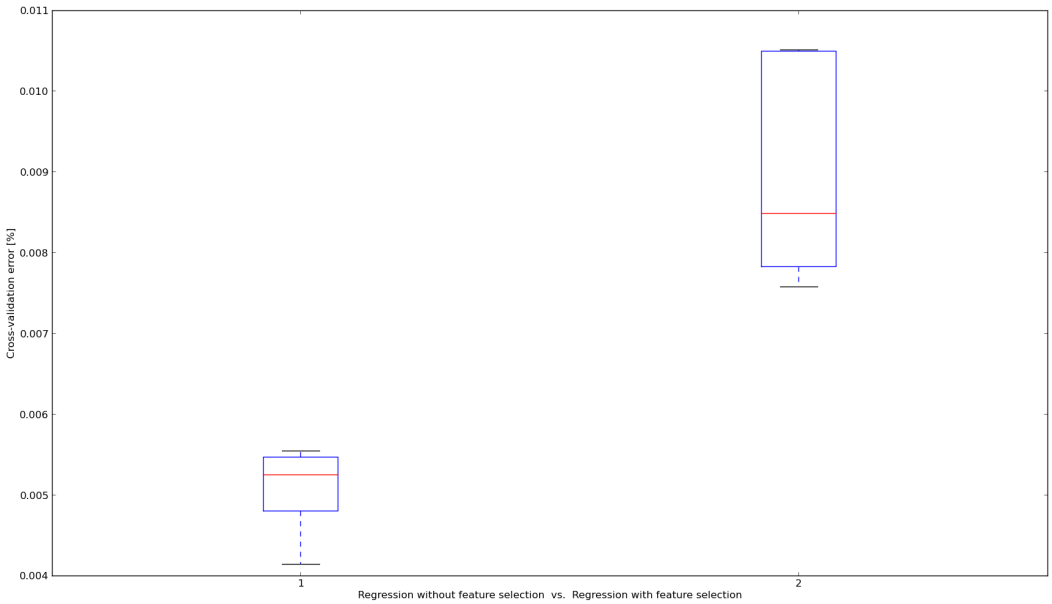

Figure 4: t-test of regression with and without feature selection

2.3 Classification of new data

New data may be predicted using the the knclassifier object of the prefered model. Similarly, the nb_classifier may be used for the Naive Bayes model.

For the K-nearest neighbours model, new data is classified by taking the majority class of the k nearest neighbours.

The error rates look very suspicious, probably due to either programming error or that the data doesn't classify easily.

2.4 T-test

Our t-test yields that the k-nearest neighbour and naive bayes classifier models are significantly different (with p=0.0007135281700851776). The other t-test comparisons gave us that the models weren't significantly different.

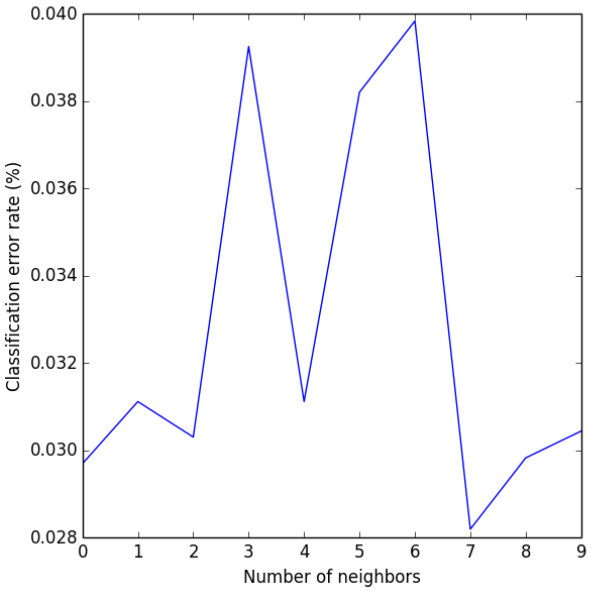

Figure 5: Error rates for Naive Bayes.

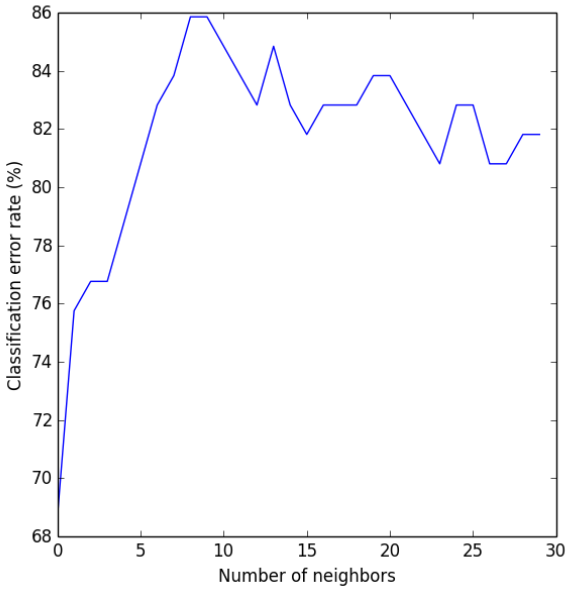

Figure 6: Error rates for K-nearest neighbours.

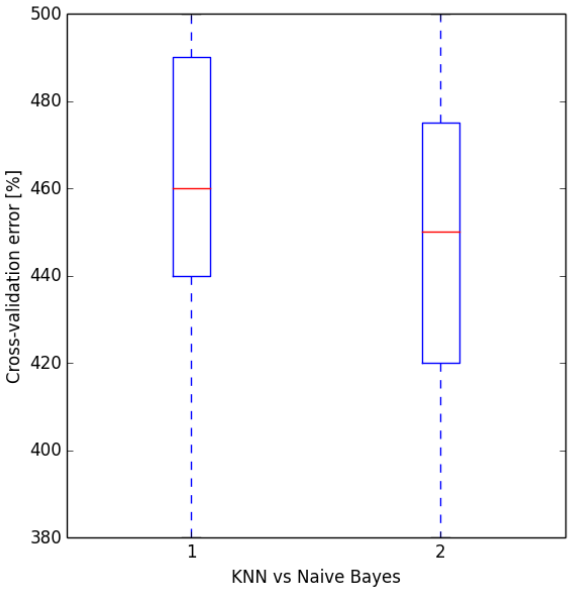

Figure 7: T test of KNN and Naive bayes.

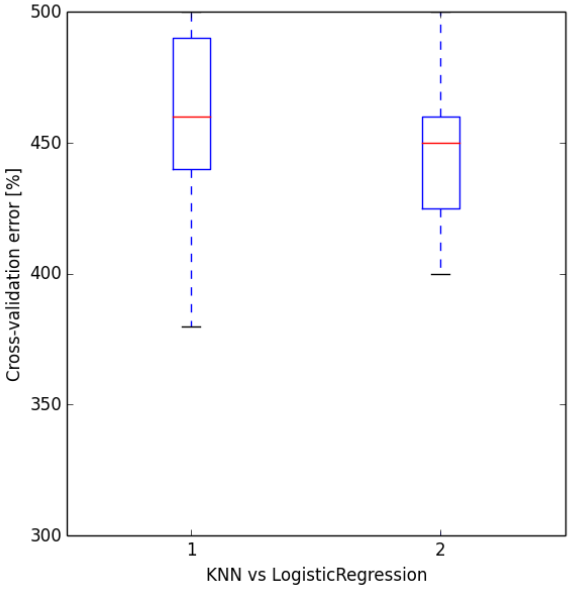

Figure 8: T test of KNN and Logistic Regression.

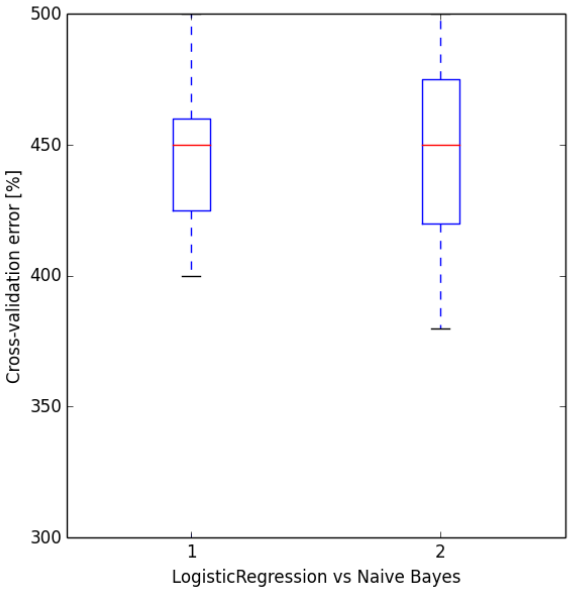

Figure 9: T test of logistic regression and Naive bayes.